Arthritis Care & Research

Aims and Scope

Arthritis Care & Research is an official journal of the American College of Rheumatology and the Association of Rheumatology Professionals, a division of the College. *Arthritis Care & Research* is a peer-reviewed journal that publishes both original research and review articles that promote excellence in the clinical practice of rheumatology. Relevant to the care of individuals with arthritis and related disorders, major topics are evidence-based practice studies, clinical problems, practice guidelines, health care economics, health care policy, educational, social, and public health issues, and future trends in rheumatology practice.

Volume 78 No. 5 May 2026

Arthritis Care & Research

An Official Journal of the American College of Rheumatology
www.arthritiscareres.org and wileyonlinelibrary.com

Cover design: Sandra Pulmano

♾ This journal is printed on acid-free paper.

Arthritis Care & Research

An Official Journal of the American College of Rheumatology
www.arthritiscareres.org and wileyonlinelibrary.com

VOLUME 78 • MAY 2026 • NO. 5

Antiphospholipid Syndrome

Development of the 2023 ACR/EULAR Antiphospholipid Syndrome Classification Criteria, Phase III-C Report: Assessment of Patient Scenarios (Derivation Cohort) and Refinement of Definitions
Medha Barbhaiya, Stephane Zuily, Deanna Jannat-Khah, Mary-Carmen Amigo, Danieli Andrade, Tadej Avcin, Maria L. Bertolaccini, D. Ware Branch, Nathalie Costedoat-Chalumeau, Guilherme Ramires de Jesús, Katrien M. J. Devreese, David Garcia, Jose A. Gomez Puerta, Francis Guillemin, Steven R. Levine, Roger A. Levy, Michael D. Lockshin, Thomas L. Ortel, Michelle Petri, Giovanni Sanna, Savino Sciascia, Surya V. Seshan, Maria Tektonidou, Denis Wahl, Rohan Willis, Cecile Yelnik, Alison Hendry, Ray Naden, Karen H. Costenbader, and Doruk Erkan, on behalf of New APS Classification Criteria Collaborators 563

Systemic Sclerosis

Systemic Sclerosis–Associated Myopathy in a Large Single-Center Cohort: Autoantibody Profiles, Histologic Features, and Independent Risk of Death
Julie J. Paik, Caoilfhionn M. Connolly, Fredrick Wigley, and Laura K. Hummers 574

Assessment of Skin in Patients With Systemic Sclerosis Using High-Frequency Ultrasound and Shear Wave Elastography: A Comparative Study With Histology, Molecular, and Clinical Parameters
Ruhani Desai, Filemon Tan, Minghua Wu, Jeffery L. Browning, Samuel Theodore, Meng Zhang, Brian Skaug, Harshdeep Singh Chawla, Manmohan Singh, Salavat Aglyamov, Kirill V. Larin, Maureen Mayes, and Shervin Assassi 582

Rheumatoid Arthritis

Exploring Factors Associated With Cognitive Impairment in Rheumatoid Arthritis
Raquelle Dawood, Hannah F. Brubeck, Kylie E. Riggles, Sebastian E. Sattui, Elena Myasoedova, Una E. Makris, Dolores M. Shoback, Jose M. Garcia, Ariela R. Orkaby, Joshua F. Baker, Patricia P. Katz, and Katherine D. Wysham 591

Financial Distress and Its Determinants in Rheumatoid Arthritis
Amber Brown Keebler, Yunju Im, Sofia Pedro, Ted R. Mikuls, Edward S. Peters, and Kaleb Michaud 601

Low Back Pain

Testing a Personalized Approach to Chronic Low Back Pain: A Randomized Controlled Trial in Older Veterans
Debra K. Weiner, Angela Gentili, Meika A. Fang, Edward Garay, Laura Lawson, Lenore Joseph, Cathy C. Lee, Michelle I. Rossi, Beverly Thorn, and Subashan Perera 609

Gout

Review Article: Climate Change, Environmental Risk Factors, and Gout
Mary L. Guan, Tamiko R. Katsumoto, Michael Toprover, Nidaa Rasheed, Shuchi Anand, Michael H. Pillinger, Michael H. Weisman, and Suzanne R. Tamang 620

Pediatrics

Diffusion-Weighted Imaging for the Evaluation of the Sacroiliac Joint in Pediatric Patients
Michael L. Francavilla, Timothy G. Brandon, Dmitry Khrichenko, Rui Xiao, Nancy A. Chauvin, Asef Khwaja, and Pamela F. Weiss 630

Other and Mixed Rheumatic Diseases

Disparities and Reproductive Health in Rheumatic Diseases: Deficits in Counseling and Contraception Use in an Urban Female Hispanic Population in Los Angeles
Eaman Alhassan, Laura Kobashigawa, Vanessa Ramos, Jack Rodman, and Leanna Wise 639

The Usefulness of the Patient-Reported Outcomes Measurement Information System Scale v1.2 Global Health for Assessing Mental and Physical Health of Individuals With Chronic Musculoskeletal Pain, Rheumatoid Arthritis, and Hip and Knee Osteoarthritis or Undergoing Physiotherapy: Results of Factor and Item Response Theory Analyses
Emanuele M. Giusti, Leonardo Pellicciari, Martine H. P. Crins, Paul Dekker, Martin van der Esch, Marike van der Leeden, Willem F. Lems, Joost Dekker, Maarten Boers, Dirkjan van Schaardenburg, Johan Joly, Patrick Verschueren, Kristien Van der Elst, Rene Westhovens, Caroline B. Terwee, and Leo D. Roorda 647

Clinical Conditions Associated With a High Antinuclear Antibody Titer in Individuals Without Autoimmune Disease
Matthew Chung, John P. Shelley, Gul Karakoc, John Still, Xiaodi Ruan, Jonathan Mosley, C. Michael Stein, and Vivian K. Kawai 662

Review Article: Accuracy of Diagnostic Codes and Algorithms Used to Identify Rheumatoid Arthritis and Juvenile Idiopathic Arthritis in Administrative Claims and Electronic Health Records: Systematic Review and Meta-Analysis
Constanza Saka-Herrán, Jessica Bennett, Yara Alkabti, Muhammad Fatir, Barbara Clyne, Caroline McCarthy, Gráinne Tynan, Nikki Dunne, Michelle Flood, Eoghan McCarthy, and Frank Moriarty 670

Brief Report: Building Consensus on the Essential Elements of the Musculoskeletal Physical Examination During Rheumatology Telehealth Encounters
Lisa Zickuhr, Alberto Sobrero, Daniel Albert, Amanda S. Alexander, Tami Bonnett-Admi, Sarah Dill, Sharon Dowell, Elizabeth D. Ferucci, Connie Herndon, Bharat Kumar, David Leverenz, Jennifer Mandal, Irene J. Tan, Swamy Venuturupalli, Tiffany Westrich-Robertson, Marcy B. Bolster, and Jason Kolfenbach, for the American College of Rheumatology Workforce Solutions Committee 683

Letter

Knee Crepitus After Anterior Cruciate Ligament Reconstruction: Not Predictive, but Still Pragmatic: Comment on the Article by Couch et al
Shuai Chen 688

Reply

Jamon L. Couch, Brooke E. Patterson, Kay M. Crossley, Danilo De Oliveira Silva, Michael A. Girdwood, Adam G. Culvenor, Ali Guermazi, Matthew G. King, and Jackie L. Whittaker 689

Cover image: The image on the cover (Desai et al; pages 582-590; Supp. Figure 1F) shows trichome staining of forearm skin for collagen density assessment in SSc.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 563–573
DOI 10.1002/acr.25599

American College of Rheumatology
Empowering Rheumatology Professionals

Development of the 2023 ACR/EULAR Antiphospholipid Syndrome Classification Criteria, Phase III-C Report: Assessment of Patient Scenarios (Derivation Cohort) and Refinement of Definitions

Medha Barbhaiya,[1] Stephane Zuily,[2] Deanna Jannat-Khah,[3] Mary-Carmen Amigo,[4] Danieli Andrade,[5] Tadej Avcin,[6] Maria L. Bertolaccini,[7] D. Ware Branch,[8] Nathalie Costedoat-Chalumeau,[9] Guilherme Ramires de Jesús,[10] Katrien M. J. Devreese,[11] David Garcia,[12] Jose A. Gomez Puerta,[13] Francis Guillemin,[14] Steven R. Levine,[15] Roger A. Levy,[16] Michael D. Lockshin,[1] Thomas L. Ortel,[17] Michelle Petri,[18] Giovanni Sanna,[19] Savino Sciascia,[20] Surya V. Seshan,[21] Maria Tektonidou,[22] Denis Wahl,[2] Rohan Willis,[23] Cecile Yelnik,[24] Alison Hendry,[25] Ray Naden,[26†] Karen H. Costenbader,[27] and Doruk Erkan,[1] on behalf of New APS Classification Criteria Collaborators

Objective. The 2023 American College of Rheumatology (ACR)/EULAR antiphospholipid syndrome (APS) classification criteria aim to identify patients with a high likelihood of APS for research. Phases I/II of our four-phase methodologic approach resulted in 27 candidate criteria organized in clinical and laboratory domains. Here, we summarize phase III efforts to reduce and refine criteria using patient scenarios.

Methods. Using standardized definitions for candidate criteria, the Steering Committee collected antiphospholipid antibody (aPL)–positive cases referred for "suspected APS." Treating physicians assessed APS case likelihood using a Likert scale. Poisson regression calculated risk ratios (RRs) and 95% confidence intervals (CIs) to quantify the direction and size of the association of candidate criteria with "highly likely" versus "equivocal or unlikely" APS, which guided Steering Committee candidate criteria refinement and organization.

Results. We collected 314 suspected APS cases (137 [44%] highly likely and 177 [56%] equivocal/unlikely APS). Provoking venous thromboembolism (VTE) or arterial thrombosis (AT) risk factors reduced the size of the association with highly likely APS (RR 4.31 [95% CI 2.11–8.78] to RR 1.56 [95% CI 0.89–2.75] for VTE and RR 3.48 [95% CI 1.91–6.32] to RR 1.64 [95% CI 0.77–3.51] for AT). Persistent lupus anticoagulant, anticardiolipin IgG antibody ≥40 U, and anti–β_2-glycoprotein-I IgG antibody ≥40 U were positively associated with highly likely APS (all $P < 0.05$). Eventually, items within eight additive and independent clinical and laboratory domains were refined.

Conclusion. Referred suspected APS cases provided insight into associations of individual candidate criteria with APS likelihood. RR analyses helped refine items and organize the draft classification system into eight additive and independent clinical and laboratory domains.

INTRODUCTION

Antiphospholipid syndrome (APS) is a systemic autoimmune disorder characterized by thrombosis and/or pregnancy morbidity in patients with persistent antiphospholipid antibodies (aPLs). Classification of APS for clinical trials and research studies was previously based on the revised Sapporo APS classification criteria published in 1999,[1] validated in 2000,[2] and revised in

Supported by the American College of Rheumatology and EULAR. Dr Barbhaiya is supported by the Rheumatology Research Foundation Investigator Award. Dr Bertolaccini is supported by the King's British Heart Foundation Centre of Research Excellence (grant RE/18/2/34213). Dr Costenbader is supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH (grant K24-AR-066109).

[1]Medha Barbhaiya, MD, MPH, Michael D. Lockshin, MD, Doruk Erkan, MD, MPH: Hospital for Special Surgery, Weill Cornell Medicine, New York, NY; [2]Stephane Zuily, MD, MPH, PhD, Denis Wahl, MD, PhD: CHRU de Nancy, University Hospital of Nancy, INSERM UMR-S 1116, University of Lorraine, Nancy, France; [3]Deanna Jannat-Khah, ••: Hospital for Special Surgery, New York, NY; [4]Mary-Carmen Amigo, MD, FACP: ABC Medical Center, Mexico City, Mexico; [5]Danieli Andrade, MD: University of São Paulo, São Paulo, Brazil; [6]Tadej Avcin, MD, PhD: Children's Hospital, University Medical Center, University of Ljubljana, Ljubljana, Slovenia; [7]Maria L. Bertolaccini, PhD: King's College London, London, UK; [8]D. Ware Branch, MD: University of Utah Health, Salt Lake City; [9]Nathalie Costedoat-Chalumeau, MD: APHP, Hôpital Cochin, Université de Paris, Sorbonne Paris Cité, Paris, France; [10]Guilherme Ramires de Jesús, MD, PhD: Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil; [11]Katrien M. J. Devreese, MD, PhD: Ghent University Hospital, Ghent

SIGNIFICANCE & INNOVATIONS

- The 2023 American College of Rheumatology (ACR)/EULAR antiphospholipid syndrome (APS) classification criteria were developed to identify patients with high likelihood of APS for research; here, we summarize our phase III efforts to reduce and refine criteria using patient cases.
- Based on standardized definitions for candidate criteria, international cases referred for "suspected APS" demonstrated varying association of individual candidate criteria with "highly likely" APS, providing relevant guidance to Steering Committee members on decision-making regarding reducing the number of candidate criteria and organizing the structure of the draft classification system within eight additive and independent clinical and laboratory domains.
- Use of international patient data to evaluate the direction and strength of the association for each potential candidate criterion with APS likelihood offers a novel advancement in rheumatic disease classification criteria development methodology.

2006.[3] Given the limitations of the revised Sapporo APS classification criteria, including the need to better capture the full spectrum of disease, distinguish APS from other comorbidities, weigh certain clinical criteria more than others, and include a strong evidence basis for definitions, the new 2023 American College of Rheumatology (ACR)/EULAR APS classification criteria were recently developed and published.[4,5]

The new APS classification criteria development involved a four-phase methodology, as previously used in other rheumatic disease classification systems.[6–8] Details of the methodology overview as well as the Steering Committee development and stakeholders involved in phases I/II/III are provided elsewhere.[9,10]

Following the development of definitions (phase III-A/B) for 27 proposed clinical and laboratory candidate criteria identified during phases I/II,[9] we collected and analyzed patient cases to guide further item reduction and refinement of definitions (phase III-C; reported here) for subsequent use in the multicriteria decision analysis (MCDA) exercises to develop a preliminary, hierarchical APS classification system and threshold for APS classification (phase III-D).[11] In particular, analysis of aPL-related disease manifestations using international patient data helped guide decision-making related to criteria definition refinement as well as hierarchical organization of criteria for use in the MCDA. We report these efforts here as they may provide a unique aspect to classification criteria development efforts moving forward.

PATIENTS AND METHODS

Patient case collection (derivation cohort). To build the derivation cohort that would ultimately be used to test and refine the draft classification system in phase III-D, and that formed the dataset used in the current analyses, we conducted an international case collection effort (approved by the Hospital for Special Surgery Institutional Review Board), during which 20 case collectors (13 Steering Committee members and 7 collaborators) from Europe as well as North and South America provided patient cases.[12] Our case collectors were specialists from hematology, neurology, obstetrics, rheumatology, and vascular medicine with clinical expertise and interest in APS.

Case collectors provided suspected APS patient scenarios reflecting the broad spectrum of clinical manifestations associated with aPL positivity or APS. Using REDCap, a secure web-based data system, we designed a standardized case collection form to represent items included in the revised Sapporo APS classification criteria[3] as well as the 27 candidate criteria identified at the end of phase II in six independent domains (macrovascular, microvascular, obstetric, cardiac valve disease, thrombocytopenia, and laboratory).[9] Using guidance from the domain subcommittees that were developed during phases I/II,[9] as well as phase III-A/B definitions, we collected patient-specific demographic and medical information as well as detailed information related to the clinical and laboratory candidate criteria definitions from phase III-A/B.[13]

Case collectors were asked to submit 20 patient cases for which they were the treating physician. For each of their own submitted cases, the case collectors provided an overall score for APS likelihood using a Likert scale of +3 to −3 (+3 = highly likely APS, 0 = not for or against APS, and −3 = very unlikely APS).

University, Ghent, Belgium; [12]David Garcia, MD: University of Washington, Seattle; [13]Jose A. Gomez Puerta, MD, PhD, MPH: Hospital Clínic and IDIBAPS, Barcelona, Spain; [14]Francis Guillemin, MD, PhD: CHRU and INSPIIRE, University of Lorraine, INSERM, Nancy, France; [15]Steven R. Levine, MD: Kings County Hospital Center and Maimonides Medical Center, State University of New York Downstate Health Sciences University, Brooklyn; [16]Roger A. Levy, MD, PhD: Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil, and GlaxoSmithKline, Philadelphia, PA; [17]Thomas L. Ortel, MD, PhD: Duke University School of Medicine, Durham, NC; [18]Michelle Petri, MD, MPH: Johns Hopkins University School of Medicine, Baltimore, MD; [19]Giovanni Sanna, MD: Guy's and St Thomas' NHS Foundation Trust, London, UK; [20]Savino Sciascia, MD: San Giovanni Bosco Hospital and University of Turin, Turin, Italy; [21]Surya V. Seshan, MBBS: Weill Cornell Medicine, New York, NY; [22]Maria Tektonidou, MD, PhD: National and Kapodistrian University of Athens, Athens, Greece; [23]Rohan Willis, MD: University of Texas Medical Branch, Galveston; [24]Cecile Yelnik, MD: Centre Hospitalier Universitaire de Lille, INSERM, UMR 1167, Lille, France; [25]Alison Hendry, ••: Middlemore Hospital, Auckland, New Zealand; [26]Ray Naden, MD, ChB, FRACP: Auckland City Hospital, Auckland, New Zealand; [27]Karen H. Costenbader, MD, MPH: Brigham and Women's Hospital, Boston, MA.

[†]Dr Naden is deceased.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25599).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25599.

Address correspondence via email to Medha Barbhaiya, MD, MPH, at barbhaiyam@hss.edu.

Submitted for publication September 11, 2024; accepted in revised form June 5, 2025.

Thus, each patient case was assigned a single Likert scale score based on the treating physician's assessment only. As the aim was to identify potential cases spanning the spectrum from highly likely to highly unlikely APS, case collectors were asked to submit cases with varying Likert scale scores (ie, four cases with +3 likelihood score, four with +2, four with +1 to −1, four with −2, and four with −3). Two Core group study team investigators and a research coordinator ensured the appropriate case distribution and completion status at the time of case submission; any discrepancies were addressed with the case collector directly (see Appendix A for the ACR/EULAR APS Classification Criteria Collaborators).

Derivation cohort analyses for candidate criteria. To evaluate the cross-sectional association of candidate criteria with APS likelihood in the derivation cohort dataset, we first categorized cases with higher Likert scores (+2 or +3) as highly likely APS and those with lower scores (+1, 0, −1, −2, or −3) as equivocal or unlikely APS. We evaluated demographic characteristics (age, sex, race, ethnicity, and region of residence) overall, as well as in the highly likely versus equivocal or unlikely APS cases. To assess the "prevalence" of each candidate criterion across collected cases, we calculated the proportion of cases in which an individual criterion occurred. Based on previous Steering Committee discussions,[9] we also evaluated the minimum duration of time between each candidate laboratory and clinical criterion for all cases. To evaluate aPL levels, we descriptively evaluated median anticardiolipin antibody (aCL) and anti–β_2-glycoprotein-I antibody (aβ_2GPI) IgG/M (enzyme-linked immunosorbent assay [ELISA]) levels in the highly likely versus the equivocal or unlikely cases.

To analyze the univariate association between each candidate criterion and APS likelihood, we stratified APS likelihood into "highly likely" (Likert score +2/+3) APS cases compared with "equivocal or unlikely" (−3 to +1) cases. Furthermore, subgroup analyses were included when relevant; for example, in the macrovascular domain, candidate criteria items were collected and analyzed individually (eg, venous thromboembolism [VTE], superficial vein thrombosis [SVT], arterial thrombosis [AT], and transient ischemic attack [TIA] events) as well as in subgroups stratified by VTE or cardiovascular disease (CVD) risk factors at the time of event when feasible. To analyze the outcome of model count variables (as opposed to dichotomous [yes/no] outcomes in logistic regression), we used Poisson regression with robust SEs to calculate risk ratios (RRs), 95% confidence intervals (CIs) and *P* values, representing the probability of the number of times that an individual candidate criterion (eg, cardiac vegetation) occurred in highly likely versus equivocal or unlikely cases. All analyses were performed using STATA version 14.0.

Discussions for domain definitions and finalization of entry criteria. To interpret these patient case collection efforts within the context of proposed clinical and laboratory criteria definitions, we held a series of teleconferences and meetings with the Steering Committee and Domain subcommittees. For each domain, the Steering Committee discussed the candidate criteria definitions, reviewed currently accepted literature on the association of each candidate criterion with APS, and evaluated the significance of the RRs in further refining or eliminating criteria before and during the phase III-D MCDA meeting. The RRs were particularly helpful in guiding Steering Committee decision-making in the absence of existing literature related to the prevalence or risk association of the candidate criterion in APS. Additionally, candidate criteria within each domain were rank-ordered based on Steering Committee agreement using clinical judgment along with the guidance of RRs by evaluating the strength and directionality of the risk association. Statistical significance was not considered in decision-making. Agreement consisted of majority approval by the Steering Committee. As part of these discussions, the Steering Committee also discussed finalization of previously proposed entry criteria from phase I/II requiring the fulfillment of at least one laboratory criterion (aPL positivity [lupus anticoagulant (LA) test, aCL IgG/M, or aβ_2GPI IgG/M positivity above the normal laboratory range] at any time) with at least one clinical criterion (identified from the clinical domains).[9]

RESULTS

Patient case collection (derivation cohort). We collected 314 potential APS cases (mean ± SD age 44.7 ± 14.6 years; 79% female) from 17 sites: Europe (n = 8 sites, 47%), North America (n = 7, 41%), and South America (n = 2, 12%) (Table 1). A total of 137 cases (44%) were rated as highly likely APS and 177 (56%) as equivocal or unlikely APS. Cases came predominantly from Europe (44%) and the United States (35%), with similar distribution in highly likely versus equivocal or unlikely APS cases (Table 1). The duration between aPL positivity and the occurrence of any of the candidate clinical criteria was ≤3 years in 89% of cases.

Derivation cohort analyses for clinical candidate criteria. *Macrovascular domain.* In the overall cohort, a similar proportion of patients experienced VTE (30%) and AT (26%), whereas few patients experienced SVT (5%) or TIA (9%) (definitions in Table 2 footnote). Compared with the unlikely or equivocal APS cases, VTE (RR 2.54 [95% CI 1.77–3.66]) or AT (RR 3.12 [95% CI 2.05–4.76]) was more likely to occur in highly likely APS. When stratified by the presence or absence of provoking risk factors, VTE cases without any VTE risk factors had a positive association (RR 4.31 [95% CI 2.11–8.78]) with highly likely APS, whereas no association was demonstrated for VTE cases with

Table 1. Baseline demographic characteristics of 314 cases, overall and by physician assessment of APS likelihood*

Patients	Total (N = 314)	Highly likely APS (n = 137)	Equivocal or unlikely APS (n = 177)
Age, mean ± SD, y	44.7 ± 14.6	45.9 ± 14.8	43.2 ± 14.3
Female, n (%)	249 (79)	105 (77)	144 (81)
Race, n (%)			
American Indian or Alaskan	1 (0.3)	0 (0)	1 (1)
Asian	6 (2)	1 (1)	5 (3)
Black	29 (9)	6 (4)	23 (13)
White	243 (77)	113 (83)	130 (74)
Another race	35 (11)	17 (12)	18 (10)
Ethnicity, n (%)			
Hispanic/Latin American	40 (13)	22 (16)	18 (10)
Not allowed to record	28 (9)	14 (10)	14 (8)
Not Hispanic/Latin American	189 (60)	73 (53)	116 (66)
Other	57 (18)	28 (20)	29 (16)
Region of residence, n (%)			
Europe	139 (44)	56 (41)	83 (47)
North America (USA and Mexico)	126 (40)	59 (43)	67 (38)
South America	49 (16)	22 (16)	27 (15)

* APS, antiphospholipid syndrome.

VTE risk factors (RR 1.56 [95% CI 0.89–2.75]). A similar positive association was demonstrated for AT cases without CVD risk factors (RR 3.48 [95% CI 1.91–6.32]), whereas no association was demonstrated in those with risk factors (RR 1.64 [95% CI 0.77–3.51]). There was no increased association for SVT (RR 0.47 [95% CI 0.15–1.45]; n = 15 cases) and TIA (RR 1.20 [95% CI 0.58–2.47]; n = 27 cases) with highly likely APS overall.

Microvascular disease domain. Microvascular domain events (including livedo racemosa, livedoid vasculopathy, adrenal hemorrhage or adrenal plexus thrombosis, acute ischemic encephalopathy, cardiac microvascular disease, pulmonary hemorrhage, or aPL nephropathy) were observed in 70 patients (22%) and were positively associated with highly likely APS (RR 1.83 [95% CI 1.20–2.78]). Based on a two-level definition of microvascular involvement, that is, suspected versus established (Table 2 footnote), we found (a) a positive association for established microvascular involvement with highly likely APS (RR 3.66 [95% CI 1.48–9.05]) and (b) a potentially positive association for suspected microvascular involvement (based on physical examination and laboratory or imaging findings) with highly likely APS (RR (1.75 [95% CI 0.97–3.14]). An inverse association was demonstrated for cases with no microvascular disease with highly likely APS (RR 0.81 [95% CI 0.72–0.92]).

Obstetric domain. Of 249 women, 188 had ever been pregnant (highly likely APS = 81; unlikely APS = 107); 45% experienced prefetal loss, 33% fetal losses, 14% severe pre-eclampsia (PEC), and 12% severe placental insufficiency (PI; not mutually exclusive). There were positive associations for fetal loss between 16 weeks and 34 weeks (RR 4.46 [95% CI 2.29–8.68]) and PEC with severe features before 34 weeks (RR 2.79 [95% CI 1.30–5.97]) with highly likely APS. Relatively few cases had PI with severe features (RR 2.01 [95% CI 0.89–4.54]).

Cardiac valve domain. Cardiac valve thickening or vegetation, observed in 27 patients (9%), was positively associated with highly likely APS (RR 2.58 [95% CI 1.20–5.58]). Cardiac valve vegetation had a positive association with highly likely APS (RR 5.81 [95% CI 1.27–26.5]), but a positive association was not demonstrated for cardiac valve thickening.

Hematology domain. Low platelet count (ever <150 × 10^9/L) was observed in 93 patients (30%) and was positively associated with a high likelihood of APS (RR 2.05 [95% CI 1.45–2.91]) compared with equivocal or unlikely APS. No association was demonstrated for a platelet count <20 × 10^9/L (n = 15) with highly likely APS cases. In contrast, for the lowest platelet count between 20 and 130 × 10^9/L, there was a positive association with highly likely APS compared with equivocal or unlikely APS (RR 2.17 [95% CI 1.35–3.51]). A similar positive association was demonstrated for the lowest platelet count of 131 to 150 × 10^9/L (ever) with highly likely APS (RR 2.79 [95% CI 1.09–7.19]).

Derivation cohort analyses for laboratory candidate criteria. *Coagulation-based functional assays: LA test.* Of 314 cases, 212 (68%) had a positive LA test (by case collector–reported commercially available tests) on at least one occasion. There were 136 patients (43%) with a history of persistent positive LA, that is, two positive tests at least 12 weeks apart, of whom 115 (37% overall) also had a positive result on most recent testing. Cases with any positive LA test result had a positive association with highly likely (vs equivocal or unlikely) APS (RR 1.63 [95% CI 1.40–1.90]); this association was higher for persistent positivity (RR 2.70 [95% CI 2.04–3.58]) and persistent positivity with the most recent test result being positive (RR 3.66 [95% CI 2.57–5.20]) (Table 3).

Solid phase assays: aCLs and aβ_2GPIs. Of 314 cases, 170 had a positive (defined as result above normal range) aCL

Table 2. Clinical manifestations of 314 international potential APS cases, overall and categorized by physician assessment*

Candidate relative criteria	Total (N = 314), n (%)	Highly likely APS (n = 137), n (%)	Equivocal or unlikely APS (n = 177), n (%)	RR (95% CI)	*P* value
Macrovascular domain[a]	183 (58)	108 (79)	75 (42)	**1.86 (1.53–2.26)**	**<0.01**
VTE	95 (30)	63 (46)	32 (18)	**2.54 (1.77–3.66)**	**<0.01**
No VTE	219 (70)	74 (54)	145 (82)	**0.66 (0.56–0.78)**	**<0.01**
VTE with any VTE risk factor	42 (83)	23 (17)	19 (11)	1.56 (0.89–2.75)	0.12
VTE without any VTE risk factor	39 (12)	30 (22)	9 (5)	**4.31 (2.11–8.78)**	**<0.01**
AT	82 (26)	58 (42)	24 (14)	**3.12 (2.05–4.76)**	**<0.01**
No AT	232 (74)	79 (58)	153 (86)	**0.67 (0.57–0.78)**	**<0.01**
AT with any CVD risk factor	25 (80)	14 (10)	11 (6)	1.64 (0.77–3.51)	0.20
AT without any CVD risk factor	48 (15)	35 (26)	13 (7)	**3.48 (1.91–6.32)**	**<0.01**
Microvascular domain[b]	70 (22)	41 (30)	29 (16)	**1.83 (1.20–2.78)**	**<0.01**
No microvascular involvement	251 (80)	97 (71)	154 (87)	**0.81 (0.72–0.92)**	**0.01**
Suspected microvascular involvement	40 (13)	23 (17)	17 (10)	1.75 (0.97–3.14)	0.06
Established microvascular involvement	23 (7)	17 (12)	6 (3)	**3.66 (1.48–9.05)**	**<0.01**
Obstetric domain[c]	127 (51)	58 (55)	69 (48)	1.12 (0.86–1.47)	0.40
Prefetal loss (<10 weeks 0 days)	85 (34)	34 (32)	51 (35)	0.93 (0.65–1.32)	0.69
Fetal loss (10 weeks 0 days to 15 weeks 6 days)	20 (8)	9 (9)	11 (8)	1.14 (0.49–2.66)	0.76
Fetal loss (16 weeks 0 days to 34 weeks 0 days)	42 (22)	32 (30)	10 (7)	**4.46 (2.29–8.68)**	**<0.01**
Severe pre-eclampsia (< 34 weeks)	27 (17)	18 (17)	9 (6)	**2.79 (1.30–5.97)**	**<0.01**
Severe placental insufficiency (<34 weeks)	22 (9)	13 (12)	9 (6)	2.01 (0.89–4.54)	0.09
Cardiac valve disease domain	27(9)	18 (13.1)	9 (5.1)	**2.58 (1.20–5.58)**	**0.02**
Thickening	11 (4)	9 (7)	2 (1)	1.72 (0.75–3.98)	0.20
Vegetation	21 (7)	12 (9)	9 (5)	**5.81 (1.27–26.5)**	**0.02**
Hematologic domain (platelet count <150 × 10^9/L)	93 (30)	57 (42)	36 (20)	**2.05 (1.46–2.91)**	**<0.01**
Lowest platelet count <20 × 10^9/L	15 (5)	7 (5)	8 (5)	1.13 (0.42–3.05)	0.81
Lowest platelet count 20–130 × 10^9/L	59 (20)	37 (28)	22 (13)	**2.17 (1.35–3.51)**	**<0.01**
Lowest platelet count 130–150 × 10^9/L	19 (6)	13 (9)	6 (3)	**2.79 (1.09–7.19)**	**<0.01**

* An RR of 1.0 represents a candidate criterion being equally probable in "highly likely" APS and "equivocal or unlikely APS" cases, whereas an RR >1.0 implies the criterion is more probable in cases of highly likely APS than in equivocal or unlikely APS. Similarly, an RR of <1.0 implied a negative association between the candidate criterion and "highly likely" APS. RRs represent the univariate association of the candidate criteria with highly likely versus equivocal/unlikely APS. $P < 0.05$ is significant and indicated with bold font. APS, antiphospholipid syndrome; AT, arterial thrombosis; CI, confidence interval; CVD, cardiovascular disease; RR, risk ratio; VTE, venous thromboembolism.

[a] Macrovascular domain: Column percentages for cases with and without any risk factors may not add up to 100 because cases with unknown risk factor status were excluded. Macrovascular events (eg, VTE and AT) were evaluated individually; however, some cases may have experienced both types of events. The overall association between superficial vein thrombosis and transient ischemic attack with APS likelihood was evaluated with the following results: superficial vein thrombosis (RR 0.47 [95% CI 0.15–1.45]; n = 15 cases) and transient ischemic attack (RR 1.20 [95% CI 0.58–2.47]; n = 27 cases). We did not stratify the superficial vein thrombosis and transient ischemic attack RR analyses by presence/absence of risk factors because of small event numbers. VTE risk factors included active malignancy, central venous catheter, estrogen-containing oral contraceptive use, hormone replacement therapy hospital admission, long-distance travel, obesity, surgery with general/spinal/epidural anesthesia, pregnancy and postpartum period, and prolonged immobilization. AT risk factors included active smoking, diabetes mellitus, family history of premature CVD, hyperlipidemia, and hypertension.

[b] Microvascular domain: (a) Suspected microvascular involvement (livedo racemosa by physical examination, livedoid vasculopathy by physical examination, pulmonary hemorrhage by clinical symptoms and imaging, and antiphospholipid antibody nephropathy by physical examination and laboratory tests) and (b) established microvascular involvement (livedoid vasculopathy by physical examination and pathology; antiphospholipid antibody nephropathy [acute or chronic] by pathology; pulmonary hemorrhage by clinical symptoms, imaging, and bronchoalveolar lavage or pathology; myocardial disease by imaging or pathology; and adrenal hemorrhage or microthrombosis by imaging or pathology).

[c] Obstetric domain: Percentages were calculated out of 249 total women in the cohort (105 with highly likely APS and 144 equivocal or unlikely APS).

IgG, 138 a positive aCL IgM, 99 a positive aβ_2GPI IgG, and 79 a positive aβ_2GPI IgM. In the highly likely APS group, the median titers (based on the most recent test) for aCL IgG, aCL IgM, aβ_2GPI IgG, and aβ_2GPI IgM were 96 U, 50 U, 90 U, and 48 U, respectively, compared with 35 U, 30 U, 46 U, and 55 U, respectively, in the equivocal or unlikely APS group.

aCLs. For the IgG isotype, any positive aCL IgG value had a positive association with highly likely APS (RR 1.46 [95% CI 1.20–1.77]); this association was stronger in those with persistent positivity (RR 2.88 [95% CI 2.07–4.01]). Among patients with persistently positive aCL IgG whose most recent test was positive (overall RR 4.78 [95% CI 3.07–7.43]), a positive association was also demonstrated for aCL IgG levels of 40 U to 79 U (RR 5.17 [95% CI 1.99–13.44]) and ≥80 U (RR 11.37 [95% CI 4.63–27.94]); no association was demonstrated for levels <40 U (RR 1.29 [95% CI 0.55–3.02]).

For the IgM isotype, any positive aCL IgM demonstrated no association with highly likely APS (RR 1.02 [95% CI 0.80–1.32]). However, persistently positive aCL IgM had a positive association with highly likely APS (RR 1.59 [95% CI 1.08–2.34]); a similar value

Table 3. Laboratory manifestations of 314 cases, overall and by physician assessment of APS likelihood*

Candidate relative criteria	Total (N = 314), n (%)	Highly likely APS (n = 137), n (%)	Equivocal or unlikely APS (n = 177), n (%)	RR (95% CI)	*P* value
LA[a]	308 (98)	137 (100)	171 (97)	**1.04 (1.01–1.06)**	**0.01**
Negative	96 (31)	17 (12)	79 (45)	**0.27 (0.17–0.43)**	**<0.01**
Any positive	212 (68)	120 (88)	92 (54)	**1.63 (1.40–1.90)**	**<0.01**
Positive: not persistent[b]	76 (24)	28 (20)	48 (27)	0.75 (0.50–1.14)	0.18
Positive: persistent[b]	136 (43)	92 (67)	44 (25)	**2.70 (2.04–3.58)**	**<0.01**
Most recent negative	21(7)	7 (5)	14 (8)	0.65 (0.27–1.56)	0.33
Most recent positive	115(37)	85 (62)	30 (17)	**3.66 (2.57–5.20)**	**<0.01**
aCL					
IgG	296 (94)[c]	129 (94)	167 (94)	1.00 (0.94–1.05)	0.94
Negative	126 (40)	39 (29)	87 (49)	**0.58 (0.43–0.78)**	**<0.01**
Any positive[d]	170 (54)	90 (66)	80 (45)	**1.46 (1.20–1.77)**	**<0.01**
Positive: not persistent[b]	39 (12)	10 (7)	29 (16)	**0.45 (0.22–0.88)**	**0.02**
Positive: persistent[b]	113 (36)	78 (57)	35 (20)	**2.88 (2.07–4.01)**	**<0.01**
Most recent negative	19 (6)	4 (3)	15 (9)	0.34 (0.12–1.02)	0.05
Most recent positive	94 (32)	74 (57)	20 (12)	**4.78 (3.07–7.43)**	**<0.01**
<40 U	20 (7)	10 (8)	10 (6)	1.29 (0.55–3.02)	0.55
40–79 U	20 (7)	25 (19)	5 (3)	**5.17 (1.99–13.44)**	**<0.01**
≥80 U	49 (17)	44 (3)	5 (3)	**11.37 (4.63–27.94)**	**<0.01**
IgM	257 (82)	109 (80)	148 (85)	0.94 (0.84–1.04)	0.24
Negative	119 (38)	48 (35)	71 (40)	0.87 (0.65–1.17)	0.36
Any positive[d]	138 (44)	61(45)	77 (44)	1.02 (0.80–1.32)	0.86
Positive: not persistent[b]	41 (13)	14 (10)	27 (15)	0.67 (0.37–1.23)	0.20
Positive: persistent[b]	78 (25)	43 (31)	39 (22)	**1.59 (1.08–2.34)**	**0.02**
Most recent negative	9 (3)	4 (3)	5 (3)	1.03 (0.28–3.78)	0.96
Most recent positive	69 (22)	39 (28)	30 (17)	**1.68 (1.10–2.56)**	**0.02**
<40	33 (11)	16 (12)	17 (10)	1.22 (0.64–2.32)	0.55
40–79	17 (5)	10 (7)	7 (4)	1.85 (0.72–4.73)	0.20
≥80	19 (6)	13 (9)	6 (3)	**2.80 (1.09–7.19)**	**0.03**
Anti-β_2-glycoprotein-I antibody					
IgG	258 (82)	113 (82)	145 (82)	1.01 (0.91–1.12)	0.90
Negative	159 (51)	47 (34)	112 (63)	**0.54 (0.42–0.70)**	**<0.01**
Any positive[d]	99 (32)	66 (48)	33 (19)	**2.58 (1.81–3.68)**	**<0.01**
Positive: not persistent[b]	13 (4)	5 (4)	8 (5)	0.81 (0.27–2.42)	0.70
Positive: persistent[b]	72 (23)	55 (40)	17 (10)	**4.18 (2.54–6.87)**	**<0.01**
Most recent negative	6 (2)	4 (3)	2 (1)	2.58 (0.48–13.9)	0.27
Most recent positive	66 (21)	51 (37)	15 (8)	**4.39 (2.58–7.47)**	**<0.01**
<40	16 (5)	9 (7)	7 (4)	1.66 (0.63–4.35)	0.30
40–79	16 (5)	12 (9)	4 (2)	**3.88 (1.28–11.77)**	**0.02**
≥80	34 (11)	30 (22)	4 (2)	**9.69 (3.49–26.89)**	**<0.01**
IgM	248 (79)	106 (77)	142 (80)	0.96 (0.86–1.08)	0.54
Negative	169 (54)	59 (43)	110 (62)	**0.69 (0.55–0.87)**	**<0.01**
Any positive[d]	79 (25)	47 (34)	32 (18)	**1.90 (1.28–2.80)**	**<0.01**
Positive: not persistent[b]	14 (9)	7 (14)	7 (4)	1.29 (0.46–3.60)	0.62
Positive: persistent[b]	52 (17)	33 (24)	19 (11)	**2.24 (1.33–3.77)**	**<0.01**
Most recent negative	6 (2)	2 (1)	4 (2)	0.65 (0.12–3.48)	0.61
Most recent positive	46 (15)	31 (23)	15 (8)	**2.67 (1.50–4.75)**	**<0.01**
<40	17 (5)	10 (7)	7 (4)	1.85 (0.72–4.73)	0.20
40–79	12 (4)	9 (66)	3 (2)	**3.88 (1.07–14.07)**	**0.04**
≥80	17 (5)	12 (9)	5 (3)	**3.10 (1.12–8.61)**	**0.03**

* An RR of 1.0 represents a candidate criterion being equally probable in "highly likely" APS and "equivocal or unlikely APS" cases, whereas an RR >1.0 implies the criterion is more probable in cases of highly likely APS than in equivocal or unlikely APS. Similarly, an RR of <1.0 implied a negative association between the candidate criterion and "highly likely" APS. RRs represent the univariate association of the candidate criteria with highly likely versus equivocal/unlikely APS. $P < 0.05$ is significant and indicated with bold font. aCL, anticardiolipin antibody; APS, antiphospholipid syndrome; CI, confidence interval; LA, lupus anticoagulant; RR, risk ratio.

[a] Among 308 patients with LA test ever performed.

[b] "Persistent" defined as positive on at least two occasions at least 12 weeks apart. Persistence status was not always known.

[c] Two patients with unknown aCL IgG results.

[d] "Positive" defined as the level above the normal range.

was found for persistent positivity with the most recent test result being positive (RR 1.68 [95% CI 1.10–2.56]). By level, in those with the most recent aCL IgM test result being positive, a positive association was only demonstrated for persistent levels ≥80 U (RR 2.80 [95% CI 1.09–7.19]), which were observed in 19 individuals. Of note, of 19 individuals with an aCL IgM level ≥80 U,

15 were concomitantly associated with LA positivity (12 of 15 with persistently positive LA); of the four without LA positivity, one was associated with persistently positive aCL IgG levels ≥80 U and two individuals had isolated IgM positivity.

aβ$_2$GPIs. For the IgG isotype, any positive aβ$_2$GPI IgG had a positive association with highly likely APS (RR 2.58 [95% CI 1.81–3.68]) as did persistent positive aβ$_2$GPI IgG (RR 4.18 [95% CI 2.54–6.87]) and the subset of persistent positivity with most recent test result being positive (RR 4.39 [95% CI 2.58–7.47]). By level, persistently positive aβ$_2$GPI IgG <40 demonstrated no association (RR 1.66 [95% CI 0.63–4.35]) with highly likely APS in individuals with the most recent test result being positive; there was however a strongly positive association for levels between 40 U to 79 U (RR 3.88 [95% CI 1.28–11.77]) and ≥80 U (RR 9.69 [95% CI 3.49–26.89]).

For the IgM isotype, any positive aβ$_2$GPI IgM had a positive association with highly likely APS (RR 1.90 [95% CI 1.28–2.80]) similar to that of persistent positive results (RR 2.24 [95% CI 1.33–3.77]). In the subset with persistent positivity and most recent aβ$_2$GPI IgM test result being positive (RR 2.67 [95 CI 1.50–4.75]), levels <40 U had no association (RR 1.85 [95% CI 0.72–4.73]) with highly likely APS. However, in this group the association was strongly positive for levels between 40 U and 79 U (RR 3.88 [95% CI 1.07–14.07]) and ≥80 U (RR 3.10 [95% CI 1.12–8.61]).

Discussions for domain definitions. Steering committee discussions and decisions in preparation for the phase III-D MCDA, based on literature review and derivation cohort analysis, are outlined in Supplementary Table 1 and summarized below.

In the clinical domains, the committee agreed to (a) separate VTE and AT into two distinct macrovascular domains, each containing two levels based on the presence or absence of a "high-risk" profile, with the higher rank order attributed to the latter consistent with recent VTE and CVD guidelines[14–16]; (b) a two-level microvascular domain (ie, "suspected" vs "established") based on the strength of objective evidence; (c) a three-level obstetric domain (recurrent prefetal deaths <10 weeks and/or early fetal death[s] between 10 weeks 0 days and 15 weeks 6 days of gestation), PEC with severe features or PI with severe features before 34 weeks of gestation, and PEC with severe features and PI with severe features before 34 weeks of gestation[17]; (d) a two-level cardiac valve disease domain with higher rank order given to cardiac valve vegetation and lower rank order assigned to cardiac valve thickening; and (e) a one-level hematology domain (ie, thrombocytopenia defined as the lowest platelet level ever between 20 and 130 × 10^9/L). In the laboratory domains, (a) a two-level LA positivity and four-level aCL/aβ$_2$GPI positivity after separating lupus anticoagulant and aCL/aβ$_2$GPI into two different domains and (b) defining a "moderate" level of aCL/aβ$_2$GPI IgG/M positivity as 40 U to 79 U and a "high" level as ≥80 U (ELISA) (Table 4).

The Steering Committee also discussed that decision-making for the MCDA should be in the context of the entry criteria, requiring at least one documented clinical criterion (from domains 1–6) in addition to a positive aPL test.[18] Based on discussions surrounding strong RR associations for aCL/aβ$_2$GPI titer levels >40 U with highly likely APS, as well as the currently available literature, the committee agreed to require a positive LA test result and/or moderate to high titers of aCL or aβ$_2$GPI IgG or IgM. Furthermore, based on previous discussions, the Steering Committee agreed with a 3-year time restriction between the clinical criterion and a positive aPL test result.[18]

DISCUSSION

In phase III-C of the 2023 ACR/EULAR APS classification criteria development, based on precise definitions for phase II candidate criteria, we used multicenter, international cases to evaluate the association of each candidate criterion with patient cases referred for suspected APS. Our cases spanned the spectrum from highly likely to unlikely APS to capture heterogeneous clinical and laboratory manifestations potentially consistent with APS. RR analyses to evaluate the association of individual aPL-related disease manifestations with APS likelihood provided relevant data to guide the Steering Committee in decision-making surrounding candidate criteria definitions and hierarchical organization within domains in preparation for the phase III-D MCDA.

During phase III, the use of international patient data to evaluate the direction and strength of the association for each potential candidate criterion with APS likelihood offers a novel advancement in rheumatic disease classification criteria development methodology. Similar examples of advancements in other recent rheumatic disease classification development include efforts toward phenotyping patients based on clinical and laboratory features in IgG4-related disease[7] and classifying early disease in rheumatoid arthritis and systemic lupus erythematosus.[6,8] In a relatively rare disease like APS, with limited published literature on the relative weight and discriminant ability of individual candidate criteria, the use of patient cases with heterogeneous manifestations provides valuable data. Using treating physician assessment to determine APS likelihood allowed for risk-stratifying candidate criteria when relevant (eg, by the presence of provoking risk factors in the macrovascular domain) and quantifying the association of less common clinical manifestations (eg, adrenal hemorrhage, cardiac microvascular disease, diffuse pulmonary hemorrhage, and aPL nephropathy, all positively associated with highly likely APS). Furthermore, the exclusion of low prevalence items with a lack of uniform definitions in the literature or inverse associations with highly likely APS in patient cases (ie, acute ischemic encephalopathy, thought to be more common in other clinical scenarios such as status epilepticus, stroke, or septic emboli) may improve the specificity of the classification criteria.

Table 4. Hierarchical organization of the candidate criteria within each domain in preparation for the phase III-D multiple criteria decision analysis*

Domains	Levels
Clinical	
1: Macrovascular (VTE)	A. No B. VTE with high-risk profile C. VTE without high-risk profile
2: Macrovascular (AT)	A. No B. AT with high-risk profile C. AT without high-risk profile
3: Microvascular	A. No B. Suspected C. Established
4: Obstetric	A. No B. ≥3 consecutive prefetal (<10 weeks) and/or early fetal deaths (10–16 weeks), or ≥1 fetal death (16–34 weeks) alone C. PEC with severe features or PI with severe features (<34 weeks) with or without fetal death D. PEC with severe features and PI with severe features (<34 weeks) with or without fetal death
5: Cardiac valve	A. No or not tested B. Valve thickening C. Valve vegetation
6: Hematology	A. No or not tested B. Thrombocytopenia (20–130 × 10^9/L)
Laboratory	
7: Antiphospholipid antibody testing: coagulation-based functional assays: lupus anticoagulant test	A: Negative or not tested B: Positive (single: one time) C: Positive (persistent)
8: Antiphospholipid antibody testing: solid phase assays: aCL and aβ_2GPI IgG/M ELISA (persistent)	A. Negative or not tested B: Moderate or high positive (IgM alone) (aCL and/or aβ_2GPI) C. Moderate positive (IgG) (aCL and/or aß$_2$GPI) D. High positive (IgG) (aCL or aβ_2GPI) E. High positive (IgG) (aCL and aβ_2GPI)

* aCL, anticardiolipin antibody; AT, arterial thrombosis; aβ_2GPI, anti–β_2-glycoprotein-I antibody; ELISA, enzyme-linked immunosorbent assay; PEC, pre-eclampsia; PI, placental insufficiency; VTE, venous thromboembolism.

In the clinical domains, our results demonstrated a positive association for certain, but not all, macrovascular criteria: VTE and AT had positive associations with highly likely cases but TIA and SVT did not and occurred less frequently. The ability to directly evaluate thrombotic events by "provoking" venous or arterial thrombotic risk factors confirmed our hypothesis that concomitant non-aPL thrombotic risk factors reduce the association of VTE or AT events with highly likely APS and provided additional rationale to incorporate this important stratification into the MCDA exercises. Furthermore, the committee acknowledged that thrombotic risk factors may be currently unknown, particularly in retrospective studies or past events, in which case the domain item should either not be scored or be applied the lowest possible weight. However, for prospective studies efforts should be made to document VTE or CVD risk factors.

Phase III-C efforts also helped support incorporating "suspected" versus "established" microvascular disease in the MCDA meeting, providing a major advancement, as microvascular disease is thought to differ from both a mechanistic and treatment point of view from moderate-to-large vessel disease in APS. Detailed discussions of each microvascular candidate criteria, including definitions, are reported elsewhere;[13] however, the Steering Committee agreed that the most challenging microvascular criterion was the definition of aPL nephropathy (new definitions by the renal pathology subcommittee have been published elsewhere for pathologists, clinicians, and researchers[19]).

During phase III discussions, under the guidance of the obstetric domain subcommittee, defining items such as fetal death, PEC, and PI were challenging at times given limited literature support; our phase III-C results helped guide decision-making. Additionally, the cardiac valve disease domain, defined as valvular thickening and/or vegetations, was added based on literature review and Steering Committee agreement; RR analyses supported the concept that vegetations may be more associated with APS than thickening. The inclusion of thrombocytopenia as a candidate criterion was also novel; decisions related to definitions and threshold levels were largely driven by Steering Committee agreement, supported whenever possible by literature and patient case collection demonstrating the lack of association of very low platelet levels (<20) with APS.

Although our sample included only a few patients with "lowest platelet count <20," our analysis did not seem underpowered. Furthermore, platelet counts <20 × 10^9/L are typically due to other etiologies (eg, idiopathic thrombocytopenic purpura [ITP]) and unlikely to be associated with aPL positivity alone. Although the correlation of aPL positivity and ITP has been examined, the frequency and clinical implications are still unknown.[20,21] Finally, within each clinical domain, the use of RR analyses supported the development of a hierarchical organization system, such that certain items may carry more weight based on a more positive association with highly likely APS cases.

In the laboratory domains, RR analyses demonstrated a similar and strongly positive association of persistent LA, persistent aCL IgG, and persistent aβ_2GPI IgG with highly likely APS cases, overall and particularly when the most recent testing result was positive. Based on this finding, the committee agreed that the "most recent" aPL result should be used for prospective studies, whereas the latest available aPL result within 3 years of the aPL-related clinical manifestation should be used for retrospective studies. Consistent with the literature demonstrating the importance of aPL persistence and the higher relative specificity IgG versus IgM isotype,[22] we found either no association or only a weakly positive association for persistent aCL/aβ_2GPI IgM, single positive LA result, and any positive aCL or aβ_2GPI IgG/M. Furthermore, the committee agreed that the relative increase in the association of aPL ELISA tests for titer levels 40 U to 79 U and ≥80 U with highly likely APS cases, particularly for IgG isotypes, further emphasized the value of assessing APS cases for classification using these aPL ELISA levels.

Our study has strengths and limitations. One of our goals was to reduce the potential bias of circular reasoning by minimizing participant overlap during classification criteria development. As the APS research community is relatively small compared with other more common systemic rheumatic diseases, and although every effort was made to ensure mutually distinct participants in phases I/II/III, some overlap in participants was unavoidable. However, in phase III-C, we ensured that the case collectors who provided patient data for the derivation cohort, which would subsequently be used to develop the draft classification criteria system, were entirely distinct from phase IV case collectors who provided data for the validation cohort. By including cases referred for suspected APS to diverse, international centers by APS physicians from varying disciplines, representative of the diversity of aPL-related manifestations thanks to a balanced design, we were able to capture a wide spectrum of heterogeneous laboratory and clinical manifestations to be studied in the next phase of APS classification criteria development. We attempted to minimize recall bias by asking APS physicians to include new patients or patients seen in the last 3 years.

Additionally, the outcome of APS "likelihood" using a Likert scale required subjective assessment by the individual treating APS physician; although they were asked to score each case for research purposes, it is possible that their scoring was influenced by the revised Sapporo classification criteria or related to diagnosis rather than classification. Although it is a strength to be able to evaluate the treating physician's assessment of each individual case, as the Likert score assigned to each case was provided only by the treating physician themselves (as opposed to consensus agreement based on chart review), we acknowledge potential subjective inconsistencies in how individual case collectors scored each case. Furthermore, although a threshold of +2 or above to define "highly likely" APS may be stringent, and the RRs may change with a lower threshold (eg, +1 or 0), given that our goal was to understand the specificity of individual candidate criteria in cases deemed to have higher APS likelihood, the Steering Committee agreed that cases with a threshold of +2 allowed for evaluation in cases where treating physicians had a relatively higher degree of confidence in the diagnosis of APS.

Although individual candidate criteria were not adjusted in multivariable regression models, our RR associations provided critical univariate data to the committee to consider in the context of other available literature or their experience. Additionally, we acknowledge the possibility of multiple comparisons given the large number of statistical associations tested within a single dataset; however, as we made the a priori decision to primarily evaluate the RR effect size and directionality to maximize clinical relevance, we did not account for *P* value spending. The resulting decisions informed the next task to iteratively assess the weight of each individual criterion within the context of other criteria, using the 1000Minds advanced computing software.

In summary, Phase III-C results were used to guide decision-making to develop and organize the structure of the draft classification system with eight additive and independent clinical and laboratory domains. These domains were refined and developed into a hierarchical and weighted classification criteria system during the subsequent Phase III-D 1000Minds™ MCDA.

ACKNOWLEDGMENTS

The APS Classification Criteria Steering Committee dedicates this work to the late Raymond P. Naden, MD. Dr Naden helped lead the development of numerous rheumatic disease classification criteria, including ours; without him, the work presented in this manuscript would not have been possible. The APS Classification Criteria Steering Committee also thanks our ACR/EULAR liaisons, Ronald van Vollenhoven, MD (Amsterdam University Medical Center, Amsterdam, the Netherlands), Ulf Muller-Ladner, MD (Immunologie, Kerckhoff-Klinik GmbH, Nauheim, Germany), and Amy Turner (ACR) for their help and guidance; Yasaman Ahmedzadeh, MD (Hospital for Special Surgery, New York, NY) for her contributions during phase I/II/III; Joann Vega, CRCC (Hospital for Special Surgery, New York, NY) and Cindy

Force (ACR) for their administrative assistance; and our patient representatives.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Barbhaiya confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Wilson WA, Gharavi AE, Koike T, et al. International consensus statement on preliminary classification criteria for definite antiphospholipid syndrome: report of an international workshop. Arthritis Rheum 1999;42(7):1309–1311.
2. Lockshin MD, Sammaritano LR, Schwartzman S. Validation of the Sapporo criteria for antiphospholipid syndrome. Arthritis Rheum 2000;43(2):440–443.
3. Miyakis S, Lockshin MD, Atsumi T, et al. International consensus statement on an update of the classification criteria for definite antiphospholipid syndrome (APS). J Thromb Haemost 2006;4(2):295–306.
4. Barbhaiya M, Zuily S, Naden R, et al; ACR/EULAR APS Classification Criteria Collaborators. The 2023 ACR/EULAR antiphospholipid syndrome classification criteria. Arthritis Rheumatol 2023;75(10):1687–1702.
5. Barbhaiya M, Zuily S, Naden R, et al; ACR/EULAR APS Classification Criteria Collaborators. 2023 ACR/EULAR antiphospholipid syndrome classification criteria. Ann Rheum Dis 2023;82(10):1258–1270.
6. Aringer M, Costenbader K, Daikh D, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Arthritis Rheumatol 2019;71(9):1400–1412.
7. Wallace ZS, Naden RP, Chari S, et al; Members of the ACR/EULAR IgG4-RD Classification Criteria Working Group. The 2019 American College of Rheumatology/European League Against Rheumatism classification criteria for IgG4-related disease. Ann Rheum Dis 2020;79(1):77–87.
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. Ann Rheum Dis 2010;69(9):1580–1588.
9. Barbhaiya M, Zuily S, Ahmadzadeh Y, et al; New APS Classification Criteria Collaborators. Development of a new international antiphospholipid syndrome classification criteria phase I/II report: generation and reduction of candidate criteria. Arthritis Care Res (Hoboken) 2021;73(10):1490–1501.
10. Barbhaiya M, Abreu MM, Amigo MC, et al. Needs-assessment survey for the update of the current antiphospholipid syndrome (APS) classifcation criteria. Ann Rheum Dis 2015;74:1087.
11. Barbhaiya M, Zuily S, Amigo MC, et al; ACR/EULAR APS Classification Criteria Collaborators. Development of the 2023 ACR/EULAR antiphospholipid syndrome classification criteria, phase III-D report: multicriteria decision analysis. Arthritis Care Res (Hoboken) 2024;76(12):1617–1625.
12. Singh JA, Solomon DH, Dougados M, et al; Classification and Response Criteria Subcommittee of the Committee on Quality Measures, American College of Rheumatology. Development of classification and response criteria for rheumatic diseases. Arthritis Rheum 2006;55(3):348–352.
13. Barbhaiya M, Zuily S, Amigo MC, et al; New APS Classification Criteria Collaborators. Development of 2023 ACR/EULAR antiphospholipid syndrome classification criteria, phase III-A/B report: defining and structuring the clinical and laboratory domains. Curr Rheumatol Rep. 2025;27(1):41.
14. Ortel TL, Neumann I, Ageno W, et al. American Society of Hematology 2020 guidelines for management of venous thromboembolism: treatment of deep vein thrombosis and pulmonary embolism. Blood Adv 2020;4(19):4693–4738.
15. Piepoli MF, Hoes AW, Agewall S, et al; ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 2016;37(29):2315–2381.
16. Mach F, Baigent C, Catapano AL, et al; ESC Scientific Document Group. 2019 ESC/EAS guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J 2020;41(1):111–188.
17. de Jesús GR, Benson AE, Chighizola CB, et al. 16th International Congress on Antiphospholipid Antibodies Task Force report on obstetric antiphospholipid syndrome. Lupus 2020;29(12):1601–1615.
18. Barbhaiya M, Zuily S, Ahmadzadeh Y, et al. Development of new international antiphospholipid syndrome classification criteria phase I/II report: generation and reduction of candidate criteria. Arthritis Care Res (Hoboken) 2021;73(10):1490–1501.
19. Barbhaiya M, Taghavi M, Zuily S, et al; New APS Classification Criteria Steering Committee and APS ACTION Collaborators. Efforts to better characterize "antiphospholipid antibody nephropathy" for the 2023 ACR/EULAR antiphospholipid syndrome classification criteria: renal pathology subcommittee report. J Rheumatol 2024;51(2):150–159.
20. Abreu MM, Danowski A, Wahl DG, et al. The relevance of "non-criteria" clinical manifestations of antiphospholipid syndrome: 14th International Congress on Antiphospholipid Antibodies Technical Task Force Report on Antiphospholipid Syndrome Clinical Features. Autoimmun Rev 2015;14(5):401–414.
21. Reynaud Q, Lega JC, Mismetti P, et al. Risk of venous and arterial thrombosis according to type of antiphospholipid antibodies in adults without systemic lupus erythematosus: a systematic review and meta-analysis. Autoimmun Rev 2014;13(6):595–608.
22. Kelchtermans H, Pelkmans L, de Laat B, et al. IgG/IgM antiphospholipid antibodies present in the classification criteria for the antiphospholipid syndrome: a critical review of their association with thrombosis. J Thromb Haemost 2016;14(8):1530–1548.

APPENDIX A: ACR/EULAR APS CLASSIFICATION CRITERIA COLLABORATORS

(1) Steering Committee – Core Planning Group: Medha Barbhaiya, Karen H. Costenbader, Doruk Erkan, Francis Guillemin, Alison Hendry*, Ray Naden, and Stephane Zuily. (*joined during phase III).
(2) Steering Committee – Additional Members: Mary-Carmen Amigo, Tadej Avcin, Maria L. Bertolaccini, D. Ware Branch, Guilherme Ramires de Jesús, Katrien M. J. Devreese, Camille Frances,

David Garcia, Steven R. Levine, Roger A. Levy, Michael D. Lockshin, Thomas L. Ortel, Savino Sciascia, Maria Tektonidou, Denis Wahl, and Rohan Willis.

(3) Phase III Case Collector Collaborators: Danieli Andrade, Nathalie Costedoat-Chalumeau, Jose A. Gomez Puerta, Giovanni Sanna, Savino Sciascia, and Cecile Yelnik.

(4) Additional Phase I/II/III Collaborators (not included in the authorship list): Nancy Agmon-Levin (Sheba Medical Center, Tel Aviv, Israel), Cassyanne Aguilar (Children's Hospital of The King's Daughters, Norfolk, Virginia, USA), Paula Alba (National University of Cordoba, Cordoba, Argentina), Oral Alpan (O&O Alpan, LLC, Tysons Corner, Virginia, USA), Ales Ambrozic (Ljubljana University Medical Center, Ljubljana, Slovenia), Luis Andrade (Universidade Federal de São Paulo, São Paulo, Brazil), Simone Appenzeller (University of Campinas, São Paulo, Brazil), Yackov Berkun (Hadassah Medical Center, Jerusalem, Israel), Antonio Cabral (University of Ottawa, Ottawa, Ontario, Canada), Guillame Canaud (Université de Paris, Paris, France), Pojen Chen (University of California San Diego, San Diego, California, USA), Cecilia Chighizola (University of Milan, Milan, Italy), Rolando Cimaz (University of Florence, Florence, Italy), Maria Jose Cuadrado (University Hospital Navarra Madrid, Madrid, Spain), Philip G. de Groot (University Medical Centre, Utrecht, the Netherlands), Philippe de Moerloose (University Hospital, Geneva, Switzerland), Ronald Derksen (University Hospital, Utrecht, the Netherlands), Thomas Dörner (Charité University, Berlin, Germany), Paul Fortin (Centre Hospitalier Universitaire de Quebec, Quebec, Canada), Bill Giannakopoulos (University of New South Wales, Sydney, Australia), Emilio B Gonzalez (University of Texas Medical Branch, Galveston, Texas, USA), Murat Inanc (Istanbul University, Istanbul, Turkey), Gili Kenet (Sheba Medical Center, Tel Aviv, Israel), Munther Khamashta (St. Thomas Hospital, London, UK), Martin Kriegel (Yale School of Medicine, New Haven, Connecticut, USA), Steven Krilis (University of New South Wales, Sydney, Australia), Danyal Ladha (University of Ottawa, Faculty of Medicine, Ottawa, Ontario, Canada), Florian Manneville (Université de Lorraine, Nancy, France), Patti Massicotte (University of Alberta, Alberta, Canada), Gale McCarty (Santa Monica, California, USA), Jamal Mikdashi (University of Maryland, College Park, Maryland, USA), Barry Myones (Texas Children's Hospital, Houston, Texas, USA), Lisa Sammaritano (Hospital for Special Surgery, New York, New York, USA), Flavio Signorelli (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), Arzu Soybilgic (University of Illinois, Chicago, Illinois, USA), Scott Woller (University of Utah, Salt Lake City, Utah, USA), and Ray Zuo (University of Michigan, Ann Arbor, Michigan, USA).

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 582–590
DOI 10.1002/acr.25658

American College of Rheumatology
Empowering Rheumatology Professionals

Assessment of Skin in Patients With Systemic Sclerosis Using High-Frequency Ultrasound and Shear Wave Elastography: A Comparative Study With Histology, Molecular, and Clinical Parameters

Ruhani Desai,[1] Filemon Tan,[2] Minghua Wu,[2] Jeffery L. Browning,[3] Samuel Theodore,[2] Meng Zhang,[2] Brian Skaug,[2] Harshdeep Singh Chawla,[4] Manmohan Singh,[4] Salavat Aglyamov,[4] Kirill V. Larin,[4] Maureen Mayes,[2] and Shervin Assassi[2]

Objective. Ultrasound (US) has been proposed as a potential tool for assessing skin fibrosis in systemic sclerosis (SSc). However, a large-scale comparison of US-based assessment with histologic markers of skin fibrosis has not been reported. We evaluated US-based skin assessments for their face validity (differentiation between involved SSc and healthy control [HC] skin), construct validity (comparison to modified Rodnan skin score [mRSS]), and criterion validity (comparison to histologic and gene expression fibrosis markers).

Methods. Twenty HCs and 52 patients with SSc underwent clinical and US assessment followed by a forearm skin biopsy. Predefined areas were assessed on the finger, hand, and forearm bilaterally. mRSS, US, histologic and molecular (reverse transcription quantitative polymerase chain reaction) evaluations were performed by blinded, independent assessors. Dermal thickness, echogenicity, and elastography were assessed using a high-frequency GE LOGIQ P9 US machine.

Results. Except in the hand area, the US variables could not differentiate between HC and clinically affected SSc skin (face validity). There was only a weak to moderate correlation between US-based measurements and mRSS in the hand and finger areas (construct validity). US-based thickness showed moderate correlation with histologic thickness ($\rho = 0.43$; $P = 0.002$) but no statistically significant correlations with other histologic or gene expression markers of fibrosis in patients with SSc (criterion validity).

Conclusion. High-frequency US assessment of SSc skin does not show consistent and strong face, construct, or criterion validity. The role of this assessment tool remains limited, underscoring the need for further development of a quantitative and accurate tool for assessing SSc skin fibrosis.

INTRODUCTION

Systemic sclerosis (SSc; scleroderma) is an autoimmune disease characterized by fibrosis and vascular abnormalities affecting multiple organs, with skin involvement being the hallmark of the disease.[1] Skin involvement undergoes sequential pathologic phases, starting with an edematous phase that coincides with inflammation in the dermis, followed by a fibrotic phase, also known as an indurative phase.[2,3] There is progressive skin thickening and tightening during this phase, which generally lasts 1 to 4 years, compared with months for the inflammatory phase. The final stage is the atrophic phase in which patients notice some improvement in skin tightening. Generally, there is no progressive worsening of the skin following this phase.[4] Many studies have shown the prognostic role of skin involvement, with rate of progression and extent of skin

Supported by the National Institutes of Health (NIH; grants R33-AR-078078, R61-AR-078078, and R01-AR-081280). Dr Skaug's work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH (grant K08-AR-081402) and by an Investigator Award from the Rheumatology Research Foundation.

[1]Macon & Joan Brock Virginia Health Sciences at Old Dominion University Eastern Virginia Medical School, Norfolk, Virginia; [2]The University of Texas Health Science Center at Houston John P. and Katherine G. McGovern Medical School, Houston; [3]Boston University, Boston, Massachusetts; [4]University of Houston, Houston, Texas.

Additional supplementary information cited in this article can be found online in the Supporting Information section (http://onlinelibrary.wiley.com/doi/10.1002/acr.25658).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25658.

Address correspondence via email to Shervin Assassi, MD, MS, at shervin.assassi@uth.tmc.edu.

Submitted for publication February 22, 2025; accepted in revised form September 16, 2025.

SIGNIFICANCE & INNOVATIONS

- Ultrasound-based methods have been suggested as a feasible tool for assessing systemic sclerosis skin involvement, but there were no prior large-scale studies comparing the validity of these methods with skin histology.
- In the present study, we assessed the performance of ultrasound-based skin assessment in a large multiethnic sample of patients with systemic sclerosis and healthy controls.
- Ultrasound-based methods of skin assessment showed only limited criterion validity. Specifically, ultrasound-based dermal thickness showed moderate correlation with the histologic dermal thickness but not with collagen density and fibrosis gene expression parameters.

involvement being a poor prognostic marker for morbidity and mortality.[5–7]

It is imperative to have a reliable and accurate method for assessment of skin involvement at presentation and follow-up visits, allowing an accurate evaluation of the progression and response to treatment. This may also serve as a surrogate marker for disease morbidity and mortality. The modified Rodnan skin score (mRSS) is the most commonly used and the only validated approach per Outcome Measures in Rheumatology (OMERACT) principles for assessing skin involvement in SSc.[8] The mRSS is obtained by assessing 17 sites by palpation. The skin is graded based on its thickness/mobility between 0 and 3, in which 0 is normal skin and 3 is very thick and immobile skin[3]. Because of the inaccuracies in the mRSS, there is significant interobserver variability.[9] Other limitations include the need for significant training and difficulty in differentiating thick from thin but tethered down skin in later stages of the disease (ie, the atrophic stage). This can limit its utility in multicenter clinical trials and ultimately lead to negative trial outcomes, contributing to the fact that there are no United States Food and Drug Administration–approved disease-modifying agents for the treatment of SSc skin involvement to date.[10,11]

Ultrasound (US) is another modality for the assessment of skin involvement, which has gained much interest in recent years. It is postulated that it can objectively measure various properties of skin such as thickness, stiffness, and echogenicity (edema/water content). Use of high-frequency (>18 MHz) US (HFUS) has provided improved resolution for assessment of skin in patients with SSc.[12,13] Moreover, US elastography has also evolved from using manual compression to provide stimulus to using shear wave technology to circumvent the measurement bias from procedural variation of manual compression.[14,15] Most previous studies have focused on differentiating SSc skin from the skin of healthy controls (face validity) or assessing correlation with local and overall mRSS (construct validity).[16] Only two studies have investigated the correlation between US measurements and skin histologic assessment (criterion validity), but both studies investigated a small number of patients with SSc. Specifically, Flower et al investigated 10 SSc skin biopsies, and Chen et al investigated 13 SSc skin samples.[17,18] Our study aims to validate the use of HFUS in the assessment of SSc skin involvement by comparing US parameters of thickness, echogenicity, and elastography with clinical (disease status, as well as local and global mRSS), histologic parameters (histologic thickness and collagen content), and fibrosis gene expression markers in a large and diverse sample of patients and matched healthy controls.

MATERIALS AND METHODS

Study participants. Participants were enrolled at The University of Texas Health Science Center Houston Rheumatology Clinic. All patients fulfilled the 2013 American College of Rheumatology/EULAR classification criteria for SSc.[19] We aimed to enroll patients from the following four disease categories to reflect the different stages and types of skin involvement in SSc: early diffuse SSc (disease duration <3 years), early limited SSc (disease duration <3 years), late-stage diffuse (disease duration >3 years), and late-stage limited SSc (disease duration >3 years). Disease duration was defined as the time since the first non-Raynaud symptom of SSc. Healthy controls did not have an autoimmune skin disease, had similar demographic background, and were not first-degree relatives of patients with SSc. They were recruited via flyers in the clinic and medical school, and most were friends or spouses of patients with SSc.

The study was conducted with the ethics approval of the institutional review board. All study procedures were in concordance with the Declaration of Helsinki. Confidentiality was maintained and written informed consent was obtained before study enrollment.

Clinical assessment. All participants underwent the same assessment in three domains: clinical, followed by US assessment, and forearm skin biopsy for histologic and fibrosis gene expression evaluation. The assessors for the clinical, US, histologic, and fibrosis gene expression evaluations were blinded to assessment results in the other domains.

To standardize the assessment, six skin sites were identified and marked. These sites were selected using previous studies and the US skin assessment guidelines.[20] They were limited to the upper extremity because the primary focus was to examine the criterion validity of US assessment using the forearm skin biopsy. The first site was the proximal phalanx of the middle finger, the next site was the dorsum of the hand (midway between the third metacarpophalangeal joint and wrist joint), and the last was the dorsum of the forearm (12 cm from the midpoint of the radial styloid and ulnar head). This region of interest (ROI) was the same for US assessment and biopsy.

The clinical assessment consisted of local (performed in the same area as the US assessment) and total mRSS, which was done by an independent experienced scleroderma expert (S. Assassi). Basic demographic data were collected by the research coordinator. In addition, patients also filled out a health assessment questionnaire (Health Assessment Questionnaire Disability Index).

US assessment. US assessment was performed using GE LOGIQ P9 US system with a shear wave elastography package. Two probes were used: L10-22 at 22 MHz in B-mode for dermal thickness and echogenicity measurements and ML 6-15 at 15 MHz for shear wave elastography. Measurements were taken from the marked ROI. A gel layer was maintained to minimize the compression of skin tissue and optimize the excitation focus in elastography measurement. The protocol for obtaining and processing images was formulated using the US skin assessment guidelines.[20] The assessment was done by a single assessor who received more than a year of training in musculoskeletal US on the same machine. The assessor also received additional training on phantom models of varying stiffness for elastography assessment.

Dermal thickness. Dermal thickness was measured in millimeters from the undersurface of the epidermis to the dermis-subcutis junction (Supplemental Figure S1). This demarcation was sometimes ambiguous in the finger and the dorsum of the hand, in which case the next best landmark, the extensor tendon, was used to define the lower limit of dermis. Three images were captured at each site, and an average of the three was used for the statistical analysis.

Echogenicity. Echogenicity was measured in decibels. Circular areas were demarcated in the middle of the dermis. Three such circular areas of 0.1 mm^2 (± 0.01 mm^2) were selected on a single image, and the average of these three measurements was used for the statistical analysis. Lower values represent tissue edema and higher values represent fibrosis.

Elastography. Similar to the echogenicity measurements, elastography measurements were performed by demarcating a circle of 0.1-mm^2 area. The captured image also had a visual scale that allowed for quality assessment. In the finger and hand, three such areas were selected and averaged to obtain the elastic modulus (Young's modulus) of the tissue (kilopascals). In the forearm, the dermal layer allowed for the measurement of three areas in the upper dermis and three areas in the lower dermis, and an average of these values was used as the elasticity of the tissue.

Histologic assessment. Two punch biopsies were obtained from all participants at the marked ROI on one of the forearms. One skin sample was immediately immersed in formalin and subsequently paraffin-embedded for histologic evaluation, and the other one was immersed in RNAlater and subsequently stored at -80°C.

For histologic evaluation, measurement of dermal thickness was performed based on the hematoxylin and eosin (H&E) staining. Specifically, the dermal thickness was measured in H&E-stained images (4× magnification) using NIS software (Nikon). The maximum straight distance (or perpendicular distance) from the dermis-epidermis junction to the dermis-adipose tissue layer junction was measured and recorded. Moreover, an automated assessment of collagen density on Masson trichrome was performed using ImageJ software (National Institutes of Health). Briefly, formalin-fixed, paraffin-embedded tissue sections were stained for collagen with Masson trichrome stain (Sigma) and quantified using ImageJ to calculate a mean intensity (grayscale, 0–255) of blue color (collagen stain) in the dermis. Lastly, a semi-quantitative assessment of skin fibrosis was completed by a SSc skin pathology expert (J.L. Browning) ranging from 0 (normal) to 3 (severely fibrotic), as previously described.[20]

For the measurement of fibrosis gene expression markers, RNA was extracted from the skin sample stored in RNAlater using miRNeasy Mini kits (Qiagen). Reverse transcription was performed with 200 ng of RNA from each sample using the High-Capacity Complementary DNA (cDNA) Reverse Transcription kit (Applied Biosystems, #4368813). Quantitative polymerase chain reaction (qPCR) was performed in the QuantStudio 6 thermocycler (Applied Biosystems) using TaqMan Gene Expression Master Mix (Applied Biosystems #4369016) with the following five TaqMan probes: 18s (Hs99999901_s1), GAPDH (Hs02758991_g1), COL1A1 (Hs00164004_m1), COL8A2 (Hs07287101_m1), and COMP (Hs00164359_m1). Samples with no reverse transcriptase during the cDNA synthesis reaction and samples with no cDNA template during the qPCR reaction were included as negative controls. Technical duplicates were included for each probe for each skin sample. Samples in which there was an SD of >0.3 in the cycle threshold (Ct) values for the technical duplicates were repeated. Transcript levels for COL1A1, COL8A2, and COMP were determined using $2^{-\Delta Ct}$ calculation, with GAPDH as a reference. COL1A1 was selected because it is the main collagen type in the skin. COL8A1 and COMP are fibrotic transcripts that showed a correlation with mRSS in our previous global gene expression studies.[21,22] The correlation of the US assessments with COL1A1 transcript levels as well as a composite score of COL1A1, COL8A1, and COMP transcript levels (termed the fibrosis gene expression score) were determined.

Statistical assessment. Statistical analysis was conducted using R (version 4.2.2). All study participants who underwent US assessment were included in the overall analysis. Descriptive statistics were performed to analyze the demographic characteristics of all participants. The Wilcoxon rank sum test was used to investigate the difference in US assessments between

Table 1. Participant demographic and clinical characteristics*

Characteristics	SSc (n = 52)	Healthy controls (n = 20)
Age, median (IQR), years	51 (17.3)	46.5 (33.8)
Female, n (%)	47 (90.4)	15 (75)
Race, n (%)		
White	43 (82.7)	17 (85)
African American	6 (11.5)	1 (5)
Asian	3 (5.8)	2 (10)
Ethnicity, n (%)		
Hispanic	17 (32.7%)	6 (30%)
SSc subtype, n (%)		
Diffuse	37 (71.2)	–
mRSS, median (IQR)	13.5 (9.2)	–
Disease duration, median (IQR) years	5.41 (6.6)	–
HAQ-DI, median (IQR)	0.8 (1.54)	–
Immunosuppressive treatment, n (%)	40 (76.9)	–
Antibody profile, n (%)		
Anticentromere	9 (17)	–
Anti-Scl70	18 (35)	–
Anti-RNA Pol III	10 (19)	–
Anti-RNP	6 (12)	–

* HAQ-DI, Health Assessment Questionnaire Disability Index; IQR, interquartile range; mRSS, modified Rodnan skin score; Pol, polymerase; SSc, systemic sclerosis.

patients with clinically affected skin (clinical local skin score >0) and healthy controls. The correlations of US assessment with clinical skin score, histologic parameters, and fibrosis gene expression markers were assessed using Spearman correlation. Only patients with SSc were included in these correlation analyses. Given the increase in Type II error with small sample sizes, further analysis based on duration of disease, type of disease, and antibody profile was not pursued.

RESULTS

Fifty-two patients with SSc (11 early diffuse, 26 late diffuse, 6 early limited, and 9 late limited) and 20 healthy controls were assessed. The demographic background of patients with SSc and controls is summarized in Table 1. The demographic characteristics of patients with SSc were consistent with epidemiologic trends, showing a higher proportion of female participants. Patients with SSc had a median mRSS of 13.5 and a median disease duration of 5.41 years. Healthy controls were of a similar demographic background.

Face validity: comparison of US parameters between patients with SSc and healthy controls. Next, the difference in US assessments between patients with SSc and healthy controls was assessed. For this purpose, we compared the measurements in healthy controls with those obtained from patients with SSc who had clinically affected skin at the site of examination (ie, local skin score > 0). As shown in Figure 1 and Table 2, the only significant differences observed in patients with SSc were increased US-based skin thickness in the right hand and higher elastography-based skin stiffness in both hands.

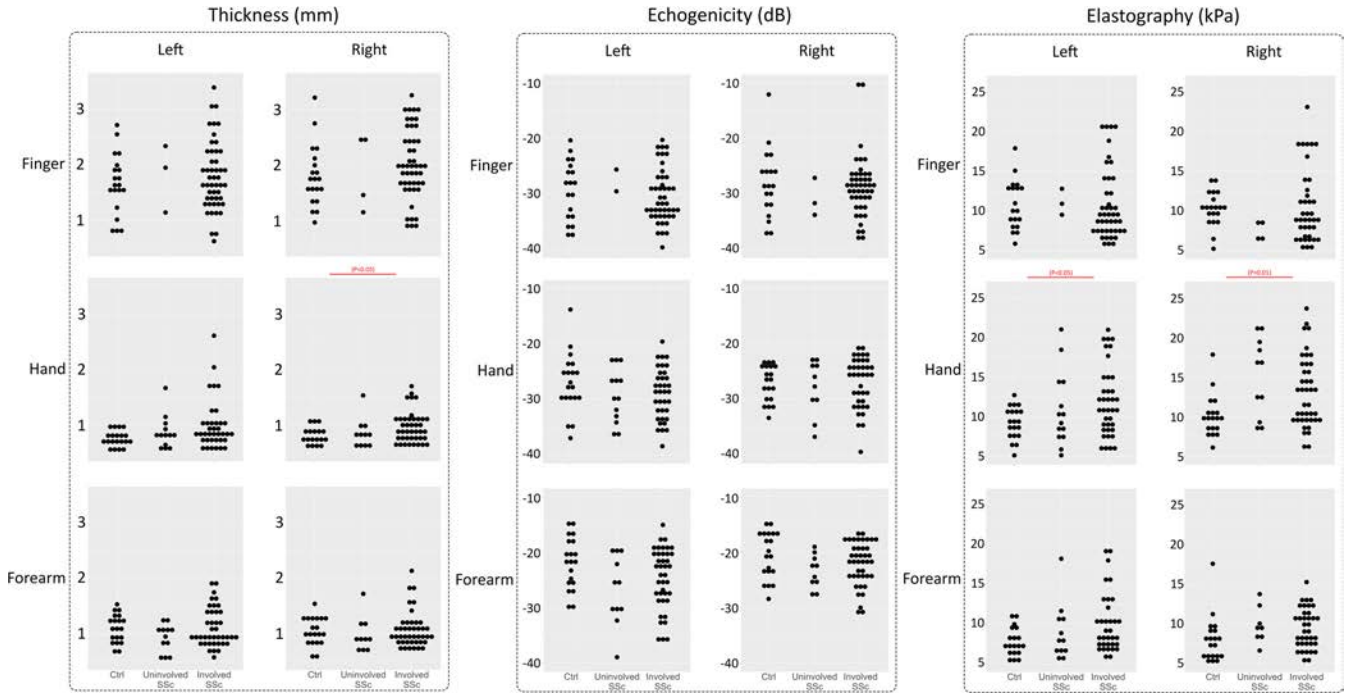

Figure 1. Ultrasound-based measurements in patients with SSc and Ctrl. Ultrasound-based measurement of dermal thickness, echogenicity, and elastography are shown in Ctrl skin, unaffected SSc skin (local skin score = 0), and affected SSc skin (local skin score ≥1) for left and right finger, hand, and forearm body areas. Ctrl, healthy controls; SSc, systemic sclerosis.

Table 2. High-frequency ultrasound measurements at each site and *P* values comparing high-frequency ultrasound measurements of SSc-involved skin (local skin score > 0) with healthy control skin*

	Side	SSc, median (IQR); n	Control, median (IQR); n	*P* value
Thickness, mm				
Finger	Right	1.95 (0.81); 48	1.70 (0.54); 20	0.138
	Left	1.68 (0.84); 49	1.63 (0.53); 20	0.569
Hand	Right	0.87 (0.36); 43	0.75 (0.23); 20	**0.015**
	Left	0.82 (0.36); 38	0.74 (0.16); 20	0.052
Forearm	Right	0.95 (0.28); 42	1.04 (0.38); 20	0.598
	Left	0.95 (0.50); 40	1.11 (0.38); 20	0.943
Echogenicity, dB				
Finger	Right	−29.0 (5.14); 48	−28.6 (6.95); 20	0.57
	Left	−32.1 (6.87); 49	−28.9 (8.87); 20	0.661
Hand	Right	−25.1 (7.63); 43	−26.4 (5.98); 20	0.143
	Left	−29.5 (7.32); 38	−27.2 (5.35); 20	0.606
Forearm	Right	−21.4 (5.94); 42	−20.1 (6.74); 20	0.098
	Left	−23.5 (7.35); 40	−21.5 (7.11); 20	0.943
Elastography, kPa				
Finger	Right	8.64 (5.91); 48	10.3 (2.39); 20	0.337
	Left	8.98 (4.88); 49	10.0 (4.85); 20	0.101
Hand	Right	13.1 (6.97); 43	9.70 (2.35); 20	**0.022**
	Left	11.1 (5.81); 38	9.31 (3.09); 20	**0.005**
Forearm	Right	8.38 (4.50); 42	7.71 (3.06); 20	0.212
	Left	8.20 (3.62); 40	7.42 (2.43); 20	0.152

* Bold font indicates significant *P* values ($P \leq 0.05$). Comparisons are made using the Wilcoxon rank sum test. IQR, interquartile range; SSc, systemic sclerosis.

Construct validity: correlation of US parameters with local and total mRSS. Next, we examined the correlation of US parameters with local skin score and the overall mRSS to assess construct validity (Table 3). The US-based dermal thickness showed only weak statistically significant positive correlations with the local skin score in the right and left forearms (ρ = 0.28 and 0.29, respectively), which was not seen with the total mRSS. There was also a weak correlation with the overall mRSS in both hands (ρ = 0.32 and 0.33, respectively). There was no significant correlation in fingers.

Table 3. Construct validity: correlation of high-frequency ultrasound measurements with local and total mRSS in patients with systemic sclerosis (n = 52)*

		Local mRSS			Total mRSS		
	Side	Spearman ρ	95% CI	*P* value	Spearman ρ	95% CI	*P* value
Thickness							
Finger	Right	0.16	−0.11 to 0.42	0.246	0.12	−0.16 to 0.38	0.405
	Left	0.084	−0.19 to 0.35	0.552	0.27	−0.003 to 0.51	0.057
Hand	Right	0.27	−0.0003 to 0.51	0.051	**0.32**	**0.06 to 0.55**	**0.019**
	Left	0.25	−0.02 to 0.49	0.069	**0.33**	**0.07 to 0.56**	**0.015**
Forearm	Right	**0.28**	**0.003 to 0.51**	**0.048**	0.20	−0.07 to 0.45	0.145
	Left	**0.29**	**0.02 to 0.52**	**0.036**	0.27	−0.002 to 0.51	0.052
Echogenicity							
Finger	Right	0.002	−0.27 to 0.25	0.991	−0.06	−0.33 to 0.21	0.65
	Left	**−0.42**	**−0.62 to -0.17**	**0.002**	**−0.44**	**−0.64 to -0.19**	**0.001**
Hand	Right	−0.02	−0.29 to 0.25	0.897	0.07	−0.21 to 0.34	0.627
	Left	−0.16	−0.41 to 0.12	0.271	−0.06	−0.33 to 0.22	0.675
Forearm	Right	−0.04	−0.31 to 0.24	0.788	−0.04	−0.31 to 0.24	0.786
	Left	0.07	−0.21 to 0.33	0.633	0.14	−0.14 to 0.40	0.324
Elastography							
Finger	Right	0.14	−0.13 to 0.40	0.308	0.00	−0.27 to 0.27	0.989
	Left	−0.06	−0.32 to 0.22	0.692	−0.04	−0.31 to 0.24	0.804
Hand	Right	−0.04	−0.31 to 0.24	0.789	−0.17	−0.42 to 0.11	0.231
	Left	**0.28**	**0.01 to 0.52**	**0.043**	**0.32**	**0.05 to 0.54**	**0.021**
Forearm	Right	−0.12	−0.38 to 0.15	0.379	−0.17	−0.42 to 0.11	0.231
	Left	0.07	−0.21 to 0.34	0.620	0.16	−0.12 to 0.41	0.256

* Bold values are significant ($P \leq 0.05$). CI, confidence interval; mRSS, modified Rodnan Skin Score.

Table 4. Criterion validity: correlation of high-frequency ultrasound parameters with histologic and gene expression fibrosis markers in patients with systemic sclerosis*

Histologic parameter	Ultrasound parameter	Spearman ρ	95% CI	*P* value
Histologic dermal thickness (n = 50)	Thickness	**0.43**	**0.18 to 0.63**	**0.002**
	Echogenicity	**0.34**	**0.06 to 0.56**	**0.017**
	Elastography	0.04	−0.24 to 0.31	0.798
Collagen density (n = 50)	Thickness	0.25	−0.03 to 0.49	0.083
	Echogenicity	0.14	−0.15 to 0.40	0.343
	Elastography	−0.04	−0.31 to 0.24	0.802
Semiquantitative skin fibrosis assessment (n = 50)	Thickness	0.23	−0.05 to 0.48	0.109
	Echogenicity	0.23	−0.05 to 0.48	0.108
	Elastography	−0.07	−0.34 to 0.22	0.645
qPCR score (n = 52)	Thickness	0.10	−0.18 to 0.36	0.477
	Echogenicity	−0.12	−0.38 to 0.16	0.385
	Elastography	0.13	−0.15 to 0.39	0.358
Col1A1 expression (n = 52)	Thickness	0.21	−0.07 to 0.45	0.142
	Echogenicity	0.13	−0.15 to 0.39	0.371
	Elastography	0.17	−0.11 to 0.42	0.239

* Bold values are significant ($P \leq 0.05$). The systemic sclerosis group contains patients with locally affected (modified Rodnan skin score > 0) and unaffected skin (modified Rodnan skin score = 0). CI, confidence interval; qPCR, quantitative polymerase chain reaction.

The echogenicity showed a negative statistically significant correlation in the left finger with the local skin score and the overall skin score (ρ = –0.41 and –0.42, respectively). Similarly, there was only a statistically significant but weak correlation between elastography and local skin score as well as overall skin score in the left hand (ρ = 0.28 and 0.32, respectively).

Criterion validity: correlations of US parameters with histologic and molecular parameters. We also examined the correlation between the US parameters with the histologic and fibrosis gene expression markers at the biopsy site (Table 4; Supplemental Table S1). There was a moderate correlation between US-based dermal thickness and histologic dermal thickness (ρ = 0.43; $P = 0.002$) and echogenicity and histologic dermal thickness (ρ = 0.34; $P = 0.017$). There were no statistically significant associations with histologic collagen density, semiquantitative skin fibrosis assessment, and the skin fibrosis gene expression markers.

Interrelationship among the US parameters. There was no correlation between echogenicity and the other two US parameters of thickness and elastography (Table 5). A weak correlation was found between US-based dermal thickness and elastography in the left hand (ρ = 0.30; $P = 0.032$).

DISCUSSION

In the present study, we conducted a comparison between US-based skin measurements with mRSS as well as histologic/molecular skin biopsy parameters in a multiethnic patient population with SSc, encompassing early and late stages and both disease categories (limited and diffuse), using a commercially available device. In our study, US-based dermal thickness

Table 5. Correlation among high-frequency ultrasound parameters*

	Right		Left	
	Spearman ρ	*P* value	Spearman ρ	*P* value
Thickness and echogenicity				
Finger	−0.05	0.732	−0.14	0.338
Hand	−0.13	0.350	−0.21	0.142
Forearm	0.09	0.518	0.22	0.113
Thickness and elastography				
Finger	0.13	0.375	−0.15	0.303
Hand	0.25	0.076	**0.30**	**0.032**
Forearm	0.00	0.980	0.20	0.147
Echogenicity and elastography				
Finger	−0.13	0.355	−0.21	0.134
Hand	−0.07	0.631	−0.09	0.541
Forearm	0.10	0.479	0.02	0.912

* Bold values are significant ($P \leq 0.05$).

showed moderate correlation with histologic dermal thickness but not with other histologic and gene expression parameters. Moreover, US-based parameters could not differentiate between affected SSc and healthy control skin in most examined anatomic areas; hence, they did not show consistent face validity. Similarly, US-based parameters did not show strong correlations with the local or overall skin score.

As expected, there was a statistically significant correlation between US-based dermal thickness and the same measurement on histology (ρ = 0.43; $P = 0.002$). However, the US-based dermal thickness did not correlate significantly with the other histologic and gene expression measurements of skin fibrosis. Moreover, there was no significant correlation between US-based dermal thickness and local clinical skin score in the forearm area. Clinicians perform skin score assessment by rolling the skin between two fingers and assessing the wrinkling of the skin in the folded skin area. This assessment is likely to be influenced by collagen density and is not only a function of dermal thickness.[23] This might explain the lack of a significant correlation between US-based skin thickness and local skin scoring in the forearm area. Similar to a previous study,[24] US-dermal thickness was highly varied even in healthy controls in the present study (Figure 1). This variability might explain why US-based skin thickness could not differentiate between SSc-affected skin and healthy control skin in the present study.

There are many challenges in determining the normal range of skin thickness, as it shows marked variation based on age, sex, and sun exposure even in healthy individuals.[24,25] In our study, assessment of the dermal thickness including distinguishing between the epidermis-dermis (E/D) junction and dermis-subcutaneous junction at finger and dorsum of the hand was challenging. We used the next best landmark to define the lower limit of the dermis. Micu et al showed that, in addition to intrinsic factors contributing to this variability, there are additional extrinsic factors that influence the identification of the E/D interface in areas such as fingers and hands. Their study showed that the E/D interface tends to be nonlinear at the finger and hand, with more irregularity in the older population. Moreover, those with sun-exposed skin have a subepidermal echogenicity band. These factors make it challenging to determine the dermal thickness accurately. Moreover, besides the frequency of the probe, the position of the probe and joint flexion can influence measurement results. For example, transverse positioning of the probe and flexion of the finger joints might lead to better visualization of the E/D interface at the digits and dorsal hand area.[25]

Two previous studies have compared the US-based skin assessment with the histologic evaluation of SSc skin. Flower et al examined the correlation of US-based measurements with histologic assessment based on forearm punch biopsies in 10 White patients with SSc.[18] In this study, skin thickness and not dermal thickness was assessed. Similar to our study, US-based skin thickness showed a moderate correlation with histologic skin thickness (ρ = 0.48).[17] However, this study reported a higher correlation of US-based skin thickness and elastography with histologic collagen density (ρ in the 0.7 range) than our study. No significant correlations were observed between echogenicity and histologic measurements. A comparison of the study population characteristics cannot be performed, as this information was not provided for the subgroup of patients who underwent skin biopsy (n = 10). Chen et al examined the correlation of US-based skin thickness and elastography with histologic skin thickness in 13 Chinese patients with early SSc (9 with diffuse SSc and 4 with limited SSc).[17] Similar to our findings, there was no correlation between US-based elastography and histologic skin thickness, but the US-based skin thickness showed a higher correlation with histologic dermal thickness (ρ = 0.69) than our study and the study by Flower et al. A comparison with histologic collagen density was not performed. Some of the higher correlations in the two aforementioned studies might be because of their substantially smaller sample size.[17,18] Technical variability might be another source of inconsistent findings. For example, different commercially available US machines were used in all three studies. Moreover, contrary to the previous studies, our larger study population was multiethnic and included a wide range of disease durations.

In our study, the US-based echogenicity measurements did not correlate with US-based dermal thickness or elastography. Similarly, US-based dermal thickness showed only statistically significant positive correlation with elastography in the hand area. These findings are consistent with previous reports. Flower et al also reported no correlation between echogenicity and other US-based domains and found a statistically significant moderate correlation between US-based skin thickness and elastography in the hand area (r = 0.45).[18] These results indicate that the investigated US-based domains do not provide overlapping information and that dermal thickness cannot be equated with dermal stiffness in investigations of SSc skin involvement.

Although we used a high-frequency probe (22 MHz) for the assessment of dermal thickness and echogenicity, there are commercially available ultra–high-frequency probes that can provide a more detailed characterization of superficial body areas such as skin. In a recent study conducted in 47 patients with SSc, Di Battista et al examined the epidermal thickness as well as epidermal and dermal echogenicity using a commercially available 70-MHz probe.[26] In this study, dermal thickness could not be assessed because the boundaries of the dermal layer were indistinguishable from the underlying hypodermis with the probe we used. Interestingly, there was no significant correlation between the US-based measurements and the local skin score or total mRSS in this study, indicating that a higher US probe frequency alone is not sufficient to provide a higher construct validity in the assessment of SSc involvement.

Our study has several strengths. To our knowledge, the present study represents the largest comparison of US-based assessment with markers of SSc skin fibrosis on histology to date. Moreover, we examined a multiethnic study population,

encompassing patients with early/late and limited/diffuse by all three domains (US, clinical, and histology/gene expression) by three blinded, independent assessors. Lastly, the US assessment was performed based on previously published, standardized protocols by a trained clinician.[20]

Our study also has limitations. The laboratory personnel performing the histologic evaluation and the assessor performing the US assessment were not aware of the clinical diagnosis of the study participants. However, the US was conducted by a trained rheumatologist who is experienced in diagnosing SSc. Hence, it is likely that the US assessor became aware of the clinical diagnosis of the study participants during the US assessment. This should not have affected the validation studies within the patient group (ie, comparison with mRSS or histologic measures in patients with SSc). This is a cross-sectional study that cannot investigate the sensitivity to change of the US-based parameters over time. Similarly, because of the cross-sectional study design, it cannot investigate the US correlates of the different phases of SSc skin involvement (eg, edematous, fibrotic, and atrophic phases) within the same patient. Generally, studies examining the sensitivity to change of US-based parameters in comparison with longitudinal changes in mRSS or histology are limited. Future longitudinal studies are needed to address this knowledge gap. Another limitation of the study is that all US assessments were performed by one assessor. Lastly, this is a single-center study with one commonly used US machine (GE LOGIQ P9). Future studies with other commercially available devices are needed to investigate the generalizability of our study results. Although US is the most commonly studied method for assessment of SSc skin involvement, other assessment tools with higher resolution, such as optical coherence tomography or elastography, are also being studied. Further studies are needed to assess whether optical coherence–based methods alone or in combination with US can provide a reliable, valid, and quantifiable assessment of the SSc skin involvement.

In conclusion, we assessed the validity of US-based measurements of dermal thickness, echogenicity, and elastography for SSc skin involvement and severity in a diverse patient population. Although there was a moderate correlation between US-based and histologic dermal thickness, we did not observe strong correlations between any US-based parameters and clinical or histologic/gene expression markers of skin fibrosis. Moreover, US-based parameters did not show consistent face validity by differentiating involved skin in patients with SSc from healthy control skin. Hence, our study does not support the wide-scale use of HFUS in SSc skin involvement and supports the need for further development of objective and accurate tools for the assessment of SSc skin fibrosis.

ACKNOWLEDGMENTS

The authors would like to thank Julio Charles and Xinjian Guo for their assistance in sample processing and carrying out laboratory experiments.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Assassi confirms that all authors have provided the final approval of the version to be published, and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988; 15(2):202–205.
2. Herrick AL, Assassi S, Denton CP. Skin involvement in early diffuse cutaneous systemic sclerosis: an unmet clinical need. Nat Rev Rheumatol 2022;18(5):276–285.
3. Khanna D, Furst DE, Clements PJ, et al. Standardization of the modified Rodnan skin score for use in clinical trials of systemic sclerosis. J Scleroderma Relat Disord 2017;2(1):11–18.
4. Pokeerbux MR, Giovannelli J, Dauchet L, et al. Survival and prognosis factors in systemic sclerosis: data of a French multicenter cohort, systematic review, and meta-analysis of the literature. Arthritis Res Ther 2019;21(1):86.
5. Shand L, Lunt M, Nihtyanova S, et al. Relationship between change in skin score and disease outcome in diffuse cutaneous systemic sclerosis: application of a latent linear trajectory model. Arthritis Rheum 2007;56(7):2422–2431.
6. Steen VD, Medsger TA Jr. Improvement in skin thickening in systemic sclerosis associated with improved survival. Arthritis Rheum 2001; 44(12):2828–2835.
7. Wyss A, Jordan S, Graf N, et al. Does regression of skin thickening predict improvement of internal organ involvement and survival in patients with diffuse cutaneous systemic sclerosis? A EUSTAR analysis. Arthritis Res Ther 2024;26(1):187.
8. Merkel PA, Clements PJ, Reveille JD, et al; OMERACT 6. Current status of outcome measure development for clinical trials in systemic sclerosis. Report from OMERACT 6. J Rheumatol 2003;30(7):1630–1647.
9. Clements P, Lachenbruch P, Siebold J, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol 1995;22(7):1281–1285.
10. Hinchcliff M, Huang CC, Wood TA, et al. Molecular signatures in skin associated with clinical improvement during mycophenolate treatment in systemic sclerosis. J Invest Dermatol 2013;133(8):1979–1989.
11. Khanna D, Lin CJF, Furst DE, et al. Tocilizumab in systemic sclerosis: a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Respir Med 2020;8(10):963–974.
12. Hesselstrand R, Carlestam J, Wildt M, et al. High frequency ultrasound of skin involvement in systemic sclerosis - a follow-up study. Arthritis Res Ther 2015;17:329.
13. Li H, Furst DE, Jin H, et al. High-frequency ultrasound of the skin in systemic sclerosis: an exploratory study to examine correlation with disease activity and to define the minimally detectable difference. Arthritis Res Ther 2018;20(1):181.
14. Iagnocco A, Kaloudi O, Perella C, et al. Ultrasound elastography assessment of skin involvement in systemic sclerosis: lights and shadows. J Rheumatol 2010;37(8):1688–1691.

15. Hou Y, Zhu QL, Liu H, et al. A preliminary study of acoustic radiation force impulse quantification for the assessment of skin in diffuse cutaneous systemic sclerosis. J Rheumatol 2015;42(3):449–455.
16. Santiago T, Santos E, Ruaro B, et al. Ultrasound and elastography in the assessment of skin involvement in systemic sclerosis: a systematic literature review focusing on validation and standardization - WSF Skin Ultrasound Group. Semin Arthritis Rheum 2022;52: 151954.
17. Chen C, Cheng Y, Zhu X, et al. Ultrasound assessment of skin thickness and stiffness: the correlation with histology and clinical score in systemic sclerosis. Arthritis Res Ther 2020;22(1):197.
18. Flower VA, Barratt SL, Hart DJ, et al. High-frequency ultrasound assessment of systemic sclerosis skin involvement: intraobserver repeatability and relationship with clinician assessment and dermal collagen content. J Rheumatol 2021;48(6):867–876.
19. van den Hoogen F, Khanna D, Fransen J, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/-European League Against Rheumatism collaborative initiative. Arthritis Rheum 2013;65(11):2737–2747.
20. Santiago T, Santos EJF, Ruaro B, et al. Recommendations for the execution and reporting of skin ultrasound in systemic sclerosis: an international collaboration under the WSF skin ultrasound group. RMD Open 2022;8(2):e002371.
21. Skaug B, Khanna D, Swindell WR, et al. Global skin gene expression analysis of early diffuse cutaneous systemic sclerosis shows a prominent innate and adaptive inflammatory profile. Ann Rheum Dis 2020; 79(3):379–386.
22. Assassi S, Wu M, Tan FK, et al. Skin gene expression correlates of severity of interstitial lung disease in systemic sclerosis. Arthritis Rheum 2013;65(11):2917–2927.
23. Assassi S, Desai R, Browning J, et al. Criterion validity of modified Rodnan skin score: does it capture skin thickness or hardening? Results of a large skin histology study. Arthritis Rheumatol 2024;76(suppl 9). https://acrabstracts.org/abstract/criterion-validity-of-modified-rodnan-skin-score-does-it-capture-skin-thickness-or-hardening-results-of-a-large-skin-histology-study/
24. Santiago T, Santos EJF, Luis M, et al. Normal percentile reference curves for skin ultrasound thickness and stiffness at Rodnan sites. RMD Open 2022;8(2):e002577.
25. Micu MC, Wortsman X, Naredo E. Skin ultrasound for rheumatologists: technical issues and challenges. J Scleroderma Relat Disord 2024;23971983241289236.
26. Di Battista M, Barsotti S, Vitali S, et al. Multiparametric skin assessment in a monocentric cohort of systemic sclerosis patients: is there a role for ultra-high frequency ultrasound? Diagnostics (Basel) 2023; 13(8):1495.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 591–600
DOI 10.1002/acr.25676

American College of Rheumatology
Empowering Rheumatology Professionals

Exploring Factors Associated With Cognitive Impairment in Rheumatoid Arthritis

Raquelle Dawood,[1,2] Hannah F. Brubeck,[1] Kylie E. Riggles,[1] Sebastian E. Sattui,[3] Elena Myasoedova,[4] Una E. Makris,[5,6] Dolores M. Shoback,[7,8] Jose M. Garcia,[1,9,10] Ariela R. Orkaby,[11,12,13] Joshua F. Baker,[14,15] Patricia P. Katz,[8] and Katherine D. Wysham[1,9,10]

Objective. Cognitive impairment is prevalent in rheumatoid arthritis (RA), yet risk factors are not well understood. We explored associations between clinical characteristics and cognitive impairment in an RA cohort.

Methods. Data were from a longitudinal RA cohort at Veterans Affairs Puget Sound Health Care System. Cognition was evaluated using the Saint Louis University Mental Status (SLUMS) examination. Demographics and health factors, objectively measured physical function, participant-reported symptoms, RA disease characteristics, and comorbidities were evaluated. Univariable linear regressions explored the association between clinical factors and SLUMS scores. Those with $P < 0.1$ in the univariable models were evaluated in separate multivariable linear regressions controlling for age, sex, and years of education.

Results. A total of 145 participants with RA were included, with a mean age of 64.5 (SD 11.6) years, and were predominantly male (74%). Using SLUMS, 42 participants (29%) had normal cognition, 83 (57%) had mild cognitive impairment, and 20 (14%) had dementia. Physical performance (aβ: 0.35, confidence interval [CI] 0.11 to 0.59), self-reported exhaustion (aβ: –2.22, CI –3.44 to –0.99), pain (aβ: –0.25, CI –0.50 to –0.00), disability (aβ: –1.79, CI –3.16 to –0.42), and trouble falling asleep (aβ: –2.57, CI –4.00 to –1.14) were all independently associated with a lower SLUMS score (all $P < 0.05$).

Conclusion. Cognitive impairment was prevalent in our cohort of veterans with RA and was associated with several modifiable clinical factors. Future longitudinal studies are needed to determine the directionality of these associations and evaluate interventions for modifiable risk factors that may mitigate cognitive dysfunction.

INTRODUCTION

Rheumatoid arthritis (RA) is a systemic autoimmune disease involving chronic inflammation, which primarily manifests as inflammatory arthritis.[1] Mild cognitive impairment (MCI) is a clinical condition representing an intermediate stage between normal cognitive aging and dementia that is characterized by noticeable cognitive deficits; however, unlike dementia, it does not interfere significantly with daily functioning.[2] Cognitive impairment has become an area of interest in RA. The prevalence of cognitive impairment in RA varies across studies; however, the majority of findings suggest an earlier onset of, as well as increased

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Supported by the VA Clinical Science Research & Development (grant ICX002351 to Dr Wysham) and the VA Puget Sound Geriatrics Research Education and Clinical Center (GRECC). Dr Makris's work was supported in part by a grant from VA Health Services Research and Development (grants IIR 20-256 and HX003350).

[1]Raquelle Dawood, BS, Hannah F. Brubeck, BS, Kylie E. Riggles, BS, Jose M. Garcia, MD, PhD, Katherine D. Wysham, MD: VA Puget Sound Health Care System, Seattle, Washington; [2]Raquelle Dawood, BS: Pacific Northwest University of Health Sciences, Yakima, Washington; [3]Sebastian E. Sattui, MD, MS: University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania; [4]Elena Myasoedova, MD: Mayo Clinic College of Medicine and Science, Rochester, Minnesota; [5]Una E. Makris, MD, MSc: VA North Texas Health Care System, Dallas; [6]Una E. Makris, MD, MSc: University of Texas Southwestern School of Medicine, Dallas; [7]Dolores M. Shoback, MD: San Francisco VA Medical Center, San Francisco, California; [8]Dolores M. Shoback, MD, Patricia P. Katz, PhD: University of California San Francisco; [9]Jose M. Garcia, MD, PhD, Katherine D. Wysham, MD: University of Washington School of Medicine, Seattle; [10]Jose M. Garcia, MD, PhD, Katherine D. Wysham, MD: VA Puget Sound GRECC, Seattle, Washington; [11]Ariela R. Orkaby, MD, MPH: VA New England GRECC, Boston, Massachusetts; [12]Ariela R. Orkaby, MD, MPH: VA Boston Healthcare System, Boston, Massachusetts; [13]Ariela R. Orkaby, MD, MPH: Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; [14]Joshua F. Baker, MD: Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania; [15]Joshua F. Baker, MD: University of Pennsylvania, Philadelphia.

Drs Dawood and Brubeck are co-first authors and contributed equally to this work.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25676).

Author disclosures and graphical abstract are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25676.

Address correspondence via email to Katherine D. Wysham, MD, at katherine.wysham@va.gov.

Submitted for publication February 20, 2025; accepted in revised form September 29, 2025.

SIGNIFICANCE & INNOVATIONS

- Cognitive impairment was present in 71% of participants with rheumatoid arthritis (RA) in our cohort.
- Several modifiable factors were associated with cognitive performance, including physical function, disability, exhaustion, pain, and difficulty falling asleep.
- Given the high prevalence of cognitive impairment in RA and its negative impact on quality of life and health outcomes, rheumatology health care professionals should consider integrating cognitive impairment screening in the care of at-risk patients.

prevalence of and risk for, cognitive impairment in people with RA compared to the general population.[3–5] It is known that aging is a major contributor to cognitive decline, and as the aging population grows, the prevalence of cognitive impairment and dementia are expected to rise,[1,6] creating an urgent need to better understand and mitigate cognitive decline.

Risk factors for cognitive impairment in RA are understudied, particularly in older cohorts. In the general population, well-established risk factors include advancing age, comorbid conditions such as poorly controlled cardiovascular disease (CVD) and mood disorders, and impaired physical function.[1,3,7] Many of these risk factors are also common in individuals with RA and may be compounded by disease-specific factors such as inflammation, pain, and medications.[8] Importantly, systemic inflammation is associated with cognitive impairment in the general population and may be an important risk factor for cognitive impairment in RA given the high inflammatory burden of the disease.[1,8] Indeed, population-based studies have shown that RA is associated with a two-fold higher risk of dementia, even after adjusting for traditional CVD risk factors and comorbidities.[9] Depression, pain, and sleep are also prevalent in participants with RA and are known to impact cognitive performance in other populations.[10,11] Determining modifiable factors for cognitive impairment in people with RA may support interventions to improve cognitive function in this at-risk population.

We aimed to explore clinical characteristics associated with cognitive impairment, with a focus on modifiable factors, in a well-phenotyped cohort of people with RA using a validated tool of evaluating cognitive status, the Saint Louis University Mental Status (SLUMS) examination.[12] Elucidating clinical factors associated with cognitive impairment is critical for developing effective preventive and therapeutic strategies in this population.

PARTICIPANTS AND METHODS

Participants. This was a cross-sectional exploratory analysis using data from a parent study focused on frailty and osteoporosis: Veterans Affairs Rheumatoid Arthritis Frailty and Osteoporosis cohort study (VA-FROst). All participants were required to provide written informed consent for the parent study. The screening and enrollment process for participants in the parent study, including inclusion and exclusion criteria that are tailored to factors relating to RA and osteoporosis as it relates to the parent study, are detailed in Supplementary Figure 1. The data for this analysis were collected at the baseline visit. This analysis was approved by the VA Puget Sound Health Care System Institutional Review Board (IRB# 1872727).

Primary outcome: cognitive status. The primary outcome measure was cognitive status, assessed by the VA SLUMS examination.[13] SLUMS is a validated cognitive screening test used to determine presence of normal cognition, MCI, or dementia.[13] Scores range from 0 to 30, and categories require adjustment for self-reported education level. If an individual's education level was ≥12 years (high-school level or higher), their scores were categorized as normal (≥27), MCI (21–26), or dementia (≤20). If the education level was <12 years, their scores were categorized as normal (≥25), MCI (20–24), or dementia (≤19).

Other variables. *Demographics and health factors.* Age, sex, race, ethnicity, education level, and smoking history were collected by participant self-report. Body mass index (BMI, kg/m^2) was calculated using height and weight measures collected during the baseline visit.

Objectively measured physical function and frailty. Fifteen-feet walk and hand grip strength were assessed. Six (4%) participants had missing fifteen-feet walk data due to physical limitations inhibiting their ability to complete the measure; they were excluded from analyses in which fifteen-feet walk was a variable. The Short Performance Physical Battery (SPPB), which is a composite measure of physical function that includes gait time, balance, and chair stand, was also performed; total score ranged from 0 to 12.[14] Physical activity was self-reported using the International Physical Activity Questionnaire.[15] Frailty was measured using the Fried Frailty Phenotype, which includes assessment of walking speed, hand grip strength, physical activity, exhaustion, and weight loss.[16] These produced scores ranging from 0 to 5. Individuals were categorized as robust (0), prefrail (1–2), or frail (3–5) based on previously established cutoffs.[16]

Participant-reported symptoms. Exhaustion was reported using the Center for Epidemiologic Studies Depression Scale,[16] defined in Supplementary Table 1. The Multidimensional Health Assessment Questionnaire (MDHAQ)[17] was used to assess pain (range: 0–10), depression (0–3.0), and disability (HAQ range: 0–3.0). Sleep (quality, refreshing, problem sleeping, falling asleep) and fatigue (overage fatigue, starting tasks, feeling rundown, and average fatigue) were evaluated with the 10 Patient-Reported Outcomes Measurement Information System (PROMIS-10) global health items[18] with each item scored 1 to 5 and listed in

Supplementary Table 1. Each of the individual items on the sleep and fatigue PROMIS measures were explored separately rather than the validated composite measures to evaluate whether the specific symptom domains of sleep and fatigue were associated with cognition.

RA disease characteristics. RA disease duration was assessed using the self-reported date of diagnosis. Biologic disease-modifying antirheumatic drug (bDMARD) and conventional synthetic DMARD (csDMARD) use as well as anti-cyclic citrullinated peptide antibody levels were all collected from the electronic health record. Prednisone dose (mg/day) was collected by self-report. RA disease activity was measured by a rheumatologist using the Disease Activity Score 28-joints with C-reactive protein (range: 0–9.4) and categorized as remission (<2.6), low (2.6–3.1), moderate (3.2–5.1), and high (>5.1) disease activity.[19]

Comorbidities. Mental health, cardiovascular, and hearing impairment comorbidities were ascertained from the active problems lists within the electronic health record at the time of the study visit. Mental health and cardiovascular comorbidities were included due to their known association with cognition in the general population.[1,20] Hearing impairment was included because hearing loss can impact performance on cognitive testing and is a risk factor for cognitive impairment.[21]

Statistical analyses. This was an exploratory cross-sectional analysis of baseline data. Baseline characteristics, including demographics and RA disease characteristics, were compared using descriptive statistics, stratified by cognitive impairment category (normal, MCI, and dementia). Frequencies and percentages were reported for categorical variables. Mean and SDs were reported for continuous variables.

Data were available on an array of variables including medications, laboratory values, and clinical factors, but we chose to focus on clinical factors including demographics and health factors, objectively measured physical function, participant-reported symptoms, RA disease characteristics, and comorbidities. We explored the association between the continuous SLUMS score and clinical factors using unadjusted linear regression models. Adjusted linear regression models were performed on each variable separately and only on those that reached a significance level of $P < 0.1$ in univariable analyses. Each multivariable model was adjusted for age, sex, and years of education.

Because SLUMS is validated in those 60 years and older, we performed sensitivity analyses repeating the primary models restricting to those 60 years and older. To evaluate associations between clinical factors and dementia, we performed additional sensitivity analyses using logistic regressions with the outcome being those with SLUMS scores meeting the dementia cutoff. Multivariable models were performed on each variable whose univariable model reached a significance level of $P < 0.1$. Multivariable linear regressions were adjusted for age, sex, and years of education. Because years of education were used to define the dementia outcome group, multivariable logistic regressions controlled only for age and sex.

Significance level was set at $P < 0.05$, and STATA version 18.0 was used to perform data analyses.[22]

RESULTS

A total of 145 participants were included, with a mean age of 64.5 (SD 11.6) years (age range 38–86 years), and 74% (n = 107) were male (Table 1). Of those, 29% (n = 42) had normal cognition, 57% (n = 83) had MCI, and 14% (n = 20) had dementia with a mean SLUMS score of 24.1 (SD 3.9; Supplementary Figure 2). Participants with MCI or dementia were older (mean 70 [SD 11] and mean 65 [SD 11] vs mean 61 [SD 12] years), were more likely to be male (75% and 81% vs 59%), had greater smoking history (mean 20 [SD 23] and mean 22 [SD 32] vs mean 12 [SD 13] pack years), and had lower BMI (mean 27 [SD 5] and mean 30 [SD 5] vs mean 31 [SD 6] kg/m^2) compared to those with normal cognition. Those with MCI or dementia also had longer RA disease duration (mean 21 [SD] 15 and mean 15 [SD 14] vs mean 11 [SD 11] years) than those with normal cognition. Individuals with worse cognitive performance had greater disability as measured by MDHAQ (mean 0.8 [SD 0.6] and mean 0.7 [SD 0.4] vs mean 0.5 [SD 0.4]), had higher pain (mean 4.7 [SD 2.6] and mean 4.6 [SD 2.4] vs mean 3.9 [SD 2.6]), and were more likely to be frail (20% and 17% vs 9%) compared to those with normal cognition.

In our univariable regression models, age, male sex, and longer disease duration were all associated with lower SLUMS score, whereas greater BMI was associated with a higher SLUMS score (Table 2). Slower walking speed, decreased lower extremity function (as measured by SPPB), and prefrail and frail status were all associated with a lower SLUMS score. Additionally, participant-reported exhaustion, pain, disability, and difficulty falling asleep were associated with a lower SLUMS score. Greater BMI and lower extremity function (as measured by SPPB) were associated with a higher SLUMS score.

In our separate multivariable regression models adjusting for age, sex, and education status, measures of physical function and disability were associated with cognition. With each point improvement in physical function as measured by SPPB (range 0–12), the SLUMS score also improved by 0.35 points (aβ: 0.35, confidence interval [CI] 0.11 to 0.59; $P = 0.005$), whereas with each point the HAQ disability score (range 0–3) worsened, the SLUMS score also worsened by 1.79 points (aβ: −1.79, CI −3.16 to −0.42; $P = 0.011$) (Table 3). We also found lower SLUMS scores in those who reported exhaustion (yes/no; aβ: −2.22, CI −3.44 to −0.99; $P < 0.001$), increasing levels of pain (range 0–10; aβ: −0.25, CI −0.50 to −0.01; $P = 0.046$), and increasing levels of difficulty falling asleep (range 1–5; aβ: −0.79, CI −1.28 to −0.34; $P = 0.002$).

Sensitivity analyses. Among participants aged ≥60 years old (n = 104), multivariable linear regression models evaluating

Table 1. Cohort demographics and clinical characteristics stratified by cognition category using SLUMS (N = 145)*

	Total, N = 145	Normal, n = 42 (29%)	MCI, n = 83 (57%)	Dementia, n = 20 (14%)
Demographics and health factors				
Age, mean (SD), y	64.5 (11.6)	61.3 (12.2)	64.9 (11.2)	69.8 (10.5)
Male, n (%)	107 (74)	25 (59)	67 (81)	15 (75)
Race, n (%)				
Black	18 (12)	5 (12)	10 (12)	3 (15)
Other/unknown[a]	9 (6)	3 (7)	5 (6)	0 (0)
White	100 (69)	30 (71)	54 (65)	16 (80)
Multiple races	18 (12)	3 (7)	14 (17)	1 (5)
Hispanic/Latino, n (%)	9 (6)	2 (5)	6 (7)	1 (5)
Schooling, mean (SD), y	14.5 (2.3)	15.0 (2.2)	14.5 (2.3)	14.8 (2.7)
Smoking, mean (SD), pack y	18.6 (26.6)	11.7 (13.3)	21.8 (31.6)	20.0 (22.8)
BMI, mean (SD), kg/m^2	30.1 (5.5)	31.0 (5.6)	30.2 (5.3)	27.4 (5.2)
RA disease characteristics				
Disease duration, mean (SD), y	14.5 (13.8)	10.8 (10.7)	14.8 (14.3)	21.2 (15.3)
Anti-CCP antibody positive, n (%)	94 (65)	26 (62)	59 (71)	9 (45)
RA disease activity[b] (0–9.4), mean (SD)	3.9 (1.3)	3.6 (1.3)	4.2 (1.2)	3.7 (1.1)
Remission (<2.6), n (%)	27 (19)	12 (29)	11 (13)	4 (20)
Low (2.6–3.1), n (%)	18 (12)	4 (10)	11 (13)	3 (15)
Moderate (3.2–5.1), n (%)	75 (52)	23 (55)	41 (49)	11 (55)
High (>5.1), n (%)	25 (17)	3 (7)	20 (24)	2 (10)
Prednisone use, mean (SD), mg/day	1.3 (3.4)	0.5 (1.8)	1.6 (3.6)	1.5 (4.6)
csDMARD use, n (%)	99 (68)	28 (67)	60 (72)	11 (55)
bDMARD use, n (%)	74 (51)	21 (50)	41 (49)	12 (60)
Objectively measured physical function, n (%)				
Fried frailty phenotype				
Robust	38 (26)	16 (38)	19 (23)	3 (15)
Prefrail	85 (59)	22 (52)	50 (60)	13 (65)
Frail	22 (15)	4 (9)	14 (17)	4 (20)
15-ft walk time, s[c]	13 (9)	2 (5)	4 (5)	7 (35)
SPPB score (0–12)	44 (30)	9 (21)	24 (30)	11 (55)
Hand grip strength, lb	58 (40)	15 (36)	35 (42.2)	8 (40)
Physical activity, met min	30 (21)	7 (16.7)	19 (22.9)	4 (20)
Participant-reported symptoms				
Exhaustion, n (%)	74 (51)	14 (33)	47 (57)	13 (65)
MDHAQ				
Depression, n (%)	14 (10)	3 (7)	8 (10)	3 (15)
Pain (0–10), mean (SD)	4.4 (2.5)	3.9 (2.6)	4.6 (2.4)	4.7 (2.6)
Disability score (0–3.0), mean (SD)	0.7 (0.5)	0.5 (0.4)	0.7 (0.4)	0.8 (0.6)
PROMIS sleep, mean (SD)	11.4 (3.6)	11.1 (3.6)	11.3 (3.7)	12.2 (3.4)
Sleep quality (1–5)	3.0 (1.0)	3.0 (0.9)	3.0 (1.1)	3.0 (0.9)
Refreshing sleep (1–5)	3.2 (1.1)	3.2 (1.1)	3.2 (1.1)	3.5 (1.2)
Problems sleeping (1–5)	2.8 (1.2)	2.8 (1.1)	2.8 (1.2)	2.9 (0.9)
Difficulty falling asleep (1–5)	2.4 (1.2)	2.1 (1.1)	2.4 (1.3)	2.9 (1.4)
PROMIS fatigue, mean (SD)	10.8 (3.8)	10.1 (4.1)	10.9 (3.7)	11.6 (3.5)
Overall fatigue (1–5)	2.9 (1.0)	2.7 (1.1)	2.9 (1.0)	3.1 (1.0)
Difficulty starting tasks (1–5)	2.5 (1.1)	2.4 (1.2)	2.5 (1.1)	2.6 (1.1)
Feeling rundown (1–5)	2.7 (1.0)	2.5 (1.2)	2.7 (1.0)	2.9 (0.7)
Fatigued on average (1–5)	2.7 (1.0)	2.5 (1.0)	2.8 (1.0)	3.0 (1.0)
Comorbidities, n (%)				
Mental health	72 (50)	18 (43)	42 (51)	12 (60)
Cardiovascular	78 (54)	20 (48)	48 (58)	10 (50)
Hearing impairment	14 (10)	3 (7)	9 (11)	2 (10)

* bDMARD, biologic disease-modifying antirheumatic drug; BMI, body mass index; CCP, cyclic citrullinated peptide; csDMARD, conventional synthetic DMARD; DAS28-CRP, disease activity score 28-joints with C-reactive protein; MCI, mild cognitive impairment; MDHAQ, Multidimensional Health Assessment Questionnaire; PROMIS, Patient-Reported Outcomes Measurement Information System; RA, rheumatoid arthritis; SLUMS, St. Louis University Mental Status; SPPB, Short Performance Physical Battery.
[a] Other/unknown race: American Indian, Alaskan Native, Pacific Islander, and Asian.
[b] Measured using DAS28-CRP.
[c] Missing data: n = 6 unable to complete walk test.

SLUMS score (Table 4) revealed that slower walking speed (aβ = –0.59, CI –1.17 to –0.01; P = 0.046) and lower SPPB scores (aβ = 0.34, CI 0.06 to 0.62; P = 0.019) were significantly associated with lower cognitive performance. Among participant-reported symptoms, exhaustion, pain, disability, difficulty falling asleep, and fatigue were all independently associated

Table 2. Univariable linear regressions evaluating the association between participant characteristics and SLUMS scores (N = 145)*

	β[a]	95% CI	*P* value
Demographics and health factors			
Age	−0.08	−0.14 to −0.03	**0.002**
Male	−1.78	−3.21 to −0.36	**0.014**
Race			
White	Ref	–	–
AI/AN/PI	0.21	−3.32 to 3.74	0.907
Black	−0.41	−2.38 to 1.56	0.680
Asian	2.14	−2.37 to 6.66	0.349
Other/unknown	5.81	−1.93 to 13.55	0.140
Multiple races	−0.80	−2.77 to 1.17	0.423
Hispanic/Latino	−1.21	−3.85 to 1.44	0.369
Schooling, y	0.18	−0.10 to 0.46	0.200
Smoking, pack y	−0.02	−0.04 to 0.00	0.106
BMI	0.12	0.00 to 0.23	**0.048**
RA disease characteristics			
Disease duration, y	−0.06	−0.10 to −0.01	**0.012**
Anti-CCP antibody positive	0.17	−1.17 to 1.51	0.800
DAS28-CRP			
Remission	Ref	–	–
Low	−1.31	−3.65 to 1.02	0.267
Moderate	−0.61	−2.33 to 1.11	0.487
High	−1.85	−3.97 to 0.28	*0.088*
Prednisone use, mg/d	−0.08	−0.27 to 0.11	0.411
csDMARD use	0.06	−1.31 to 1.44	0.926
bDMARD use	0.06	−1.22 to 1.34	0.922
Objectively measured physical function			
Fried frailty phenotype			
Robust	Ref	–	–
Prefrail	−1.72	−3.19 to −0.25	**0.022**
Frail	−2.20	−4.22 to −0.18	**0.033**
15-ft walk time, s[b]	−0.73	−1.24 to −0.22	**0.005**
SPPB score (0–12)	0.44	0.20 to 0.67	**<0.001**
Hand grip strength, lb	0.01	−0.02 to 0.03	0.580
Physical activity, met min	−0.00	−0.00 to 0.00	0.673
Participant-reported symptoms			
Exhaustion	−1.81	−3.06 to −0.57	**0.005**
MDHAQ			
Depression	−1.25	−3.41 to 0.90	0.253
Pain	−0.22	−0.47 to 0.03	*0.084*
Disability	−2.13	−3.47 to −0.78	**0.002**
PROMIS sleep			
Sleep quality	0.22	−0.42 to 0.86	0.497
Refreshing sleep	0.21	−0.35 to 0.77	0.458
Problems sleeping	0.07	−0.49 to 0.62	0.810
Difficulty falling asleep	−0.67	−1.18 to −0.17	**0.009**
PROMIS fatigue			
Overall fatigue	−0.31	−0.92 to 0.30	0.318
Difficulty starting tasks	−0.04	−0.63 to 0.54	0.887
Feeling rundown	−0.37	−0.99 to 0.24	0.234
Fatigued on average	−0.49	−1.13 to 0.15	0.135
Comorbidities			
Mental health	−0.43	−1.70 to 0.85	0.511
Cardiovascular	−0.59	−1.87 to 0.69	0.364
Hearing impairment	−0.22	−2.39 to 1.94	0.838

* Bolded values = significant at $P < 0.05$. Italicized values = significant at $P < 0.1$. AI/AN/PI, American Indian, Alaskan Native, Pacific Islander; bDMARD, biologic disease-modifying antirheumatic drug; BMI, body mass index; CI, confidence interval; CCP, cyclic citrullinated peptide; csDMARD, conventional synthetic DMARD; DAS28-CRP, disease activity score 28-joints with C-reactive protein; PROMIS, Patient-Reported Outcomes Measurement Information System; RA, rheumatoid arthritis; SLUMS, St. Louis University Mental Status; SPPB: Short Performance Physical Battery.

[a] Values are beta coefficients for the change in SLUMS score per unit increase in the participant characteristic variable.

[b] n = 139 included in model. Six participants were unable to complete the walk test.

Table 3. Multivariable linear regressions evaluating the association between participant characteristics and SLUMS scores (N = 145)*

	aβ[a]	95% CI	*P* value
Sociodemographics			
BMI	0.08	−0.04 to 0.19	0.198
RA disease characteristics			
Disease duration, y	−0.03	−0.08 to 0.02	0.186
DAS28-CRP			
Remission	Ref	–	–
Low	−1.40	−3.67 to 0.86	0.223
Moderate	−1.14	−2.86 to 0.57	0.189
High	−1.95	−4.01 to 0.11	0.064
Objectively measured physical function			
Fried frailty phenotype			
Robust	Ref	–	–
Prefrail	−1.41	−2.87 to 0.06	0.060
Frail	−1.58	−3.62 to 0.45	0.126
15-ft walk time, s[b]	−0.51	−1.05 to 0.04	0.071
SPPB score (0–12)	**0.35**	**0.11 to 0.59**	**0.005**
Participant-reported symptoms			
Exhaustion	**−2.22**	**−3.44 to −0.99**	**<0.001**
MDHAQ			
Pain	**−0.25**	**−0.50 to −0.01**	**0.046**
Disability	**−1.79**	**−3.16 to −0.42**	**0.011**
PROMIS sleep			
Difficulty falling asleep	**−0.79**	**−1.28 to −0.31**	**0.002**

* Bolded values indicate statistical significance of $P < 0.05$. BMI, body mass index; CI, confidence interval; DAS28-CRP, disease activity score 28-joints with C-reactive protein; MDHAQ, Multidimensional Health Assessment Questionnaire; PROMIS, Patient-Reported Outcomes Measurement Information System; RA, rheumatoid arthritis; SLUMS, St. Louis University Mental Status; SPPB, Short Performance Physical Battery.
[a] Values are beta coefficients for the change in SLUMS score per unit increase in the participant characteristic variable. Models were adjusted for age, sex, and education.
[b] n = 139 included in model. Six participants were unable to complete the walk test.

with worse cognitive performance. Moderate (aβ = −2.32, CI −4.32 to −0.32; $P = 0.023$) and high disease activity (aβ = −2.72, CI −5.12 to −0.31; $P = 0.027$) were also associated with significantly lower SLUMS scores in those aged ≥60 years.

Multivariable logistic regression models using dementia as the outcome found that greater physical function by SPPB score and anti–cyclic citrullinated peptide antibody positivity were associated with lower odds of dementia (adjusted odds ratio [aOR] 0.83, CI 0.71–0.98; $P = 0.024$ and aOR 0.37, CI 0.14–0.99; $P = 0.049$, respectively; Table 5). Greater difficulty falling asleep was associated with three times the odds of dementia (aOR 3.05, CI 1.09–8.58; $P = 0.034$).

DISCUSSION

In this single-center study of people with RA, cognitive impairment as measured by SLUMS examination was present in 71% of participants. Decreased physical function, participant-reported outcomes, and RA disease characteristics were all associated with lower cognitive testing scores, both in our primary analyses and in sensitivity analyses. Importantly, the majority of clinical factors associated with decreased cognitive function may be modifiable through targeted interventions.

Poor physical performance, sleep, and pain are known risk factors for poor cognitive performance in the general population yet have various levels of evidence in RA.[5,20,23,24] We found physical function to be associated with cognitive function. Increased self-reported physical activity has been previously found to be associated better self-reported cognitive function in a large RA cohort.[25] A predominantly female RA cohort also found an association between poorer cognition scores and objective physical function by the SPPB, closely aligning with our findings.[26] Further supporting our findings, studies in the general population also suggest that physical activity in the form of consistent exercise enhances cognition and decreases risk of developing cognitive disabilities.[27]

Poor sleep quality has been found to be associated with cognitive dysfunction in RA.[23,24] We found difficulty falling asleep, but not overall poor sleep quality, to have significant associations both with lower SLUMS scores and dementia, which contrasts with prior work that demonstrated associations between sleep quality and cognition.[23,24] Although exploratory, our findings suggest that specific sleep disturbance symptom domains, namely difficulty falling asleep, may be particularly relevant in adults with RA. This is supported by evidence from the general population, in which impaired sleep initiation and other sleep domains may predict cognitive decline.[28] Further work is needed to explore these specific symptom domains of sleep with respect to cognition in RA. Given that poor sleep can result in impaired cognitive function, which can in turn lead to poor sleep, special attention is needed to ensure sleep hygiene is optimized in RA.

Table 4. Multivariable linear regressions evaluating the association between participant characteristics and SLUMS scores if ≥60 years old (n = 104)*

	β[a]	95% CI	*P* value	aβ[b]	95% CI	*P* value
Sociodemographics						
BMI	0.16	0.03 to 0.30	**0.019**	0.12	−0.02 to 0.27	0.099
RA disease characteristics						
DAS28-CRP						
Remission	Ref	–	–	Ref	–	–
Low	−1.95	−4.55 to 0.64	0.139	−2.31	−4.86 to 0.23	0.074
Moderate	−1.75	−3.74 to 0.25	0.086	−2.32	−4.32 to −0.32	**0.023**
High	−2.67	−5.14 to −0.21	**0.034**	−2.72	−5.12 to −0.31	**0.027**
Objectively measured physical function						
Fried frailty phenotype						
Robust	Ref	–	–	Ref	–	–
Prefrail	−1.51	−3.26 to 0.24	0.089	−0.99	−2.82 to 0.83	0.283
Frail	−2.53	−4.92 to −0.13	**0.039**	−1.77	−4.29 to 0.75	0.166
15-ft walk time, s[c]	−0.67	−1.24 to −0.11	**0.020**	−0.59	−1.17 to −0.01	**0.046**
SPPB score (0–12)	0.41	0.13 to 0.68	**0.004**	0.34	0.06 to 0.62	**0.019**
Participant-reported symptoms						
Exhaustion	−2.10	−3.58 to −0.61	**0.006**	−2.08	−3.54 to −0.61	**0.006**
MDHAQ						
Pain	−0.34	−0.63 to −0.05	**0.024**	−0.37	−0.66 to −0.08	**0.012**
Disability	−2.54	−4.21 to −0.88	**0.003**	−2.33	−4.03 to −0.63	**0.008**
PROMIS sleep						
Difficulty falling asleep	−0.73	−1.33 to −0.14	**0.017**	−0.82	−1.41 to −0.24	**0.006**
PROMIS fatigue						
Feeling rundown	−0.94	−1.74 to −0.14	**0.022**	−1.10	−1.89 to −0.32	**0.006**
Fatigued on average	−0.80	−1.65 to 0.06	0.068	−0.99	−1.83 to −0.14	**0.022**

* Bolded values indicate statistical significance of $P < 0.05$. BMI, body mass index; CI, confidence interval; DAS28-CRP, disease activity score 28-joints with C-reactive protein; MDHAQ, Multidimensional Health Assessment Questionnaire; PROMIS, Patient-Reported Outcomes Measurement Information System; RA, rheumatoid arthritis; SLUMS, St. Louis University Mental Status; SPPB, Short Performance Physical Battery.
[a] Values are beta-coefficients for the change in SLUMS score per unit increase in the participant characteristic variable.
[b] Models were adjusted for age, sex, and education.
[c] n = 139 included in model. Six participants were unable to complete the walk test.

Lastly, research has shown that chronic pain can negatively impact cognitive function in both the general population[6,29] and in RA,[20,23] which aligns with our findings, though the mechanisms are unknown. Existing research suggests pain may be a mediator of other cognition-related factors including sleep[30] and physical function.[23] Previous studies also suggest that alleviating pain could help preserve cognitive function; however, the direct impact of pain management on mitigating cognitive decline has not been clearly established.[6,29] These characteristics of physical function, sleep disturbance, and pain have all been shown to improve cognition in the general population when intervened upon[6,27,29,31,32] and are also all impacted by RA disease activity.[33–35] Although

Table 5. Multivariable logistic regressions evaluating the association between participant characteristics and presence of dementia as assessed by SLUMS scores (N = 145)*

	OR[a]	95% CI	*P* value	aOR[b]	95% CI	*P* value
Sociodemographics						
BMI	0.89	0.81 to 0.98	0.023	0.91	0.82 to 1.01	0.072
RA disease characteristics						
Disease duration, y	1.04	1.00 to 1.07	0.024	1.02	0.99 to 1.06	0.132
Anti-CCP antibody positive	0.39	0.15 to 1.00	0.051	0.37	0.14 to 0.99	**0.049**
Objectively measured physical function						
15-ft walk time, s[c]	1.47	1.02 to 2.10	0.038	1.29	0.88 to 1.88	0.186
SPPB score (0–12)	0.80	0.69 to 0.94	0.006	0.83	0.71 to 0.98	**0.024**
Participant-reported symptoms						
PROMIS sleep						
Difficulty falling asleep	1.39	0.96 to 2.02	0.085	1.51	1.02 to 2.23	**0.034**

* Bolded values indicate statistical significance. aOR, adjusted OR; BMI, body mass index; CCP, cyclic citrullinated peptide; CI, confidence interval; OR, odds ratio; RA, rheumatoid arthritis; SLUMS, St. Louis University Mental Status; SPPB, Short Performance Physical Battery.
[a] Values are ORs for presence of dementia per unit increase in the participant characteristic variable.
[b] Models were adjusted for age and sex.
[c] n = 139 included in model. Six participants were unable to complete the walk test.

our primary analysis did not find a significant association between disease activity categories and cognition, a sensitivity analysis restricted to those 60 years and older did. Existing literature has demonstrated a relationship between disease activity and cognitive dysfunction that supports our findings[36,37] and suggests that controlling disease activity may directly improve cognition as well as indirectly by improving physical function, sleep, and pain, an important direction for future work. The association observed among participants over 60 years may reflect the cumulative burden of inflammation and longer disease duration, which should be explored in future longitudinal studies.[1,3,8] However, a minimal clinical difference has not yet been determined for the SLUMS measure, impacting the ability to comment on the clinical implications of our findings.

Our results are also consistent with prior studies in RA that report a high prevalence of cognitive impairment.[5,38–41] McDowell et al reported that 72% of their cohort with established RA (mean age: 67 years) exhibited cognitive impairment using the Montreal Cognitive Assessment (MoCA),[40] which aligns closely with our findings, whereas Simos et al only found a 20% prevalence of cognitive impairment in a relatively younger (mean age: 54 years) cohort of 100 people with early RA (mean disease duration: 1.4 years).[42] Varied prevalence of cognitive impairment in prior studies may also be attributed to differences in cognitive assessment tools, demographics, and age of cohort. Importantly, cognitive impairment may be underestimated in our cohort due to the requirement to provide informed consent to participate in the study, which excludes those with the most severe cognitive performance.

Our study adds to the literature by evaluating objectively measured factors as well as multiple participant-reported factors associated with cognition in a well-phenotyped cohort of people with RA. Additionally, our cohort adds to the paucity of literature evaluating men with RA. Despite the male predominance, we do have a significant proportion of female participants (26%). Importantly, the veteran population has unique exposure to stressors not encountered by the general population (eg, combat exposure), which therefore may amplify their risk of cognitive decline due to the interaction of these factors with RA.[43] This heightened vulnerability highlights the importance of studying cognitive health in this population. The long disease duration seen in this cohort may allow for the observation of chronic impacts of RA disease and treatment on cognition that a cohort of early RA otherwise would be unable to demonstrate.[1,3]

Our findings are also subject to limitations. Our analyses were both cross-sectional and exploratory, which both limit causal inferences and may increase the risk of Type 2 error. Our study collected multiple measures of physical performance and symptoms, which may increase the likelihood toward finding associations with these measures. Additionally, our sleep and fatigue variables were the individual components of validated measures, rather than the validated composite measures, to explore and identify particular symptom domains that may have relationships with cognition. This cohort was designed to evaluate the association of frailty and bone mineral density; therefore, more detailed measures related to cognition, such as comprehensive neuropsychological testing, brain imaging studies, or biomarker analysis (eg, blood-based markers for neurodegeneration), were not available. The SLUMS examination, although a validated screening tool, does not provide a definitive diagnosis of MCI or dementia, introducing the potential for misclassification. However, the SLUMS has importantly been found to have adequate validity and reliability,[12,44] as well as show similar high sensitivity to detecting MCI and dementia as compared to the MoCA and other cognitive testing tools.[12,13,45,46] In our study, the SLUMS examination was administered by research staff rather than clinical professionals, potentially introducing errors in measurement, though all staff were trained to administer the SLUMS which should limit this concern. In addition, although the use of SLUMS has been validated for individuals aged 60 years and older, its use in younger populations has not been widely studied, which may impact the generalizability of our findings to younger individuals.

In conclusion, we identified a high prevalence of cognitive impairment in a well-phenotyped cohort of people with RA and found physical function, participant-reported outcomes, and RA disease characteristics to have significant associations with lower cognitive performance. Our findings suggest cognitive screening is likely to be important in RA, particularly in those 60 years and older. Our findings also highlight intervenable targets (sleep, pain, and physical function), which are well-established in the general population, that may mitigate cognitive decline in RA. Although longitudinal studies are needed to evaluate the impact of improving these outcomes on cognition, improving sleep, physical function, and pain leads to improvements in quality of life and therefore should be treated alongside RA management. Given the high prevalence of cognitive impairment in RA and its negative impact on quality of life and health outcomes, rheumatology health care professionals should consider integrating cognitive impairment screening in the care of at-risk patients.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Wysham confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Basile MS, Ciurleo R, Bramanti A, et al. Cognitive decline in rheumatoid arthritis: insight into the molecular pathogenetic mechanisms. Int J Mol Sci 2021;22(3):1185.

2. Petersen RC, Lopez O, Armstrong MJ, et al. Practice guideline update summary: Mild cognitive impairment [RETIRED]: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. Neurology 2018;90(3): 126–135.
3. Meade T, Manolios N, Cumming SR, et al. Cognitive impairment in rheumatoid arthritis: a systematic review. Arthritis Care Res (Hoboken) 2018;70(1):39–52.
4. Hale JM, Schneider DC, Mehta NK, et al. Cognitive impairment in the U.S.: lifetime risk, age at onset, and years impaired. SSM Popul Health 2020;11:100577.
5. Shin SY, Katz P, Wallhagen M, et al. Cognitive impairment in persons with rheumatoid arthritis. Arthritis Care Res (Hoboken) 2012;64(8): 1144–1150.
6. Chen J, Wang X, Xu Z. The relationship between chronic pain and cognitive impairment in the elderly: a review of current evidence. J Pain Res 2023;16:2309–2319.
7. Martins VF, Peyré-Tartaruga LA, Haas AN, et al. Observational evidence of the association between physical and psychological determinants of aging with cognition in older adults. Sci Rep 2024;14(1): 12574.
8. Myasoedova E, Sattui SE, Lee J, et al. Cognitive impairment in individuals with rheumatic diseases: the role of systemic inflammation, immunomodulatory medications, and comorbidities. Lancet Rheumatol 2024;6(12):e871–e880.
9. Kodishala C, Hulshizer CA, Kronzer VL, et al. Risk factors for dementia in patients with incident rheumatoid arthritis: a population-based cohort study. J Rheumatol 2023;50(1):48–55.
10. Katchamart W, Narongroeknawin P, Chanapai W, et al. Health-related quality of life in patients with rheumatoid arthritis. BMC Rheumatol 2019;3(1):34.
11. Lwin MN, Serhal L, Holroyd C, et al. Rheumatoid arthritis: the impact of mental health on disease: a narrative review. Rheumatol Ther 2020;7(3):457–471.
12. Spencer RJ, Noyes ET, Bair JL, et al. Systematic review of the psychometric properties of the Saint Louis University Mental Status (SLUMS) examination. Clin Gerontol 2022;45(3):454–466.
13. Tariq SH, Tumosa N, Chibnall JT, et al. Comparison of the Saint Louis University mental status examination and the mini-mental state examination for detecting dementia and mild neurocognitive disorder–a pilot study. Am J Geriatr Psychiatry 2006;14(11):900–910.
14. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994;49(2):M85–M94.
15. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35(8):1381–1395.
16. Fried LP, Tangen CM, Walston J, et al; Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001;56(3):M146–M156.
17. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999; 42(10):2220–2230.
18. Hays RD, Bjorner JB, Revicki DA, et al. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009;18(7):873–880.
19. England BR, Tiong BK, Bergman MJ, et al. 2019 Update of the American College of Rheumatology recommended rheumatoid arthritis disease activity measures. Arthritis Care Res (Hoboken) 2019; 71(12):1540–1555.
20. Brown SC, Glass JM, Park DC. The relationship of pain and depression to cognitive function in rheumatoid arthritis patients. Pain 2002; 96(3):279–284.
21. Thomson RS, Auduong P, Miller AT, et al. Hearing loss as a risk factor for dementia: a systematic review. Laryngoscope Investig Otolaryngol 2017;2(2):69–79.
22. Stata. Version 18. StataCorp LLC; 2023. Accessed June 6 2025. https://www.stata.com/stata18/
23. Chaurasia N, Singh A, Singh IL, et al. Cognitive dysfunction in patients of rheumatoid arthritis. J Family Med Prim Care 2020;9(5):2219–2225.
24. Kim SH, Kim H, Kim S, et al. Sleep quality independently affects health-related quality of life and cognitive function in Korean female patients with rheumatoid arthritis: a case-control study. J Korean Med Sci 2018;33(35):e216.
25. Shadick NA, Katz P, Iannaccone CI, et al. The impact of exercise, lifestyle, and clinical factors on perceived cognitive function in patients with rheumatoid arthritis: results from a prospective cohort study. ACR Open Rheumatol 2019;1(10):620–626.
26. Shin SY, Julian L, Katz P. The relationship between cognitive function and physical function in rheumatoid arthritis. J Rheumatol 2013;40(3): 236–243.
27. Sewell KR, Erickson KI, Rainey-Smith SR, et al. Relationships between physical activity, sleep and cognitive function: a narrative review. Neurosci Biobehav Rev 2021;130:369–378.
28. Dzierzewski JM, Perez E, Ravyts SG, et al. Sleep and cognition: a narrative review focused on older adults. Sleep Med Clin 2022;17(2): 205–222.
29. Khera T, Rangasamy V. Cognition and pain: a review. Front Psychol 2021;12:673962.
30. Power JD, Perruccio AV, Badley EM. Pain as a mediator of sleep problems in arthritis and other chronic conditions. Arthritis Rheum 2005;53(6):911–919.
31. Torossian M, Fiske SM, Jacelon CS. Sleep, mild cognitive impairment, and interventions for sleep improvement: an integrative review. West J Nurs Res 2021;43(11):1051–1060.
32. Middleton LE, Manini TM, Simonsick EM, et al. Activity energy expenditure and incident cognitive impairment in older adults. Arch Intern Med 2011;171(14):1251–1257.
33. Katz P, Pedro S, Michaud K. Sleep disorders among individuals with rheumatoid arthritis. Arthritis Care Res (Hoboken) 2023;75(6):1250–1260.
34. Ibrahim F, Ma M, Scott DL, et al. Defining the relationship between pain intensity and disease activity in patients with rheumatoid arthritis: a secondary analysis of six studies. Arthritis Res Ther 2022;24(1):218.
35. Feng J, Yu L, Fang Y, et al. Correlation between disease activity and patient-reported health-related quality of life in rheumatoid arthritis: a cross-sectional study. BMJ Open 2024;14(5):e082020.
36. Katchamart W, Narongroeknawin P, Phutthinart N, et al Disease activity is associated with cognitive impairment in patients with rheumatoid arthritis. Clin Rheumatol 2019;38(7):1851–1856.
37. Lee JH, Kim GT, Kim YK, et al. Cognitive function of patients with rheumatoid arthritis is associated with disease activity but not carotid atherosclerotic changes. Clin Exp Rheumatol 2018;36(5):856–861.
38. Gwinnutt JM, Toyoda T, Jeffs S, et al. Reduced cognitive ability in people with rheumatoid arthritis compared with age-matched healthy controls. Rheumatol Adv Pract 2021;5(2):rkab044.
39. Kronzer VL, Crowson CS, Davis JM III, et al. Trends in incidence of dementia among patients with rheumatoid arthritis: a population-based cohort study. Semin Arthritis Rheum 2021;51(4):853–857.

40. McDowell B, Marr C, Holmes C, et al. Prevalence of cognitive impairment in patients with rheumatoid arthritis: a cross sectional study. BMC Psychiatry 2022;22(1):777.
41. Appenzeller S, Bertolo MB, Costallat LT. Cognitive impairment in rheumatoid arthritis. Methods Find Exp Clin Pharmacol 2004;26(5):339–343.
42. Simos P, Ktistaki G, Dimitraki G, et al. Cognitive deficits early in the course of rheumatoid arthritis. J Clin Exp Neuropsychol 2016;38(7):820–829.
43. Yaffe K, Lwi SJ, Hoang TD, et al. Military-related risk factors in female veterans and risk of dementia. Neurology 2019;92(3):e205–e211.
44. Noyes ET, Major S, Wilson AM, et al. Reliability and factor structure of the Saint Louis University Mental Status (SLUMS) examination. Clin Gerontol 2023;46(4):525–531.
45. Lee YC, Lin YT, Chiu EC. A comparison of test-retest reliability of four cognitive screening tools in people with dementia. Disabil Rehabil 2022;44(15):4090–4095.
46. Stewart S, O'Riley A, Edelstein B, et al. A preliminary comparison of three cognitive screening instruments in long term care: the MMSE, SLUMS, and MoCA. Clin Gerontol 2011;35(1):57–75.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 601–608
DOI 10.1002/acr.25670

American College of Rheumatology
Empowering Rheumatology Professionals

Financial Distress and Its Determinants in Rheumatoid Arthritis

Amber Brown Keebler,[1] Yunju Im,[2] Sofia Pedro,[3] Ted R. Mikuls,[4,5] Edward S. Peters,[6] and Kaleb Michaud[3,4]

Objective. To quantify the degree of financial distress and identify its determinants in adults with rheumatoid arthritis (RA) given the frequent prolonged use of expensive disease-modifying therapies.

Methods. We identified adults enrolled in the FORWARD databank with either RA or noninflammatory musculoskeletal disease (NIMSKD) completing the Functional Assessment of Chronic Illness Therapy–Comprehensive Score for Financial Toxicity (FACIT-COST) questionnaire. In this cross-sectional study, FACIT-COST was analyzed as a continuous (higher score indicates less financial distress) and binary variable (presence of financial distress with threshold <26). Least Absolute Shrinkage and Selection Operator (LASSO) was applied to linear and logistic regression to select covariates for inclusion in multivariable models.

Results. Participants with RA (n = 2,277) had lower FACIT-COST scores, indicating greater financial distress, than those with NIMSKD (n = 1,340) (mean of 30.2 ± SD 9.4 vs mean of 34.0 ± SD 8.4; unadjusted $P < 0.001$). Assessed as a binary outcome, financial distress was more frequent in participants with RA than participants with NIMSKD (29% vs 15%; unadjusted $P < 0.001$). Differences in financial distress by diagnosis persisted following multivariable adjustment. Among those with RA, determinants identified in multivariable models included depression (adjusted odds ratio 1.12; 95% confidence interval 1.09–1.16) and disease severity.

Conclusion. Financial distress is prevalent in adults with RA and appears to be greatest in those with comorbidities, specifically depression, identifying a potential area for intervention. Notably, expensive biologic or targeted synthetic disease-modifying antirheumatic drugs were not associated with FACIT-COST scores.

INTRODUCTION

Health care costs and financial distress are recognized as significant barriers to successfully managing chronic health conditions. Financial distress, often referred to as financial toxicity, is defined as the adverse effect of health care costs on a patient's financial well-being.[1] Health outcomes of decreased medication and treatment adherence in chronic disease are associated with financial distress. In a study by Khera et al,[2] financial distress was described among patients with cardiovascular disease. In a study by Patel et al,[3] the Comprehensive Score for Financial Toxicity was validated in patients with diabetes. It has been most robustly studied in patients undergoing treatment for cancer. Among young adults with cancer, it is also associated with unhealthy coping such as skipping treatments and lower levels of psychological well-being.[4] Overall, the perception of difficulty paying bills is associated with lower health care use.[5] Moreover, higher levels of financial distress are also associated with an increased frequency of depression and anxiety, which may serve to mediate the relationship between financial distress and poor patient outcomes.[4,6]

Although recognized as important in facilitating shared decision making, financial hardship, socioeconomic status, and financial distress are not routinely evaluated or discussed in day-to-day clinical practice. Though health insurance status is routinely evaluated, insurance coverage alone does not protect against financial distress.[6] In a study by Perez et al,[7] only 50% of US internists who responded to a survey reported having frequent cost

Dr Mikuls's work was supported by the NIH (grant U54GM115458), a VA Merit (I01-BX003635-05), and the Department of Defense (PR200793). Dr Michaud's work was supported by research funding from the Rheumatology Research Foundation.

[1]Division of General Internal Medicine, Department of Internal Medicine, University of Nebraska Medical Center, Omaha; [2]Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha; [3]FORWARD, The National Databank for Rheumatic Diseases, Wichita, Kansas; [4]Division of Rheumatology, Department of Internal Medicine, University of Nebraska Medical Center, Omaha; [5]Medicine and Research Services, VA Nebraska-Western Iowa Health Care System, Omaha, Nebraska; [6]Department of Epidemiology, College of Public Health, University of Nebraska Medical Center, Omaha.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25670).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25670.

Address correspondence via email to Kaleb Michaud, PhD, at kmichaud@unmc.edu.

Submitted for publication March 27, 2025; accepted in revised form September 23, 2025.

SIGNIFICANCE & INNOVATIONS

- To our knowledge, this report is among the first to describe the prevalence of financial distress in adults with rheumatic disease in the US.
- Approximately one-third of participants with rheumatoid arthritis (RA) demonstrate evidence of financial distress with a prevalence that is approximately twice that of those with noninflammatory musculoskeletal disease.
- Disease severity, depression, and other comorbidities are associated with financial distress in RA.
- Receipt of expensive advanced therapies in RA are not strongly associated with the presence or severity of financial distress.

conversations with their patients, though 75% reported using cost of care as a factor in guiding clinical decision-making. Strikingly, patients report high levels of interest in having these conversations with health care providers. In a study by Henrikson et al,[8] most respondents preferred their provider to have a cost conversation with them, and 75% felt comfortable with having these discussions.

Despite the importance of cost of care and its communication to both patients and health care providers, comprehensive examinations of financial distress in rheumatic disease, including rheumatoid arthritis (RA), have been limited to date. This knowledge gap is highly relevant as RA accounts in the US for 257,884 disability-adjusted life-years, and RA-related health care costs have risen markedly in recent years.[9] For a patient with RA receiving advanced therapies, for example, total annual per-person direct medical costs have been estimated to exceed $36,000.[10]

Given these prevailing knowledge gaps, we sought to examine the impact of financial distress in adults with rheumatic disease, including individuals with RA, and in those with noninflammatory musculoskeletal disease (NIMSKD). Using data from a well-characterized population of rheumatic disease participants, we sought to test the hypotheses that financial distress is more prevalent and more severe in RA than NIMSKD. Moreover, we sought to identify factors predictive of financial distress beyond rheumatic disease status. We tested the additional hypotheses that among those with RA, the use of advanced biologic or targeted synthetic therapies along with comorbid conditions, particularly the presence of depression and obesity, would be associated with higher levels of financial distress independent of other factors.

PATIENTS AND METHODS

Study design and participants. This cross-sectional study was performed using data from the FORWARD databank, previously known as The National Databank for Rheumatic Diseases.[11] FORWARD is an ongoing longitudinal observational registry that collects patient-reported outcomes on multiple rheumatic conditions (eg, RA, NIMSKD, and other inflammatory/autoimmune conditions) using biannual questionnaires covering a wide range of information such as sociodemographic factors, rheumatologic diagnoses, comorbidities, treatments, symptoms, health care utilization, and other long-term outcomes as previously detailed.[12–14] Participating US rheumatology clinics aid in the recruitment of registry participants and confirmation of primary rheumatologic diagnoses.[11]

Inclusion and exclusion criteria. This study included databank participants with either RA or NIMSKD and was limited to individuals with data available for the primary outcome as detailed below. Diagnoses considered as NIMSKD included predominantly chronic musculoskeletal diseases including Dupuytren's contracture (68%) and osteoarthritis (32%) but excluding fibromyalgia.[15,16] Participants with other forms of systemic inflammatory disease (eg, systemic lupus erythematosus, spondyloarthritis, crystal arthropathy) were excluded.

Outcomes. Financial distress experienced by participants was measured using the validated 12-item Comprehensive Score for Financial Toxicity–Functional Assessment of Chronic Illness Therapy (FACIT-COST) instrument and served as the primary outcome for this analysis.[1] This tool includes measurement of effect on dominant themes of finance, resources, affect, and coping, as well as effects on family relationships. The score has established correlations with perceived support from family and friends, age, savings, and household income.[17,18] Although originally utilized in oncology, the FACIT-COST measure has been validated in other chronic disease populations, demonstrating both internal consistency and predictive validity.[19] Response options for each item are provided on a 5-point Likert scale with each response scored from 0 to 4 based on responses of "not at all," "a little bit," "somewhat," "quite a bit," and "very much," respectively. The sum of individual item scores is then used as a severity score (range of 0 to 48), with lower scores indicating higher financial distress. For study purposes, the FACIT-COST score was computed if at least 6 of the 12 questions were answered per developer scoring guidance.[20] In addition to examining the severity score as a quantitative measure, we also examined a binary categorization defining the presence of financial stress by scores <26 and absence of financial distress by scores of ≥26 as utilized in other studies.[20–22]

Covariables. In addition to rheumatologic disease status (RA vs NIMSKD), other patient-reported factors were examined as possible determinants of treatment-related financial distress. Depression was assessed using the eight-item Patient Health Questionnaire (PHQ-8) depression scale, a standardized, well-validated diagnostic and severity measure for depressive

disorders.[23] This scale includes the criteria for depression as defined by the Diagnostic and Statistical Manual of Mental Disorders V, excluding suicidal ideation, as related questions are not included on FORWARD questionnaires. PHQ-8 scores range from 0 to 24 with scores ≥10 indicative of moderate-to-severe depression.[23]

Obesity was assessed using body mass index (BMI: kg/m^2) calculated from self-reported height and weight values at the time of completing the FACIT-COST questionnaire. BMI categories included underweight (<20 kg/m^2), normal (20–24.9 kg/m^2, referent category), overweight (25–29.5 kg/m^2), class 1 obesity (30–34.9 kg/m^2), class 2 obesity (35–39.5 kg/m^2), and class 3 obesity (≥40 kg/m^2).

Additional variables examined as possible determinants of financial distress included the following self-reported sociodemographic factors: sex, race (White vs non-White), age, college education or higher (yes/no), total annual household income, marital status (married vs not married), insurance coverage (Medicaid or Medicare: yes/no), and employment status (employed/retired vs not employed/retired). Clinical variables included Rheumatic Disease Comorbidity Index (RDCI),[24] physical disability quantified using the Health Assessment Questionnaire Disability Index (HAQ-DI),[25] Patient Global Assessment (PGA; 0–10),[26,27] and patient-reported disease activity (Patient Activity Scale [PAS], range from 0 to 10).[26] Individual comorbidities composing the RDCI (cancer, stroke, cardiovascular disease, diabetes, fracture, gastrointestinal conditions, neurologic disease, respiratory disease, renal impairment, liver disease, and depression), glucocorticoid use, and disease-modifying antirheumatic drug (DMARD) treatment were also included in analyses. Where relevant, DMARDs were categorized as conventional synthetic DMARD (csDMARD), tumor necrosis factor inhibitor (TNFi), non-TNFi biologic, or JAK inhibitor (JAKi). Participants with a NIMSKD diagnosis were excluded if they were on a DMARD or prednisone to reduce the risk of selection bias.

Statistical analysis. Descriptive statistics were used to describe characteristics of the overall sample and make comparisons between the RA and NIMSKD groups. Diagnosis groups were further subdivided by presence of financial distress (threshold of FACIT-COST score of <26); within group comparisons were performed using *t*-tests and chi-square tests as appropriate.

Univariate linear and logistic regression analyses were initially undertaken to examine marginal associations between financial distress (both continuous and binary outcomes, respectively) and each covariate. Least Absolute Shrinkage and Selection Operator (LASSO) was then used to select statistically significant variables associated with financial distress. LASSO is a linear regression method that estimates the regression coefficients by maximizing the log-likelihood function with a constraint imposed on the sum of the absolute values of regression coefficients.[28] To adequately account for confounding effects and given our focus on disease status, age, sex, total annual household income, RDCI, and RA status were fixed and not subject to LASSO variable selection. All other covariates were available for LASSO selection. As LASSO does not provide the SEs required for computing confidence intervals (CIs), we subsequently used a double-LASSO approach in the overall model to provide 95% CIs.[28,29] For RA specific models, estimates and 95% CIs were obtained by following initial LASSO variable selection with standard regression.

For missing values, where available, a most recent value carried forward or backward approach was used. Missing sex status was also imputed using an available registry variable for "menstrual cycles." Missing HAQ-DI and PAS values were replaced using HAQ-II or PAS-II scores, respectively, when available in the same phase questionnaire. Beyond these, a complete-case approach was used, excluding individuals with missing covariate data from relevant analyses. Less than 10% of participants were excluded from multivariable models due to missing covariates.

All data analyses were conducted using Stata version 17[30] (StataCorp) and R (R Foundation) glmnet.[31] Values of $P < 0.05$ were considered statistically significant.

Ethical approval. This study was considered exempt by the University of Nebraska Institutional Review Board (IRB; number 0093-07-EX) based on previous approval from the Ascension Via Christi Hospitals Wichita, Inc, IRB (IRB number: IRB00001674). Individuals participating in FORWARD undergo informed consent before data collection.

RESULTS

Participant characteristics. A total of 3,617 participants were included in the sample and analyzed in two diagnosis groups, RA (n = 2,277) and NIMSKD (n = 1,340). Characteristics of those with a diagnosis of RA or NIMSKD are shown in Table 1. Both RA and NISMKD similarly comprised participants with older ages and predominantly of White race, reflective of the overall FORWARD participant population. Though both groups were predominantly female, the proportion was higher in RA (84%) compared to NIMSKD (67%). Retirement status as well as Medicaid and Medicare insurance status were also similar between the two diagnosis groups. Compared to NIMSKD, RA participants had a lower total household income, were less likely to have a college education, and were less likely to be employed. Among participants with RA, 49.7% were on biologic treatment modalities and 59.5% were on csDMARDs.

FACIT-COST scores in RA vs NIMSKD. In unadjusted analyses, RA participants had lower FACIT-COST scores (mean ± SD: 30.2 ± 9.4) compared to participants with NIMSKD (34.0 ± 8.4) indicative of greater financial distress in RA ($P < 0.001$ for comparison; Figure 1). The presence of high financial distress

Table 1. Characteristics of study participants*

Variables	RA (n = 2,277)	NIMSKD (n = 1,340)
Demographics		
Age, y	69.8 (10.7)	68.8 (10.2)
Total household income, US $1,000	73.4 (42.1)	90.8 (41.5)
College education, %	54.1	73.2
Employed, %	20.1	28.2
Married, %	64.1	71.3
Retired, %	60.1	62.0
Female, %	83.9	66.6
White, %	90.8	96.8
Medicaid, %	4.3	2.6
Medicare, %	70.1	66.0
Quality of life and disease-related severity		
HAQ-II (0–3)	0.9 (0.7)	0.4 (0.6)
Patient Global (0–10)	3.6 (2.5)	2.1 (2.4)
PAS (0–10)	3.5 (2.2)	1.9 (2.1)
Comorbidities		
RDCI (0–9)	2.0 (1.7)	1.7 (1.6)
BMI, kg/m^2	28.4 (7.0)	26.6 (6.6)
Cancer, %	6.0	6.4
Stroke, %	4.5	5.8
Cardiovascular, %	14.5	12.4
PHQ-8 (0–24)	4.4 (4.4)	3.1 (3.7)
Depression, %	14.2	10.2
Diabetes, %	10.3	8.4
Liver disease, %	3.3	3.4
Treatments		
csDMARD, %	59.5	–
Biologic, %	49.7	1.1
TNFi, %	28.5	0.5
Non-TNFi, %	22.1	0.6
JAKi, %	9.4	0.1

* Values are mean (SD) unless otherwise indicated. Mean standard difference is >0.1 for all variables except for the following: age, retired, Medicaid, Medicare, cancer, stroke, cardiovascular, diabetes, and liver disease. Stroke includes transient ischemic attack and other prior cerebrovascular accidents; cardiovascular includes heart failure and myocardial infarction, but not high blood pressure. BMI, body mass index; csDMARD, conventional synthetic disease-modifying antirheumatic drugs; HAQ-II, Health Assessment Questionnaire II; JAKi, JAK inhibitor; NIMSKD, noninflammatory musculoskeletal disease; PAS, Patient Activity Scale; PHQ-8, Patient Health Questionnaire 8; RA, rheumatoid arthritis; RDCI, Rheumatic Disease Comorbidity Index; TNFi, tumor necrosis factor inhibitor.

(defined as a FACIT-COST score <26) was also more frequent in RA than NIMSKD (29% vs 15%; unadjusted $P < 0.001$) (Figure 1). After application of LASSO and accounting for covariates, associations of RA diagnosis (vs NIMSKD) was significant in the multivariable linear regression model (–β 0.9; 95% CI 0.11–1.69) though not in the logistic model (adjusted odds ratio [aOR] 1.38; 95% CI 1.02–1.88) (Figure 2). In these models, covariates associated with increased prevalence and/or severity of financial distress included RDCI, PAS, and PHQ-8, whereas higher household income and older age were independently associated with lower financial distress (Supplemental Table 1). Higher BMI and being retired were significantly associated with higher severity but not prevalence.

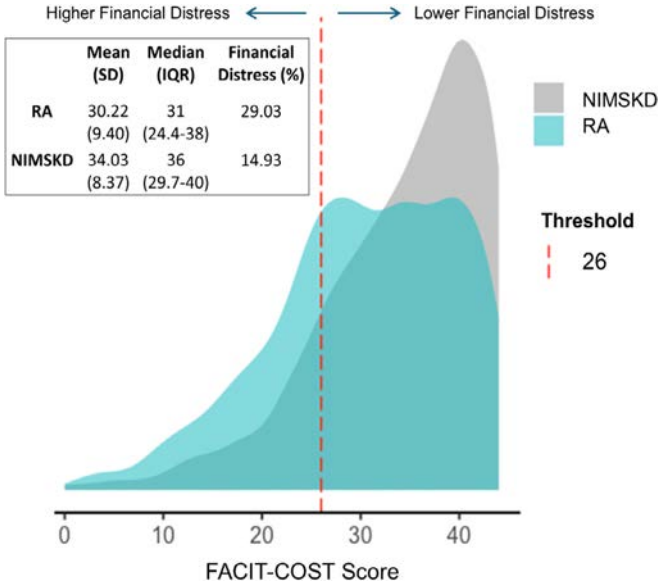

Figure 1. Distribution of FACIT-COST scores for people with RA and NIMSKD. Lower scores as seen in RA are indicative of higher financial distress. FACIT-COST, Comprehensive Score for Financial Toxicity–Functional Assessment of Chronic Disease; IQR, interquartile range; NIMSKD, noninflammatory musculoskeletal disease; RA, rheumatoid arthritis.

Determinants of financial distress in RA. Subsequent univariable and multivariable analyses were conducted to identify determinants of financial distress only in individuals with RA. In univariable analysis, Table 2 presents the characteristics that were associated with the presence of financial distress (Supplemental Table 2 similarly shows parallel analyses for NIMSKD). Consistent with our a priori hypothesis, those with financial distress had a higher prevalence of multiple comorbid conditions (including depression) and were slightly more likely to be receiving JAKi advanced therapy and less likely to be receiving csDMARDs. Similar results were observed examining univariable associations of clinical factors with FACIT-COST score.

After applying LASSO, independent determinants of financial distress in RA included younger age (aOR 0.97; 95% CI 0.96–0.98), lower total household income (aOR 0.89; 95% CI 0.86–0.92), higher RDCI score (aOR 1.11; 95% CI 1.02–1.21), higher PAS score (aOR 1.23; 95% CI 1.05–1.43), and higher PHQ-8 score (aOR 1.12; 95% CI 1.09–1.16). Similar determinants were again identified in separate multivariable analyses examining FACIT-COST score with LASSO selection of the additional variable of being retired (–β –2.2; 95% CI –3.02 to 1.37; Figure 3).

DISCUSSION

In this large observational study, we describe the prevalence and determinants of US-wide financial distress in adults with RA. As anticipated, a significant proportion of participants have

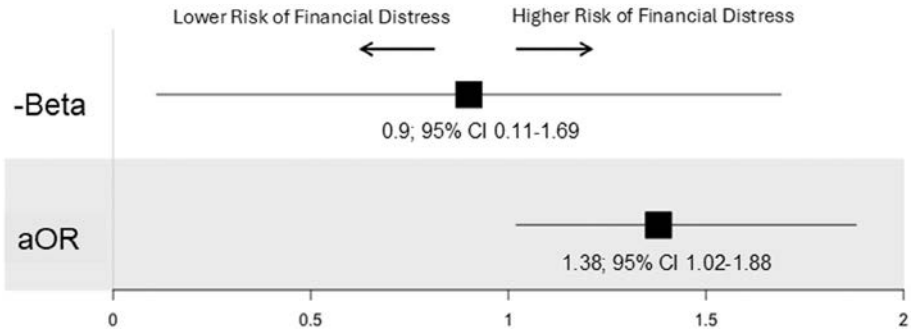

Figure 2. Diagnosis of RA associated with risk of financial distress in both linear and logistic models. Linear model represented as –beta. Logistic model represented as OR; both models indicate that a diagnosis of RA is associated with higher risk of financial distress. aOR, adjusted OR; CI, confidence interval; OR, odds ratio; RA, rheumatoid arthritis.

financial distress, emphasizing the need for identifying associated modifiable risk factors, though given the adjusted regression models, it was unlikely due to the higher use of expensive treatment modalities as originally hypothesized. This is notable as previous studies have clearly shown medication expenses compose a significant proportion of steep medical costs in the treatment of RA.[10] Interestingly, this suggests that factors other than potential treatment costs are contributing to financial distress.

Table 2. RA with and without FD*

Variables	FD present: score <26 (n = 661)	FD NOT present: score ≥26 (n = 1,616)	*P* value
Demographics			
Age, y	67.2 (10.6)	70.8 (10.6)	**
Income, US $1,000	57.7 (39.5)	79.9 (41.4)	**
College education, %	44.7	57.9	**
Employed, %	22.6	19.1	0.08
Married, %	58.7	66.3	*
Retired, %	45.1	66.3	**
Female, %	89.9	81.5	**
Caucasian, %	87.5	92.2	*
Medicaid, %	9.6	2.1	**
Medicare, %	61.6	73.5	**
Quality of life and disease-related severity			
HAQ-II (0–3)	1.25 (0.71)	0.81 (0.71)	**
Patient Global (0–10)	4.7 (2.3)	3.1 (2.4)	**
PAS (0–10)	4.6 (2.1)	2.0 (2.1)	**
Comorbidities			
RDCI (0–9)	2.6 (1.8)	1.8 (1.6)	**
BMI, kg/m^2	29.0 (7.9)	27.7 (6.4)	**
Cancer, %	7.6	5.3	<0.05
Stroke, %	6.6	3.7	*
Cardiovascular, %	17.4	13.4	<0.05
PHQ-8 (0–24)	7.2 (5.3)	3.3 (3.4)	**
Depression, %	26.7	9.1	**
Diabetes, %	16.8	7.6	**
Liver disease, %	4.7	2.7	<0.05
Treatment			
csDMARD, %	55.5	61.1	<0.05
Biologic, %	48.1	50.4	0.35
TNFi, %	26.8	29.1	0.28
Non-TNFi, %	22.2	22.0	0.96
JAKi, %	12.1	8.2	*

* The values are mean (SD) unless otherwise indicated. Stroke includes transient ischemic attack and other prior cerebrovascular accidents; cardiovascular includes heart failure and myocardial infarction, but not high blood pressure. BMI, body mass index; csDMARD, conventional synthetic disease-modifying antirheumatic drugs; FD, financial distress; HAQ-II, Health Assessment Questionnaire II; JAKi, JAK inhibitor; PAS, Patient Activity Scale; PHQ-8, Patient Health Questionnaire 8; RA, rheumatoid arthritis; RDCI, Rheumatic Disease Comorbidity Index; TNFi, tumor necrosis factor inhibitor.
* $P < 0.01$.
** $P < 0.001$.

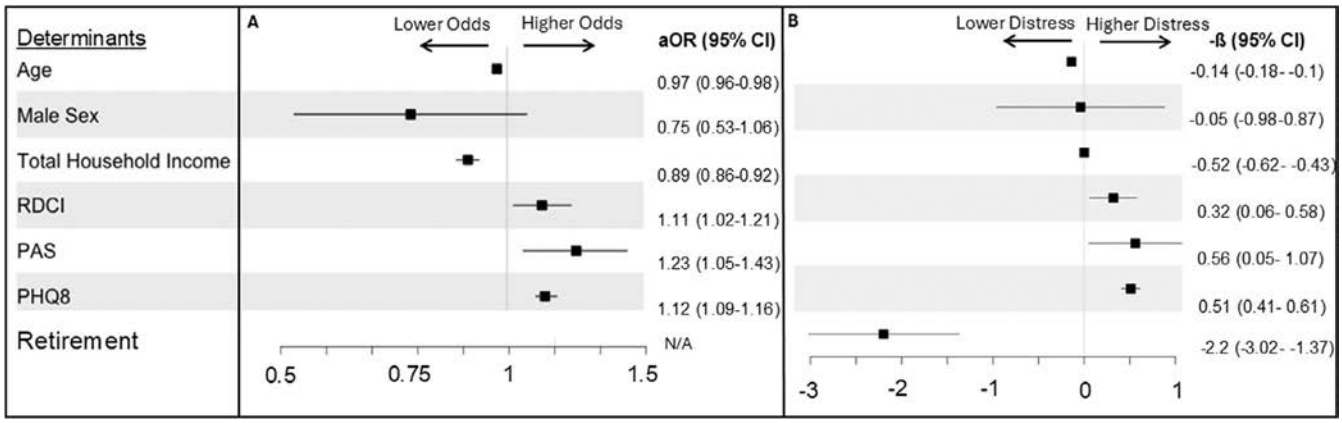

Figure 3. Determinants of financial distress in people with RA. (A) Logistic model with aOR found RDCI, PAS, and PHQ8 scores were associated with higher odds of financial distress. (B) Linear model with beta coefficients found that RDCI, PAS, and PHQ8 scores were associated with a higher financial distress. Retirement had a significant association with lower financial distress. aOR, adjusted odds ratio; CI, confidence interval; N/A, not applicable; PAS, Patient Activity Scale; PHQ8, Patient Health Questionnaire 8; RDCI, Rheumatic Disease Comorbidity Index; RA, rheumatoid arthritis.

Use of a FACIT-COST threshold score of 26 has been reported in oncology literature in patients with breast, lung, and colorectal cancer.[21,22] Prevalence of financial distress using this method demonstrated proportions of 45%[21] and 77%.[22] In comparison, approximately 29% of participants with RA in our study met a comparable threshold for financial distress; this lower prevalence can be explained by the significant disparity in medical costs in cancer treatment versus RA treatment. RA specific direct costs compose a substantial proportion of medical costs ranging between $3,723 and $20,262 per annum.[10] However, cancer far exceeds RA in out-of-pocket costs, ranging $2,160 to $31,200[32] versus $4,801[33] per annum, respectively.

We observed among the FORWARD participants determinants of financial distress including decreased quality of life, comorbidities including obesity, and depression. Similar to other studies, multiple comorbidities and decreased quality of life have been consistently associated with increased costs and predictive of more financial distress.[34] Previous studies have also demonstrated a pervasive association of financial distress with mental health diagnoses including depression and anxiety.[4,6] Strikingly, although unsurprising, depression is the most consistent association with financial distress in our analysis, regardless of diagnosis; the relationship between the two should strongly be considered for future longitudinal studies.

Measures indicative of financial status are associated with financial distress as demonstrated in the use of the FACIT-COST measure in both cancer and chronic disease.[3,19,21,22] Our findings are consistent with these previous studies, as we identified a significant inverse association between total household income and financial distress. Unexpectedly, retirement status had a significant negative association with financial distress. In contrast, employment status had a significant positive association with financial distress when FACIT-COST score was analyzed as a continuous variable. This is in contrast to a study by Weismann et al,[35] who reported that employment was associated with lower financial distress; in addition, older age was associated with higher financial distress. These differences are likely due to the participants' widely diverse medical diagnoses and that their study purposefully excluded those aged 65 years or greater, leaving a gap in knowledge of the experience of financial distress for those more likely to be retired. In our study, younger age was more likely to be associated with financial distress, though participants were on average older than 65 years of age. Weismann et al also found, similarly to our results, that lower income and chronic health conditions were associated with financial distress. In our study, secondary analyses demonstrated retired patients had a significantly higher income than those who were not retired. Thus, retirement is a surrogate marker for financial security and wealth, subsequently demonstrating a negative association with financial distress.

Given the prevalence of financial distress in RA, there are significant implications for studying clinical practice implementation methods. Sloan et al[36] recommended routinely including out-of-pocket costs in patient encounters, going so far as to screen for financial hardship even if not suspected. Assessment of financial hardship would appropriately broaden our clinical approach to social determinants of health in routine clinical care. Some studies have further shown that the use of a financial navigator significantly decreased health care costs and financial distress.[37,38]

Though our study provides insight into financial distress in people with rheumatic disease and its associated determinants, there are some limitations. There is less generalizability to non-White, less educated, and younger affected populations, an issue somewhat common to research registries that have participation/selection bias. As a patient-reported database, FORWARD is

subject to missing data. However, given the low overall rate of missingness and the use of complete-case analysis for all inferential results, we believe the findings are robust and not meaningfully affected by missing data. As this is a cross-sectional observational study, we are not able to establish causal inference and some unmeasured confounding may be present. The relationship across depression, financial distress, and disease severity is complex with multiple mediating factors. Future studies should be designed to prospectively determine causal direction and include a more diverse population with respect to race and ethnicity as well as socioeconomic status. Finally, it is unclear if the threshold score of the FACIT-COST measure is most appropriate, as it was determined in other nonrheumatic cohorts. However, a threshold score could be more meaningful and clinically useful due to ease of interpretation and ability to take immediate actionable steps.[21,22]

In conclusion, we observed that financial distress is prevalent in individuals with RA. Comorbidities including depression and measures of disease severity are significantly associated with financial distress. To our knowledge, this is one of the first studies describing financial distress in a large sample of adults with RA in the US. It highlights the relationship of financial distress with overall health and mental well-being. These findings should motivate future prospective studies to elucidate causality and evaluate the utility for clinical implementation.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Michaud confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. de Souza JA, Yap BJ, Hlubocky FJ, et al. The development of a financial toxicity patient-reported outcome in cancer: the COST measure. Cancer 2014;120(20):3245–3253.
2. Khera R, Valero-Elizondo J, Nasir K. Financial toxicity in atherosclerotic cardiovascular disease in the United States: current state and future directions. J Am Heart Assoc 2020;9(19):e017793.
3. Patel MR, Zhang G, Heisler M, et al. Measurement and validation of the Comprehensive Score for Financial Toxicity (COST) in a population with diabetes. Diabetes Care 2022;45(11):2535–2543.
4. Thom B, Benedict C. The impact of financial toxicity on psychological well-being, coping self-efficacy, and cost-coping behaviors in young adults with cancer. J Adolesc Young Adult Oncol 2019;8(3):236–242.
5. Hanmer J, Cherepanov D. A single question about a respondent's perceived financial ability to pay monthly bills explains more variance in health utility scores than absolute income and assets questions. Qual Life Res 2016;25(9):2233–2237.
6. Zafar SY, Peppercorn JM, Schrag D, et al. The financial toxicity of cancer treatment: a pilot study assessing out-of-pocket expenses and the insured cancer patient's experience. Oncologist 2013;18(4):381–390.
7. Perez SL, Weissman A, Read S, et al. U.S. internists' perspectives on discussing cost of care with patients: structured interviews and a survey. Ann Intern Med 2019;170(9_Suppl):S39–S45.
8. Henrikson NB, Chang E, Ulrich K, et al. Communication with physicians about health care costs: survey of an insured population. Perm J 2017;21(2):16–070.
9. Nguyen AT, Aris IM, Snyder BD, et al. Musculoskeletal health: an ecological study assessing disease burden and research funding. Lancet Reg Health Am 2024;29:100661.
10. Hresko A, Lin TC, Solomon DH. Medical care costs associated with rheumatoid arthritis in the US: a systematic literature review and meta-analysis. Arthritis Care Res (Hoboken) 2018;70(10):1431–1438.
11. Michaud K. The National Data Bank for Rheumatic Diseases (NDB). Clin Exp Rheumatol 2016;34(5 Suppl 101):S100–S101.
12. Ogdie A, Grewal SK, Noe MH, et al. Risk of incident liver disease in patients with psoriasis, psoriatic arthritis, and rheumatoid arthritis: a population-based study. J Invest Dermatol 2018;138(4):760–767.
13. Ogdie A, Hwang M, Veeranki P, et al. Health care utilization and costs associated with functional status in patients with psoriatic arthritis. J Manag Care Spec Pharm 2022;28(9):997–1007.
14. Ogdie A, Yu Y, Haynes K, et al. Risk of major cardiovascular events in patients with psoriatic arthritis, psoriasis and rheumatoid arthritis: a population-based cohort study. Ann Rheum Dis 2015;74(2):326–332.
15. Wilburn J, McKenna SP, Perry-Hinsley D, et al. The impact of Dupuytren disease on patient activity and quality of life. J Hand Surg Am 2013;38(6):1209–1214.
16. Yan H, Guo J, Zhou W, et al. Health-related quality of life in osteoarthritis patients: a systematic review and meta-analysis. Psychol Health Med 2022;27(8):1859–1874.
17. D'Rummo KA, Nganga D, Chollet-Hinton L, et al. Comparison of two validated instruments to measure financial hardship in cancer survivors: comprehensive score for financial toxicity (COST) versus personal financial wellness (PFW) scale. Support Care Cancer 2022;31(1):12.
18. Meeker CR, Wong YN, Egleston BL, et al. Distress and financial distress in adults with cancer: an age-based analysis. J Natl Compr Canc Netw 2017;15(10):1224–1233.
19. Pavela G, Fifolt M, Tison SE, et al. Re-validation of the COmprehensive Score for Financial Toxicity (COST): assessing the scale's utility in chronic disease populations. Health Serv Insights 2021;14:11786329211057352.
20. The FACIT Group. FACIT-COST. Accessed August 16, 2024. https://www.facit.org/measures/facit-cost.
21. Liu M, Huang D, Liu Y. Financial toxicity of patients with lung cancer in China: results from a national survey study. Cancer Med 2023;12(4):4751–4760.
22. Rivers Z, Shankaran V, Porter LS, et al. Discordance between patient with cancer and caregiving partner self-reported financial toxicity [abstract]. J Clin Oncol 2022;40(28 suppl):182. doi:https://doi.org/10.1200/JCO.2022.40.28_suppl.182
23. Kroenke K, Strine TW, Spitzer RL, et al. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009;114(1–3):163–173.
24. England BR, Sayles H, Mikuls TR, et al. Validation of the rheumatic disease comorbidity index. Arthritis Care Res (Hoboken) 2015;67(6):865–872.
25. Bruce B, Fries JF. The Health Assessment Questionnaire (HAQ). Clin Exp Rheumatol 2005;23(5)(suppl 39):S14–S18.

26. Anderson JK, Zimmerman L, Caplan L, et al. Measures of rheumatoid arthritis disease activity: Patient (PtGA) and Provider (PrGA) Global Assessment of Disease Activity, Disease Activity Score (DAS) and Disease Activity Score with 28-Joint Counts (DAS28), Simplified Disease Activity Index (SDAI), Clinical Disease Activity Index (CDAI), Patient Activity Score (PAS) and Patient Activity Score-II (PAS-II), Routine Assessment of Patient Index Data (RAPID), Rheumatoid Arthritis Disease Activity Index (RADAI) and Rheumatoid Arthritis Disease Activity Index-5 (RADAI-5), Chronic Arthritis Systemic Index (CASI), Patient-Based Disease Activity Score With ESR (PDAS1) and Patient-Based Disease Activity Score without ESR (PDAS2), and Mean Overall Index for Rheumatoid Arthritis (MOI-RA). Arthritis Care Res (Hoboken) 2011;63(suppl 11):S14–S36.

27. Nikiphorou E, Radner H, Chatzidionysiou K, et al. Patient global assessment in measuring disease activity in rheumatoid arthritis: a review of the literature. Arthritis Res Ther 2016;18(1):251.

28. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58(1):267–288.

29. Fitzgerald Sice J, Lattimore F, Robinson T, et al. Practical insights into applying double LASSO. SSRN. Preprint posted online June 4, 2023. doi:https://doi.org/10.2139/ssrn.44527

30. Stata Statistical Software: Release 17. StataCorp; 2021. https://www.stata.com/

31. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1–22.

32. Iragorri N, de Oliveira C, Fitzgerald N, et al. The out-of-pocket cost burden of cancer care-a systematic literature review. Curr Oncol 2021;28(2):1216–1248.

33. Erath A, Dusetzina SB. Assessment of expected out-of-pocket spending for rheumatoid arthritis biologics among patients enrolled in Medicare Part D, 2010-2019. JAMA Netw Open 2020;3(4):e203969.

34. Wang J, Mann F, Lloyd-Evans B, et al. Associations between loneliness and perceived social support and outcomes of mental health problems: a systematic review. BMC Psychiatry 2018;18(1):156.

35. Weissman J, Russell D, Mann JJ. Sociodemographic characteristics, financial worries and serious psychological distress in U.S. adults. Community Ment Health J 2020;56(4):606–613.

36. Sloan CE, Ubel PA. The 7 habits of highly effective cost-of-care conversations. Ann Intern Med 2019;170(9_Suppl):S33–S35.

37. Shankaran V, Leahy T, Steelquist J, et al. Pilot feasibility study of an oncology financial navigation program. J Oncol Pract 2018;14(2):e122–e129.

38. Yezefski T, Steelquist J, Watabayashi K, et al. Impact of trained oncology financial navigators on patient out-of-pocket spending. Am J Manag Care 2018;24(5 suppl):S74–S79.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 609–619
DOI 10.1002/acr.25671

American College of Rheumatology
Empowering Rheumatology Professionals

Testing a Personalized Approach to Chronic Low Back Pain: A Randomized Controlled Trial in Older Veterans

Debra K. Weiner,[1,2] Angela Gentili,[3,4] Meika A. Fang,[5,6] Edward Garay,[2] Laura Lawson,[3,4] Lenore Joseph,[3,4] Cathy C. Lee,[6,7] Michelle I. Rossi,[1,8] Beverly Thorn,[9] and Subashan Perera[2,10]

Objective. We aimed to test the efficacy of personalized treatment of older veterans with chronic low back pain (CLBP) delivered by Aging Back Clinics (ABCs) as compared with usual care (UC).

Methods. Two hundred ninety-nine veterans aged 65 to 89 with CLBP from three Veterans Affairs (VA) medical centers underwent baseline testing, randomization to ABC or UC, and 12 months of follow-up. ABC care was guided by trained physicians and published algorithms targeting key conditions contributing to CLBP (eg, hip osteoarthritis, depression, fibromyalgia). UC was guided by the participant's primary care provider. The primary outcome was six-month change in the Oswestry Disability Index (ODI). Among multiple secondary outcomes were pain intensity, quality of life (Patient-Reported Outcomes Measurement Information System [PROMIS]-Global Health), self-reported physical function (PROMIS-29), falls, life space, and health care utilization collected at 3, 6, 9, and 12 months. Analyses were conducted according to intention-to-treat.

Results. There were no significant group differences in ODI change. Greater improvement with ABC in the PROMIS-29 physical function scale was observed at 12 months (1.7 vs −0.4 points), PROMIS Global physical health at 6 (1.3 vs −1.2) and 12 months (0.7 vs −1.5), PROMIS Global mental health at 6 months (0.2 vs −2.3), and present and prior week average/worst pain over 12 months (all $P < 0.05$). There were marginally fewer falls over 12 months ($P = 0.0527$).

Conclusion. We did not find confirmatory evidence that personalized care (ABC) was superior with respect to ODI. We did find preliminary evidence that ABC was superior in other respects including improved self-reported physical health, less pain, and fewer falls.

INTRODUCTION

Back and other musculoskeletal pains are the most common reasons that United States veterans seek medical care,[1] and approximately half of such patients are aged 65 and older.[2] The prevalence of low back pain in those 85 and older, the most vulnerable and fastest growing segment of society, is estimated at 44%.[3] Chronic low back pain (CLBP; ie, occurring on at least half the days for at least six months[4]) is associated with the overwhelming majority of low back pain–associated health care resource utilization and personal suffering, including physical disability, depression, anxiety, insomnia, cognitive dysfunction, loss of sleep and appetite, and social isolation.[5–7] In older veterans (ie, aged ≥65), physical and emotional suffering may be compounded by a heightened risk of iatrogenic adversity associated with commonly employed interventions such as spinal injections, surgery, and potentially harmful medications (eg, gastrointestinal bleeding and renal failure with nonsteroidal anti-inflammatory drugs, falls and hip fractures with opioids[8]). Furthermore, although substantial resources continue to target treating patients with back pain, treatment outcomes are typically modest and have remained stagnant.[9] What accounts for this health care crisis?

The contents of this report do not represent the views of the Department of Veterans Affairs or the United States government.

Supported by the Office of Research and Development, Veterans Health Administration, Department of Veterans Affairs (grant 101RX002808).

[1]Geriatric Research, Education and Clinic Center, VA Pittsburgh Healthcare System, Pittsburgh, Pennsylvania; [2]VA Pittsburgh Healthcare System, Pittsburgh, Pennsylvania; [3]Central Virginia VA Health Care System, Richmond; [4]Virginia Commonwealth University Health System, Richmond; [5]VA Greater Los Angeles Healthcare System, Los Angeles, California; [6]David Geffen School of Medicine, University of California, Los Angeles; [7]Geriatric Research, Education and Clinic Center, VA Greater Los Angeles Healthcare System, Los Angeles, California; [8]School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania; [9]University of Alabama, Tuscaloosa; [10]Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25671).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25671.

Address correspondence via email to Meika A. Fang, MD, at meika.fang@va.gov.

Submitted for publication April 23, 2025; accepted in revised form September 16, 2025.

SIGNIFICANCE & INNOVATIONS

- This is the first clinical trial testing the efficacy of a personalized approach to managing chronic low back pain (CLBP) in older adults, guided by published algorithms that operationalize the evaluation and treatment of conditions that contribute to pain and disability and reside outside the lumbar spine.
- Although we did not find confirmatory evidence of reducing self-reported CLBP disability, we did find some evidence supporting reduction of pain intensity and falls as well as improvement in self-reported physical health in participants who received personalized as compared with usual care.

CLBP is referred to in the medical literature as a "nonspecific" condition evaluated and managed using imaging-guided spine-specific approaches (eg, spinal injections) or generic treatments (eg, pain medications, physical therapy, cognitive behavioral therapy). Spine-specific approaches may lead to suboptimal outcomes in older adults because imaging-identified pathology such as degenerative disc disease and spinal stenosis are common in older individuals, even those who are pain free[10,11]; thus, such approaches may or may not target actual contributors to pain. Generically applied first-line treatments such as physical therapy, cognitive behavioral therapy, and/or oral analgesics typically result in only modest improvement in pain and function,[12] perhaps not surprising because CLBP is a biopsychosocial syndrome in which disability may be caused by a combination of physical and psychosocial factors for which optimal treatment requires a comprehensive approach.[13] In older adults, there may be a combination of specific CLBP-associated disability contributors such as hip osteoarthritis, myofascial pain, fibromyalgia, and depression that cannot be detected by spinal imaging but for which evidence-based treatments exist.

The present manuscript reports the results of a clinical trial based on the premise that CLBP in older adults is a geriatric syndrome, that is, a final common pathway for the expression of multiple specific contributors.[14] The lumbar spine is considered the weakest link (ie, the most vulnerable part of the system), and multiple conditions outside it are considered potential treatment targets.[15] We have published a series of evidence and expert opinion-based clinical evaluation and treatment algorithms for key conditions that are commonly overlooked or misdiagnosed in older adults with CLBP (ie, hip osteoarthritis, fibromyalgia, myofascial pain, depression, anxiety, insomnia, maladaptive coping, sacroiliac joint syndrome, leg length inequality, lateral hip and thigh pain, lumbar spinal stenosis, dementia)[16–24]; we have explored their prevalence in older veterans[25] and conducted a pilot study demonstrating feasibility and preliminary efficacy of algorithm-guided care.[26] We hypothesized that older veterans with CLBP who are randomized to receive care in Aging Back Clinics (ABCs) that conduct evaluations and prescribe treatment using a comprehensive, algorithm-guided approach will have greater 6-month (primary endpoint) and 12-month reductions in pain-related disability (primary outcome) and improvements in other outcomes as compared with those who receive usual care (UC).

PARTICIPANTS AND METHODS

Study participants. The parent study methodology is described in detail elsewhere.[27] The study was approved by the Veterans Affairs (VA) Central Institutional Review Board (no. 18-23, approved 9-13-2018). Briefly, 299 veterans aged 65 to 89 with CLBP, on average, of at least moderate severity (based on a verbal descriptor scale) were recruited from three VA sites (VA Pittsburgh Healthcare System, Pittsburgh, Pennsylvania; Central Virginia VA Health Care System, Richmond, Virginia; West LA VA Medical Center, Los Angeles, California). Participants were excluded if they did not speak English or had other communication difficulties (severe uncorrected vision or hearing loss), screened positive for possible dementia based on a Folstein mini-mental state examination (MMSE)[28] score less than 24 of 30 (because the psychometric properties of our outcome measures have not been established in those with dementia), had psychotic symptoms, were acutely ill, had previously undergone lumbar surgery, had pain in other body locations more severe than that in the lower back, reported active illicit substance use, or had "red flags" (ie, signs or symptoms indicative of serious underlying illness requiring urgent care for their low back pain). Race and ethnicity were obtained from the participant's electronic medical record.

Testing procedures. *Baseline.* The following parameters were assessed on all participants at baseline: (1) the Oswestry Disability Index (ODI)[29], our primary outcome measure, assesses interference of pain with function; (2) the Minimal Data Set (MDS), recommended by the National Institutes of Health Task Force on research standards for CLBP,[4] measures pain severity and interference with daily activities, widespread pain, prior CLBP treatments, overall physical function, depressive symptoms, sleep, psychological maladaptation (ie, fear-avoidance beliefs and catastrophizing), alcohol/drug use, cigarette smoking, demographics (age, race, ethnicity, gender, education, marital status), height, and weight; (3) the Patient-Reported Outcomes Measurement Information System (PROMIS)-29 items not already included in the MDS—anxiety symptoms, fatigue, and participation in social roles and activities[30]; (4) other key cofactors that could contribute to disability—medical comorbidity (Duke comorbidity index[31]), pain medications (opioids and nonopioids), social support (Medical Outcomes Study Social Support Scale[32]), opioid difficulties in those taking opioids (Prescribed Opioids Difficulties Scale[33]) and smoking status; (5) pain severity

rated 0 to 10 for pain now, average pain over the past week, and worst pain over the past week; (6) the Pain Self-Efficacy Questionnaire[34] assesses the confidence people with ongoing pain have in performing activities while in pain; (7) quality of life was assessed with (a) the PROMIS Global Health scale and (b) the Veterans RAND 12 Item Health Survey (VR-12), a measure of health-related quality of life comprised of 12 items and collapsed into two summary variables—physical health and mental health; (8) the Quick Mild Cognitive Impairment (QMCI) screen is a validated measure that screens for the presence of mild cognitive impairment[35] (participants with more significant cognitive impairment were already excluded using the MMSE); (9) balance confidence was evaluated with the Falls Efficacy Scale-international short form[36]; (10) falls were assessed as the number of self-reported falls over the prior three months; and (11) mobility was assessed with gait speed and life space, the spatial extent of a person's mobility.[37]

Intervention phase. Procedures conducted during the intervention phase are described in detail elsewhere.[27] Following collection of baseline measures, participants were randomized in equal proportions to receive either UC or algorithm-guided personalized CLBP care in ABCs stratified by VA site and using random block sizes of four and six (chosen with equal likelihood). Those randomized to UC were instructed to discuss their CLBP care with their primary care provider. Those randomized to ABC care received a personalized treatment regimen by ABC providers.

ABC providers were VA medical center physicians. At the Pittsburgh VA, there were three providers—two geriatricians and one physiatrist/pain medicine physician. At the Richmond VA, there were three providers—two geriatricians and one neurologist. At the West LA VA there were two providers—one geriatrician and one rheumatologist.

ABC providers received structured training from the principal investigator (PI) on the application of our published algorithms on diagnostic criteria and management guidelines of 11 separate conditions contributing to CLBP. The PI also assessed treatment implementation fidelity (ie, adherence to treatment algorithms) by conducting a monthly review of structured research notes and providing ABC providers with written feedback.

At the initial ABC visits, the study coordinators asked the veterans to complete a self-report tool on an electronic tablet, *Take Back Your Back*, to screen for conditions contributing to their CLBP and related disability (eg, hip osteoarthritis, depression, fibromyalgia) and help guide the clinical assesment.[38] The ABC providers reviewed the results of the screening tool and conducted a structured history and physical examination to identify the specific treatment targets. As an example, if the participant screened positive for depression on the Patient Health Questionnaire-2, they would conduct a more detailed assessment for depression. ABC providers educated veterans on the clinical findings and the conditions contributing to the CLBP and then collaborated with veterans to design a multifaceted personalized care plan with specific treatment recommendations for each condition as guided by our algorithms. An educational booklet was provided to patients detailing their personalized treatment targets and plans.

As the management plan was based on a collaborative approach, veterans were free to choose some and decline other treatment recommendations. The ABC providers referred participants to the appropriate services (eg, physical therapy, chiropractic, acupuncture, behavioral health) and conducted follow-up visits as they felt appropriate. The number of treatments was determined by the providers to which veterans were referred. Veterans were permitted to continue identified treatments for the duration of the follow-up period or as long as deemed appropriate by the receiving service. If acute issues arose before their scheduled follow-up appointments with ABC providers, veterans were permitted to reach out to schedule an earlier appointment.

Follow-up. In the ABC group, the ABC provider scheduled follow-up visits as clinically indicated, with the last visit scheduled at the end of the study. The primary outcome point was at six months following randomization. The follow-up period lasted up to 12 months. In both the ABC and UC groups, measures were collected over the telephone by a research coordinator masked to randomization assignment, either monthly (pain intensity and health care utilization), every 3 months (ODI, PROMIS-29, falls and falls efficacy, PROMIS global health, VR-12), or at 12 months (medications, global impression of change). Health care utilization was assessed as numbers of office visits, emergency room visits, hospitalizations, opioid prescriptions, and back injections in both intervention arms. At the final study visit, participants were asked to evaluate the perceived value of the various treatments that they received, from 0 = not helpful to 10 = extremely helpful.

For the 9-month and 12-month data collection time points, data were collected on a subset of participants. The clinical trial was conducted during the peak of the COVID-19 pandemic and recruitment was often challenging, with recruitment extending beyond the planned period. The final participants in the trial were asked to volunteer to be observed for only six months, as it was the primary study end point. This is reflected in Figure 1, which annotates some participants as "optional."

Statistical analysis. Prior data and assumptions include a between-subject SD of 18 points for pre- to postintervention change in the primary outcome ODI[39] and a conservative dropout rate of 15% from other back pain trials.[40] With 330 participants randomized and 280 anticipated completers, we sought be able to detect an observed primary outcome difference as small as seven points with 90% power in a two-tailed test at the $\alpha = 0.05$ level. Due to the aforementioned challenges due to the COVID-19 pandemic, we were able to randomize only 299 participants,

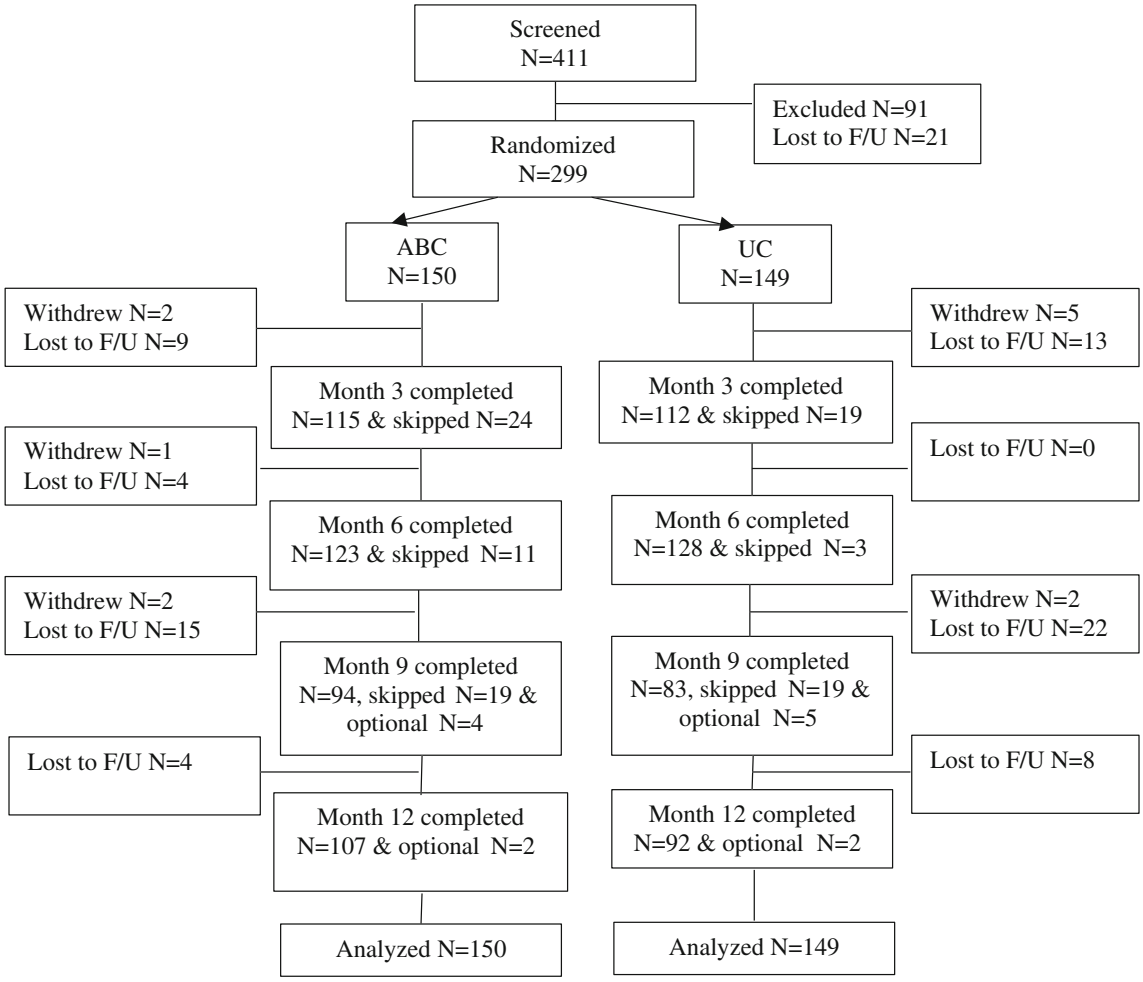

Figure 1. Study participant flow. ABC, Aging Back Clinic; F/U, follow-up; UC, usual care.

which was estimated to afford 87% statistical power under the same assumptions.

We used appropriate descriptive statistics to summarize participant characteristics and outcomes by intervention group and time point. Participant baseline characteristics were compared between intervention arms using independent samples *t*- and chi-square tests, as appropriate. The main analyses were conducted on an intention-to-treat basis and multiple imputation for missing data in continuous variables from the first six months. Specifically, for continuous outcomes (including the primary outcome), we fitted a series of linear mixed models with change from baseline as the dependent variable; intervention group, follow-up time point, and their interaction as fixed effects of interest; baseline value of the outcome as a fixed effect covariate; and participant as a random effect. Means contrasts were constructed to estimate the adjusted between-group difference at each follow-up time point. For count outcomes representing health care utilization and falls, we fitted negative binomial regression models with logarithm of reporting months as an offset (exposure) and intervention group as the independent variable, and we estimated incident rate ratios to quantify the intervention effects. SAS software version 9.4 (SAS Institute, Inc.) was used for statistical analysis.

RESULTS

Participant flow is shown in Figure 1. A total of 411 veterans were screened and 299 were randomized. Supplemental Table 1 lists reasons for exclusion. All follow-up data were collected by telephone. There were no reportable adverse events.

Participant baseline characteristics are displayed in Table 1. Mean age was approximately 73 years, most participants were male, more than two-thirds were White, approximately a quarter were Black, and majority had attended at least some college. There were no significant between-group differences, with the exception of a greater percentage of participants in the ABC group reporting refreshing sleep and more participants in the UC group with sleep disturbance and worse QMCI performance on the clock drawing and logical memory tests.

Table 1. Participant baseline characteristics*

Participant characteristic	ABC (n = 150)	UC (n = 149)	*P* value[a]
Age	73.5 ± 5.1	73.3 ± 5.0	0.7956
Male gender	136 (90.7)	143 (96.0)	0.0663
Hispanic ethnicity	11 (7.3)	14 (9.4)	0.5194
Race			0.5653
Asian	0 (0.0)	3 (2.0)	
Black	41 (27.3)	41 (27.5)	
Native American	2 (1.3)	2 (1.3)	
Pacific Islander	1 (0.7)	1 (0.7)	
White	105 (70.0)	99 (66.4)	
Unspecified	1 (0.7)	3 (2.0)	
Education			0.4162
High school or less	50 (33.6)	40 (26.7)	
At least some college	75 (50.3)	85 (56.7)	
Graduate/professional degree	24 (16.1)	25 (16.7)	
Oswestry Disability Index	30.5 ± 14.7	33.1 ± 15.3	0.1239
Duke Comorbidity Index	3.4 ± 1.3	3.6 ± 1.3	0.2165
Cardiac	35 (23.3)	33 (22.2)	0.8068
Neurologic	14 (9.3)	22 (14.8)	0.1490
Musculoskeletal	140 (93.3)	136 (91.3)	0.5043
General	90 (60.0)	91 (61.1)	0.8494
Visual/hearing	114 (76.0)	122 (81.9)	0.2126
Diabetes	44 (29.3)	49 (32.9)	0.5070
Cancer	39 (26.0)	43 (28.9)	0.5795
Lung	35 (23.3)	40 (26.9)	0.4836
Minimum data set			
Height, in.	68.5 ± 3.4	69.5 ± 3.3	0.0119
Weight, lb	209.2 ± 42.7	212.4 ± 38.2	0.4930
Body mass index	31.5 ± 6.5	31.0 ± 5.4	0.5009
Average pain	5.6 ± 1.8	5.6 ± 1.9	0.1213
Pain duration 5+ y	118 (78.7)	120 (80.5)	0.6882
Pain every day	133 (88.7)	128 (85.9)	0.4736
Leg pain	55 (36.7)	52 (34.9)	0.7499
Bothersome stomach pain	111 (74.0)	110 (73.8)	0.9726
Bothersome joint pain	33 (22.0)	28 (18.8)	0.4913
Bothersome headache	109 (72.7)	97 (65.1)	0.1576
Bothersome widespread pain	76 (50.7)	79 (53.0)	0.6838
Prior low back pain treatments			
Opioid painkiller use (ever used)	63 (42.0)	57 (38.3)	0.5089
Current opioid use	29 (19.3)	33 (22.2)	0.5484
Injections (ever used)	45 (30.0)	45 (30.2)	0.9697
Exercise therapy (ever used)	98 (65.3)	94 (63.1)	0.6854
Psychological counseling (ever used)	18 (12.0)	19 (12.8)	0.8435
Unemployed 1+ mo	6 (4.0)	11 (7.4)	0.2066
Disability/worker's compensation	10 (6.7)	15 (10.1)	0.2882
Felt worthless (sometimes/often/always)	21 (14.0)	21 (14.1)	0.9813
Felt helpless (sometimes/often/always)	23 (15.3)	22 (14.8)	0.8907
Felt depressed (sometimes/often/always)	29 (19.3)	35 (23.5)	0.3809
Felt hopeless (sometimes/often/always)	12 (8.0)	15 (10.1)	0.5329
Sleep quality (good/very good)	52 (34.7)	50 (33.6)	0.8396
Refreshing sleep (somewhat/quite a bit/very much)	106 (70.7)	86 (57.7)	0.0195
Problems with sleep (somewhat/quite a bit/very much)	83 (55.3)	92 (61.7)	0.2605
Difficulty falling asleep (somewhat/quite a bit/very much)	49 (32.7)	60 (40.3)	0.1721
Unsafe to be physically active (fear-avoidance belief)	39 (26.0)	43 (28.9)	0.5795
Pain is terrible and never going to get better (catastrophizing)	55 (36.7)	61 (40.9)	0.4484
Drinking/drug use	12 (8.0)	18 (12.1)	0.2403
Need to reduce drinking	11 (7.3)	13 (8.7)	0.6579
Smoking status:			0.5145
Never smoker	48 (32.0)	40 (26.9)	
Current smoker	19 (12.7)	24 (16.1)	
Quit smoking	83 (55.3)	85 (57.1)	
Pain medications	103 (68.7)	100 (67.1)	0.7737
Opioid pain medications	17 (11.3)	14 (9.4)	0.5827

(Continued)

Table 1. (*Cont'd*)

Participant characteristic	ABC (n = 150)	UC (n = 149)	*P* value[a]
MOS social support			
Emotional/informational	76.3 ± 24.3	75.7 ± 23.5	0.8344
Tangible	78.4 ± 25.6	78.6 ± 24.4	0.9360
Affectionate	75.3 ± 24.9	72.1 ± 26.7	0.2869
Positive social interaction	78.6 ± 24.2	75.5 ± 24.6	0.2808
Overall	77.2 ± 23.7	76.0 ± 23.3	0.6605
Prescribed opioids difficulties scale			
Psychological problems	0.4 ± 2.2	0.7 ± 3.0	0.2415
Opioid control concerns	0.3 ± 1.7	0.5 ± 2.2	0.3333
Overall	0.7 ± 2.8	1.2 ± 4.2	0.1684
Opioids helpful (at least a little)	21 (14.0)	25 (16.8)	0.5055
MMSE	28.6 ± 1.6	28.5 ± 1.6	0.6186
Pain scale			
At the moment	4.0 ± 2.7	4.2 ± 2.6	0.4736
Prior week average	5.7 ± 2.0	6.0 ± 2.0	0.2691
Worst	8.0 ± 1.8	8.1 ± 2.1	0.7005
Pain self-efficacy scale	44.6 ± 11.6	43.5 ± 12.6	0.4574
PROMIS-29			
Physical function	39.2 ± 7.2	39.0 ± 7.5	0.8491
Anxiety scale	46.8 ± 8.4	48.1 ± 8.9	0.1716
Depression scale	46.7 ± 7.7	46.9 ± 7.7	0.8679
Fatigue	51.6 ± 8.8	51.1 ± 9.6	0.6559
Sleep disturbances	51.1 ± 7.6	53.0 ± 8.4	0.0419
Social roles	50.1 ± 8.7	48.7 ± 9.8	0.2027
Pain interference	59.0 ± 6.4	59.7 ± 7.6	0.4421
Average pain	5.6 ± 1.8	5.9 ± 1.9	0.2232
PROMIS Global Health			
General health	3.3 ± 0.9	3.0 ± 0.9	0.3561
Roles	3.3 ± 0.9	3.4 ± 1.1	0.7132
Physical health	41.4 ± 6.0	41.6 ± 7.3	0.7220
Mental health	48.8 ± 8.3	49.2 ± 78.7	0.7061
QMCI			
Orientation	10.0 ± 0.3	9.9 ± 0.4	0.6143
Word registration	4.7 ± 0.6	4.6 ± 0.6	0.1109
Clock drawing	13.3 ± 2.4	12.5 ± 2.9	0.0186
Delayed recall	13.1 ± 5.2	12.6 ± 5.3	0.4940
Verbal fluency	14.1 ± 5.0	14.3 ± 5.0	0.7972
Logical memory	19.1 ± 5.3	17.6 ± 5.7	0.0190
Total score	74.2 ± 12.1	71.5 ± 12.5	0.0647
VR-12			
Physical summary	35.2 ± 10.0	35.2 ± 10.5	0.9613
Mental summary	53.8 ± 10.2	53.6 ± 9.8	0.8579
Gait speed, m/s	0.72 ± 0.25	0.70 ± 0.23	0.3233
Life Space Assessment	69.5 ± 27.3	69.0 ± 27.5	0.8939
Falls Efficacy Scale	12.3 ± 5.5	12.7 ± 5.9	0.5466
Falls prior 3 mo			
Any	40 (26.7)	46 (30.9)	0.4072
Multiple	23 (15.3)	25 (16.8)	0.6953

* Values are in mean ± SD or n (%). Computed using independent samples *t* tests for continuous variables, and chi-square and Fisher's exact tests for categorical variables. ABC, Aging Back Clinic; MMSE, mini-mental state examination; MOS, Medical Outcomes Study; PROMIS, Patient-Reported Outcomes Measurement Information System; QMCI, Quick Mild Cognitive Impairment; UC, usual care; VR-12, Veterans RAND 12 Item Health Survey.
[a] Using independent samples *t*- or chi-square tests.

Table 2 and Figure 2 display change of the ODI, the primary outcome. There were no significant differences in ODI change (the primary outcome) at any of the time points. Significant differences in several of the secondary outcome changes were observed, as displayed in Supplemental Table 2. ABC group showed greater improvement in the PROMIS-29 physical function scale at 12 months (1.7 vs –0.4 points), PROMIS Global general health at 12 months (–0.2 vs –0.4), PROMIS Global physical health scale at 6 (1.3 vs –1.2) and 12 months (0.7 vs –1.5), PROMIS Global mental health at 6 months (0.2 vs –2.3), and present and prior week average/worst pain over 12 months (all $P < 0.05$; Figure 2).

Table 3 displays effect on count outcomes representing health care utilization and falls. There were more office visits in

Table 2. Continuous primary outcome changes from baseline by intervention group*

Outcome and period of change, Oswestry Disability Index	ABC (n = 150), mean ± SD	UC (n = 149), mean ± SD	Adjusted difference,[a] mean ± SE	*P* value
3 mo	−0.4 ± 13.0	2.7 ± 11.3	−2.7 ± 1.6	0.0982
6 mo[b]	−0.8 ± 12.0	1.2 ± 11.5	−2.0 ± 1.5	0.1820
9 mo	1.3 ± 14.7	−0.2 ± 12.6	−0.3 ± 1.7	0.8738
12 mo	0.2 ± 14.0	2.8 ± 12.1	−2.8 ± 1.6	0.0844

* ABC, Aging Back Clinic; UC, usual care.
[a] Adjusted for preadministration value of the physiologic measure as a covariate using linear mixed models and multiple imputation for missing data.
[b] Primary outcome and endpoint.

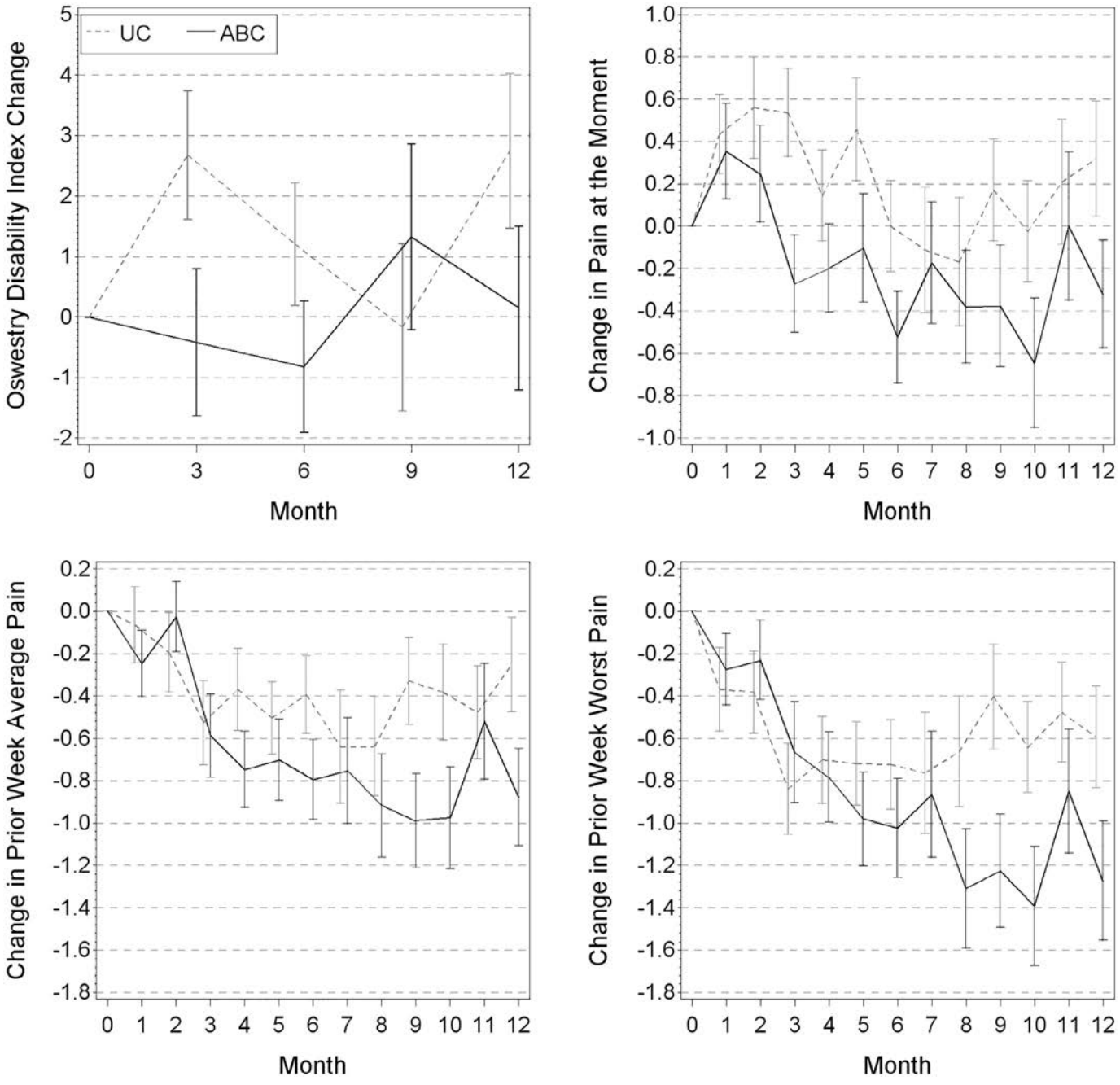

Figure 2. Changes in pain and related disability between usual care (UC) group and aging back clinic (ABC) group.

Table 3. Effect on healthcare utilization outcome (6–12 months) and falls (over 12 months)*

Outcome	ABC, rate per person-month	UC, rate per person-month	Incident rate ratio[a] (95% CI)	*P* value
Office visits	2.30	1.78	1.59 (1.05-2.41)	0.0287
Emergency room visits	0.08	0.06	1.78 (0.64-4.95)	0.2658
Hospital visits	0.03	0.02	1.59 (0.52-4.84)	0.4172
Opioid prescriptions	0.01	0.02	0.74 (0.23-2.35)	0.6064
Back injections	0.00	0.01	0.24 (0.01-10.5)	0.4561
Falls	0.23	0.36	0.39 (0.15-1.01)	0.0527

* ABC, Aging Back Clinic; CI, confidence interval; UC, usual care.
[a] Estimated using a negative binomial regression model unless otherwise noted.

the ABC group (2.3 vs 1.78 per person-month; $P = 0.0287$) and marginally significantly fewer falls (0.23 vs 0.36 per person-month; $P = 0.0527$).

Supplemental Table 3 describes the conditions identified at baseline by the ABC provider and the services to which participants were triaged for the length of their study participation. Each condition screened was present in at least 20% participants, with the exception of fibromyalgia (present in 2.7%) and leg length discrepancy (present in 11.3%). One or more central nervous system (CNS) conditions (anxiety, depression, fibromyalgia, insomnia, maladaptive coping) were present in 67.3% of participants. One or more physical conditions (hip osteoarthritis, lateral hip/thigh pain, leg length discrepancy, lumbar spinal stenosis, myofascial pain, sacroiliac joint syndrome) was identified in 90% of participants. Both CNS and physical conditions were present in 64.7% of participants. Services most often referred to varied by condition, and overall were behavioral health, physical therapy, hip injection, and acupuncture. Supplemental Table 4 repeats Table 2 and Supplemental Table 2 analyses with ABC group restricted to participants that were highly compliant with treatment recommended by the ABC provider (assessed by the ABC provider at the final study visit), showing a strengthening of the magnitudes of intervention effect.

Supplemental Table 5 displays data on the frequency of various treatments received during the course of the study and participants perceived value of these treatments. For the majority of treatments queried, more participants in the ABC group received treatment than those in the UC group. For shots in the spine/muscle and massage, more participants in the UC group received these treatments than those in the ABC group. Although a similar percentage of participants were taking opioids in the ABC group (19.3%) and the UC group (17.4%), their perceived value was greater in the UC group (mean rating 6.6) as compared with the ABC group (mean rating 5.3). For the acupuncture and two education treatments, their perceived value was more than one point greater in the ABC group (6.3, 4.9, 5.2) than in the UC group (4.4, 3.5, 3.4). For yoga, its perceived value was greater in the UC group (6.8) than in the ABC group (4.5). For each of the other treatments, their perceived value was similar in both groups.

DISCUSSION

Older veterans with CLBP who were treated using a personalized, algorithm-guided approach did not show confirmatory evidence of greater improvement in CLBP-associated disability as compared with those who received UC. Several secondary outcomes including at-the-moment and prior week pain and physical health demonstrated statistically significant evidence of greater improvement, and physical health improvement was also clinically significant based on a two-point change, the lower end of the published range.[41] Also, the fall rate demonstrated somewhat weaker evidence of greater improvement in the ABC group compared to the UC group.

We did not demonstrate evidence of a greater improvement in our primary outcome, the ODI, in contrast to other clinical trials that have focused on older adults with CLBP. One trial tested the efficacy of 12 weeks versus 36 weeks of chiropractic spinal manipulation and exercise in 182 older adults with back pain for an average of 18 weeks and demonstrated significant reduction in ODI scores in both groups.[42] A small study (n = 31) conducted in older women with low back pain for at least 3 months of the prior 12 (mean low back pain duration ~8 years) demonstrated efficacy of core stability training with significant improvement in the ODI.[43] Another small study (n = 20) that compared two types of physical therapy in older adults with a mean low back pain duration of 12 to 16 months also demonstrated significant ODI improvement immediately after an eight-week intervention.[44] However, we do also note the preliminary evidence of efficacy with respect to other secondary outcomes and strengthening of the intervention effects in the per-protocol post hoc analysis as aspects of consistency with prior studies.

The reason for difference between our findings and those of others is likely multifaceted. In each of the above studies, the frequency and content of the interventions were standardized, but ours were not. Although our algorithms recommend a specific triage path for participants with contributing conditions, the content, frequency, and duration of the treatments themselves were left to the discretion of treating clinicians and collaborative discussions with participants with a view toward pragmatism. In addition, participants were referred to different clinicians and therapists with variable skill sets. Participant compliance (eg, with

treatment session attendance and/or follow-through with a home program) also was very likely quite variable.

Approximately 80% of our participants reported back pain duration of at least five years. In two of the three studies referenced above, low back pain duration was significantly shorter, and it may have influenced treatment response.[42,44] Although one of the pilot studies included participants with a mean back pain duration of eight years, its small sample size (n = 31) and restriction to only female participants prevent a direct comparison with our study. Another study had 63 participants, women with CLBP randomized to Pilates twice a week for 10 weeks or to a home exercise program. This study demonstrated significant improvement in the well-validated Roland Morris Disability Questionnaire (RMDQ), but only 3 of 63 participants reported a back pain duration of at least 12 months.[45]

Participants' duration of back pain in two fully powered efficacy trials was similar to ours, thus permitting more direct comparison of study findings. Morone and colleagues conducted a study of mindfulness meditation in 282 older adults with CLBP and demonstrated significant improvement in pain, but not in the RMDQ.[40] Karp and colleagues conducted a randomized controlled trial of older adults with comorbid CLBP and depression in which participants were randomized to either pharmacotherapy and supportive therapy or pharmacotherapy and problem-solving therapy.[46] Although there was no between-groups difference in outcomes, both groups improved significantly on the RMDQ.[46] The findings highlight the importance of treating key nonphysical factors in ameliorating CLBP-associated disability in older adults and the value of delivering structured, intensive treatment toward demonstrating an impact on outcomes.

Another factor that potentially threatened our ability to demonstrate a significant impact of ABC care on pain-associated disability was that UC was our control group. Because pain initiatives have been implemented in the VA health care system nationally, usual pain care in the VA is likely distinct from that in general community settings. In 2009 the Veterans Health Administration (VHA) published a directive (2009-053) that included a stepped-care model for managing pain in veterans, spanning from foundational services (ie, pain self-management) to primary care services (step 1), to specialty pain care services (step 2), to tertiary pain care services (step 3). Subsequent research in veterans with chronic pain demonstrated the efficacy of this model in reducing pain-related disability, pain interference, and pain severity compared with UC as it existed at the time, that is, care not guided by stepped-care management.[47] In 2013 the VHA launched another major pain initiative, the Opioid Safety Initiative, that was designed to mitigate high risk prescribing behavior. Because of such initiatives, it is likely that UC in our clinical trial incorporated more robust pain management strategies than that in non-VA settings.

The impact of ABC care on pain intensity and falls is worth highlighting. Participants in the ABC group reported less present and average pain over the prior week at 6 and 12 months, and marginally fewer falls over the 12 months. The link between pain and falls has been demonstrated by others.[48] It has been previously demonstrated that a decline in the PROMIS-Global physical health summary score is associated with increased falls,[49] and it is noteworthy that our study participants randomized to ABC care had greater improvement in the PROMIS-Global physical health summary score than those randomized to UC.

The only difference between groups in health care utilization were more office visits in the ABC group than the UC group. This was likely related to the multiple visits associated with many pain interventions in the ABC group, such as acupuncture, chiropractic, physical therapy, and behavioral therapy. It should be highlighted that utilization of health care services was determined by self-report, which may be less accurate than direct query of electronic medical records, an assessment modality that was not employed in our study. Future studies should consider including more distal measurement of health care utilization (eg, emergency room visits, performance of invasive procedures, opioid prescriptions) as a way to assess an intervention's ultimate impact.

Our findings regarding the prevalence and perceived value of various treatments received are worth highlighting. The prevalence of referral to many pain treatments was greater in the ABC group than the UC group. A key aspect of providing care for patients with chronic pain is active self-management. It is noteworthy that participants randomized to UC received passive intervention modalities (ie, spinal injections and massage) more often than did those in the ABC group. In addition, it is noteworthy that participants in the UC group rated the perceived value of opioids higher (mean 6.6 on a 10-point scale) than those randomized to ABC care (mean 5.3 on a 10-point scale). As we only queried participants about the perceived value of various treatments at the end of the study, we cannot determine the impact of ABC care itself on modifying the perceived value of opioids. It is conceivable that exposure of the ABC participants to more education on the limited role of opioids and the value of nonopioid treatments (per Supplemental Table 5) played a role and should be taken into consideration in future studies.

ABC participants rated the perceived value of two educational components more highly than UC participants—education about causes of their back pain by the doctor (4.9 in ABCs versus 3.5 in UC), and the educational booklet on how to think about pain differently (5.2 in ABCs versus 3.4 in UC). Ample evidence supports the value of education in treating patients with CLBP,[50] and practitioners should evaluate the quality of educational materials provided before their distribution.

Although our study's sample size, geographic diversity, and incorporation of a comprehensive biopsychosocial approach were strengths, its limitations also should be highlighted. Conducting a study exclusively in veterans naturally limits potential for generalizability to nonveterans. Also, even though ours was the first substantial randomized controlled trial conducted using

our algorithms, we used a pragmatic approach. Typically, initial efficacy trials test an intervention that is administered using a rigorous protocol, often overtreating to demonstrate an effect. The nature of our multifaceted intervention, however, did not allow for such an approach. Although beneficial in terms of ease of implementation and application of study findings to the care of veterans in real-world settings, our intervention's pragmatic nature likely required a significantly larger sample size as well as strategies to promote and measure compliance to demonstrate a significant effect on ODI. As noted above, however, others who have implemented protocolized interventions also have been unable to demonstrate an impact on CLBP-associated disability.

Given the value of pragmatism as well as protocolization, perhaps future clinical trials might consider implementing elements of both, for example, a structured physical activity program and/or protocolized treatment of key conditions (eg, depression, per the study of Karp et al) in addition to algorithm-guided care.[46] Investigators may also wish to consider restricting participants to those with only one or two key disability-causing conditions. Another limitation of our study is the subjective measure of treatment adherence (ie, provider assessed). A future large study may wish to consider an objective assessment of adherence (eg, attendance of visits to referred providers through collection of electronic medical record data). Additionally, we had a large number of secondary outcomes, and the efficacy with respect to those should be interpreted with caution and as preliminary findings. Given the burgeoning of older adults worldwide, additional research targeting this vulnerable population with CLBP is needed.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Fang confirms that all authors have provided the final approval of the version to be published, and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Kerns RD, Otis J, Rosenberg R, et al. Veterans' reports of pain and associations with ratings of health, health-risk behaviors, affective distress, and use of the healthcare system. J Rehabil Res Dev 2003; 40(5):371–379.
2. National Center for Veterans Analysis and Statistics. U.S. Department of Veterans Affairs. Accessed June 1, 2016. https://www.va.gov/vetdata/Veteran_Population.asp
3. Selim AJ, Fincke G, Berlowitz DR, et al. Comprehensive health status assessment of centenarians: results from the 1999 large health survey of veteran enrollees. J Gerontol A Biol Sci Med Sci 2005;60(4): 515–519.
4. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. J Pain 2014; 15(6):569–585.
5. Rundell SD, Sherman KJ, Heagerty PJ, et al. Predictors of persistent disability and back pain in older adults with a new episode of care for back pain. Pain Med 2017;18(6):1049–1062.
6. Weiner DK, Rudy TE, Morrow L, et al. The relationship between pain, neuropsychological performance, and physical function in community-dwelling older adults with chronic low back pain. Pain Med 2006;7(1):60–70.
7. Finan PH, Goodin BR, Smith MT. The association of sleep and pain: an update and a path forward. J Pain 2013;14(12):1539–1552.
8. By the 2023 American Geriatrics Society Beers Criteria® Update Expert Panel. American Geriatrics Society 2023 updated AGS Beers Criteria® for potentially inappropriate medication use in older adults. J Am Geriatr Soc 2023;71(7):2052–2081.
9. Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. JAMA 2008;299(6): 656–664.
10. Hicks GE, Morone N, Weiner DK. Degenerative lumbar disc and facet disease in older adults: prevalence and clinical correlates. Spine 2009; 34(12):1301–1306.
11. Jarvik JJ, Hollingworth W, Heagerty P, et al. The longitudinal assessment of imaging and disability of the back (LAIDBack) study: baseline data. Spine 2001;26(10):1158–1166.
12. Qaseem A, Wilt TJ, McLean RM, et al; Clinical Guidelines Committee of the American College of Physicians. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2017; 166(7):514–530.
13. Karayannis NV, Sturgeon JA, Chih-Kao M, et al. Pain interference and physical function demonstrate poor longitudinal association in people living with pain: a PROMIS investigation. Pain 2017;158(6):1063–1068.
14. Inouye SK, Studenski S, Tinetti ME, et al. Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept. J Am Geriatr Soc 2007;55(5):780–791.
15. Resnick NM, Marcantonio ER. How should clinical care of the aged differ? Lancet 1997;350(9085):1157–1158.
16. Weiner DK, Fang M, Gentili A, et al. Deconstructing chronic low back pain in the older adult--step by step evidence and expert-based recommendations for evaluation and treatment: part I: Hip osteoarthritis. Pain Med 2015;16(5):886–897.
17. Polsunas PJ, Sowa G, Fritz JM, et al. Deconstructing chronic low back pain in the older adult – step by step evidence and expert-based recommendations for evaluation and treatment: part X: sacroiliac Joint Syndrome. Pain Med 2016;17(9):1638–1647.
18. Karp JF, DiNapoli EA, Wetherell J, et al. Deconstructing chronic low back pain in the older adult – step by step evidence and expert-based recommendations for evaluation and treatment: part IX: anxiety. Pain Med 2016;17(8):1423–1435.
19. DiNapoli EA, Craine M, Dougherty P, et al. Deconstructing chronic low back pain in the older adult—step by step evidence and expert-based recommendations for evaluation and treatment. Part V: maladaptive coping. Pain Med 2016;17(1):64–73.
20. Lisi AJ, Breuer P, Gallagher RM, et al. Deconstructing chronic low back pain in the older adult - step by step evidence and expert-based recommendations for evaluation and treatment: part II: myofascial pain. Pain Med 2015;16(7):1282–1289.
21. Carley JA, Karp JF, Gentili A, et al. Deconstructing chronic low back pain in the older adult - step by step evidence and expert-based recommendations for evaluation and treatment. Part IV: depression. Pain Med 2015;16(11):2098–2108.

22. Fatemi G, Fang MA, Breuer P, et al. Deconstructing chronic low back pain in the older adult--Step by step evidence and expert-based recommendations for evaluation and treatment part III: fibromyalgia syndrome. Pain Med 2015;16(9):1709–1719.
23. Rho M, Camacho-Soto A, Cheng A, et al. Deconstructing chronic low back pain in the older adult – step by step evidence and expert-based recommendations for evaluation and treatment: part VIII: lateral hip and thigh pain. Pain Med 2016;17(7):1249–1260.
24. Bramoweth AD, Renqvist JG, Germain A, et al. Deconstructing chronic low back pain in the older adult - step by step evidence and expert-based recommendations for evaluation and treatment. Part VII: insomnia. Pain Med 2016;17(5):851–863.
25. Weiner DK, Gentili A, Coffey-Vega K, et al. Biopsychosocial profiles and functional correlates in older adults with chronic low back pain: a preliminary study. Pain Med 2019;20(7):1300–1310.
26. Weiner DK, Gentili A, Rossi M, et al. Aging Back Clinics - a geriatric syndrome approach to treating chronic low back pain in older adults: results of a preliminary randomized controlled trial. Pain Med 2020; 21(2):274–290.
27. Weiner DK, Gentili A, Fang MA, et al. Caring for older veterans with chronic low back pain using a geriatric syndrome approach: rationale and methods for the aging back clinics (ABC) trial. Contemp Clin Trials 2020;95:106077.
28. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12(3):189–198.
29. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine 2000; 25(22):2940–2952.
30. Deyo RA, Katrina Ramsey, Buckley DI, et al. Performance of a Patient Reported Outcomes Measurement Information System (PROMIS) short form in older adults with chronic musculoskeletal pain. Pain Med 2016;17(2):314–324.
31. Rigler SK, Studenski S, Wallace D, et al. Co-morbidity adjustment for functional outcomes in community-dwelling older adults. Clin Rehabil 2002;16(4):420–428.
32. Sherbourne CD, Stewart AL. The MOS social support survey. Soc Sci Med 1991;32(6):705–714.
33. Banta-Green CJ, Von Korff M, Sullivan MD, et al. The prescribed opioids difficulties scale: a patient-centered assessment of problems and concerns. Clin J Pain 2010;26(6):489–497.
34. Nicholas MK. The pain self-efficacy questionnaire: taking pain into account. Eur J Pain 2007;11(2):153–163.
35. O'Caoimh R, Gao Y, McGlade C, et al. Comparison of the quick mild cognitive impairment (Qmci) screen and the SMMSE in screening for mild cognitive impairment. Age Ageing 2012;41(5):624–629.
36. Kempen GI, Yardley L, van Haastregt JC, et al. The Short FES-I: a shortened version of the falls efficacy scale-international to assess fear of falling. Age Ageing 2008;37(1):45–50.
37. Peel C, Sawyer Baker P, Roth DL, et al. Assessing mobility in older adults: the UAB Study of Aging Life-Space Assessment. Phys Ther 2005;85(10):1008–1119.
38. Madill E, Samuels R, Newman D, et al. Development of an evaluative, educational, and communication-facilitating app for older adults with chronic low back pain: patient perceptions of usability and utility. Pain Med 2019;20(11):2120–2128.
39. Delitto A, Piva SR, Moore CG, et al. Surgery versus nonsurgical treatment of lumbar spinal stenosis: a randomized trial. Ann Intern Med 2015;162(7):465–473.
40. Morone NE, Greco CM, Moore CG, et al. A mind body program for older adults with chronic low back pain: a randomized clinical trial. JAMA Intern Med 2016;176(3):329–337.
41. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 2021;30(10):2729–2754.
42. Maiers M, Hartvigsen J, Evans R, et al. Short-or long-term treatment of spinal disability in older adults with manipulation and exercise. Arthritis Care Res (Hoboken) 2019;71(11):1516–1524.
43. Ge L, Huang H, Yu Q, et al. Effects of core stability training on older women with low back pain: a randomized controlled trial. Eur Rev Aging Phys Act 2022;19(1):10.
44. Kim J-H, Park H-Y. Effects of combined hip exercise and passive stretching on muscle stiffness, pain perception and pain related disability, and physical function in older adults with low back pain. Phys Act Nutr 2022;26(3):16–24.
45. Lytras D, Iakovidis P, Sykaras E, et al. Effects of a tailored mat-Pilates exercise program for older adults on pain, functioning, and balance in women with chronic non-specific low back pain: a randomized controlled trial. Aging Clin Exp Res 2023;35(12):3059–3071.
46. Karp JF, Gao X, Wahed AS, et al. Effect of problem-solving therapy versus supportive management in older adults with low back pain and depression while on antidepressant pharmacotherapy. Am J Geriatr Psychiatry 2018;26(7):765–777.
47. Bair MJ, Ang D, Wu J, et al. Evaluation of stepped care for chronic pain (ESCAPE) in veterans of Iraq and Afghanistan conflicts: a randoized clinical trial. JAMA Intern Med 2015;175(5):682–689.
48. Leveille SG, Jones RN, Kiely DK, et al. Chronic musculoskeletal pain and the occurrence of falls in an older population. JAMA 2009; 302(20):2214–2221.
49. Allen J, Alpass FM, Stephens CV. The sensitivity of the MOS SF-12 and PROMIS® global summary scores to adverse health events in an older cohort. Qual Life Res 2018;27(8):2207–2215.
50. Piano L, Audasso P, Benzi L, et al. Individual education for patients with chronic low back pain: likely a clinically relevant effect for long-term disability compared to noneducational interventions. A systematic review with meta-analysis. J Orthop Sports Phys Ther 2025; 55(5):331–343.

American College of Rheumatology
Empowering Rheumatology Professionals

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 620–629
DOI 10.1002/acr.25673

REVIEW ARTICLE

Climate Change, Environmental Risk Factors, and Gout

Mary L. Guan,[1,2] Tamiko R. Katsumoto,[3] Michael Toprover,[1,2] Nidaa Rasheed,[4,5] Shuchi Anand,[6] Michael H. Pillinger,[1,2] Michael H. Weisman,[3] and Suzanne R. Tamang[3,7]

Worldwide, gout is increasing at a rapid rate. Although genetics and diet play primary roles in gout development, emerging evidence suggests that environmental risk factors may also play a significant contributory role. In this review, we examine the evidence linking environmental exposures to gout risk, summarize potential pathophysiologic mechanisms, and highlight key knowledge gaps and underexplored areas. In particular, we highlight the impact of air pollution, ambient temperature, water contamination with heavy metals, chronic kidney disease (as it relates both to gout and climate change), dietary factors such as ultraprocessed food consumption, and various other pollutants in increasing the risk of gout.

Introduction

The prevalence of gout, the most common inflammatory arthritis, has increased alarmingly. Globally, the total gout population doubled from 1990 to 2020.[1] A recent forecasting model estimated that gout cases will increase from 55.8 million in 2020 to 95.8 million in 2050, which is related to population growth but also rising rates of obesity, metabolic syndrome, chronic kidney disease (CKD), and other comorbidities.[1] Gout is characterized by monosodium urate (MSU) crystal deposition in the joints; acute flares represent an inflammatory response to deposited MSU crystals. Hyperuricemia is the primary risk factor for gout, and gout incidence increases with serum urate (SU) levels.[2] Both overproduction of urate and renal underexcretion of uric acid due to CKD or medications, promote hyperuricemia.

Although genetics and diet are key drivers of gout incidence, mounting evidence suggests that environmental factors play an important role. Climate change, characterized by long-term shifts in temperature leading to extreme weather events, droughts, increasing air pollution, and environmental degradation, is a pressing societal concern and a potential contributor to gout prevalence. Environmental exposures to multiple contaminants, some widespread and persistent, have been implicated. The impacts of climate change and environmental pollution on rheumatic diseases are understudied. A white paper by the American College of Rheumatology reviewing published literature on climate-related effects on rheumatic conditions, along with a scoping review on the topic, found that air pollution was the most studied exposure followed by atmospheric temperature.[3,4] However, no papers were identified focusing on extreme weather events (eg, hurricanes, floods, drought, or wildfires). Systemic lupus erythematosus and rheumatoid arthritis are the most studied diseases to date, with few papers examining gout.[4] Concepts whereby climate change and environmental pollution may provoke inflammatory diseases have been incompletely described; mechanisms may include epigenetic changes, oxidative stress, and the expression of inflammatory cytokines.[3] Here, we review emerging evidence linking climate change and environmental exposures to gout, discuss pathophysiologic mechanisms that may explain how environmental factors can trigger gout (Figure 1 and Table 1), and highlight knowledge gaps deserving further investigation (Supplemental Table 1).

Air pollution and gout

Greenhouse gas emissions (GHGs) from fossil fuels and agricultural sources (eg, methane) drive climate change and contribute to air pollution, which is anticipated to worsen with rising

[1]Department of Medicine, Division of Rheumatology, New York University School of Medicine, New York; [2]New York Harbor Health Care System, New York Campus, US Department of Veterans Affairs, New York; [3]Stanford University School of Medicine, Stanford, California; [4]University of California, San Francisco-Fresno, Fresno; [5]Fresno Healthcare System, US Department of Veterans Affairs, Fresno, California; [6]Center for Tubulointerstitial Kidney Disease, Stanford University, Stanford, California; [7]Center for Population Health Sciences, Stanford University, Stanford, California.

Drs Guan and Katsumoto contributed equally as co-first authors and contributed equally to this work. Drs Weisman and Tamang contributed equally as co-last authors and contributed equally to this work.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25673).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25673.

Address correspondence via email to Mary L. Guan, MD, at mary.guan@nyulangone.org.

Submitted for publication May 2, 2025; accepted in revised form September 29, 2025.

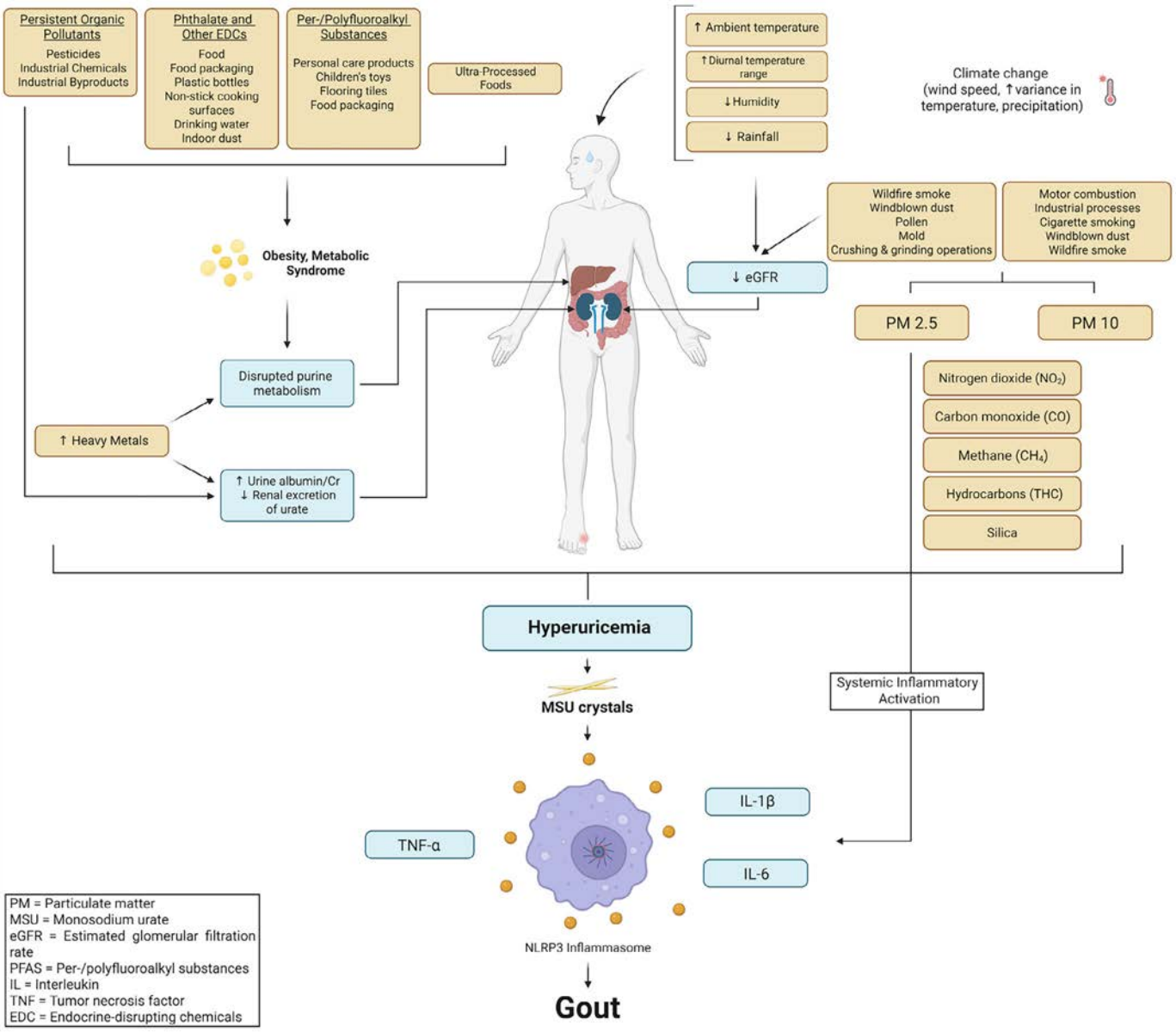

Figure 1. Relationships among environmental perturbations, hyperuricemia, and gout. Environmental exposures—including air pollution, meteorological parameters (ambient temperature, diurnal temperature range, humidity, and rainfall), water pollution, and endocrine-disrupting chemicals—and their associations with hyperuricemia and gout. Potential mechanisms are shown where applicable. EDC, endocrine-disrupting chemical; eGFR, estimated glomerular filtration rate; IL-1β, interleukin-1β; MSU, monosodium urate; PM, particulate matter; TNF-α, tumor necrosis factor-α.

temperatures. To date, most studies on air pollutants and gout have been conducted in Asia where pollution is more severe than in the United States, with concentration limits for particulate matter (PM; 2.5 and 10) in China that are significantly higher than World Health Organization global air quality guidelines.[5] Among air pollutants, PM2.5 may impair regulatory T cells, activate Toll-like receptors, initiate NLRP3 signaling, and stimulate mediators such as interleukin-1α (IL-1α), IL-1β and IL-6, contributing to a proinflammatory state that may lower the activation threshold for gout.[6] A Taiwanese cohort study examined exposure to air pollutants (2000–2011), finding that persons exposed to higher levels of total hydrocarbons, methane, and PM2.5 were more likely to develop incident gout.[7] Interestingly, gout risk was higher in those exposed to hydrocarbons and methane compared with PM2.5, which has been well established to drive systemic inflammatory states. In a study of UK Biobank data of 443,587 individuals enrolled in 2006 to 2010 and followed until 2023, increased concentrations of PM2.5, PM10, nitrogen dioxide (NO_2), and nitrogen oxides were associated with higher risk of incident gout, particularly among older adults.[8] Among Chinese patients with

Table 1. Climate and environmental perturbations and their effects on gout and/or hyperuricemia*

Perturbation	Exposure	Examples of pathways leading to ↑gout and/or ↑hyperuricemia
Temperature	↑ Cold temperatures and diurnal variation	↑Cold → ↑uric acid precipitation → ↑gout
	↑ Hot temperatures	↑Heat →↑dehydration →↑AKI and/or CKDu → ↑gout ↑Heat → ↑drought → ↑wildfires → ↑air pollution → ↑gout ↑Heat → ↑drought → ↑heavy metal concentration and contamination of groundwater → ↑CKD → ↑hyperuricemia and/or gout
Pollution	↑Air pollution ↑Water pollution ↑Chemical pollution	↑PM2.5, PM10, methane, hydrocarbons, or other pollutants → ↑gout ↑Heavy metals, ↑PFAS, ↑pesticides (groundwater contamination) → ↑hyperuricemia ↑Contaminants (eg, EDCs including POPs, PFAS, plastics, and other CECs) → ↑obesity and/or metabolic syndrome → ↑hyperuricemia ↑Emerging contaminants (eg, heavy metals, POPs, chemical fertilizers, and other endocrine disruptors) → ↑CKD → ↑hyperuricemia
Food	Temperature impacts	↑Extreme heat →↑ drought → ↑crop failures and/or food spoilage → ↑UPF consumption → ↑gout flares ↑Extreme weather → need for increased pesticide use → ↑hyperuricemia
	Pollution impacts	↑Environmental pollution (water, air, and land) → issues with food supply and food contamination (including POPs, pesticides, and heavy metals) → ↑hyperuricemia
Lifestyle	Lifestyle impacts	↑Extreme temperatures → ↑sedentary behavior → ↑obesity, metabolic syndrome, T2DM, or hyperuricemia → ↑gout ↑Western diet → ↑UPF consumption, ↑animal product consumption → ↑hyperuricemia and/or gout ↑Extreme weather events → ↑climate anxiety → ↑EtOH consumption → ↑hyperuricemia and/or gout

* AKI, acute kidney injury; CEC, contaminants of emerging concern; CKD, chronic kidney disease; CKDu, chronic kidney disease of uncertain etiology; EDC, endocrine-disrupting chemical; EtOH, ethanol; PFAS, perfluoroalkyl substances; PM, particulate matter; POP, persistent organic pollutant; T2DM, type 2 diabetes mellitus; UPF, ultraprocessed food.

established gout, hospitalizations for gout flares increased with higher NO_2 and carbon monoxide (CO) exposures based on a time-series analysis.[9] In Incheon City, South Korea, increased PM10 levels led to increased risk of emergency department visits for acute gout flares based on a time-series analysis.[10] Importantly, only one study reporting an independent association between air pollutants and incident gout accounted for confounding variables[8]; other studies did not address confounders.[7,9,10] Exposure to inorganic dusts (eg, silica) has also been associated with incident gout.[11]

Even transient exposures to pollutants may lower inflammatory activation thresholds, with implications for gout. One meta-analysis found that days-to-weeks exposure to PM2.5 and/or PM10 was associated with higher blood levels of IL-6 and tumor necrosis factor.[12] Exposure to ozone over just two to seven days can lead to changes in the serum levels of inflammatory cytokines.[13] Additionally, exposures to PM2.5, sulfur dioxide, NO_2, and CO correlated with increased IL-1β levels.[13]

In gout, MSU crystals activate the NLRP3 inflammasome, leading to production of active IL-1β, promoting the inflammatory state seen in gout flares. In theory, exposure to elevated levels of air pollutants could promote gout flares via an increase in baseline IL-1β levels. Studies suggest that PM2.5 may cause both intracellular and extracellular changes in ATP levels and that increased extracellular ATP levels can induce alveolar macrophages to produce reactive oxygen species that activate the NLRP3 inflammasome.[14] In vitro, PM2.5 ingestion by alveolar macrophages triggers inflammasome activation.[15] Once PM2.5 is within the alveolar macrophages, various intracellular signals occur including lysosomal disruption, potassium efflux, mitochondrial dysfunction, and enhanced superoxide anion production. These signals lead to NLRP3 inflammasome activation and IL-1β release.[15] In murine models, pulmonary PM10 exposure also induces NLRP3 inflammasome expression and a proinflammatory state.[16] Air pollution may also cause higher frequencies of other conditions driven by NLRP3 inflammasome activation (eg, calcium pyrophosphate deposition arthropathy and juvenile idiopathic arthritis), but few studies exist.

Ambient temperature and gout

Climate change is leading to an increasing frequency of temperature extremes.[17] Lower temperatures decrease urate solubility, potentially increasing crystal precipitation risk in joints to trigger gout flares.[18] Conversely, exposure to excessive temperature (eg, heat waves) may increase proinflammatory cytokines and activate neutrophils, leading to increased inflammatory disease activity and potentially increased susceptibility to gout. Higher ambient temperatures, greater diurnal temperature ranges, and a sharp drop in temperature between neighboring days have been associated with increased risk of gout flare and hospitalization.[19,20] A combination of high temperature and low humidity had the greatest association with gout flares.[19] In one South Korean prospective study, change in humidity and diurnal change in temperatures were associated with increased risk of gout flare.[21] Higher temperatures may lead to the clinical impact

of dehydration and metabolic acidosis, which can reduce kidney excretion of uric acid and increase SU levels. Additionally, higher temperatures have been associated with decreased renal function (estimated glomerular filtration rate [eGFR]), which is a known risk factor for development of gout.[22] The association of higher temperatures with gout may be multifactorial and include behavioral changes. For example, in one study in Mexico, decreased rainfall and increased temperature, particularly in tropical regions, were shown to correspond to lower consumption of unprocessed and higher consumption of ultraprocessed foods (UPFs), which is associated with gout risk.[23,24] Increased intake of alcohol and fructose-sweetened beverages may also occur on hot days, leading to transient increases of SU levels.

Climate change, CKD, and gout

Gout and CKD are closely linked. Urate is first filtered in the glomeruli, and then 91% to 95% of filtered urate is reabsorbed by the proximal tubule. The remaining 5% to 9% of filtered urate is excreted into the urine, representing two-thirds of total urate elimination. Gain-of-function mutations in urate reabsorption transporters (eg, GLUT1) and loss of function mutations in urate efflux transporters (eg, ABCG2) have been associated with hyperuricemia and gout.[25] Additionally, declining GFR or transient insults such as intravascular volume depletion and hypoperfusion can cause increased urate reabsorption and decreased excretion, leading to hyperuricemia.[26] Approximately 20% of individuals with gout have CKD Stage 3 or greater compared with 5% of those without gout; 15% of individuals with hyperuricemia have CKD Stage 3 or greater compared with 3% of individuals without gout.[27] In a UK cohort of 1,775,505 individuals, 24,768 of whom developed gout over the course of seven years, after adjusting for confounding variables, individuals with gout were more than twice as likely to have CKD Stage 3 or more advanced compared with those without gout.[28] In this same cohort, chronic renal impairment and chronic renal failure were associated with higher risk of incident gout (relative risks of 2.55 and 3.82, respectively), likely due to impaired urate excretion.[28] Reduced creatinine clearance has also been associated with early development of gout-related subcutaneous tophi.

Climate change can have multiple adverse effects on kidney health, with potential implications for gout. Water shortages, drought, and rising ambient temperatures leading to dehydration all increase the likelihood of developing kidney stones, a known risk factor for CKD. Low urine output and increased urinary concentration—worsened by high ambient temperatures – lead to higher concentrations of products that form kidney stones such as calcium oxalate and uric acid. In five US metropolitan areas, higher mean daily temperature was associated with higher risk of kidney stone presentation,[29] particularly when the temperature exceeded 30°C. The "kidney stone belt" in the United States is predicted to shift northward and westward with rising temperatures.[30]

The potential impact of climate change on kidney disease extends beyond stone formation, to what some experts term "heat stress nephropathy." Agricultural workers are especially vulnerable to acute and chronic changes in kidney function from heat exposure. In specific areas of the world, "hotspots" of kidney disease (termed "CKDu," or CKD of uncertain etiology) have been described among agricultural workers, and heat is felt to be a major risk factor for CKDu progression. One study of 105 Guatemalan sugarcane workers who started the harvest season with normal eGFR found a high prevalence of acute cross-shift (preshift versus postshift) declines in eGFR. Increased wet bulb globe temperature, decreased urine pH, and, most strongly, high SU levels were all risk factors for cross-shift declines in kidney function.[31] Similarly, among 283 California agricultural workers in summer 2014, 12.3% developed incident acute kidney injury (AKI) over the course of a work shift.[32] Among Nicaraguan sugarcane workers, more physically demanding shifts were correlated with increased incident kidney injury and decline in eGFR.[33] In Central America, the most consistent risk factor for development of CKDu is working in sugarcane agriculture; a secondary risk factor is working in hot temperatures. As temperatures rise, a greater proportion of agricultural workers may be at risk for developing CKDu.

Heat stress may not be the sole cause of CKDu. Contamination of groundwater and drinking water by heavy metals such as cadmium and arsenic from stormwater runoff and chemical fertilizers, studied in Sri Lanka, is another proposed mechanism.[34] Uric acid handling has been hypothesized to be linked to CKDu in two ways: (1) hyperuricemia may be an early marker of disease in this subgroup, and (2) hyperuricemia may provide a clue to site of injury, with at least one study indicating a link between biopsy-proven CKDu and a genetic polymorphism in a proximal tubular channel.[35]

Increasing evidence suggests that the effects of heat on kidney function, although most evident in highly exposed populations such as agricultural workers, may also have implications for the general population. Among 1,354,675 AKI episodes across England between September and April during the years 2017 to 2021, there was a 62.3% increase in odds of AKI on a day with a maximum temperature[36] of 32°C compared with days with maximum temperatures of 17°C. A seven-day heatwave in July 2021 in England was associated with a 28.6% increase in AKI episodes.[36] Extreme heat and extreme weather patterns are therefore expected to increase CKD incidence. A 9.5-year retrospective cohort study of Swiss primary care patients observed that higher ambient temperatures were associated with lower eGFR across all age groups in an unadjusted analysis.[22] A post hoc analysis of the dapagliflozin CKD study found that within a 120-day window, each 30 days' heat index of more than 30°C was associated with a –0.6% change in eGFR across 4,017

participants in 21 countries.[37] Globally, an increase in CKD from increasing temperatures will likely result in a greater population of persons with hyperuricemia and gout. Although no studies to date have reported on gout incidence in the context of AKI and heat stress nephropathy, such cases may go undetected because the time from hyperuricemia to gout due to heat stress nephropathy could be delayed.

Water pollution and heavy metals and gout

Environmental contamination, particularly water pollution, includes toxic heavy metal exposure to surface water and groundwater due to industrial wastewater, stormwater, and agricultural runoff. Heavy metals found in contaminated water supplies include chromium, cadmium, arsenic, nickel, copper, and lead. Climate change exacerbates the impact of heavy metal contamination in the environment. Rising temperatures accelerate the release of heavy metals from soil, flooding spreads toxic sediments, and drought conditions concentrate heavy metals in water bodies and soils, which enter our drinking water and food supply.[38]

Due to profligate use before an appreciation of its health impact, lead has been ubiquitous in the environment, including the drinking water supply and paints. Many municipalities in the United States continue to rely on lead pipes. Until recently, the city of Flint, Michigan had lead pipes and high population blood lead levels, which were compounded when the city decided to use water from the untreated Flint River, whose corrosive properties caused additional lead to leach out of pipes. Lead release from water pipes is increased at higher water temperatures due to increased solubility of lead stabilizers at higher temperatures.[39]

Historically, lead toxicity–associated gout has been described in the setting of lead-induced CKD ("saturnine gout"). Even before declines in eGFR are noted, lead interferes with the kidney's ability to excrete uric acid, suggesting a defect in kidney tubular secretion due to direct tubular damage, tubulointerstitial fibrosis, and increased production of reactive oxygen radicals.[40,41] Lead also interferes with the metabolism of purines, leading to accumulation of guanine in tissues and promoting MSU crystal formation.[41] Using the National Health and Nutrition Examination Survey (NHANES) database, studies have shown that blood levels of lead lower than the current US national standard (<1.21 umol/L) can be associated with increased gout prevalence and hyperuricemia.[42] An observational study in China found that among individuals living in a lead-polluted area between 2013 and 2014, blood lead levels were positively associated with higher prevalence of hyperuricemia, particularly among female participants.[43] A cross-sectional study using the NHANES database (2011–2018) found that increased blood levels of mercury, lead, and selenium were associated with gout-related outcomes.[44]

Blood levels of other heavy metals (eg, cadmium) have not been associated with increased gout prevalence. Nonetheless, other heavy metal exposures can worsen renal function and hyperuricemia. Among adults >60 years old in the NHANES database, high blood concentrations of lead, cadmium, and cobalt were all associated with elevated urinary albumin/creatinine ratio.[45] A South Korean study of 2,682 participants found a dose-response relationship between hyperuricemia and elevated blood lead and cadmium levels, with exposure differences observed between sexes.[46] Similarly, an analysis of the NHANES database showed that arsenic, cadmium, and lead were independent risk factors for hyperuricemia.[47] In the United States, the percentage of community water systems with average arsenic levels above the standard level was 2.3% in 2009 to 2011. In a large cross-sectional study of 5,632 participants from 2003 to 2010 using the NHANES database, low-level arsenic exposure was associated with higher risk of hyperuricemia in men and prevalence of gout in women.[48]

Pesticides can also be found in soil or surface water and be transferred to groundwater; climate change–induced temperature and rainfall patterns can influence pesticide distribution and degradation.[49] A study of avian physiologic parameters reported a positive correlation between pesticide exposure and SU levels.[50] The proposed mechanism was thought to involve glomerular and tubular cell damage, with studies indicating an inverse correlation between serum organic pollutant levels and eGFR. Similar changes in humans could lead to impaired uric acid clearance and chronic nephrotoxicity.[51] Additionally, organic pollutants may influence uric acid metabolism by promoting reactive oxygen species generation, which suppresses the antioxidant defense system. Overall, these findings are preliminary, and further research is needed to establish a clear understanding of the underlying pathophysiology.

Dietary factors and gout

Dietary choices directly impact climate.[52,53] Conversely, dietary factors also play a prominent role affecting hyperuricemia and the risk of gout. Significant associations of gout with red meat, seafood, alcohol, and fructose have long been observed, corroborated by a recent meta-analysis.[54] Interestingly, vegetables with a high purine content showed a negative association with gout,[54] and plant-based diets also have much lower GHGs and other environmental impacts (eg, water usage, eutrophication),[52] suggesting a potential virtuous cycle. In prospective cohort studies, vegetarian diets have been associated with a lower risk of gout.[55] One recent study showed that the healthy plant-based diet index was inversely associated with gout, in contrast with the unhealthy plant-based diet index.[56] In countries such as China, diets rich in fat and animal-based foods have outstripped plant-based diets, mirroring the rapid increase in gout prevalence.[57]

UPFs are energy-dense, industrially prepared foods containing refined grains, added sugars, or fats. UPFs may also contain preservatives, emulsifiers, and other chemicals that may or may not be reflected in the packaging label. Although cheaper per calorie than minimally processed foods (eg, fruit, vegetables, whole grains, meat, and fish), UPFs are associated with intensive agriculture practices such as deforestation and pesticide and plastic use and have been estimated to have greater impact on GHGs.[58] In the United States, the consumption of UPFs increased by 3.4% between 2001 and 2018, with the largest increases for ready-to-eat meals. The trend was especially prominent among older adults (aged ≥60 years) and was found across all income groups.[59] In an analysis from 2017 to 2018, 58% of total daily energy intake in US adults came from UPFs.[60] A prospective study of 18,444 adults in China found that increased UPF consumption was an independent risk factor for hyperuricemia.[61] In a population-based cohort study using UK Biobank data, high UPF consumption was associated with a 16% increased risk for developing gout, particularly among those with a genetic predisposition, in whom the risk increased 1.9-fold.[24] Replacing 20% of UPF in daily intake led to 13% lower gout risk.[24] The mechanism of the UPF effect is unknown but could reflect biologic changes, lifestyle choices, and higher consumption of fructose. In one Italian cohort of 21,315 patients, higher UPF intake was associated with increased low-grade inflammation.[62] The authors opined that inflammation is not only related to the proinflammatory capacity of UPF but also due to nonnutritional factors such as cosmetic additives, food contact materials, and degradation of the food matrix.

Fructose is a major constituent of UPFs. The consumption of fructose rose dramatically in the 1970s with the substitution of high-fructose corn syrup (HFCS) for sucrose (cane sugar) in prepared foods. HFCS is stable and inexpensive compared with sucrose and is a regular additive to UPFs that are not recognized as sweetened by the consumer. Studies have shown that the increased prevalence of hyperuricemia and gout in developed countries paralleled the increase in consumption of total fructose and HFCS.[63] Biologically, phosphorylation of fructose facilitates ATP depletion and leads to increased urate levels.[64] Acute intake of fructose in humans leads to increased degradation of purine nucleotides and increased purine synthesis, thus also contributing to hyperuricemia; however, it is unclear if these acute effects are sustained long term.[65]

A meta-analysis examining the impact of UPFs on health outcomes found that the highest UPF consumption was associated with a significant increase[66] in risk of overweight and/or obesity (by 39%) and metabolic syndrome (by 79%). Obesity and weight gain are both linked to hyperuricemia and gout, and HFCS is a known dietary risk factor for hyperuricemia.[67] However, the exact linkages between metabolic syndrome, type 2 diabetes, and gout remain unclear. In a Mendelian randomization analysis study, genetically predicted fasting insulin levels led to significant increases in SU level and gout risk.[68] Higher levels of insulin resistance may also reduce renal excretion of urate, leading to hyperuricemia. However, a different study found that a diagnosis of diabetes was associated with lower future risk of gout, independent of other risk factors.[69]

Persistent organic pollutants and other contaminants of emerging concern

Persistent organic pollutants (POPs) can remain in the environment for decades after production has ceased. POPs include pesticides such as organochlorine pesticides (OCPs), industrial chemicals such as polychlorinated biphenyls (PCBs) and perfluoroalkyl and polyfluoroalkyl substances (PFAS, also known as "forever chemicals"), and industrial byproducts such as dioxins and polyaromatic hydrocarbons (PAHs). POPs are resistant to biodegradation and have high lipophilicity, leading to bioaccumulation in fat and biomagnification up the food chain. Consequently, consumption of high-fat animal-derived foods is a major source of exposure.[70] PFAS exposure can occur via food packaging, plastic bottles, nonstick cooking surfaces, as well as drinking water and indoor dust. Although some POPs are banned under the Stockholm Convention global environmental treaty, many developing nations have only recently begun to restrict their use. Certain POPs (such as PFAS) are still widely used and remain prevalent as unintentional byproducts of industrial processes. Climate change itself (due to increased variance in temperature, wind speed, and precipitation) can increase the transport and distribution of POPs—in particular, temperature increases volatilization and impacts partitioning among soil, sediment, water, and atmosphere.[71] Most POPs are endocrine disruptors: prior studies have reported positive relationships between POP concentrations and type 2 diabetes, insulin resistance, hypertension, and metabolic syndrome.[72] POP concentrations have been correlated with unhealthy metabolic phenotypes, not only in obese and overweight individuals but also in normal-weight individuals.

Although no studies to date have directly investigated POPs and gout, a number have examined the link between POPs and hyperuricemia. The risk of hyperuricemia was higher in US individuals with higher serum concentrations of POPs (OCPs, polychlorinated dibenzodioxins, and dioxin-like PCBs).[73] A systematic review and meta-analysis similarly found the risk of hyperuricemia to be associated with three particular PFAS compounds as well as PCBs, DDT, and its breakdown product dichlorodiphenyldichloroethylene.[74] These observations were substantiated by cohort studies from Spain[75] and China (only women).[76] Kidney function decline may mediate this effect.[77] Using causal mediation analysis, a Chinese study examining 2003 to 2018 NHANES data found that decreased eGFR may mediate between PFAS exposure and the risk of resultant hyperuricemia,[77] with a mediated proportion ranging from 19% to 57%. The biologic mechanism

remains unknown—perhaps accumulation in the systemic circulation may lead to kidney injury and impaired uric acid excretion.

PAHs are ubiquitous environmental contaminants that are listed as top pollutants in the substance priority lists by the Agency for Toxic Substances and Disease Registry.[78] PAHs generated from both natural sources (eg, forest fires) and anthropogenic activities (eg, incomplete combustion of fossil fuels, cigarette smoking, and high-temperature home cooking) can lead to exposure via air inhalation, diet, and dermal contact. PAH exposure has been associated with hyperuricemia in a study of 12,287 US adults in the NHANES registry[79] as well as a study of 4,812 residents in China.[80]

A number of contaminants of emerging concern (CECs) are not typically monitored but may pose known or suspected threats to humans and the environment. For example, phthalates are synthetic endocrine-disrupting chemicals that are considered CECs and are used to improve durability of polyvinyl chloride plastics. Phthalates can be found in many daily necessities, including personal care products, food packaging, floor tiles, and children's toys. The global market of phthalic anhydride (the principal component in phthalate manufacture) is rapidly increasing, especially in developing countries, for construction and automotive purposes.[81] Phthalates can be absorbed orally (from food packaging), by inhalation (volatilization from flooring), and dermally (cosmetics).[82] Studies using NHANES data have found a linear relationship between urinary phthalate metabolites and risk of hyperuricemia.[83] Alteration of the systemic immune-inflammation index, an inflammatory index including serum markers such as platelets, lymphocytes, and neutrophils, may play a role in mediating risk of hyperuricemia with phthalate exposure.[83] Mounting evidence suggests that bisphenols, a group of industrial compounds used in plastics, are also linked with hyperuricemia, possibly mediated through renal structural damage and impairment of liver purine metabolism.[83] There are many other CECs deserving of further evaluation, which may contribute to hyperuricemia and risk for gout.

Discussion and future research priorities

The rising global incidence and prevalence of gout is both a public health problem and a tocsin for the adverse impact of environmental threats. The interactions between climate change, environmental pollutants, and lifestyle are complex, interdependent, and often synergistic. Here, we have reviewed the available data identifying key issues for consideration (Supplemental Table 1). In addition, we propose a research agenda to further our understanding of the impacts of climate change and environmental exposures on gout, focusing on defining pathogenic mechanisms, promoting surveillance activities, especially among vulnerable populations, and identifying mitigations at the individual and population levels (Supplemental Table 2).

Despite the well-documented health harms of climate change and environmental contaminants, their impact on gout has been inadequately explored. Existing research has primarily focused on air pollution and extreme heat, with much of the literature centering on hyperuricemia rather than gout itself. However, critical knowledge gaps remain (Supplemental Table 2). It is unclear which specific component(s) of PM2.5 pose the greatest risk to patients with gout and what types of exposures—such as acute exposures from wildfires versus prolonged, lower-level exposure—are most clinically significant. Other potential environmental factors, such as ambient temperature, diet, and water pollution, have been less frequently examined, independently or in combination with PM, highlighting opportunities for future research. No significant evidence to date was found linking extreme weather events, rural air pollutants, or indoor air pollutants to the development or exacerbation of gout. However, a substantial body of research has established the relationship between CKD, dehydration, and extreme temperatures, underscoring the impact of environmental stressors on kidney function.

Several illustrative examples highlight the interplay between climate change, environmental pollution, and lifestyle factors (Figure 1). More frequent droughts driven by climate change may concentrate heavy metal pollutants in groundwater, increasing exposure risk and potentially triggering more gout cases. Concurrently, the rising incidence of obesity and metabolic syndrome, largely driven by sedentary lifestyles, Western diets rich in meats and UPFs, and exposure to endocrine disruptors, may further accelerate the incidence of hyperuricemia and gout. Climate change–driven extreme weather events could exacerbate these trends by promoting more sedentary behavior and increasing reliance on UPFs due to food scarcity from crop failures, spoilage, and contamination. Moreover, the global rise of CKD, which is closely linked to gout, remains an area of concern. The exact etiology of CKDu remains poorly understood but likely involves a combination of heat stress and environmental toxins.[33,34]

Environmental pollutants represent a largely unexamined threat in the context of gout. The US Environmental Protection Agency has identified more than 85,000 chemicals under the Toxic Substances Control Act,[84] with an estimated 350,000 chemical substances registered for large-scale production and use in recent decades.[84] Although much attention has been directed toward well-known contaminants, these may represent only the tip of the iceberg. The potential role of various air pollutants (eg, PM2.5 and PM10), water contaminants, food additives, and other chemical exposures in hyperuricemia and gout remains underexplored, particularly regarding their underlying mechanisms.

Several limitations in the existing research on environmental factors and gout deserve mention. Much of the literature remains in its early stages, with many studies focusing on association rather than causal mechanisms. Although some evidence links air pollution to gout, many studies emphasize hyperuricemia, an

outcome that does not always translate to clinical gout, as only approximately 20% of patients with hyperuricemia develop the disease.[85] Additionally, much of the current research relies on cross-sectional studies, making it difficult to infer causal relationships. Larger prospective cohort studies, time-series analyses, and population-based research in the United States could provide more definitive insights. Future studies should prioritize gout-specific outcomes rather than using hyperuricemia as a surrogate.

Accurate exposure assessment of environmental factors poses several methodologic challenges, including regional and temporal variations in environmental factors. In observational studies, the critical time window during which an environmental exposure is most likely to trigger a gout flare remains uncertain, complicating lagged exposure modeling and causal inference. Additionally, many environmental exposures are highly correlated with socioeconomic factors, making it difficult to disentangle independent effects. Persons of low socioeconomic status may live closer to highways or factories or have poor ventilation in their homes, increasing their risk of exposure to air pollution. They may be unable to seek shelter on hotter days, leading to exposure to higher ambient temperatures. Food deserts are more common in lower-income communities, limiting supply of fresh and healthy food options and promoting consumption of UPF. Finally, low-income communities may have less access to clean drinking water, walkable areas, and community gyms, causing sedentary lifestyles and increasing risk of toxic exposures. Future research should leverage high-resolution spatiotemporal data to refine exposure assessments and develop electronic health record–based flare detection algorithms. Robust study designs, such as case-crossover and time-series approaches, will help minimize bias from individual-level confounders and improve causal inference. In addition, with the advent of multiomic technologies, systems epidemiology approaches can help identify linkages between environmental exposures and gout, probing specific metabolomic, transcriptomic, proteomic, and/or epigenomic signatures that may underlie these associations (Supplemental Table 2).

Clinicians need to recognize the increasing incidence of gout in the face of these evolving environmental threats as urate-lowering therapy is used in fewer than half of patients with gout.[85] In addition, lifestyle approaches such as whole-food, plant-centered diets should be promoted as measures that would have salutary effects on gout as well as on planetary health.[56] Identifying environmental risk factors for gout with more precision will allow for more effective clinical risk mitigation strategies and public awareness campaigns regarding the links between climate change, environmental exposures, and gout. More importantly, such data could inform advocacy efforts for greater regulatory oversight of environmental pollutants and for policy change to support human and planetary health.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Guan confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Cross M, Ong KL, Culbreth GT, et al; GBD 2021 Gout Collaborators. Global, regional, and national burden of gout, 1990-2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2024;6(8):e507–e517.
2. Campion EW, Glynn RJ, DeLabry LO. Asymptomatic hyperuricemia. Risks and consequences in the Normative Aging Study. Am J Med 1987;82(3):421–426.
3. Dellaripa PF, Sung LH, Bain PA, et al; American College of Rheumatology Committee on Research. American College of Rheumatology white paper: the effects of climate change on rheumatic conditions-an evolving landscape and a path forward. Arthritis Rheumatol 2024;76(10):1459–1466.
4. Katsumoto TR, Stolyar L, Dandeniya CL, et al. Impact of climate change on rheumatic diseases: a scoping review. J Clim Change Health 2024;19:100338.
5. Goshua A, Akdis CA, Nadeau KC. World Health Organization global air quality guideline recommendations: executive summary. Allergy 2022;77(7):1955–1960.
6. Mpakosi A, Cholevas V, Tzouvelekis I, et al. Autoimmune diseases following environmental disasters: a narrative review of the literature. Healthcare (Basel) 2024;12(17):1767.
7. Hu WS, Lin CL. Effect of air pollution on gout development: a nationwide population-based observational study. QJM 2021;114(7):471–475.
8. Li A, Zhang Q, Zhou L, et al. Long-term exposure to ambient air pollution and incident gout: a prospective cohort study in the UK Biobank. Environ Pollut 2024;345:123540.
9. He Y-S, Wang G-H, Wu Q, et al. The relationship between ambient air pollution and hospitalizations for gout in a humid subtropical region of China. J Inflamm Res 2021;14:5827–5835.
10. Ryu HJ, Seo MR, Choi HJ, et al. Particulate matter (PM10) as a newly identified environmental risk factor for acute gout flares: a time-series study. Joint Bone Spine 2021;88(2):105108.
11. Sigurdardottir V, Jacobsson L, Schiöler L, et al. Occupational exposure to inorganic dust and risk of gout: a population-based study. RMD Open 2020;6(2):e001178.
12. Tang H, Cheng Z, Li N, et al. The short- and long-term associations of particulate matter with inflammation and blood coagulation markers: a meta-analysis. Environ Pollut 2020;267:115630.
13. Kim JH, Woo HD, Lee JJ, et al. Association between short-term exposure to ambient air pollutants and biomarkers indicative of inflammation and oxidative stress: a cross-sectional study using KoGES-HEXA data. Environ Health Prev Med 2024;29(0):17.
14. Zeng X, Liu D, Wu W, et al. PM2.5 exposure inducing ATP alteration links with NLRP3 inflammasome activation. Environ Sci Pollut Res Int 2022;29(17):24445–24456.
15. Caceres L, Abogunloko T, Malchow S, et al. Molecular mechanisms underlying NLRP3 inflammasome activation and IL-1β production in

air pollution fine particulate matter (PM2.5)-primed macrophages. Environ Pollut 2024;341:122997.
16. Marín-Palma D, Tabares-Guevara JH, Taborda N, et al. Coarse particulate matter (PM10) induce an inflammatory response through the NLRP3 activation. J Inflamm (Lond) 2024;21(1):15.
17. Blackport R, Fyfe JC. Amplified warming of North American cold extremes linked to human-induced changes in temperature variability. Nat Commun 2024;15(1):5864.
18. Ahn H, Lee G, Lee G-S. Lower temperatures exacerbate NLRP3 inflammasome activation by promoting monosodium urate crystallization, causing gout. Cells 2021;10(8):1919.
19. Neogi T, Chen C, Niu J, et al. Relation of temperature and humidity to the risk of recurrent gout attacks. Am J Epidemiol 2014;180(4): 372–377.
20. He Y-S, Wang G-H, Wu Z-D, et al. Association between non-optimal temperature and hospitalizations for gout in Anqing, China: a time-series analysis. Environ Sci Pollut Res Int 2022;29(10):13797–13804.
21. Choi HJ, Moon KW, Kim HO, et al. Seasonal variations and associated factors of gout attacks: a prospective multicenter study in Korea. J Korean Med Sci 2020;35(20):e133.
22. Schietzel S, Zechmann S, Valeri F, et al. Ambient temperature and kidney function in primary care patients. J Nephrol 2024;37(1):95–105.
23. López-Olmedo N, Diez-Roux AV, Pérez-Ferrer C, et al. Climate trends and consumption of foods and beverages by processing level in Mexican cities. Front Nutr 2021;8:647497.
24. Zhang T, Xu X, Chang Q, et al. Ultraprocessed food consumption, genetic predisposition, and the risk of gout: the UK Biobank study. Rheumatology (Oxford) 2024;63(1):165–173.
25. Benn CL, Dua P, Gurrell R, et al. Physiology of hyperuricemia and urate-lowering treatments. Front Med (Lausanne) 2018;5:160.
26. Maesaka JK, Fishbane S. Regulation of renal urate excretion: a critical review. Am J Kidney Dis 1998;32(6):917–933.
27. Zhu Y, Pandya BJ, Choi HK. Comorbidities of gout and hyperuricemia in the US general population: NHANES 2007–2008. Am J Med 2012; 125:679–687.e1.
28. Cea Soriano L, Rothenbacher D, Choi HK, et al. Contemporary epidemiology of gout in the UK general population. Arthritis Res Ther 2011; 13(2):R39.
29. Tasian GE, Pulido JE, Gasparrini A, et al; Urologic Diseases in America Project. Daily mean temperature and clinical kidney stone presentation in five U.S. metropolitan areas: a time-series analysis. Environ Health Perspect 2014;122(10):1081–1087.
30. Brikowski TH, Lotan Y, Pearle MS. Climate-related increase in the prevalence of urolithiasis in the United States. Proc Natl Acad Sci USA 2008;105(28):9841–9846.
31. Sorensen CJ, Butler-Dawson J, Dally M, et al. Risk factors and mechanisms underlying cross-shift decline in kidney function in Guatemalan sugarcane workers. J Occup Environ Med 2019;61(3):239–250.
32. Moyce S, Mitchell D, Armitage T, et al. Heat strain, volume depletion and kidney function in California agricultural workers. Occup Environ Med 2017;74(6):402–409.
33. Hansson E, Glaser J, Weiss I, et al. Workload and cross-harvest kidney injury in a Nicaraguan sugarcane worker cohort. Occup Environ Med 2019;76(11):818–826.
34. Wimalawansa SJ. Escalating chronic kidney diseases of multi-factorial origin in Sri Lanka: causes, solutions, and recommendations. Environ Health Prev Med 2014;19(6):375–394.
35. Nanayakkara S, Senevirathna ST, Abeysekera T, et al. An integrative study of the genetic, social and environmental determinants of chronic kidney disease characterized by tubulointerstitial damages in the North Central Region of Sri Lanka. J Occup Health 2014;56(1):28–38.
36. Hajat S, Casula A, Murage P, et al. Ambient heat and acute kidney injury: case-crossover analysis of 1 354 675 automated e-alert episodes linked to high-resolution climate data. Lancet Planet Health 2024;8(3):e156–e162.
37. Zhang Z, Heerspink HJL, Chertow GM, et al. Ambient heat exposure and kidney function in patients with chronic kidney disease: a post-hoc analysis of the DAPA-CKD trial. Lancet Planet Health 2024;8(4): e225–e233.
38. Xiao W, Zhang Y, Chen X, et al. The easily overlooked effect of global warming: diffusion of heavy metals. Toxics 2024;12(6):400.
39. Zhang Y, Lin YP. Leaching of lead from new unplasticized polyvinyl chloride (uPVC) pipes into drinking water. Environ Sci Pollut Res Int 2015;22(11):8405–8411.
40. Baki AE, Ekiz T, Öztürk GT, et al. The effects of lead exposure on serum uric acid and hyperuricemia in young adult workers: a cross-sectional controlled study. Arch Rheumatol 2016;31(1):71–75.
41. Dalvi SR, Pillinger MH. Saturnine gout, redux: a review. Am J Med 2013;126(5):450.e1–450.e8.
42. Krishnan E, Lingala B, Bhalla V. Low-level lead exposure and the prevalence of gout: an observational study. Ann Intern Med 2012;157(4): 233–241.
43. Dai H, Huang Z, Deng Q, et al. The effects of lead exposure on serum uric acid and hyperuricemia in Chinese adults: a cross-sectional study. Int J Environ Res Public Health 2015;12(8):9672–9682.
44. Xu J, Zhu X, Hui R, et al. Associations of metal exposure with hyperuricemia and gout in general adults. Front Endocrinol (Lausanne) 2022; 13:1052784.
45. Pan S, Niu Y, Duan S, et al. Uric acid mediates the relationship between mixed heavy metal exposure and renal function in older adult people. Front Public Health 2024;12:1403878.
46. Jung W, Kim Y, Lihm H, et al. Associations between blood lead, cadmium, and mercury levels with hyperuricemia in the Korean general population: a retrospective analysis of population-based nationally representative data. Int J Rheum Dis 2019;22(8):1435–1444.
47. Ma Y, Hu Q, Yang D, et al. Combined exposure to multiple metals on serum uric acid in NHANES under three statistical models. Chemosphere 2022;301:134416.
48. Kuo C-C, Weaver V, Fadrowski JJ, et al. Arsenic exposure, hyperuricemia, and gout in US adults. Environ Int 2015;76:32–40.
49. Saha P, Paul B. Assessment of heavy metal toxicity related with human health risk in the surface water of an industrialized area by a novel technique. Hum Ecol Risk Assess 2019;25(4):966–987.
50. Sonne C, Bustnes JO, Herzke D, et al. Relationships between organohalogen contaminants and blood plasma clinical-chemical parameters in chicks of three raptor species from Northern Norway. Ecotoxicol Environ Saf 2010;73(1):7–17.
51. Siddharth M, Datta SK, Bansal S, et al. Study on organochlorine pesticide levels in chronic kidney disease patients: association with estimated glomerular filtration rate and oxidative stress. J Biochem Mol Toxicol 2012;26(6):241–247.
52. Willett W, Rockström J, Loken B, et al. Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. Lancet 2019;393(10170):447–492.
53. Willett WC, Hu FB, Rimm EB, et al. Building better guidelines for healthy and sustainable diets. Am J Clin Nutr 2021;114(2):401–404.
54. Li R, Yu K, Li C. Dietary factors and risk of gout and hyperuricemia: a meta-analysis and systematic review. Asia Pac J Clin Nutr 2018; 27(6):1344–1356.
55. Chiu THT, Liu C-H, Chang C-C, et al. Vegetarian diet and risk of gout in two separate prospective cohort studies. Clin Nutr 2020;39(3): 837–844.
56. Rai SK, Wang S, Hu Y, et al. Adherence to healthy and unhealthy plant-based diets and the risk of gout. JAMA Netw Open 2024;7(5): e2411707.

57. Zhu B, Wang Y, Zhou W, et al. Trend dynamics of gout prevalence among the Chinese population, 1990-2019: a joinpoint and age-period-cohort analysis. Front Public Health 2022;10:1008598.
58. Hendrie GA, Baird D, Ridoutt B, et al. Overconsumption of energy and excessive discretionary food intake inflates dietary greenhouse gas emissions in Australia. Nutrients 2016;8(11):690.
59. Juul F, Parekh N, Martinez-Steele E, et al. Ultra-processed food consumption among US adults from 2001 to 2018. Am J Clin Nutr 2022;115(1):211–221.
60. Steele EM, O'Connor LE, Juul F, et al. Identifying and estimating ultra-processed food intake in the US NHANES according to the nova classification system of food processing. J Nutr 2023;153(1):225–241.
61. Zhang T, Gan S, Ye M, et al. Association between consumption of ultra-processed foods and hyperuricemia: TCLSIH prospective cohort study. Nutr Metab Cardiovasc Dis 2021;31(7):1993–2003.
62. Mignogna C, Costanzo S, Di Castelnuovo A, et al; Moli-sani Study Investigators. The inflammatory potential of the diet as a link between food processing and low-grade inflammation: an analysis on 21,315 participants to the Moli-sani study. Clin Nutr 2022;41(10):2226–2234.
63. Jamnik J, Rehman S, Blanco Mejia S, et al. Fructose intake and risk of gout and hyperuricemia: a systematic review and meta-analysis of prospective cohort studies. BMJ Open 2016;6(10):e013191.
64. Herman MA, Birnbaum MJ. Molecular aspects of fructose metabolism and metabolic disease. Cell Metab 2021;33(12):2329–2354.
65. Choi HK, Curhan G. Soft drinks, fructose consumption, and the risk of gout in men: prospective cohort study. BMJ 2008;336(7639):309–312.
66. Pagliai G, Dinu M, Madarena MP, et al. Consumption of ultra-processed foods and health status: a systematic review and meta-analysis. Br J Nutr 2021;125(3):308–318.
67. Singh JA, Gaffo A. Gout epidemiology and comorbidities. Semin Arthritis Rheum 2020;50(3S):S11–S16.
68. Yokose C, McCormick N, Choi HK. The role of diet in hyperuricemia and gout. Curr Opin Rheumatol 2021;33(2):135–144.
69. Rodríguez G, Soriano LC, Choi HK. Impact of diabetes against the future risk of developing gout. Ann Rheum Dis 2010;69(12):2090–2094.
70. Guo W, Pan B, Sakkiah S, et al. Persistent organic pollutants in food: contamination sources, health effects and detection methods. Int J Environ Res Public Health 2019;16(22):4361.
71. Nadal M, Marquès M, Mari M, et al. Climate change and environmental concentrations of POPs: a review. Environ Res 2015;143(Pt A):177–185.
72. Gasull M, Castell C, Pallarès N, et al. Blood concentrations of persistent organic pollutants and unhealthy metabolic phenotypes in normal-weight, overweight, and obese individuals. Am J Epidemiol 2018;187(3):494–506.
73. Lee YM, Bae SG, Lee SH, et al. Persistent organic pollutants and hyperuricemia in the U.S. general population. Atherosclerosis 2013;230(1):1–5.
74. Yu Y, Quan X, Wang H, et al. Assessing the health risk of hyperuricemia in participants with persistent organic pollutants exposure - a systematic review and meta-analysis. Ecotoxicol Environ Saf 2023;251:114525.
75. Arrebola JP, Ramos JJ, Bartolomé M, et al; BIOAMBIENT.ES. Associations of multiple exposures to persistent toxic substances with the risk of hyperuricemia and subclinical uric acid levels in BIOAMBIENT.ES study. Environ Int 2019;123:512–521.
76. Feng Y, Fu M, Guan X, et al. Associations of exposure to perfluoroalkyl substances with serum uric acid change and hyperuricemia among Chinese women: results from a longitudinal study. Chemosphere 2022;308(Pt 3):136438.
77. Niu Z, Duan Z, He W, et al. Kidney function decline mediates the adverse effects of per- and poly-fluoroalkyl substances (PFAS) on uric acid levels and hyperuricemia risk. J Hazard Mater 2024;471:134312.
78. Agency for Toxic Substances and Disease Registry. ATSDR. Accessed March 20, 2025. https://www.atsdr.cdc.gov/index.html
79. Liu Y, Ding X, Yu L, et al. Associations of polycyclic aromatic hydrocarbons exposure with serum uric acid and hyperuricemia in US adults: the role of systemic inflammation. J Environ Sci (China) 2025;150:412–421.
80. Ding X, Liu Y, Wan S, et al. Cross-sectional and longitudinal associations of PAHs exposure with serum uric acid and hyperuricemia among Chinese urban residents: the potential role of oxidative damage. Environ Pollut 2024;360:124664.
81. Benjamin S, Masai E, Kamimura N, et al A. Phthalates impact human health: epidemiological evidences and plausible mechanism of action. J Hazard Mater 2017;340:360–383.
82. Chang W-H, Herianto S, Lee C-C, et al. The effects of phthalate ester exposure on human health: a review. Sci Total Environ 2021;786:147371.
83. Fu W, Guang Y, Zhang Z, et al. Epidemiological perspectives on emerging contaminants and gout or hyperuricemia. Emerging Contam 2025;11(2):100485.
84. Toxic Substances Control Act, 15 USC §2601 (1976).
85. Chen-Xu M, Yokose C, Rai SK, et al. Contemporary prevalence of gout and hyperuricemia in the United States and decadal trends: The National Health and Nutrition Examination Survey, 2007-2016. Arthritis Rheumatol 2019;71(6):991–999.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 630–638
DOI 10.1002/acr.25661

American College of Rheumatology
Empowering Rheumatology Professionals

Diffusion-Weighted Imaging for the Evaluation of the Sacroiliac Joint in Pediatric Patients

Michael L. Francavilla,[1] Timothy G. Brandon,[2] Dmitry Khrichenko,[2] Rui Xiao,[3] Nancy A. Chauvin,[4] Asef Khwaja,[2] and Pamela F. Weiss[2,3]

Objective. Maturational signal in the sacroiliac joint (SIJ) of skeletally immature youth is often misinterpreted as inflammation. Diagnostic tools that improve specificity are greatly needed. Apparent diffusion coefficient (ADC) values from diffusion-weighted imaging (DWI), when used with standard imaging, may enhance diagnostic accuracy. We aimed to define normative pediatric ADC values and establish thresholds to distinguish normal from inflammatory SIJ signals.

Methods. ADC values were measured using circular regions of interest (ROIs) on the anterior, central, and posterior slices of the cartilaginous SIJs (36 total ROIs). Mean ADCs were analyzed by age group, bone (iliac or sacral), and joint height (superior, mid, inferior), accounting for within-patient clustering. In sacroiliitis cases, ROIs were placed on DWI at sites of increased signal on fluid-sensitive sequences. Thresholds differentiating normal and inflammatory signals were derived by age, bone, and joint height (ilium only) and assessed by area under the receiver operating characteristic (AUROC) and specificity.

Results. The reference group included 86 youth. Inferior ilium ADC values were higher than mid and superior regions in all immature age groups (all $P < 0.0001$) and decreased with age ($P = 0.0001$). Sacral ADCs also declined with age ($P = 0.0001$). No age trend was observed in the superior or mid ilium ($P = 0.14$). ADC thresholds distinguished normal from inflammatory signals with AUROC ≥0.90 in most iliac regions, except the peripubertal inferior ilium (AUROC 0.78). Sacral thresholds performed acceptably (AUROC ≥ 0.77), though they were lower in the prepubertal group (AUROC 0.68).

Conclusion. Age- and bone-specific ADC reference values were established and effectively differentiated normal from inflammatory SIJ signals.

INTRODUCTION

The moment sacroiliitis is detected in juvenile spondyloarthritis (enthesitis-related arthritis [ERA] or psoriatic arthritis) is a key therapeutic decision point—axial disease does not respond to first-line disease-modifying agents such as methotrexate and warrants consideration of biologic medications.[1] Up to one-third of children with spondyloarthritis develop sacroiliitis within several years of diagnosis.[2–5] Magnetic resonance imaging (MRI) is the most used imaging modality to diagnose sacroiliitis in children. However, it has variable sensitivity and specificity in everyday practice and is reliant on the experience of the interpreting radiologist. Recent work across larger tertiary care centers in the US demonstrated that when leveraging dedicated pelvic MRI sequences and a central imaging team as a reference standard, sensitivity of local radiologists to identify sacroiliitis is high (>90%), but the specificity (median: 68%, interquartile range [IQR]: 53%–86%) and positive predictive values (median: 57%, IQR: 41%–57%) are moderate and low, respectively.[6] The most common error is misinterpretation of the normal physiologic metaphyseal-equivalent signal as an inflammatory signal.

In pediatric imaging, the signal of the metaphyseal-equivalent part beneath the articular surface may mimic the inflamed bone marrow signal in children with unfused or partially fused sacroiliac

Dr Weiss's work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH (K24-AR-078950).

[1]Whiddon College of Medicine, University of South Alabama, Mobile; [2]Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; [3]Perelman School of Medicine, University of Pennsylvania, Philadelphia; [4]The Cleveland Clinic, Cleveland, Ohio.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25661).

Author disclosures and graphical abstract are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25661.

Address correspondence via email to Pamela F. Weiss, MD, MSCE, at weisspa@chop.edu.

Submitted for publication July 14, 2025; accepted in revised form September 18, 2025.

SIGNIFICANCE & INNOVATIONS

- Apparent diffusion coefficient reference ranges were established by age group along the iliac and sacral sacroiliac joint.
- Cut-offs to differentiate normal signal from pathologic inflammation had excellent and moderate area under the receiver operating characteristics at the ilium and sacrum, respectively.

apophyses; immature bone marrow has higher water content than mature bone marrow.[7] There is no gold standard for the determination of normal versus pathologic inflammatory signal. As such, the reference standard for published studies is the consensus of musculoskeletal radiologists experienced in not only the diagnosis of sacroiliitis but also in the interpretation of the maturing sacroiliac joint (SIJ).[6,8–12] Diffusion-weighted imaging (DWI) is a widely available MRI technique that reflects the diffusion of water within intra- and extra-cellular tissue compartments. The apparent diffusion coefficient (ADC) is a quantitative assessment of directionally averaged water diffusion and can provide a quantitative measure of inflammation. Although not completely objective, ADC from DWI may facilitate interpretation. In adults, ADC has been shown to distinguish active from inactive sacroiliitis and can be followed over time to assess response to treatment.[13–17] Small studies in children have confirmed similar findings.[10,18,19] In adults, the addition of DWI to standard fluid-sensitive sequences significantly enhanced the specificity of the imaging assessment from 67% to 80%, although the authors concluded the addition of DWI did not improve overall accuracy, sensitivity, or confidence in the ascertainment of sacroiliitis.[20] In pediatrics, the issue is not failure to detect sacroiliitis; it is primarily misinterpretation of normal maturational signals for inflammation, so tools that can provide incremental improvements in specificity, not sensitivity, are greatly needed. We hypothesize that ADC from DWI may be a useful tool to improve the specificity in the interpretation of the pediatric SIJ imaging evaluation, in addition to standard sequences, while maintaining the sensitivity. This project aimed to define normative ADC values in a pediatric reference population and establish data-driven thresholds to distinguish normal from inflammatory SIJ signal.

PATIENTS AND METHODS

This was a retrospective cross-sectional study consisting of two phases. In phase 1, standards of ADC interpretation in a reference population of control patients were generated, and in phase 2, empirical data-driven thresholds to differentiate normal and inflammatory signals were developed.

Ethics. This study met criteria for exemption and was granted a waiver of informed consent by the Children's Hospital of Philadelphia Institutional Review Board (20-018172).

Patients. *Reference control group.* The reference control group included 87 youth. Twenty-nine participants were prospectively recruited from a primary care practice as part of a prior imaging study; they had no personal history of back pain, trauma, surgery, or known endocrine, oncologic, or rheumatologic disease.[21] An additional 58 patients were retrospectively identified through a query of the picture archiving and communication system (PACS). These individuals underwent standard MRI, including DWI, as part of routine clinical evaluation for concerns such as stress fractures, infection, or mechanical back pain—but none were ultimately found to have underlying pathology. All imaging studies were rereviewed by a central review team to confirm the absence of abnormalities, ensuring their suitability as controls.

Established sacroiliitis group. The sacroiliitis group included 44 patients. Thirty-four were prospectively recruited from rheumatology clinics at three centers as part of a separate study.[22] All had a clinical diagnosis of sacroiliitis and met criteria for ERA or psoriatic arthritis based on the International League of Associations for Rheumatology (ILAR)[23] or European Spondyloarthropathy Study Group (ESSG)[24] classification criteria. An additional 10 patients with ERA and sacroiliitis were identified retrospectively through a PACS query. In all cases, sacroiliitis was confirmed by clinical assessment along with standard MRI, demonstrating periarticular bone marrow edema-like signal. Demographic information was obtained directly from participants or extracted from the electronic medical record.

Imaging protocol. Oblique coronal small field-of-view T1-weighted, fluid-sensitive STIR sequences and axial DWI sequences of the SIJs were acquired according to either (a) the research study protocol in which the patient enrolled in or (b) the institutional pelvic imaging protocol used for clinical care. Both protocols used three-directional DWI with single-shot spin-echo echo-planar imaging. Diffusion gradient b values varied, with most sequences using 50, 400/500, and 800/1000 s/mm^2. All scans (N = 123) were performed at 3.0 Tesla. MRI manufacturers and models are detailed in Supplementary Tables 1 and 2. Slice thicknesses for T1, STIR, and DWI sequences ranged from 3 to 5 mm.

Imaging interpretation. *Reference control group.* The DWI and STIR images were anonymized and reviewed independently by two experienced pediatric musculoskeletal radiologists using custom software (Parametric MRI; https://www.parametricmri.com/). Both radiologists had expertise in imaging evaluation of sacroiliitis and interpreting imaging of the maturing SIJ. They completed real-time iterative calibration modules to optimize scoring accuracy of SIJ MRI lesions.[25,26] Axial DWI sequences were reformatted by the software into a semicoronal

plane to facilitate comparison with fluid-sensitive STIR sequences and placement of regions of interest (ROIs).

Interpretation standards, terminology, and ROI placements were established through a consensus review of 10 reference control cases. To define normative ADC reference ranges, up to 12 circular ROIs were placed per anterior, central, and posterior slice of the SIJ in each patient, following a method similar to Bozgeyik et al.[13] On each slice, ROIs were placed on both iliac and sacral sides along the cartilaginous SIJ, divided into superior, mid, and inferior portions of the joint (Figure 1). The central coronal slice was selected to include the S1 and S2 vertebral bodies, choosing the slice showing the greatest portion of both if multiple slices qualified. ROIs were omitted if part of the joint was not visible. Each control patient had up to 36 ROIs evaluated.

ROI volumes ranged from 0.15 to 0.22 cm^2 and were placed on subchondral bone marrow, including any apophyseal growth cartilage but excluding the articular bone cortex. All reference control studies were evaluated with the protocol for ROI placement described in the Imaging Protocol section. ADC values were calculated by the software using a best-fit linear logarithmic regression of all b values.

Established sacroiliitis group. For patients with confirmed sacroiliitis, ROIs were identified through two distinct exercises. First, two radiologists independently placed ROIs on areas demonstrating high signal intensity on STIR images, simulating how ADC might be applied in routine clinical practice. Second, the central review team conducted a consensus session to collaboratively place ROIs across all cases. This second step aimed to optimize placement for establishing diagnostic thresholds. The sequence of these exercises was intentionally designed to reduce bias during the independent ROI placement.

In both exercises, subchondral bone marrow and apophyseal cartilage were evaluated for signal intensity and appearance on coronal oblique fluid-sensitive STIR images, along with corresponding signal on the ADC map. ROIs were placed in regions showing increased signal (see Figure 2). Similar to ROI placement in the reference control group, care was taken to avoid areas of erosion, adjacent cortical bone, sclerosis, or fat metaplasia to prevent volume averaging and inaccurate ADC values. For patients with sacroiliitis, multiple ROIs were allowed to capture a representative range of signal intensities. However, placement was limited to one ROI each in the superior, mid, and inferior regions on the right and left iliac and sacral bones for every anterior, central, and posterior slice of the SIJs. The number of ROIs per case ranged from 1 to 18. Sample case in which edema signal was present in the central SIJ imaging plane on the (a) STIR image and (b) the corresponding placement of the ROI on the DWI sequence.

Statistical analysis. Demographics and clinical characteristics were summarized using medians with IQRs or counts with percentages.

Phase 1: Reference limit estimation and group comparisons. Scatter plots and histograms were created to visually inspect the

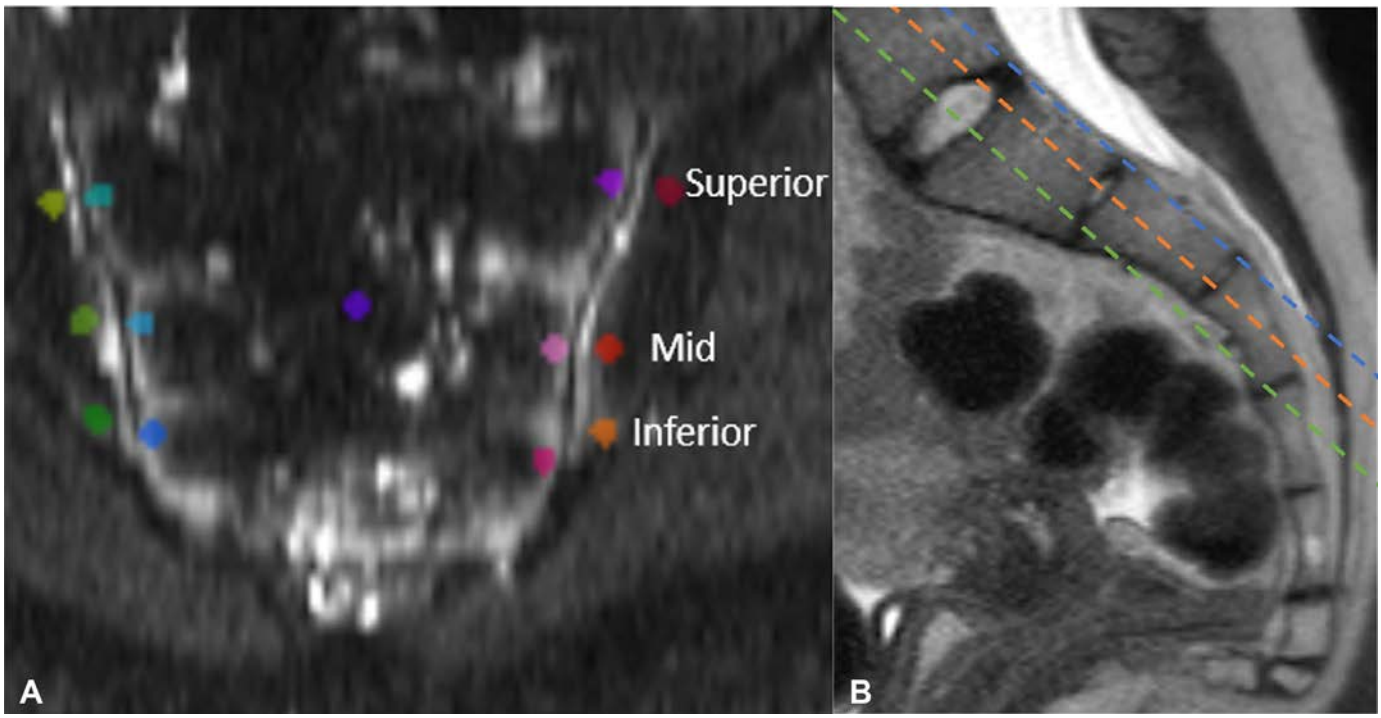

Figure 1. ROI placement on control patients. (a) Sample placement of ROIs in the SIJ on the central slice of the coronal oblique imaging plane. Analogous ROI placement was performed on the SIJs in one anterior slice and one posterior slice. (b) Sagittal view of the SIJ showing the anterior (orange), central (green), and posterior (blue) SIJ acquisition planes. ROI, region of interest; SIJ, sacroiliac joint.

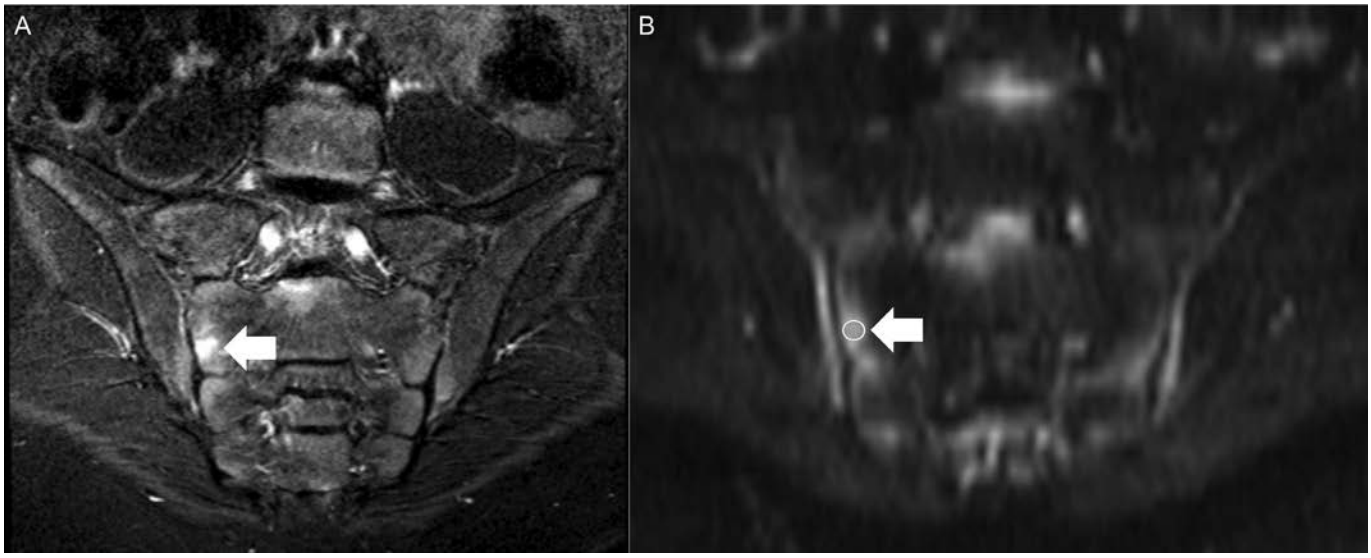

Figure 2. ROI placement on patients with established sacroiliitis. ROI, region of interest.

data. Reference limits (2.5th and 97.5th percentiles) were estimated using the mean ADC from both central reviewers for each ROI by age group (prepubertal: 8–10 years, peripubertal: 11–13 years, postpubertal: 14–16 years, and mature: ≥17 years), bone (iliac or sacral), and joint height (superior, mid, inferior), accounting for within-patient clustering. Interrater agreement for all reference control studies was visualized using Bland–Altman plots, with limits of agreement defined as the mean difference ±1.96 SDs. Differences among ADC values across the inferior, mid, and superior regions of the anterior, central, and posterior slices of the iliac and sacral bones were assessed using the Kruskal–Wallis tests; Dunn's test was performed as a post-hoc procedure following rejection of a Kruskal–Wallis test. A test for trend evaluated changes in ADC values with increasing age.

Phase 2: Threshold selection and performance. Thresholds differentiating normal and inflammatory signals were generated by age, bone, and joint height (ilium only) using the Liu method.[27] This method optimizes the threshold for a diagnostic test using the product of the sensitivity and specificity. Each threshold was assessed by area under the receiver operating characteristic (AUROC), specificity, and sensitivity implemented in an R package "fastAUC."[28] AUROC of less than 0.60 was considered poor; 0.60 to 0.75, moderate; and more than 0.75, excellent discrimination.[29,30] Analyses were conducted using Stata version 17 (StataCorp LLC) and R version 4.4.1.[31]

RESULTS

A total of 87 youth were included in the reference group to develop normative ADC values, whereas an additional 44 youth with confirmed sacroiliitis were used to establish data-driven thresholds for distinguishing normal from inflammatory SIJ signal. The reference control group consisted of youth in the 8 to 10 (n = 23), 11 to 13 (n = 19), 14 to 16 (n = 27), and 17+ (n = 18) years age groups. In comparison to the control population, there were fewer patients who were female (36.4% vs 43.7%), ages 8 to 10 years old (13.6% vs 26.4%), and 17 years old or older (15.9% vs 20.7%) and more patients who were 11 to 13 years old (34.1% vs 21.8%) and 14 to 16 years old (36.4% vs 31.0%), but those differences were not statistically significant.

Establishment of normative ADC values and evaluation of age-related trends. Supplementary Table 3 shows the range and median (IQR) of ADC values in the reference population, grouped by age, bone (iliac or sacral), and joint height (superior, mid, or inferior). In the ilium, ADC values at the superior and mid joint heights did not differ significantly within age groups and were therefore combined. In contrast, the inferior ilium showed significantly higher ADC values than the mid or superior region in the pre-, peri-, and post-pubertal age groups (age < 17 years, all $P < 0.0001$). ADC in the inferior ilium decreased significantly with age ($P = 0.0001$), whereas no significant age-related trend was observed in the combined superior and mid regions ($P = 0.14$). In the sacrum, ADC values did not differ significantly by joint height (superior, mid, inferior), but overall sacral ADC values declined significantly with age ($P = 0.0001$).

Interobserver agreement. Interrater agreement, as assessed by a Bland–Altman plot, is shown in the Supplementary Figure 1. The mean difference in ADC assessments between raters was –36.2 units, which was relatively close to zero given the range of values, suggesting no significant bias between the two raters. The 95% limit of agreement was –719 to 647. The heteroscedasticity or spread of points increased with average rater

ADC values, with a variance for the ROIs with an average rater ADC greater than 800 (18.5% of the ROI data points) over 2.5 times that of the variance for ROIs with an average rater ADC less than 800, indicating less reliable agreement at extreme values.

Threshold selection. Figure 3 shows the distribution of median raw ADC values for the reference and sacroiliitis groups. Table 1 presents empirically derived thresholds for distinguishing normal maturational signals from inflammatory signals, along with their corresponding AUROC, specificity, and sensitivity in two populations: (1) controls plus consensus ROI cases and (2) controls plus the average independent ROI placement cases (performance of the independent rater ROI placements on cases is available in Supplementary Table 4). These thresholds were established using normative values from phase 1 and consensus ROI placement in the sacroiliitis cases.

For the ilium, thresholds achieved AUROC values ≥0.90 across all age groups (pre-, peri-, and post-pubertal) in both the superior/mid and inferior regions—except in the peripubertal group, in which the inferior ilium had a lower AUROC of 0.78. Specificity was ≥0.92 for the superior and mid ilium and ≥0.77 in the inferior ilium across all age groups. In the sacrum, AUROC and specificity were ≥0.77 and ≥0.76, respectively, for all age groups—except in the prepubertal group, in which AUROC dropped to 0.68 and specificity to 0.52.

When thresholds were applied to data from an independent ROI placement exercise (Figure 4)—in which each reviewer independently placed ROIs in areas of high STIR signal to simulate real-world clinical application—the superior and mid ilium showed AUROC values ≥0.84 across all age groups. In the inferior ilium, AUROC values were ≥0.72 across all age groups, except in the prepubertal group, in which the AUROC was 0.68. Notably, specificity values for the ADC thresholds in the ilium in all age groups were equal to or higher than those obtained from the consensus ROI exercise.

DISCUSSION

Differentiating inflammatory sacroiliitis from its mimics based on symptoms and clinical examination alone is challenging. As a result, MRI findings increasingly guide treatment decisions.

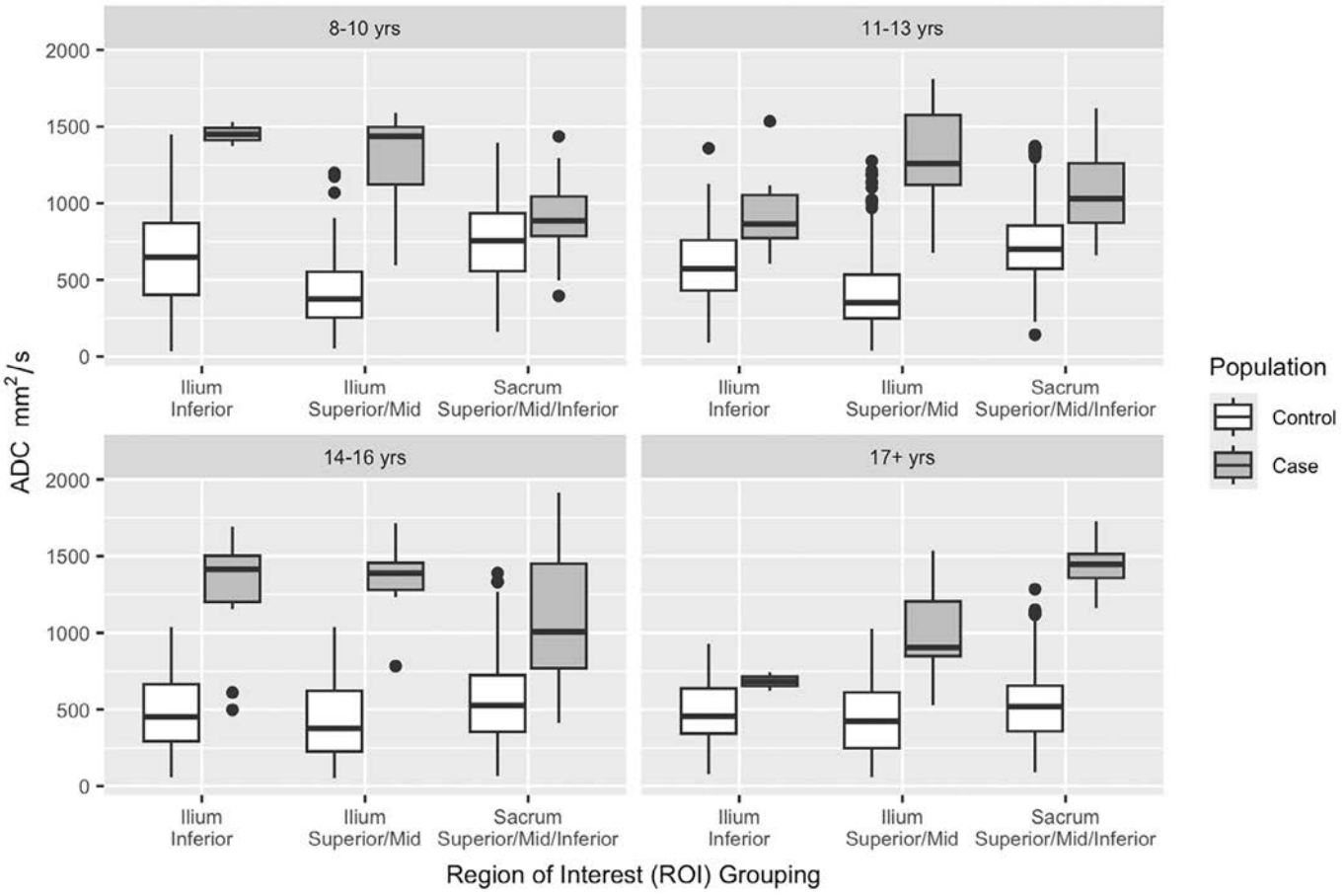

Figure 3. Boxplots of ADC values of controls and patients with sacroiliitis by ROI groupings. ADC, apparent diffusion coefficient; ROI, region of interest.

Table 1. Empiric thresholds and test properties to differentiate normal from inflammatory signal*

ROI location	Age group, y	n[a]	Cut point (95% CI)	Case ROI source	AUROC	Specificity	Sensitivity
Ilium							
Superior/Mid	8–10	260	854 (601–1,271)	Consensus	0.91 (0.84–0.99)	0.97 (0.95–0.99)	0.84 (0.69–1)
				Independent	0.84 (0.63–0.93)	0.99 (0.97–1)	0.69 (0.45–0.93)
Inferior	8–10	86	1,338 (1,332–1,489)	Consensus	0.98 (0.95–1)	0.8 (0.66–0.95)	0.56 (0.19–0.92)
				Independent	0.68 (0.52–0.91)	0.95 (0.88–1.03)	0.4 (0.09–0.71)
Superior/Mid	11–13	222	849 (678–1,195)	Consensus	0.93 (0.84–0.99)	0.92 (0.84–1)	0.95 (0.84–1.05)
				Independent	0.9 (0.83–0.96)	0.93 (0.86–1.01)	0.87 (0.77–0.98)
Inferior	11–13	100	794 (593–967)	Consensus	0.78 (0.62–0.9)	0.78 (0.68–0.88)	0.79 (0.66–0.91)
				Independent	0.72 (0.55–0.86)	0.76 (0.64–0.88)	0.68 (0.4–0.96)
Superior/Mid	14–16	267	772 (764–1,211)	Consensus	0.94 (0.9–0.98)	0.94 (0.87–1.01)	0.88 (0.71–1.05)
				Independent	0.84 (0.74–0.94)	0.89 (0.82–0.97)	0.79 (0.6–0.98)
Inferior	14–16	129	1,097 (1,072–1,272)	Consensus	0.9 (0.5–1)	0.77 (0.65–0.9)	0.77 (0.57–0.97)
				Independent	0.77 (0.5–0.88)	0.96 (0.92–1)	0.57 (0.29–0.86)
Superior/Mid	17+	216	839 (836–980)	Consensus	0.91 (0.79–0.98)	0.94 (0.9–0.99)	0.87 (0.77–0.98)
				Independent	0.87 (0.7–0.95)	0.96 (0.92–1)	0.79 (0.57–1)
Inferior	17+	82	613 (610–1,269)	Consensus	0.86 (0.78–0.93)	0.88 (0.76–0.99)	0.74 (0.53–0.95)
				Independent	0.81 (0.6–0.91)	0.7 (0.52–0.88)	0.92 (0.74–1.1)
Sacrum							
Superior/Mid/Inferior	8–10	342	767 (763–890)	Consensus	0.68 (0.57–0.76)	0.52 (0.41–0.62)	0.83 (0.73–0.94)
				Independent	0.69 (0.62–0.78)	0.52 (0.42–0.63)	0.86 (0.76–0.96)
Superior/Mid/Inferior	11–13	299	860 (788–1,136)	Consensus	0.78 (0.66–0.87)	0.76 (0.68–0.83)	0.81 (0.64–0.99)
				Independent	0.71 (0.64–0.8)	0.73 (0.65–0.8)	0.69 (0.54–0.84)
Superior/Mid/Inferior	14–16	435	741 (609–969)	Consensus	0.77 (0.63–0.88)	0.76 (0.67–0.85)	0.79 (0.56–1.02)
				Independent	0.7 (0.58–0.79)	0.76 (0.68–0.85)	0.64 (0.46–0.82)
Superior/Mid/Inferior	17+	295	1,156 (1,150–1,332)	Consensus	1 (0.99–1)	0.79 (0.64–0.95)	0.8 (0.44–1.16)
				Independent	0.81 (0.65–0.93)	0.79 (0.64–0.95)	0.8 (0.44–1.16)

* AUROC of less than 0.60 was considered poor discrimination, 0.60 to 0.75 moderate discrimination, and more than 0.75 excellent discrimination. AUROC, area under the receiver operating characteristic; CI, confidence interval; ROI, region of interest.
[a] n = ROIs assessed at each location per maturational group.

However, interpreting sacroiliac imaging in skeletally immature patients remains difficult,[6,11,12,21,32–34] and there is a need to avoid unnecessary use of biologics, which carry significant risks and costs. Therefore, additional tools that enhance imaging accuracy—that are simple, accessible, and complementary to standard sequences—are needed. This study achieved two key goals. First, it established normative age- and bone-specific ADC values in the SIJs of youth, revealing that the inferior ilium and sacral ADCs decline with age, whereas superior and mid ilium values remain stable. Second, it derived empirical thresholds to differentiate normal from inflammatory signals, demonstrating strong performance (AUROC ≥0.90 in most iliac regions and ≥0.77 in most sacral regions). A recent multicenter study found, with dedicated MRI sequences, that radiologists at large tertiary care academic pediatric hospitals achieved high sensitivity (>90%) for detecting sacroiliitis but only moderate specificity (median 68%).[6] The current findings suggest that the ADC from DWI sequences may meaningfully enhance the specificity and overall accuracy of MRI evaluation for inflammatory sacroiliitis, particularly in the pediatric population.

A growing body of evidence supports the use of DWI sequences for assessing sacroiliitis in both adults[13,15,35] and children.[9,10,19] DWI is most commonly applied to highly vascularized tissues or those with high water content, such as the brain. At the SIJ, DWI may be useful in detecting inflammation by identifying increased water content in the subchondral bone marrow. However, in adults, it remains debated whether quantitative measures from DWI—specifically ADC values—provide added diagnostic value, especially in cases in which inflammation is subtle and the signal may be difficult to accurately quantify.[36] In pediatric patients, the challenge is different. Maturational bone marrow signal is often pronounced and can easily mimic pathology.[6,11,12,21,32–34] Unlike adults with longstanding axial disease, who frequently exhibit structural lesions that complicate ROI placements and ADC interpretation, children are typically evaluated early in the disease course—often before structural changes are detectable. This presents a unique opportunity to leverage DWI more effectively in the pediatric setting, specifically, to aid in improved specificity and overall accuracy of MRI evaluation.

Several studies have explored the use of DWI sequences to evaluate sacroiliitis.[9,13,18,20,35,37] We leveraged key insights from these studies to inform the design of our ROI placement strategy, including the number, size, and positioning. Coronal oblique sequences were selected for ROI placement rather than axial sequences, as they provide better visualization of the SIJ and facilitate more accurate ROI positioning. Vendhan et al used the mid-sacrum to normalize the ADC values.[9] However, their study showed that normalized ADC values of patients with mild

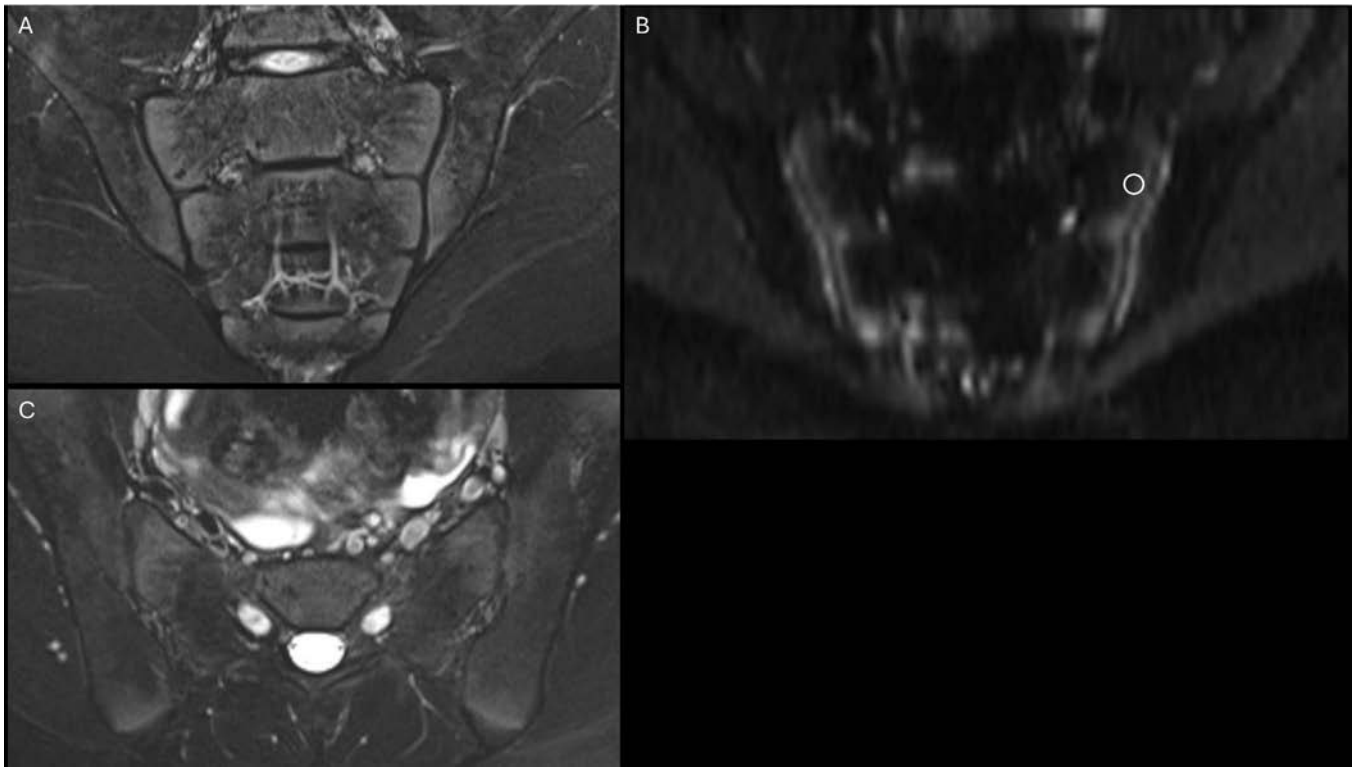

Figure 4. Representative application of the DWI technique. (A) Coronal oblique STIR showing normal type I maturational signal. (B) Coronal oblique DWI showing ROI placement in the superior left sacrum. Measured ADC was 666 mm^2/s. (C) Axial oblique T2 fat-saturated image showing normal maturational signal in the iliac and sacral bones about the SIJs and iliac apophyses. ADC, apparent diffusion coefficient; DWI, diffusion-weighted imaging; ROI, region of interest; SIJ, sacroiliac joint.

inflammation were similar to those in controls with immature SIJs. Additionally, the variability in reference ADC values contributed substantially to the spread of normalized results. Normalization also reduced the intraclass correlation coefficient and widened the Bland–Altman limits of agreement, suggesting that the use of uncorrected ADC values might yield better interobserver reproducibility than normalized ones. These findings indicate that normalizing ADC may not be ideal when attempting to define empirical ADC thresholds.[9] For this reason, we chose not to normalize our ADC data to the mid-sacrum.

Previous DWI studies have several limitations, including not accounting for age or sex differences,[38] placing ROIs across the joint instead of within bone,[9,18] using few ROIs, and introducing subjectivity in ROI placement. This study addresses age-related variation by stratifying results by age, though it was underpowered to assess sex differences. Although ROI placement on areas of edema remains somewhat subjective, standardization was prioritized—defining the central coronal slice, setting subchondral ROI criteria, and providing reproducible guidance for less experienced users. Circular ROIs were used to ensure clinical feasibility, as they can be generated in standard PACS systems. Larger ROIs were selected to capture a range of signals and better define cut-off values. Although custom semi-automated software offers more precision, it is not widely accessible. Our software also adjusts for differences in MRI scanners (manufacture and model), imaging protocols, and ADC calculations[39] across institutions, improving generalizability. Reference values are tied to joint position, allowing flexibility in image acquisition. If axial images are obtained, our program can reformat them to a semicoronal plane—though at the cost of some spatial resolution.

The relatively wide Bland–Altman limits and high variability in bone ADC due to low signal and noise sensitivity could be considered limitations. Signal-poor structures like bone have an expected high amount of variance due to noise. Therefore, small changes in ROI placement result in substantial variation in measured mean ADC. However, the differences in mean ADC between controls and cases were generally large, reflecting the power of DWI to differentiate between noise in controls and abnormal signal in cases. Although ROI selection was based on STIR images—introducing potential bias—interpreters were experienced and well calibrated. The sample size was modest, but AUROC values for derived thresholds were strong. This study did not assess the fat content within the bone marrow, which may influence ADC values,[40] or the additive diagnostic value of DWI. This study also did not assess the impact of physical activity on ADC, as physical activity was not recorded for the reference control group or the patients with sacroiliitis. The reference control group also underwent imaging for a range of clinical reasons.

However, the intent of this tool is to be applied to children undergoing evaluation for back pain stemming from diverse potential causes. In this context, the heterogeneity of our control group is actually a strength, as it enhances the real-world clinical relevance and applicability of the tool. Lastly, variability in pediatric SIJ interpretation is widespread[6] and multifactorial. Although no single solution exists, DWI and improved training and/or automated image analysis may enhance diagnostic accuracy.

In summary, this study establishes normative age- and bone-specific ADC values for the pediatric SIJ and demonstrates that ADC thresholds from DWI sequences can distinguish normal maturation from inflammatory signal. By addressing limitations of prior studies—such as variation by age, ROI placement inconsistency, and normalization issues—this work supports the use of DWI as a potential practical tool to improve specificity in the MRI assessment of inflammatory sacroiliitis in youth. DWI is not intended to replace standard imaging; rather, it may offer meaningful incremental value, particularly given the unique challenges of interpreting skeletally immature joints and the clinical importance of avoiding unnecessary biologic treatment. Future research should further evaluate the impact of marrow fat content, assess diagnostic value in conjunction with standard sequences, and explore automated approaches to enhance reproducibility and diagnostic accuracy.

In summary, this study established normative age- and bone-specific ADC reference ranges for the pediatric SIJ and demonstrated that ADC thresholds from DWI can distinguish normal maturational signal from inflammatory signal with high AUROC and specificity. This work supports the continued exploration of DWI as a tool, in addition to standard sequences, to improve specificity in the MRI assessment of inflammatory sacroiliitis in youth in routine clinical care. Future work should further evaluate the impact of marrow fat content, assess diagnostic value in conjunction with standard sequences, and explore automated approaches to enhance reproducibility and diagnostic accuracy.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Francavilla confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Haibel H, Brandt HC, Song IH, et al. No efficacy of subcutaneous methotrexate in active ankylosing spondylitis: a 16-week open-label trial. Ann Rheum Dis 2007;66(3):419–421.
2. Bollow M, Braun J, Biedermann T, et al. Use of contrast-enhanced MR imaging to detect sacroiliitis in children. Skeletal Radiol 1998; 27(11):606–616.
3. Stoll ML, Bhore R, Dempsey-Robertson M, et al. Spondyloarthritis in a pediatric population: risk factors for sacroiliitis. J Rheumatol 2010; 37(11):2402–2408.
4. Flatø B, Hoffmann-Vold AM, Reiff A, et al. Long-term outcome and prognostic factors in enthesitis-related arthritis: a case-control study. Arthritis Rheum 2006;54(11):3573–3582.
5. Pagnini I, Savelli S, Matucci-Cerinic M, et al. Early predictors of juvenile sacroiliitis in enthesitis-related arthritis. J Rheumatol 2010;37(11): 2395–2401.
6. Weiss PF, Brandon TG, Bohnsack J, et al. Variability in interpretation of magnetic resonance imaging of the pediatric sacroiliac joint. Arthritis Care Res (Hoboken) 2021;73(6):841–848.
7. Herregods N, Anisau A, Schiettecatte E, et al. MRI in pediatric sacroiliitis, what radiologists should know. Pediatr Radiol 2023;53(8):1576–1586.
8. Weiss PF, Xiao R, Biko DM, et al. Assessment of sacroiliitis at diagnosis of juvenile spondyloarthritis by radiography, magnetic resonance imaging, and clinical examination. Arthritis Care Res (Hoboken) 2016;68(2):187–194.
9. Vendhan K, Bray TJ, Atkinson D, et al. A diffusion-based quantification technique for assessment of sacroiliitis in adolescents with enthesitis-related arthritis. Br J Radiol 2016;89(1059):20150775.
10. Tasar S, Ciraci S, Yilmaz PD, et al. Efficacy of diffusion weighted imaging in sacroiliac joint MRI in children. North Clin Istanb 2023;10(2): 131–138.
11. Herregods N, Maksymowych WP, Jans L, et al. Atlas of MRI findings of sacroiliitis in pediatric sacroiliac joints to accompany the updated preliminary OMERACT pediatric JAMRIS (Juvenile Idiopathic Arthritis MRI Score) scoring system: part I: active lesions. Semin Arthritis Rheum 2021;51(5):1089–1098.
12. Herregods N, Maksymowych WP, Jans L, et al. Atlas of MRI findings of sacroiliitis in pediatric sacroiliac joints to accompany the updated preliminary OMERACT pediatric JAMRIS (Juvenile Idiopathic Arthritis MRI Score) scoring system: part II: structural damage lesions. Semin Arthritis Rheum 2021;51(5):1099–1107.
13. Bozgeyik Z, Ozgocmen S, Kocakoc E. Role of diffusion-weighted MRI in the detection of early active sacroiliitis. AJR Am J Roentgenol 2008; 191(4):980–986.
14. Gaspersic N, Sersa I, Jevtic V, et al. Monitoring ankylosing spondylitis therapy by dynamic contrast-enhanced and diffusion-weighted magnetic resonance imaging. Skeletal Radiol 2008;37(2):123–131.
15. Gezmis E, Donmez FY, Agildere M. Diagnosis of early sacroiliitis in seronegative spondyloarthropathies by DWI and correlation of clinical and laboratory findings with ADC values. Eur J Radiol 2013;82(12): 2316–2321.
16. Pedersen SJ, Poddubnyy D, Sørensen IJ, et al. Course of magnetic resonance imaging-detected inflammation and structural lesions in the sacroiliac joints of patients in the randomized, double-blind, placebo-controlled Danish multicenter study of adalimumab in spondyloarthritis, as assessed by the Berlin and Spondyloarthritis Research Consortium of Canada Methods. Arthritis Rheumatol 2016;68(2):418–429.
17. Sieper J, van der Heijde D, Dougados M, et al. A randomized, double-blind, placebo-controlled, sixteen-week study of subcutaneous golimumab in patients with active nonradiographic axial spondyloarthritis. Arthritis Rheumatol 2015;67(10):2702–2712.
18. Bray TJ, Vendhan K, Roberts J, et al. Association of the apparent diffusion coefficient with maturity in adolescent sacroiliac joints. J Magn Reson Imaging 2016;44(3):556–564.

19. J P Bray T, Vendhan K, Ambrose N, et al. Diffusion-weighted imaging is a sensitive biomarker of response to biologic therapy in enthesitis-related arthritis. Rheumatology (Oxford) 2017;56(3):399–407.
20. Beltran LS, Samim M, Gyftopoulos S, et al. Does the addition of DWI to fluid-sensitive conventional MRI of the sacroiliac joints improve the diagnosis of sacroiliitis? AJR Am J Roentgenol 2018;210(6):1309–1316.
21. Chauvin NA, Xiao R, Brandon TG, et al. MRI of the sacroiliac joint in healthy children. AJR Am J Roentgenol 2019;212(6):1303–1309.
22. Brandon TG, Xiao R, Lovell DJ, et al. Core set of responsive and discriminatory measures for use in pragmatic trials of youth with axial juvenile spondyloarthritis. Arthritis Care Res (Hoboken) 2025;77(10):1229–1239.
23. Petty RE, Southwood TR, Manners P, et al; International League of Associations for Rheumatology. International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. J Rheumatol 2004;31(2):390–392.
24. Dougados M, van der Linden S, Juhlin R, et al. The European Spondylarthropathy Study Group preliminary criteria for the classification of spondylarthropathy. Arthritis Rheum 1991;34(10):1218–1227.
25. Maksymowych W, Hadsbjerg AEFEF, Østergaard M, et al. Validation of SPARCC MRI-RETIC e-tools for increasing scoring proficiency of MRI sacroiliac joint lesions in axial spondyloarthritis. RMD Open 2024;10(1):e003923.
26. Weiss PF, Maksymowych WP, Xiao R, et al. Spondyloarthritis Research Consortium of Canada sacroiliac joint inflammation and structural scores: change score reliability and recalibration utility in children. Arthritis Res Ther 2020;22(1):58.
27. Liu H, Tang T, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data Anal 2009;53(4):853–856.
28. *fastAUC*. Version 2.1. Patrick Hilden; 2016. Accessed August 2024.
29. Assessing the fit of the model. In: Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. John Wiley & Sons; 2000:143–202.
30. D'Agostino RB, Nam B-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. Handb Stat 2003;23:1–25.
31. R. Version 4.4.1. R Foundation; 2024. Accessed August 2024.
32. Herregods N, Jans LBO, Chen M, et al. Normal subchondral high T2 signal on MRI mimicking sacroiliitis in children: frequency, age distribution, and relationship to skeletal maturity. Eur Radiol 2021;31(5):3498–3507.
33. Storey T, Borgia RE, Merritt P, et al. Normal magnetic resonance imaging appearance of marrow adjacent to the sacroiliac joints in children during development. J Comput Assist Tomogr 2022;46(1):91–96.
34. Herregods N, Jans LBO, Paschke J, et al. Magnetic resonance imaging findings in the normal pediatric sacroiliac joint space that can simulate disease. Pediatr Radiol 2021;51(13):2530–2538.
35. Bradbury LA, Hollis KA, Gautier B, et al. Diffusion-weighted imaging is a sensitive and specific magnetic resonance sequence in the diagnosis of ankylosing spondylitis. J Rheumatol 2018;45(6):771–778.
36. Lambert RG, Maksymowych WP. Diffusion-weighted imaging in axial spondyloarthritis: a measure of effusion or does it elicit confusion? J Rheumatol 2018;45(6):729–730.
37. Dewan K, El-saadany H. Diagnostic value of diffusion weighted magnetic resonance image in early ankylosing spondylitis. Egypt J Radiol Nucl Med 2014;45(4):1147–1152.
38. Lavdas I, Rockall AG, Castelli F, et al. Apparent diffusion coefficient of normal abdominal organs and bone marrow from whole-body DWI at 1.5 T: the effect of sex and age. AJR Am J Roentgenol 2015;205(2):242–250.
39. Sasaki M, Yamada K, Watanabe Y, et al; Acute Stroke Imaging Standardization Group-Japan (ASIST-Japan) Investigators. Variability in absolute apparent diffusion coefficient values across different platforms may be substantial: a multivendor, multi-institutional comparison study. Radiology 2008;249(2):624–630.
40. Liu YJ, Tsai TS, Li YH, et al. Understanding ADC variation by fat content effect using a dual-function MRI phantom. Eur Radiol Exp 2024;8(1):19.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 639–646
DOI 10.1002/acr.25677

American College of Rheumatology
Empowering Rheumatology Professionals

Disparities and Reproductive Health in Rheumatic Diseases: Deficits in Counseling and Contraception Use in an Urban Female Hispanic Population in Los Angeles

Eaman Alhassan,[1,2] Laura Kobashigawa,[1] Vanessa Ramos,[3] Jack Rodman,[3] and Leanna Wise[1]

Objective. Systemic lupus erythematosus (SLE) and inflammatory arthritis disproportionately affect reproductive-age Hispanic women, who experience more severe disease and worse outcomes. Certain factors may contribute to disparities in reproductive health counseling. This survey-based study examined the association between reproductive health counseling and contraceptive use, considering a variety of demographics and teratogenic medication use in a predominantly Hispanic population in urban Los Angeles.

Methods. An anonymous survey, based on the 2020 American College of Rheumatology Reproductive Health Guideline, was administered to 304 female patients of childbearing age with SLE, rheumatoid arthritis (RA), or juvenile idiopathic arthritis (JIA). Analyses included two-sample *t*-tests, Pearson's chi-square test, multivariable logistic regression, and Fisher's exact tests, with significance set at $P \leq 0.05$.

Results. Of 304 patients, English speakers were significantly more likely to receive contraception counseling than Spanish speakers (odds ratio [OR] 3.36; 95% confidence interval [CI] 1.52–7.40; $P < 0.01$). Older age was associated with lower odds of receiving counseling (OR 0.95; 95% CI 0.91–0.98; $P = 0.01$). Among those on teratogenic medications not desiring pregnancy, contraception counseling was linked to higher odds of using long-acting reversible contraception (OR 5.12; 95% CI 1.26–20.71; $P = 0.02$). Younger patients, those with SLE (vs RA or JIA), and English speakers had higher odds of perceiving their physician as knowledgeable about reproductive health (all $P < 0.05$).

Conclusion. Reproductive health counseling was inadequate for Spanish-speaking and older patients. However, reproductive health counseling was positively associated with effective contraception use among patients on teratogenic medications. Clinicians should be aware of these disparities and the potential benefit of counseling to improve highly effective contraception use.

INTRODUCTION

Systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) are chronic autoimmune diseases predominantly affecting women during their reproductive years. Juvenile idiopathic arthritis (JIA) is a chronic autoimmune disease diagnosed before the age of 16 that persists during adulthood.[1] These autoimmune diseases are relatively prevalent in the Hispanic population, who often experience a more severe disease course and worse clinical outcomes.[2–6] Overall, the literature documents that patients of Hispanic ethnicity face significant reproductive health challenges, including limited health care access and language barriers.[7–10] The Rheumatology Informatics System for Effectiveness (RISE) registry study found that less than one-third of patients had documented contraception use, with markedly lower rates among Black and Hispanic women, as well as those receiving potentially teratogenic medications, such as methotrexate.[11] These factors can delay pregnancy planning and increase the risk of adverse outcomes, emphasizing the need for culturally sensitive and more comprehensive and tailored approaches to care. Further, as the US Hispanic population consistently experiences higher rates of unintended pregnancies compared

Biostatistical support was funded by the National Center for Advancing Translational Sciences of the US NIH (grants UL1-TR-001855 and UL1-TR-000130). Compensation for patients was provided by the Committee of Interns and Residents Service Employees International Union Healthcare Patient Care quality improvement fund.

[1]Keck School of Medicine, University of Southern California, Los Angeles; [2]University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; [3]University of Southern California, Los Angeles.

Dr Alhassan and Dr Kobashigawa contributed equally to this work.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25677).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25677.

Address correspondence via email to Leanna Wise, MD, MPH, at Leanna.Wise@med.usc.edu.

Submitted for publication April 21, 2025; accepted in revised form October 2, 2025.

SIGNIFICANCE & INNOVATIONS

- This study highlights significant disparities in contraception counseling among Hispanic women with systemic lupus erythematosus and inflammatory arthritis, with English-speaking patients having an over three-fold higher odds of receiving counseling than Spanish speakers.
- Patients on teratogenic medications who received contraception counseling were significantly more likely to report use of highly effective contraception, emphasizing the impact of provider discussions on patient safety.
- Older patients were less likely to receive contraception counseling or perceive their physician as knowledgeable on reproductive health and rheumatic disease, suggesting provider bias or misconceptions about reproductive intentions in this population.

to non-Hispanic White women, this is a crucial issue requiring attention from specialists who treat these diseases.[12,13] Finally, to complicate the matter, many medications used in the treatment of SLE, RA, and JIA are indisputably teratogenic, necessitating effective use of contraception and effective employment of conversations about such issues. As the US Hispanic population is the second largest racial or ethnic group after the non-Hispanic White population and one of the fastest growing groups over the past 25 years, the issues mentioned are particularly salient.[14,15]

In 2020, the American College of Rheumatology (ACR) published the Guideline for the Management of Reproductive Health in Rheumatic and Musculoskeletal Diseases (2020 ACR Reproductive Health Guideline), which addressed best practices for multiple reproductive health issues, such as optimal contraceptive options, contraception counseling, and careful use versus discontinuation of teratogenic medications in female patients.[16] However, how these guidelines are "experienced" downstream by patients (especially by those at highest risk of poor outcomes) and physicians in clinical practice remains to be understood—especially in a medically underserved population. In our study to address this question, we administered an anonymous survey based on the 2020 ACR Reproductive Health Guideline to women with SLE, RA, or JIA at the Los Angeles General Medical Center (LAGMC; formerly Los Angeles County + University of Southern California) outpatient rheumatology clinic, a public medical center serving urban east Los Angeles County in California. We examined the association between contraception counseling and contraceptive practices in a predominantly Hispanic, urban, publicly-insured population, using multivariable models adjusted for key sociodemographic factors, language, and teratogenic medication use. The results are reported herein.

PATIENTS AND METHODS

This is a cross-sectional study conducted at the LAGMC outpatient rheumatology clinic from March 2023 to December 2023. As the purpose of this study was to gather an extensive, first-time, unbiased examination into the needs of our patient population, no electronic medical record information was gathered from patients at the time of the survey, including concomitant disease activity scores or verification of prescribed medications. University of Southern California Institutional Review Board approval was obtained (UP-23-00099).

Study population and clinic setting. A total of 304 participants completed the survey, including 153 patients diagnosed with SLE and 151 patients with either RA or JIA (combined as inflammatory arthritis [IA]). Participants were required to be female by birth, within the age range of 18 to 50 years at the time of the clinic visit based on medical records, have a confirmed diagnosis of SLE, RA, or JIA diagnosed by a rheumatologist, and be able to read in either English or Spanish. Participants must have attended at least one outpatient rheumatology clinic visit at LAGMC within the year before the study (ie, patients who were being seen for their first visit [new patient visits] were excluded). Potential participants were identified before each clinic session by LW, who reviewed electronic medical records to determine initial eligibility. Study team members or the research coordinator at the Division of Rheumatology approached eligible patients in the clinic at LAGMC. These potential patients were then informed about the survey study and given the chance to take the survey before their clinic visit.

Surveys were administered in the patient's preferred language, either English or Spanish, and certified medical interpreters assisted when study staff were not fluent in Spanish. The majority (>80%) of surveys were completed on an electronic tablet while patients were waiting for their rheumatology appointment, and some were completed immediately after their visit. For those that completed it immediately after, their treating clinicians were not told about the details of the study, so as not to bias the participants' responses. Participants received $30 as compensation for their time. The average response rate for eligible patients in the survey time period was 94% for patients with SLE and 91% for patients with IA.

At the time of survey distribution, the rheumatology clinic at LAGMC was staffed by a team comprising five to six attending physicians, five to six rheumatology fellows, and four to seven internal medicine residents. The patient population was predominantly Hispanic, with approximately 85% identifying as Mexican or of Central American descent and a similar proportion (85%–90%) covered by Medi-Cal, California's Medicaid program. Among the attending rheumatologists, one was a native Spanish–English bilingual speaker, two had learned Spanish as a second language, and the remaining three relied on telephone-based

medical translation services when caring for non-English–speaking patients. At the time of survey implementation, all the rheumatology fellows relied on medical translation for the care of Spanish-speaking patients. Most trainees (internal medicine residents, medical students from the affiliated Keck School of Medicine) also use medical translation services to communicate with the clinic's Spanish-speaking patients.

Survey design and implementation. The survey was designed with questions based on core clinical topics from the 2020 ACR Reproductive Health Guideline. Both the English and Spanish versions of the SLE patient survey were pilot tested with three individuals representative of our target population for each language, and feedback was used to modify some of the questions for comprehension and clarity. The survey gathered key clinical and demographic information, including self-reported race (categorized as American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, White, or mixed) and ethnicity (Hispanic or non-Hispanic), in alignment with the NIH's terminology for race and ethnicity classifications. Additional data included a patient-reported list of SLE or IA medications, reproductive health history, current sexual history and practices, and details about patient-reported discussions with rheumatologists regarding contraception and family planning. The survey also assessed patient-reported knowledge on topics related to rheumatic disease and reproductive health. The English IA survey questions are provided in Supplement 1 for reference, and the questions for patients with SLE were similar except for substituting "lupus" for every incidence of "arthritis."

Statistical analyses. Demographic and survey data were separated by SLE or IA and summarized using the mean (SD) for continuous variables and frequency (percentage) for categorical variables. Tests for associations between SLE and IA were estimated using Pearson's chi-square test or Fisher's test (depending on counts) for categorical variables and independent samples *t*-test or Wilcoxon rank sum test (depending on normality).

Multivariable logistic regression was used to evaluate the associations between reported contraception counseling and independent variable of interest. Firth logistic regression was used when issues of small cell sizes and data separation arose. Covariates entered in the final model were based on the presence of confounding (as determined by any covariate that changes the effect of contraception counseling on the variable of interest ≥15%), a cutoff of 15%, univariate significance with the contraception counseling, and clinical relevance. Model fit was evaluated via inspection of collinearity among independent variables via variance inflation factor, leverage, deviance, and normality of residuals. No model fit issues arose for any analyses. All statistical tests were determined significant at $P \leq 0.05$ and conducted within SAS Enterprise version 8.3 (SAS Institute Inc).

RESULTS

Demographics. Results are shown in Table 1. A total of 304 patients with SLE or IA completed the survey. The mean age was 39.31 years (SD 8.87). In total, 274 patients (91.03%) identified as Hispanic, with 190 patients (62.91%) taking the survey in Spanish. Over three-quarters of patients (226 patients, 75.33%) were born outside the United States.

Sexual activity, contraception use, and counseling. Regarding sexual activity, in total, 118 patients (40.97%) were not currently sexually active at the time of the survey (Table 2). Among those who reported being sexually active, 59 patients (20.49%) were neither trying to conceive nor avoiding pregnancy, and 90 patients (31.25%) were avoiding pregnancy. No patients were pregnant at the time of survey collection.

Overall, 133 patients (51.15%) reported discussing contraception with their doctor within the past year, whereas 127 patients (48.85%) did not; additional contraception and sexual activity practices are shown in Table 2. In multivariable regression assessing patient-reported contraception counseling, survey language was the strongest predictor of receiving contraception counseling (Table 3). English survey takers reported a *3.36-fold* increase in odds of receiving counseling in the past one year compared to Spanish survey participants (odds ratio [OR] 3.36; 95% confidence interval [CI] 1.52–7.40; $P < 0.01$). Age was also a significant predictor, with every one year in age being associated with a 5% lower odds of reporting contraception counseling (OR 0.95, 95% CI 0.91–0.98; $P = 0.01$). There was not enough evidence to reject the null hypothesis of no association across disease type (SLE vs IA), marital status, and nativity with reported counseling.

Teratogenic medications and contraceptive use. Proportions of teratogenic medication use are shown in Supplemental Table 1. Among sexually active patients on teratogenic medications who were not actively trying to conceive, patients who reported receiving contraception counseling in the past one year were significantly more likely to use very highly effective birth control (long-acting reversible birth control), though the magnitude of this effect remains uncertain given large CIs (OR 5.12; 95% CI 1.26–20.71; $P = 0.02$; Table 4). Similarly, for sexually active patients who reported a history of contraception counseling and were on teratogenic medications, there was a trend toward more likely use of barrier or spermicidal methods of contraception, as opposed to no use of contraception (OR 4.14; 95% 0.87–19.75; $P = 0.07$). Short-acting birth control and natural methods also demonstrated increased odds of receiving contraception counseling in the past one year, though wide CIs suggest further exploration. Contraception use by those on teratogenic medications is shown in Supplemental Table 2, and notably,

Table 1. Descriptive statistics of total population and by disease type*

	Total population (N = 304)[a]	Arthritis (n = 151)[a]	Lupus (n = 153)[a]
Age, y	39.31 (8.87)	41.87 (7.23)	36.75 (9.62)
Missing, n	4	2	2
What is your race?			
White	171 (62.64%)	114 (82.01%)	57 (42.54%)
Black/African American	6 (2.20%)	0 (0.00%)	6 (4.48%)
Asian American	10 (3.66%)	1 (0.72%)	9 (6.72%)
American Indian/Alaska Native	1 (0.37%)	1 (0.72%)	0 (0.00%)
Native Hawaiian/Pacific Islander	1 (0.37%)	1 (0.72%)	0 (0.00%)
Mixed race	7 (2.56%)	2 (1.44%)	5 (3.73%)
Other	56 (20.51%)	13 (9.35%)	43 (32.09%)
Missing, n	50	19	31
Ethnicity			
Hispanic	274 (91.03%)	147 (98.00%)	127 (84.11%)
Non-Hispanic	25 (8.31%)	3 (2.00%)	22 (14.57%)
Missing, n	3	1	2
Survey language			
English	112 (37.10%)	26 (17.22%)	86 (56.95%)
Spanish	190 (62.91%)	125 (82.78%)	65 (43.05%)
Missing, n	0	0	0
Education level			
Less than high school	102 (34.11%)	75 (50.00%)	27 (18.12%)
High school diploma or equivalent	129 (43.14%)	53 (35.33%)	76 (51.01%)
Technical/vocational training	38 (12.71%)	13 (8.67%)	25 (16.78%)
Bachelor's degree	29 (9.70%)	9 (6.00%)	20 (13.42%)
Master's degree or above	1 (0.33%)	0 (0.00%)	1 (0.67%)
Missing, n	3	1	2
Marital status			
Never married	106 (35.57%)	39 (26.35%)	67 (44.67%)
Married or in a domestic partnership	125 (41.95%)	72 (48.65%)	53 (35.33%)
Widowed	4 (1.34%)	2 (1.35%)	2 (1.33%)
Divorced or separated	45 (15.10%)	26 (17.57%)	19 (12.67%)
Missing, n	22	12	10
Religious affiliation			
Catholic	183 (61.41%)	100 (66.67%)	83 (56.08%)
Any denomination of Christianity other than Catholic	56 (18.79%)	24 (16.00%)	32 (21.62%)
Other religious group	16 (5.37%)	4 (2.67%)	12 (8.11%)
No religious affiliation	43 (14.43%)	22 (14.67%)	21 (14.19%)
Missing, n	4	1	3
Nativity			
Born in the United States	74 (24.67%)	23 (15.33%)	51 (34.00%)
Born outside of the United States	226 (75.33%)	127 (84.67%)	99 (66.00%)
Missing, n	2	1	1

* Numbers represent frequency (column percent) for categorical variables and mean (SD) for continuous. Significant at $P \leq 0.05$.
[a] Values in each column may not add up to total sample/100% due to missingness.

44 patients (27.3%) did not use any contraception while on teratogenic medications.

Clinician and patient relationships. In multivariable logistic regression to assess participants' perception of their rheumatologists' knowledge about reproductive health (Question #22 on survey; Supplementary Appendix 1), patients with SLE had significantly higher odds of reporting that their rheumatologist had sufficient information to discuss contraception and pregnancy compared to patients with IA (OR 2.35, 95% CI 1.30–4.24; $P < 0.01$; Table 5). However, those who took the Spanish survey reported lower odds of their rheumatologist having sufficient knowledge to address reproductive health and rheumatic disease (OR 0.46, 95% CI 0.22–0.95; $P = 0.04$). Additionally, older age was associated with lower odds of responding in the affirmative to the same question, with each additional 1 year reducing the odds by 5% (OR 0.95, 95% CI: 0.91–0.99; $P < 0.01$).

Regarding reproductive health discussions, patients with IA were more likely than patients with SLE to report that their doctor initiated these conversations (52.52% vs 37.86%), whereas patients with SLE were more likely to report that both they and their doctor equally initiated discussions (32.14% vs 11.51%) (Supplemental Table 3). Notably, 17.14% of patients with SLE and 11.51% of patients with IA reported never having discussed reproductive health-related topics with their rheumatology doctor.

Table 2. Patient-reported reproductive health characteristics*

Variable	Total population[a]	Arthritis[a]	Lupus[a]
Please select the following that best describes you:			
I have never been sexually active with a male partner	26 (9.03)	11 (7.75)	15 (10.27)
I have been sexually active with a male partner, but I am not currently sexually active with a male partner	92 (31.94)	44 (30.99)	48 (32.88)
I am sexually active with a male partner, and I am currently trying to get pregnant	21 (7.29)	16 (11.27)	5 (3.42)
I am sexually active with a male partner, and I am currently trying to AVOID a pregnancy	90 (31.25)	38 (26.76)	52 (35.62)
I am sexually active with a male partner, and am neither trying to get pregnant nor trying to avoid it	59 (20.49)	33 (23.24)	26 (17.81)
I am currently pregnant	0 (0.00)	0 (0.00)	0 (0.00)
In the past one year, my arthritis/lupus doctor has had at least one discussion regarding contraception with me			
No	127 (48.85)	84 (61.31)	43 (34.96)
Yes	133 (51.15)	53 (38.69)	80 (65.04)
My arthritis/lupus doctor has explained what type of contraception I should or should not use in light of my arthritis/lupus			
No	112 (45.90)	79 (59.40)	33 (29.73)
Yes	132 (54.10)	54 (40.60)	78 (70.27)
In the past one year, my arthritis/lupus doctor has asked what my plans are, if any, for pregnancy			
No	106 (40.00)	71 (51.45)	35 (27.56)
Yes	159 (60.00)	67 (48.55)	92 (72.44)
My arthritis/lupus doctor has told me that it's important that my arthritis is in remission before becoming pregnant			
No	34 (35.05)	31 (57.41)	3 (6.98)
Yes	63 (64.95)	23 (42.59)	40 (93.02)
My arthritis/lupus doctor has explained which medications I should continue when I'm trying to become pregnant or when I'm pregnant			
No	40 (42.11)	32 (57.14)	8 (20.51)
Yes	55 (57.89)	24 (42.86)	31 (79.49)

* Numbers represent frequency (column percent) for categorical variables.
[a] Values in each column may not add up to total sample/100% due to missingness.

DISCUSSION

Our study highlights the importance of contraception counseling among all patients, regardless of language preference or age, particularly as 27.3% of those who were sexually active on teratogenic medications denied any form of contraception whatsoever. Our study results also revealed that several factors influenced the dynamics of contraception counseling conversations among this patient population with their rheumatology providers at LAGMC. Language preferences appear to play a

Table 3. Multivariable analysis of influences on patient-reported contraception counseling (n = 268)*

Variable	Adjusted odds ratio[a] (95% CI)	*P* value
Disease type: arthritis vs lupus (ref)	0.88 (0.49, 1.56)	0.65
Age	0.95 (0.91, 0.98)	0.01[b]
Survey language: English vs Spanish (ref)	3.36 (1.52, 7.40)	<0.01[b]
Marital status (married ref)		
Never married	1.21 (0.62, 2.36)	0.53
Widowed	0.79 (0.09, 7.11)	0.81
Divorced or separated	0.91 (0.43, 1.93)	0.87
Place of birth: United States vs not in the United States (ref)	0.59 (0.26, 1.35)	0.24

* All covariates in the initial model were included in the final model.
CI, confidence interval.
[a] Each factor is adjusted for all covariates in the model.
[b] Indicates significant *P* value.

Table 4. Patient-reported contraception counseling conversations and contraception use (only survey respondents on teratogenic medication, n = 75)*

Variable[a]	Odds ratio (95% CI)	*P* value
Natural methods (coitus interuptus, fertility awareness)[a]	11.94 (0.35, 406.89)	0.17
Barrier or spermicidal methods (condoms, spermicide, cervical cap, or cervical diaphragm)[a]	4.14 (0.87, 19.75)	0.07
Short-acting birth control (oral contraceptive pills, vaginal ring, injection, or patch)[a]	2.46 (0.53, 11.37)	0.24
Long-acting reversible birth control (intrauterine device, hormonal implant)[a]	5.12 (1.26, 20.71)	0.02[b]
Sterilization (tubal ligation, hysterectomy, vasectomy, or postmenopausal)[a]	0.63 (0.14, 2.94)	0.55

* CI, confidence interval.
[a] Each category is compared against the "None—I don't use any kind of contraception" category.
[b] Indicates significant *P* value.

Table 5. Multivariable analysis of influences on patient-reported perception of physician knowledge on reproductive health and rheumatic disease (n = 280)*

	Adjusted odds ratio[a] (95% CI)	*P* value
Disease type: lupus vs arthritis (ref)	2.35 (1.30, 4.24)	<0.01[b]
Age	0.95 (0.91, 0.99)	<0.01[b]
Survey language: Spanish vs English (ref)	0.46 (0.22, 0.95)	0.04[b]

* All covariates in the initial model were included in the final model.
CI, confidence interval.
[a] Each factor is adjusted for all covariates in the model.
[b] Indicates significant *P* value.

significant role in shaping contraception counseling discussions between patients and providers, as those patients who chose to take the Spanish survey noted markedly lower odds of contraception counseling in the past one year, as well as less confidence in the knowledge base of their physicians. These findings indicate a significant foundational disparity that may have major downstream consequences. We hypothesize that the reason for this discrepancy is likely multifactorial, with clinician–patient language discordancy being a potential concern, as many clinicians are not fluent in Spanish. The use of telephone-based medical interpreters (although an absolute necessity if the clinician is not fluent) may hinder reproductive communication between patients and clinicians, contribute to possible medical misunderstanding, and/or exacerbate time constraints.[17–20]

Ultimately, there were findings in our study that display an increased level of responsibility for an accountable approach to health planning in this population. Namely, among those participants not actively seeking to conceive and also taking teratogenic medications, having a history of contraception counseling was significantly associated with using highly effective methods of contraception (long-active reversible contraceptives, such as an intrauterine device or hormonal implant). Nevertheless, a large proportion of patients on teratogenic medications (and who were also sexually active) also reported not using any form of birth control. The reason for this discrepancy may have a multiplicity of causes, including lack of access to birth control, misunderstanding regarding the need for contraception despite counseling, or personal or cultural preferences. It should be noted that within the care model at LAGMC, there are no types of birth control that are restricted or "higher tier" compared to others—that is, all birth control options are offered to patients regardless of insurance status. Ultimately, these findings shed light on the importance of effective communication between provider and patient to underscore the need for contraception, particularly while on teratogenic medications.

Interestingly, our study found that individuals with SLE, when compared to individuals with IA, were significantly more likely to feel that their doctor had enough information to answer their reproductive health questions, and this was adjusted for the greater proportion of younger and English-speaking individuals in the SLE population. We hypothesize that this difference may reflect how providers perceive the severity of disease for patients diagnosed with SLE compared to those diagnosed with IA, with the risks of uncontrolled SLE in pregnancy being linked to increased risk of miscarriage and preeclampsia in addition to disease flares.[21] Nevertheless, it should still be noted that poorly controlled IA is linked to poor maternal-fetal outcomes.[22]

Reported rates of disease flares among patients with SLE during pregnancy range widely in the literature from 13% to 74%, with a recent multicenter study by Davis et al reporting that 20.8% of patients with SLE experienced mild and moderate flares during pregnancy, whereas 6.25% had severe flares, as assessed by the SELENA-SLEDAI Flare Index.[23] Additionally, the multi-organ involvement in SLE, including lupus nephritis, may make providers more attuned to the severity of the disease, underscoring the importance of discussing pregnancy planning to protect vital organs. Our study highlights the need for comprehensive reproductive health counseling among individuals with IA as well, particularly given the frequent use of teratogenic medications, such as methotrexate and leflunomide, as well as the potential risks of poor maternal-fetal outcomes with poorly controlled IA activity, including high risk of spontaneous abortions, preterm birth, intrauterine growth restriction, stillbirth, and eclampsia.[22,24–27]

Other studies comparing reproductive health counseling among those with SLE versus IA have found contrasting results. A report (abstract) by Kompa et al found that among female patients with SLE or RA between the ages of 18 and 45 years who were taking either mycophenolate or methotrexate, patients with RA were significantly more likely to receive contraception counseling compared to patients with SLE ($P < 0.01$).[28] In addition, a study by Clowse et al that evaluated 9,826 women with SLE and 19,009 with RA, all of childbearing age, from the RISE registry found that rates of contraceptive documentation were significantly lower for women with SLE compared to women with RA (OR 0.84, 95% CI 0.76–0.92).[29] The authors hypothesized that the higher rate of contraceptive documentation for patients with RA occurred because of physicians' concerns about the risks of pregnancy while taking methotrexate.

Our study also observed that older participants appeared to have reproductive issues addressed less thoroughly than younger participants. This is evidenced by the findings of lower proportions of contraception counseling and evidence of a lack of confidence in their physician's ability to handle the intersection of reproductive health in rheumatic disease in this older group. This raises the concern that perhaps providers falsely assume that older individuals may not require as much reproductive health counseling compared to younger individuals. For example, providers may incorrectly assume that older patients are no longer interested in or capable of pregnancy. This may not be the case, particularly for women who desire multiple pregnancies or who

may have had to delay pregnancy due to active disease. As many young people with either SLE or IA delay childbearing until later years (ie, after their disease is well-controlled), it is just as imperative to approach contraception counseling and pregnancy planning for these women, regardless of their age. This is supported by similar results regarding age and reproductive health counseling, including an abstract from Weisleder et al, which showed that at an urban medical center, women above the age of 35 years were significantly less likely to receive contraception counseling compared to younger women ($P = 0.007$).[30] This was also seen in the study by Chandramouli et al with only patients with SLE (N = 478; primarily African American and White), which also demonstrated that, based on chart review documentation only, increasing age was associated with lower odds of contraception counseling.[31]

Our study has several strengths. We were able to gain insight into contraception counseling from the patient perspective, using a survey that was intentionally derived from the 2020 ACR Reproductive Health Guideline and pilot tested with individuals of the target population. To our knowledge, this is the first study to directly compare influence on patient-reported contraception counseling among patients with SLE versus IA at an urban medical center serving a predominantly Hispanic population that has high language discordancy with their treating clinicians. Although a review of the literature does reveal other studies comparing contraception use and/or counseling in participants with SLE vs RA, some of these studies are based on electronic medical record review rather than direct patient-reported data or derived from much smaller numbers of patients.[29,32] We focused on a southern Californian female population at a relatively high risk for unintended pregnancies as well as poor maternal-fetal outcomes, given well-documented data that have demonstrated worse SLE and IA disease activity as well as higher rates of unintended pregnancies in US Hispanic populations compared to non-Hispanic White populations.[12,13]

Limitations of our study include the anonymous nature of our survey, which prevented us from verifying information such as teratogen use, pregnancy history, and types of birth control methods. Further, the patients in our study were recruited from an urban medical center in Los Angeles, in which the patient population is predominantly Hispanic and Spanish-speaking, which may limit the generalizability of our findings to other practice settings. We also did not gather clinician-based data—that is, how rheumatologists or trainees perceive the completeness or appropriateness of providing contraception counseling. The survey also did not query about last menstrual period or whether patients perceived themselves to be peri- or post-menopausal, which may affect whether contraception is offered by a clinician. Further, for the subset of patients who were not sexually active, a richer history about whether or not contraception was offered and declined, or not discussed at all, was not obtained. Finally, we focused only on obtaining data from female patients, as active female rheumatic disease activity and use of teratogenic medications or immunosuppressive medications have been clearly documented to have untoward effects on the fetus or neonate. Nevertheless, we acknowledge that reproductive health topics are also important issues for male patients, with key concepts such as sexual health, sperm count and fertility, and medication safety being of similar concern in the male rheumatic disease population.[33,34] We anticipate that these are key areas for a future study to understand and subsequently address gaps in care.

In conclusion, our study highlights critical shortcomings in reproductive health counseling among women with SLE and IA, particularly regarding contraception counseling conversations. English-speaking participants were significantly more likely to receive counseling compared to Spanish-speaking participants, whereas younger age was associated with increased odds of receiving contraception counseling. Participants with SLE were more likely to perceive their rheumatologist as knowledgeable on reproductive health topics; older individuals, as well as Spanish-speaking individuals, were more likely to report the opposite opinion. An encouraging finding was that a reported history of contraception counseling did appear to translate into greater odds of using highly effective contraception methods for survey takers on teratogenic medications and not actively trying to conceive. Our findings underscore the need for more proactive and integrated discussions on reproductive health within rheumatology care, particularly for Spanish-speaking and older patients. Future studies will be necessary to dig deeper (particularly with direct interview methodologies) into what is behind these counseling disparities in order to optimize contraceptive use among underserved populations.

ACKNOWLEDGMENTS

We thank Dr Michael Weisman and Dr William Stohl for their critical review of the manuscript before publication submission and Ms Sara Madrigal for her diligent assistance with patient recruitment.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Wise confirms that all authors have provided the final approval of the version to be published, and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Martini A, Lovell DJ, Albani S, et al. Juvenile idiopathic arthritis. Nat Rev Dis Primers 2022;8(1):5.

2. Shen G, Swaminathan M, Huang I, et al. Racial disparities in pregnancy outcomes among women with rheumatic diseases: a systematic literature review. Semin Arthritis Rheum 2023;60:152193.
3. Clowse ME, Grotegut C. Racial and ethnic disparities in the pregnancies of women with systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2016;68(10):1567–1572.
4. Kaplowitz ET, Ferguson S, Guerra M, et al. Contribution of socioeconomic status to racial/ethnic disparities in adverse pregnancy outcomes among women with systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2018;70(2):230–235.
5. Pelajo CF, Angeles-Han ST, Prahalad S, et al. Evaluation of the association between Hispanic ethnicity and disease activity and severity in a large cohort of patients with juvenile idiopathic arthritis. Rheumatol Int 2013;33(10):2549–2554.
6. Kumar N, Ramphul K, Ramphul Y, et al. Children hospitalized for juvenile arthritis in the United States. Reumatologia 2021;59(4):270–272.
7. Hathaway EC, Lee J, Wojeck RK. Socioeconomic disparities associated with healthcare utilization among adults with rheumatoid arthritis. Nurs Res 2023;72(3):167–174.
8. Divi C, Koss RG, Schmaltz SP, et al. Language proficiency and adverse events in US hospitals: a pilot study. Int J Qual Health Care 2007;19(2):60–67.
9. Brener S, Jiang S, Hazenberg E, Herrera D. A cyclical model of barriers to healthcare for the Hispanic/Latinx population. J Racial Ethn Health Disparities 2024;11(2):1077–1088.
10. Komorowski AS, Jain T. A review of disparities in access to infertility care and treatment outcomes among Hispanic women. Reprod Biol Endocrinol 2022;20(1):1.
11. Clowse MEB, Li J, Birru Talabi M, et al. The frequency of contraception documentation in women with systemic lupus erythematosus and rheumatoid arthritis within the Rheumatology Informatics System for Effectiveness Registry. Arthritis Care Res (Hoboken) 2023;75(3):590–596.
12. Finer LB, Zolna MR. Declines in unintended pregnancy in the United States, 2008-2011. N Engl J Med 2016;374(9):843–852.
13. Kim TY, Dagher RK, Chen J. Racial/Ethnic differences in unintended pregnancy: evidence from a national sample of U.S. women. Am J Prev Med 2016;50(4):427–435.
14. New estimates highlight differences in growth between the U.S. Hispanic and non-Hispanic populations. United States Census Bureau. June 27, 2024. Accessed April 1, 2025. https://www.census.gov/newsroom/press-releases/2024/population-estimates-characteristics.html
15. Frey WH. Census shows America's post-2020 population is driven by diversity especially among the young. July 24, 2024. Accessed April 1, 2025. https://www.brookings.edu/articles/census-shows-americas-post-2020-population-is-driven-by-diversity-especially-among-the-young/?utm_campaign=Metro%20Update&utm_medium=email&utm_content=318370993&utm_source=hs_email
16. Sammaritano LR, Bermas BL, Chakravarty EE, et al. 2020 American College of Rheumatology Guideline for the management of reproductive health in rheumatic and musculoskeletal diseases. Arthritis Rheumatol 2020;72(4):529–556.
17. Cooke P, Morales-Alemán MM, Ferreti G. Healthcare provider perceptions of the use of medical interpretation in primary care. South Med J 2024;117(5):221–225.
18. Karliner LS, Jacobs EA, Chen AH, et al. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. Health Serv Res 2007;42(2):727–754.
19. Flores G. The impact of medical interpreter services on the quality of health care: a systematic review. Med Care Res Rev 2005;62(3):255–299.
20. Lee JS, Pérez-Stable EJ, Gregorich SE, et al. Increased access to professional interpreters in the hospital improves informed consent for patients with limited English proficiency. J Gen Intern Med 2017;32(8):863–870.
21. Tarter L, Bermas BL. Expert perspective on a clinical challenge: lupus and pregnancy. Arthritis Rheumatol 2024;76(3):321–331.
22. Lv J, Xu L, Mao S. Association between disease activity of rheumatoid arthritis and maternal and fetal outcomes in pregnant women: a systematic review and meta-analysis. BMC Pregnancy Childbirth 2023;23(1):724.
23. Davis-Porada J, Kim MY, Guerra MM, et al. Low frequency of flares during pregnancy and post-partum in stable lupus patients. Arthritis Res Ther 2020;22(1):52.
24. Smith CJF, Förger F, Bandoli G, et al. Factors associated with preterm delivery among women with rheumatoid arthritis and women with juvenile idiopathic arthritis. Arthritis Care Res (Hoboken) 2019;71(8):1019–1027.
25. Gerosa M, Chighizola CB, Pregnolato F, et al. Pregnancy in juvenile idiopathic arthritis: maternal and foetal outcome, and impact on disease activity. Ther Adv Musculoskelet Dis 2022;14:1759720X221080375.
26. Bobircă A, Simionescu AA, Mușetescu AE, et al. Outcomes of prospectively followed pregnancies in rheumatoid arthritis: a multicenter study from Romania. Life (Basel) 2023;13(2):359.
27. Huang W, Wu T, Jin T, et al. Maternal and fetal outcomes in pregnant women with rheumatoid arthritis: a systematic review and meta-analysis. Clin Rheumatol 2023;42(3):855–870.
28. Kompa K, Kyttaris V. Rheum for improvement: contraception counseling and adherence in patients with lupus and rheumatoid arthritis on teratogenic medications [abstract]. Arthritis Rheumatol 2024;76-(suppl 9). https://acrabstracts.org/abstract/rheum-for-improvement-contraception-counseling-and-adherence-in-patients-with-lupus-and-rheumatoid-arthritis-on-teratogenic-medications/
29. Clowse MEB, Li J, Talabi MB, et al. Frequency of contraception documentation in women with systemic lupus erythematosus and rheumatoid arthritis within the Rheumatology Informatics System for Effectiveness Registry. Arthritis Care Res (Hoboken) 2023;75(3):590–596.
30. Weisleder H, Osuorji C, Fazzari M, et al. Improving age disparities in contraceptive counseling in women with lupus: a quality improvement initiative. Arthritis Rheumatol 2024;76(suppl 9). https://acrabstracts.org/abstract/improving-age-disparities-in-contraceptive-counseling-in-women-with-lupus-a-quality-improvement-initiative/
31. Chandramouli S, Alvarez C, Englund TR, et al. Teratogenic medication use associated with favourable odds of contraception counselling in a cohort of women with systemic lupus erythematosus at a large tertiary academic medical centre. Lupus Sci Med 2022;9(1):e000823.
32. Skinner-Taylor CM, Perez-Barbosa L, Barriga-Maldonado ES, et al. Reproductive health counseling and contraceptive use in Mexican women with rheumatic diseases: a cross-sectional study. Rheumatol Int 2021;41(2):409–414.
33. Bechman K, Subesinghe S, Norton S, et al. A systematic review and meta-analysis of infection risk with small molecule JAK inhibitors in rheumatoid arthritis. Rheumatology (Oxford) 2019;58(10):1755–1766.
34. Zhu J, Zhu Q, Li X, et al. Systemic lupus erythematosus and male reproductive health: a systematic review and meta-analysis. Autoimmun Rev 2025;24(4):103742.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 647–661
DOI 10.1002/acr.25674

American College of Rheumatology
Empowering Rheumatology Professionals

The Usefulness of the Patient-Reported Outcomes Measurement Information System Scale v1.2 Global Health for Assessing Mental and Physical Health of Individuals With Chronic Musculoskeletal Pain, Rheumatoid Arthritis, and Hip and Knee Osteoarthritis or Undergoing Physiotherapy: Results of Factor and Item Response Theory Analyses

Emanuele M. Giusti,[1] Leonardo Pellicciari,[2] Martine H. P. Crins,[3] Paul Dekker,[3,4] Martin van der Esch,[3,5] Marike van der Leeden,[3,6] Willem F. Lems,[3,6] Joost Dekker,[6] Maarten Boers,[6] Dirkjan van Schaardenburg,[3,6] Johan Joly,[7] Patrick Verschueren,[7,8] Kristien Van der Elst,[7] Rene Westhovens,[7,8] Caroline B. Terwee,[6,9] and Leo D. Roorda[3]

Objective. This study aims to evaluate whether the 10-item Patient-Reported Outcomes Measurement Information System Scale v1.2-Global Health (PROMIS-GH) is useful to assess global mental health (GMH) and global physical health (GPH) in individuals with musculoskeletal disorders.

Methods. PROMIS-GH was administered to 4,295 individuals (mean ± SD age 56 ± 14 years, 70% female, chronic musculoskeletal pain [n = 1,142], rheumatoid arthritis [n = 1,987], hip/knee osteoarthritis [n = 418], and undergoing physiotherapy [n = 947]). We investigated the legitimacy of calculating a GMH and a GPH subscale score (confirmatory factor analyses) and of converting raw ordinal subscale scores to interval scores (checking item response theory [IRT] assumptions [unidimensionality, local independence, and monotonicity] and graded response model fit), the ability of the subscales to discriminate different levels of health (items' discrimination parameters α) to cover the relevant range of GMH and GPH (range of difficulty parameters β) and their precision (item- and subscale-level information), and to compare demographic and clinical subgroups (differential item functioning [DIF]).

Results. It is legitimate to calculate a GMH and a GPH subscale score (comparative fit index = 0.98/0.97, Tucker-Lewis index = 0.97/0.95, root mean square error of approximation = 0.15/0.19, and standardized root mean square residual = 0.05/0.07) and to convert raw scores to interval scores (IRT assumptions met and adequate model fit). Discrimination (items' $\alpha \geq 2$), coverage (lowest $\beta \leq -2$) and precision (reliability ≥ 0.70 for large portions of the health domain) were adequate for all items, except item Global10, underscoring that they can distinguish individuals with different levels of GMH and GPH with precision. The subscales can be used to compare demographic and clinical subgroups (no DIF).

Conclusion. PROMIS-GH can be used to measure GMH and GPH in individuals with musculoskeletal disorders. Replacement or improvement of the Global10 item could be considered.

INTRODUCTION

Musculoskeletal disorders (MSDs), such as osteoarthritis (OA), rheumatoid arthritis (RA), and low back and neck pain, are highly prevalent, represent one of the most important causes of global disability,[1] and often impair health-related quality of life (HR-QoL) in the physical, mental, and social domains.[2] Therefore, using valid instruments to measure these domains is increasingly

Data collection for this project was financially supported by the Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam the Netherlands.

[1]University of Insubria, Varese, Italy; [2]IRCCS Istituto Delle Scienze Neurologiche di Bologna, Bologna, Italy; [3]Reade, Center for Rehabilitation and Rheumatology, Amsterdam, The Netherlands; [4]Leiden University Medical Center, Leiden, The Netherlands; [5]Amsterdam University of Applied Sciences, Amsterdam, The Netherlands; [6]Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam, The Netherlands; [7]University Hospitals Leuven, Leuven, Belgium; [8]Katholieke Universiteit Leuven, Leuven, Belgium; [9]Amsterdam Public Health Research Institute, Methodology, Amsterdam, The Netherlands.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25674).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25674.

SIGNIFICANCE & INNOVATIONS

- The 10-item Patient-Reported Outcomes Measurement Information System Scale v1.2 Global Health can be used to assess global mental health (GMH) and global physical health (GPH) in individuals diagnosed with chronic musculoskeletal pain, rheumatoid arthritis, or hip/knee osteoarthritis or those undergoing physiotherapy (77% musculoskeletal disorders).
- Ordinal scores from the GMH and GPH subscales can be converted into interval scores (*t*-scores), enabling easy interpretation and facilitating analysis of change; both subscales show adequate discrimination, coverage, and precision and can be used to compare demographic and clinical subgroups.
- Item Global10 (emotional problems) could be improved or replaced as it exhibits the highest misfit to a graded response model and does not improve the precision of the GMH subscale.

important for clinical and research applications, including drug approvals.[3–5] Several measurement instruments are available for these purposes.

Within the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative,[6] the PROMIS Scale Global Health (PROMIS-GH) aims to measure global mental health (GMH) and global physical health (GPH).[7] It is a 10-item self-report questionnaire that represents five core health domains (physical health, pain, fatigue, emotional distress, and social health) and generates GMH and GPH scores. The PROMIS-GH has been translated into multiple languages[8] and is recommended by the International Consortium for Health Outcomes Measurement for assessing HR-QoL in adults[9] and in core outcome sets for individuals with low back pain[10] and stroke.[11]

Although the PROMIS-GH has been recommended for use in clinical practice and research, surprisingly little psychometric evidence supports its validity in these contexts. Psychometric studies in individuals with MSDs mainly focused on the responsiveness[12–14] and construct validity[15] of the GMH and GPH subscales. Katzan and Lapin[16] found adequate internal consistency and construct validity in patients with stroke. However, important psychometric properties of the GMH and GPH subscales are understudied, including their scoring and interpretation, presence of floor/ceiling effects, precision, and universal applicability and suitability for subgroup comparisons (eg, comparing people with different MSDs).

Many researchers describing the validation of measurement instruments report their properties from the psychometrician's perspective. However, for the potential users of the measurement instrument, the decision on whether it is fit for clinical and research purposes depends on the practical implications of these properties and the ease of interpretation of the scores.[17] Framing research questions from the user's perspective is therefore important to overcome barriers for the instrument's use. This study aims to investigate, in individuals with MSDs, the necessity of calculating separate GMH and GPH scores from the PROMIS-GH, the legitimacy of converting ordinal to interval scores, the capability of the GMH and GPH subscales to discriminate between individuals with small differences in their mental and physical health, the precision of the subscale scores, and their ability to make valid comparisons between demographic and clinical subgroups.

PATIENTS AND METHODS

Participants. Data from four cohorts were used. The first cohort (Amsterdam pain cohort [AMS-PAIN]) included 1,247 Dutch individuals with chronic musculoskeletal pain who were treated at the rehabilitation outpatient department of Reade, a center for rehabilitation and rheumatology in the Netherlands.[18] The inclusion criteria were having at least one chronic musculoskeletal pain condition for ≥3 months before enrollment and being aged ≥21 years. The second cohort (RA) included 2,029 Dutch and Flemish individuals with RA who were treated at the rheumatology outpatient department of Reade and the University Hospitals Leuven, Flanders, Belgium, respectively.[19] The eligibility criteria were having RA according to the American College of Rheumatology (ACR)/EULAR 2010 criteria[20] and being aged ≥18 years. The third cohort (AMS-OA) included 425 individuals with hip or knee OA who were treated at the rehabilitation outpatient department of Reade.[21] The eligibility criteria included being diagnosed with knee or hip OA according to the ACR 1991 criteria.[22] The fourth cohort (physiotherapy [PT]) included 805 Dutch individuals who underwent any kind of physiotherapy (77% MSDs) in primary care in the year before completing the questionnaire.[23] Data were collected in several physical therapy practices across the Netherlands. Finally, data from the Dutch general population (n = 4,370)[24] were used to study measurement invariance between the clinical and general populations.

Procedures. Individuals were invited by email or letter to complete a digital or paper questionnaire, including questions on demographic and clinical characteristics, and the PROMIS-GH. Local institutional review boards approved each study protocol.

Address correspondence via email to Leonardo Pellicciari, PT, PhD, at leonardo.pellicciari@gmail.com.

Submitted for publication April 14, 2025; accepted in revised form October 2, 2025.

Measurement instrument. The PROMIS-GH consists of 10 items[7]: items Global02 (overall quality of life), Global04 (mental health), Global05 (satisfaction with social activities), and Global10 (emotional problems) make up the GMH subscale, and items Global03 (physical health), Global06 (physical function), Global07 (pain intensity), and Global08 (fatigue) make up the GPH subscale. The other two items, Global01 (general health) and Global09 (ability to carry out social activities), do not contribute to subscale scores but can be used as single items. Each item is scored on a five-point scale, except Global07, which is scored on an 11-point scale. Subscale scores are expressed as *t*-scores, which can be computed with the online scoring service provided by the US Assessment Center[25] or by calculating raw summed (ordinal) scores and converting them to (interval) *t*-scores (ie, scores with a mean of 50 and an SD of 10) based on reference values derived from a Dutch population.[26] Higher GMH and GPH scores indicate better health. The Dutch-Flemish v1.2 PROMIS-GH version was used in this study.[8]

Statistical analysis. We excluded data from n = 211 of 4,506 individuals (5%) because of missing PROMIS-GH responses. Details on missing data analysis are reported in the Supplementary Materials. Analyses were based on the PROMIS analysis plan.[27] A summary of the studied psychometric properties, the research questions underpinning them framed both from a user's and from a psychometrician's perspectives, and the statistical criteria employed to evaluate whether the PROMIS-GH achieved those psychometric properties is reported in Table 1.

The use of the PROMIS-GH in clinical populations for clinical and research purposes requires addressing the following questions.

1. Is it legitimate and necessary to calculate separate GMH and GPH scores? It is important to know whether the GMH and GPH assess different health domains. If not, one total score can be used.

From a psychometric perspective, this requires that a two-factor structure shows a good fit to the data, meaning that items reflecting the same health domain (eg, GMH or GPH) are correlated with each other and that after accounting for this health domain, very low residual correlations with items of the other health domain are present. We addressed this for the total sample by performing a confirmatory factor analysis (CFA).

Additionally, it is required that this two-factor structure is invariant across the four cohorts. We addressed this point by performing a multigroup CFA in which we evaluated across cohorts the presence of configural invariance (ie, items reflect the same health domains across cohorts) and metric invariance (ie, the item loadings, indicating the strength of the correlation of each item to the corresponding health domain, are equal across cohorts).

2. Is it legitimate to convert GMH and GPH ordinal scores to interval scores? Like all Likert-based subscales, raw GMH and GPH scores are ordinal, meaning the intervals between scores (eg, between 1 and 2 vs 2 and 3) may not be equal. Establishing that these scores can be converted to an interval scale (eg, *t*-scores) is essential for allowing meaningful comparisons between and within individuals.

From a psychometric perspective, this involves, firstly, ascertaining whether the GMH and GPH subscales meet the assumptions of an item response theory (IRT) model, that is, unidimensionality (items assess only one health domain), local independence (after accounting for the health domain, there is no residual correlation between items), and monotonicity (the probabilities of choosing higher response categories increase with increasing levels of the health domain). We checked unidimensionality for both subscales by assessing the fit of a unidimensional CFA model (ie, all items are correlated, and after accounting for the health domain there is low residual correlation between items) and by calculating two unidimensionality indices, explained common variance (ECV) and omega-hierarchical (ω_H), via an exploratory bifactor analysis.[28] We investigated local independence by ensuring there were low residual correlations between items in the unidimensional CFAs. We studied monotonicity through the items' H_i coefficients via Mokken scale analysis[29] and by visually inspecting the IRT-based item characteristic curves (ICCs) to ensure there were no disordered response categories. The ICCs represent the model-implied probability of endorsing each category across the health domains. Disordered categories indicate that individuals with higher health domain values are more likely to choose a lower category or vice versa.

After having checked these assumptions we evaluated the fit to the graded response model (GRM), an IRT model applicable to ordinal items, through Bonferroni-corrected S-χ^2 tests.[30] In case misfit was significant for an item, we assessed the magnitude of this misfit by inspecting the associated item-level root means square error of approximation (RMSEA) and empirical plots contrasting the GRM-based ICCs and the observed probabilities of endorsing each category by groups of individuals with similar values of the health domain. Note that the GRM, like all IRT models, operates on a latent trait metric, which represents an individual's position on the underlying latent trait (latent variable, construct, concept, or domain) being measured, which in this study are either the GMH or the GPH health domains. This metric is scaled with a mean of 0 and an SD of 1, where higher values indicate greater levels of the latent trait.

3. Can the GMH and GPH subscales discriminate between individuals with small differences in GMH and GPH? It is important that individuals with different levels of health get different subscale scores and that floor/ceiling effects, which occur when a considerable portion of respondents score at the lowest/highest possible levels, are absent.

In the IRT framework, a questionnaire's ability to distinguish between individuals with small differences in the measured health domains is quantified by the discrimination parameter α of each item. High discrimination indicates that individuals with small differences in the health domain have a different probability of endorsing a specific response category. Floor/ceiling effects are

Table 1. Summary of research questions from the user's and the psychometrician's perspectives, psychometric properties, statistical analyses, statistical indices, and criteria considered for the analysis of the PROMIS-GH*

Research question from the user's perspective	Research question from the psychometrician's perspective	Psychometric property	Analysis	Statistical index	Criteria with reference
1. Is it legitimate and necessary to calculate separate GMH and GPH scores?	Does the PROMIS-GH fit a two-factor structure?	Health domain structure	CFA[a,b]	CFI	≥0.95[27]
				TLI	≥0.95[27]
				RMSEA	≤0.06[27]
				SRMR	≤0.08[27]
	Is the two-factor structure invariant across cohorts?	Measurement invariance	Multigroup CFA (configural invariance model)[a]	CFI	≥0.95[27]
				TLI	≥0.95[27]
				RMSEA	≤0.06[27]
				SRMR	≤0.08[27]
			Multigroup CFA (metric invariance model vs configural invariance model)[a]	ΔRMSEA	≤0.05[45]
2. Is it legitimate to convert GMH and GPH ordinal scores to interval scores?	Do the items of both subscales assess only one health domain?	Unidimensionality[a]	CFA[b,c]	CFI	≥0.95[27]
				TLI	≥0.95[27]
				RMSEA	≤0.06[27]
				SRMR	≤0.08[27]
			Exploratory bifactor analysis[c]	ECV	>0.70[28]
				ω_H	>0.80[28]
	Do the items of both subscales relate only to the health domains being measured?	Local independence	Computation of the residual matrix from the unidimensional CFA[c]	Residual correlations between item pairs	≤0.20[27]
	Do the probabilities of affirmative responses to the items of both subscales increase with increasing levels of the health domains?	Monotonicity	Mokken scale analysis[c]	Item scalability (H_i)	≥0.50[29]
			Analysis of disordered categories[c]	Item characteristic curves	Graphical display[46]
	Can the relationship between the items of both subscales and the health domains be described using an IRT model?	Fit to an IRT model	GRM model fit[c]	Items' S-χ^2 and associated Bonferroni-corrected *P* value	$P \geq 0.05$[30]
				S-χ^2–based item-level RMSEA	≤0.06[27]
				Empirical plot	Graphical display[47]
3. Can the GMH and GPH subscales discriminate between individuals with small differences in GMH and GPH?	Do the items of both subscales and the total scores discriminate between different levels of the health domain?	Item discrimination coefficients	Analysis of GRM coefficients[c]	Items' α	≥1.35[48]
		Floor or ceiling effects	Analysis of scale scores	Percentage of participants obtaining the lowest or highest scale score	<15%
4. Do the GMH and GPH subscales cover the relevant (clinical) range of GMH and GPH?	Do the items of both subscales cover the relevant (clinical) range of the health domains?	Range of item difficulty coefficients	Analysis of GRM coefficients[c]	Range of items' β	Minimum β ≤ −2 and maximum β ≥ 2
5. Can the GMH and GPH subscales measure with high precision?	How much information do these subscales provide at different levels of the health domains?	Information	Analysis of total and item information curves[c]	Graphical display	Portion of the health domain with information ≥ 3.45[49]

(Continued)

Table 1. (*Cont'd*)

Research question from the user's perspective	Research question from the psychometrician's perspective	Psychometric property	Analysis	Statistical index	Criteria with reference
6. Can the GMH and GPH subscales be used to discriminate physical and mental health across demographic and clinical subgroups unbiased?	Can these measures be used to compare groups in terms of demographic and clinical variables?	Measurement invariance	Analysis of uniform DIF[c,d]	McFadden pseudo-R^2	<0.02[50]
			Analysis of nonuniform DIF[c,e]	McFadden pseudo-R^2	<0.02[50]
			Analysis of test characteristic curves across subgroups[c]	Test characteristic curves	Graphical display[34]

* The numbers next to the questions refer to the numbers of the measurement property reported in the methods. CFA, confirmatory factor analysis; CFI, comparative fit index; DIF, differential item functioning; ECV, explained common variance; GMH, general mental health; GPH, general physical health; GRM, graded response model; IRT, item response theory, PROMIS-GH, Patient-Reported Outcomes Measurement Information System Scale v1.2-Global Health; SRMR, standardized root mean square residual; TLI, Tucker-Lewis index; ΔRMSEA, change in root mean square error of approximation; ω_H, omega-hierarchical.

[a] The analysis was performed on the eight GMH and GPH items.

[b] Estimator: mean- and variance-adjusted weighted least squares.

[c] The analysis was performed separately on the GMH and GPH subscales.

[d] Comparison between ordinal logistic regression models in which item responses are predicted by the health domain versus by the health domain and group membership.

[e] Comparison between ordinal logistic regression models in which item responses are predicted by the health domain and group membership versus by the health domain, group membership, and their interaction.

considered present if >15% of the respondents get the lowest/highest subscale score.[31]

4. Do the GMH and GPH subscales cover the relevant (clinical) range of GMH and GPH? For both subscales, it is important that the entire range of clinically relevant variations in the health domain results in different scores.

Within the IRT framework, this aspect is quantified by the range of the difficulty parameter β, identifying the levels of the health domain where two adjacent score categories are equally probable. Adequate coverage of the health domain is achieved when the minimum and maximum difficulty parameters across the items are ≤ -2 and ≥ 2, respectively. This implies that at least one item can distinguish between individuals 2 SDs below the average and at least one can do so 2 SDs above.

5. Can the GMH and GPH subscales measure with high precision? Using GMH and GPH scores to support clinical decisions requires precise (accurate) estimation of the scores.

Within the IRT framework, this aspect is quantified by the item and subscale-level information, that is, the reciprocal of the SE at a given point on the health domain. We investigated the portion of the health domain with information >3.45, equivalent to a reliability coefficient of 0.70.

6. Can the GMH and GPH subscales be used to discriminate physical and mental health across demographic and clinical subgroups unbiased? Differences between groups in GMH or GPH subscales should reflect true differences rather than different functioning of the subscales across groups.

Within the IRT framework, this requires that the IRT coefficients are similar across groups, meaning that individuals from different populations with similar levels of health respond similarly to the items. This is assessed via differential item functioning (DIF) analysis, which evaluates the presence of uniform (constant magnitude of DIF across the entire health domain range) and nonuniform DIF (varying magnitude of DIF across the health domain).[32] Using a logistical regression framework,[33] we explored DIF for cohort (AMS-PAIN, RA, AMS-OA, and PT), sex (male or female), age (<57 years and ≥57 years; 57 being the sample median), education (none/primary school, secondary school/college, and university/higher), disease duration (<1 year, 1–5 years, and >5 years), and administration modality (digital or paper). In addition, also using the data from the Dutch general population, DIF was investigated between clinical and general populations. If DIF was detected, we evaluated its impact on the total score by inspecting the test characteristic curves (TCCs), representing the relationship between the health domain and the expected subscale score for a given value of the health domain for each group.[34] All the analyses were repeated using data from each cohort. The analyses were performed using R (version 4.2.1).

RESULTS

Participant characteristics. Table 2 summarizes the demographic and clinical characteristics of the participants. The mean age was different across the cohorts, with the AMS-PAIN cohort being younger and the AMS-OA cohort being older. The AMS-OA also had the lowest percentage of workers. More than half of the individuals in the AMS-PAIN, RA, and AMS-OA cohorts had a disease duration >5 years, whereas in the PT cohort more than half of the individuals had a disease duration <1 year. The results of the psychometric analyses are summarized in Table 3.

1. Is it legitimate and necessary to calculate both GMH and GPH scores in the considered clinical populations? The two-factor model showed three adequate fit indices (comparative fit index [CFI] = 0.98, Tucker-Lewis index [TLI] = 0.97, and standardized root mean square residual [SRMR] = 0.05) and one suboptimal index (RMSEA = 0.15) (Table 3, question 1). A one-factor model applied to the eight GMH and GPH items together showed worse fit (CFI = 0.97, TLI = 0.95, RMSEA = 0.19, and SRMR = 0.07). Both analyses support that GMH and GPH subscales measure distinct health domains.

The two-factor structure of the PROMIS-GH showed configural and metric invariance across the four cohorts (Table 3, question 1). Similar to the one-group CFA, the model evaluating configural invariance had adequate CFI values and suboptimal RMSEA (RMSEA = 0.15). Constraining the item loadings to be equal across the cohorts improved the fit indices, indicating the presence of metric invariance (ΔCFI = 0.01 and RMSEA = −0.03). Overall, these results support the use of separate GMH and GPH scores in all four cohorts.

2. Is it legitimate to convert the GMH and GPH ordinal subscale scores to interval scores? IRT assumptions. The unidimensional models applied to the GMH and GPH subscales (Table 3, question 2) showed three adequate fit indices (CFI ≥ 0.95, TLI ≥ 0.95, and SRMR ≤ 0.08) and one suboptimal index (RMSEA = 0.15 and 0.10, respectively). The exploratory bifactor analysis indicated that both indices of the GMH subscale were above the respective cutoff, whereas for the GPH subscale the ECV was above but the ω_H was slightly below the cutoff (ω_H = 0.79). Considering these results, we deemed both subscales as unidimensional. All residual correlations between item pairs were <0.20, indicating sufficient local independence. The Mokken scale analysis revealed that the items in both subscales had high scalability (Table 4). The visual inspection of the ICCs showed no disordered categories (Supplementary Figures S1 and S2). Therefore, the monotonicity assumption was met by both subscales.

IRT model fit. Considering the S-χ^2 values, all items revealed misfits to the GRM. However, item-level RMSEAs were below the prespecified cutoff, suggesting that model-predicted item response probabilities and observed response probabilities were close. Visual inspection of the empirical plots indicated that Global10 within the GMH subscale (Supplementary Figures S3 and S4) showed the highest misfit, as the probability of individuals with similar health domain values endorsing each category did not

Table 2. Demographic and clinical characteristics of the total analyzed sample and patients of the four cohorts*

Variable	Total sample (N = 4,295)	Cohorts			
		AMS-PAIN (n = 1,142)	RA (n = 1,987)	AMS-OA (n = 418)	PT (n = 748)[a]
Age, mean ± SD, y	56 ± 14	49 ± 13	59 ± 13	67 ± 9	54 ± 14
Sex, n (%)					
Male	1,296 (30)	254 (22)	618 (31)	116 (28)	308 (41)
Female	2,999 (70)	888 (78)	1,369 (69)	302 (72)	440 (59)
Education, n (%)					
None or primary	445 (10)	178 (16)	226 (11)	21 (5)	20 (3)
High school	2,429 (57)	565 (49)	1,222 (62)	251 (60)	391 (52)
University or higher	1,206 (28)	202 (18)	525 (26)	142 (34)	337 (45)
Missing	215 (5)	197 (17)	14 (1)	4 (1)	0 (0)
Employment status, n (%)					
Student	45 (1)	23 (2)	9 (0)	1 (0)	12 (2)
Worker (full- or part-time)	1,702 (40)	433 (38)	755 (38)	83 (20)	431 (57)
Unemployed, volunteer, or household	700 (16)	293 (26)	272 (14)	61 (15)	74 (10)
Retired	1,180 (28)	89 (8)	688 (35)	245 (59)	158 (21)
Other or more than one status	609 (14)	279 (24)	231 (12)	26 (6)	73 (10)
Missing	59 (1)	25 (2)	32 (1)	2 (0)	0 (0)
Social status, n (%)					
Single	1,065 (25)	386 (34)	434 (22)	115 (27)	130 (17)
Married or living together	2,903 (68)	607 (53)	1,452 (73)	277 (66)	567 (76)
Living apart together	141 (3)	54 (5)	46 (2)	18 (4)	23 (3)
Living with parents	56 (1)	23 (2)	21 (1)	0 (0)	12 (2)
Other or more than one status	121 (2)	66 (6)	31 (12)	8 (2)	16 (2)
Missing	9 (0)	6 (0)	3 (0)	0 (0)	0 (0)
Disease duration, n (%)					
<1 y	495 (12)	67 (6)	51 (3)	2 (0)	375 (50)
1–5 y	1,240 (29)	475 (42)	451 (23)	99 (24)	215 (29)
≥5 y	2,549 (59)	597 (52)	1,479 (74)	315 (75)	158 (21)
Missing	11 (0)	3 (0)	6 (0)	2 (0)	0 (0)
Administration modality, n (%)					
Digital	2,740 (64)	487 (43)	1,189 (60)	316 (76)	748 (100)
Paper	1,555 (36)	655 (57)	798 (40)	102 (24)	0 (0)
Global mental health, mean ± SD, *t*-scores	45 ± 10	38 ± 9	476 ± 10	48 ± 10	50 ± 9
Global physical health, mean ± SD, *t*-scores	41 ± 10	34 ± 8	43 ± 9	43 ± 10	48 ± 9

* AMS-OA, Amsterdam osteoarthritis cohort; AMS-PAIN, Amsterdam pain cohort; PT, physiotherapy cohort; RA, rheumatoid arthritis cohort.
[a] Includes 77% musculoskeletal disorders.

follow the expected functional form under the GRM. Considering the overall evaluation of these results, the fit of the IRT model was considered adequate enough for both subscales. The estimated IRT parameters are reported in Table 4.

3. Can the GMH and GPH subscales be used to discriminate individuals with small differences in their levels of GMH and GPH? Inspection of the IRT parameters showed that all items had discrimination parameters $\alpha \geq 2$, indicating adequate ability to discriminate between individuals with small differences in GMH and GPH (Table 4). Floor/ceiling effects were not detected.

4. Do the GMH and GPH subscales cover the relevant range of GMH and GPH? Inspection of the IRT parameters showed that the GMH items had adequate coverage in the lower end of the health domain but not in its higher end (highest $\beta < 2$), whereas the GPH subscale had adequate coverage in both ends (Table 4).

5. Can the GMH and GPH subscales be used to assess individuals with different levels of the relevant health domain with precision? Figure 1 represents the test information curves of the GMH and GPH subscales and the item information curves. The information of both subscales was higher than the cutoff of 3.45 (equivalent to reliability ≥ 0.70) for large portions of the health domain (GMH subscale: from −2.3 to 2.0; GPH subscale: from -2.8 to 2.2). Inspection of the item information curves of the GMH subscale revealed that Global10 provided little information along the health domain but provided more information than the other items at very low health domain values.

6. Can the GMH and GPH subscales be used to compare physical and mental health across demographic and clinical subgroups? Supplementary Table S1 reports the results of the DIF analyses. Only uniform DIF was detected for Global10. Visual inspection of the TCCs for the GMH subscale across cohorts revealed that the impact of Global10 DIF on the total GMH score was minimal (Figure 2). No other DIF was detected.

Results of the analyses performed separately on data from each cohort. Subgroup analyses in the four cohorts are reported in the Supplementary Materials. Overall, the IRT

Table 3. Summary of the results of the CFA and IRT analysis of the PROMIS-GH for the total sample*

Research question from a user's perspective	Psychometric property	Analysis	Statistical index	Results			Psychometric conclusion	Practical implications
				Total scale	GMH	GPH		
1. Is it legitimate and necessary to calculate separate GMH and GPH scores?	Health domains structure	CFA[a]	CFI	0.98	–	–	The two-factor model showed suboptimal fit, but the unidimensional model fit worse, indicating that GMH and GPH subscales assess distinct health domains	It is legitimate and necessary to calculate two scale scores (GMH and GPH)
			TLI	0.97	–	–		
			RMSEA	**0.15**	–	–		
			SRMR	0.05	–	–		
	Measurement invariance	Multigroup CFA (configural invariance model)[a]	CFI	0.97	–	–	The two-factor structure meets configural and metric invariance	It is legitimate and necessary to calculate two scale scores (GMH and GPH) in all four clinical cohorts
			RMSEA	**0.15**	–	–		
		Multigroup CFA (metric vs configural invariance model)[a]	CFI	0.98	–	–		
			RMSEA	0.12	–	–		
			ΔRMSEA	-0.03	–	–		
2. Is it legitimate to convert GMH and GPH ordinal scores to interval scores?	Unidimensionality[a]	CFA[b]	CFI	–	0.99	0.99	The GMH and the GPH show minor departures from unidimensionality	It is legitimate to convert the GMH and GPH scores to true numeric values
			TLI	–	0.99	0.99		
			RMSEA	–	**0.15**	**0.10**		
			SRMR	–	0.02	0.01		
		Exploratory Bifactor Analysis[b]	ECV	–	0.91	0.86		
			ω_H	–	0.84	**0.79**		
	Local independence	Computation of residual matrix from the unidimensional CFA[b]	Residual correlations between item pairs	–	All ≤ 0.20	All ≤ 0.20	The GMH and GPH show sufficient local independence	
	Monotonicity	Mokken scale analysis[b]	Item scalability (H_i)	–	Range: 0.68–0.77 (Table 4)	Range: 0.66–0.70 (Table 4)	The GMH and GPH items are monotonic	
			Item characteristic curves	–	No disordered categories (Supplementary Figure S1)	No disordered categories (Supplementary Figure S2)		
	Fit to an IRT model	GRM model fit[b]	S-χ^2 and *P* value of the items	–	***P* < 0.05** for all items (Table 4)	***P* < 0.05** for all items (Table 4)	Overall, the fit to the IRT model was considered adequate, with item Global10 displaying the highest misfit	
			S-χ^2–based item-level RMSEA	–	All ≤ 0.06	All ≤ 0.06		
			Empirical plots	–	Global10 had the highest misfit (Supplementary Figure S3)	Negligible misfits (Supplementary Figure S4)		

(Continued)

Table 3. (*Cont'd*)

Research question from a user's perspective	Psychometric property	Analysis	Statistical index	Results			Psychometric conclusion	Practical implications
				Total scale	GMH	GPH		
3. Can the GMH and GPH subscales discriminate between individuals with small differences in GMH and GPH?	Range of item discrimination coefficients	Analysis of GRM coefficients[b]	α_i	–	≥ 1.35 (Table 4)	≥ 1.35 (Table 4)	The ability to discriminate individuals with small differences in GMH or GPH was adequate	The GMH and GPH scales can distinguish well between different levels of the relevant health domain
	Floor and ceiling effect	Analysis of scale scores	Percentages of participants obtaining the lowest or highest subscale score	–	Lowest score: 1.2%; highest score: 4.1%	Lowest score: 0.3%; highest score: 1.3%	No floor or ceiling effects	
4. Do the GMH and GPH subscales cover the relevant (clinical) range of GMH and GPH?	Range of item difficulty coefficients	Analysis of GRM coefficients[b]	β_i	–	Range: −2.62 to 1.57	Range: −2.74 to 2.24	The GMH subscale is not targeted to measure the health domain in those with high values, indicating good health. The GPH has adequate coverage	The GMH and GPH subscales cover the relevant (clinical) range of the health domain
5. Can the GMH and GPH subscales measure with high precision?	Information	Analysis of total information curves[b]	Graphic display	–	Information ≥ 3.45 between −2.28 and 2.03	Information ≥3.45 between −2.80 and 2.18	Adequate information along the health domain	The GPH and GMH assess precise (or reliable)
6. Can the GMH and GPH subscales be used to discriminate physical and mental health across demographic and clinical subgroups unbiased?	Measurement invariance	DIF analysis (ordinal logistic regression framework)[b]	Change in McFadden R^2	–	McFadden R^2 = **0.32** for Global10, indicating uniform DIF across cohorts (Supplementary Table S1)	No DIF was detected	DIF across the considered subgroups was minimal or absent	The GMH and GPH can be used to compare groups across the considered pain conditions, sex, age, education, disease duration, administration modality, and across the clinical/general population
			Test characteristic curves	–	The impact of DIF of Global10 on the total score is minimal (Figure 2)	–		

* Statistical values beyond the recommended cutoff are presented in bold. CFA, confirmatory factor analysis; CFI, comparative fit index; DIF, differential item functioning; ECV, explained common variance; GMH, general mental health; GPH, general physical health; GRM, graded response model; IRT, item response theory; PROMIS-GH, Patient-Reported Outcomes Measurement Information System Scale v1.2 Global Health; SRMR, standardized root mean square residual; TLI, Tucker-Lewis index; ΔRMSEA, change in root mean square error of approximation; ω_H, omega-hierarchical.

[a] The analysis was performed on the 8 PROMIS-GH items intended to measure GMH and GPH.

[b] The analysis was performed separately on the GMH and GPH scales.

Table 4. Descriptive statistics, scalability coefficients, fit to a graded response model and IRT parameters of the Patient-Reported Outcomes Measurement Information System Scale v1.2 Global Health for the total sample*

Item	Content	Item descriptive statistics Mean (SD)	Mokken scale analysis H (SE)	Graded response model fit[a] S-χ^2	*P* value	IRT parameters[b] α (SE)	β_1 (SE)	β_2 (SE)	β_3 (SE)	β_4 (SE)
Global mental health										
Global02	Quality of life	2.92 (1.07)	0.73 (0.01)	62.34	<0.001	3.28 (0.09)	−1.58 (0.03)	−0.40 (0.02)	0.64 (0.02)	1.57 (0.03)
Global04	Mental health	3.09 (1.13)	0.77 (0.01)	105.27	<0.001	4.09 (0.14)	−1.60 (0.03)	−0.50 (0.02)	0.42 (0.02)	1.22 (0.03)
Global05	Satisfaction with social activities	2.97 (1.13)	0.74 (0.01)	85.62	<0.001	3.84 (0.12)	−1.45 (0.03)	−0.40 (0.02)	0.50 (0.02)	1.39 (0.03)
Global10	Emotional problems	3.61 (1.11)	0.68 (0.01)	183.09	<0.001	2.06 (0.06)	−2.62 (0.07)	−1.19 (0.03)	−0.12 (0.02)	0.83 (0.03)
Global physical health										
Global03	Physical health	2.47 (0.99)	0.68 (0.01)	51.98	<0.001	2.54 (0.08)	−1.28 (0.03)	0.20 (0.02)	1.27 (0.03)	2.24 (0.05)
Global06	Physical activities	3.61 (1.19)	0.70 (0.01)	88.7	<0.001	2.72 (0.08)	−2.13 (0.05)	−1.02 (0.03)	−0.13 (0.02)	0.61 (0.02)
Global07	Pain	3.22 (1.02)	0.67 (0.01)	78.89	<0.001	2.68 (0.08)	−2.74 (0.08)	−0.65 (0.03)	0.25 (0.02)	1.47 (0.04)
Global08	Fatigue	3.14 (1.06)	0.66 (0.01)	42.92	0.03	2.48 (0.07)	−2.00 (0.05)	−0.71 (0.03)	0.42 (0.02)	1.51 (0.04)

* IRT, item response theory.
[a] The fit of the models was evaluated using a generalization of the Orlando and Thissen S-χ^2 test for polytomous data.
[b] "α" indicates item slopes, that is, the ability of an item to discriminate between people with different values on the health domain. "β" indicates item difficulty parameters, that is, the points in the health domain where two adjacent score categories are equally probable.

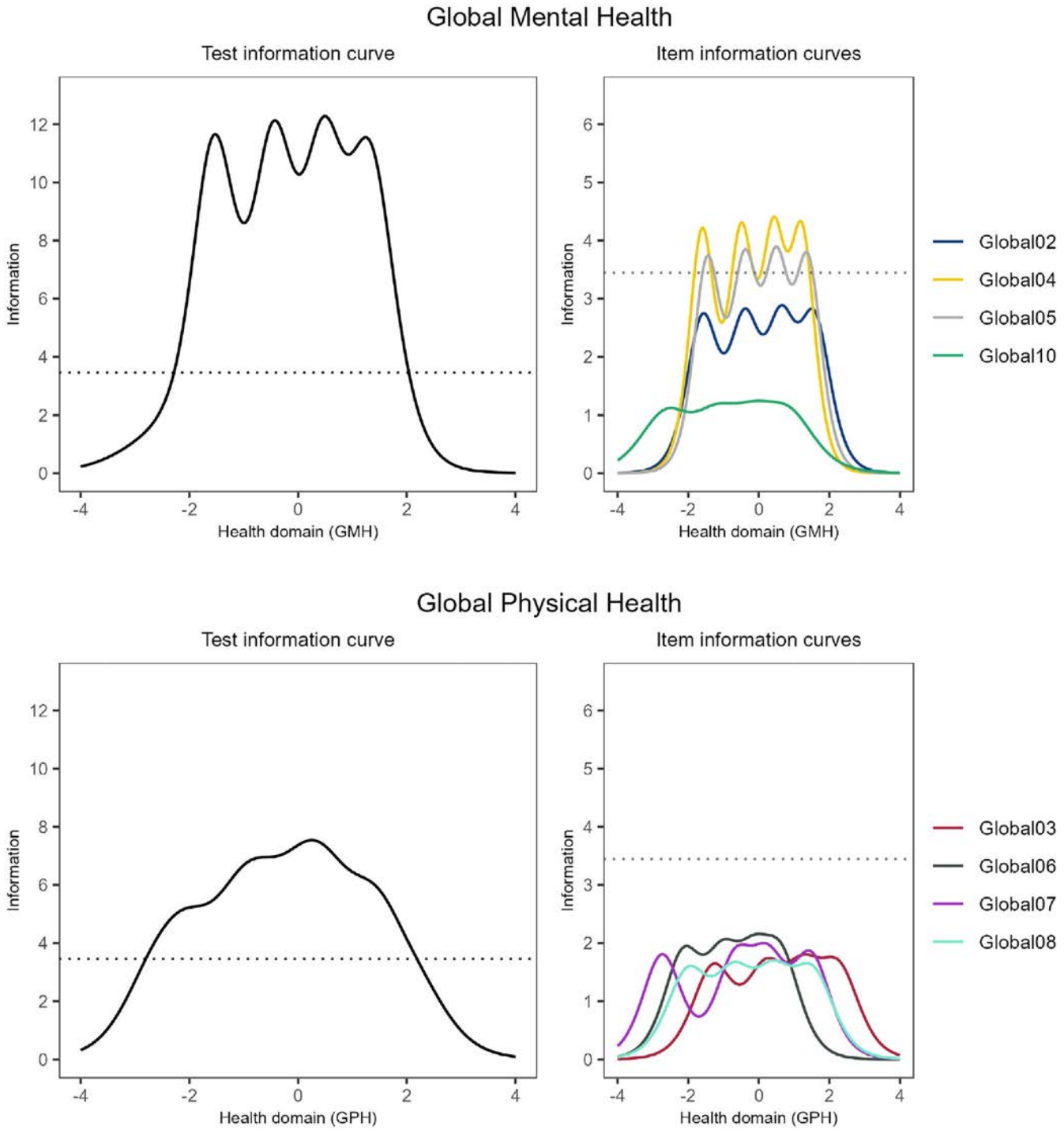

Figure 1. Test and item information curves of the GMH and the GPH for the total sample. The dotted line represents an information value of 3.45, corresponding to a reliability coefficient of 0.70. Health domain values have a mean of 0 and an SD of 1. For both scales, information exceeds this cutoff across large portions of the health domain. Global10 contributes minimal information across the health domain but provides the highest information at very low levels of GMH. GMH, global mental health; GPH, global physical health.

assumptions were met in each cohort. The analysis of the fit to the GRM model showed nonnegligible misfit for Global10 in all the cohorts and for Global07 in the PT cohort. The discrimination, coverage, and precision were adequate, except for the GPH subscale in the PT cohort, where the information drops below the cutoff for health domain values exceeding 1.29. Only a small DIF for age in Global06 was found in the RA cohort and small DIF for education in the same item in the AMS-OA cohort.

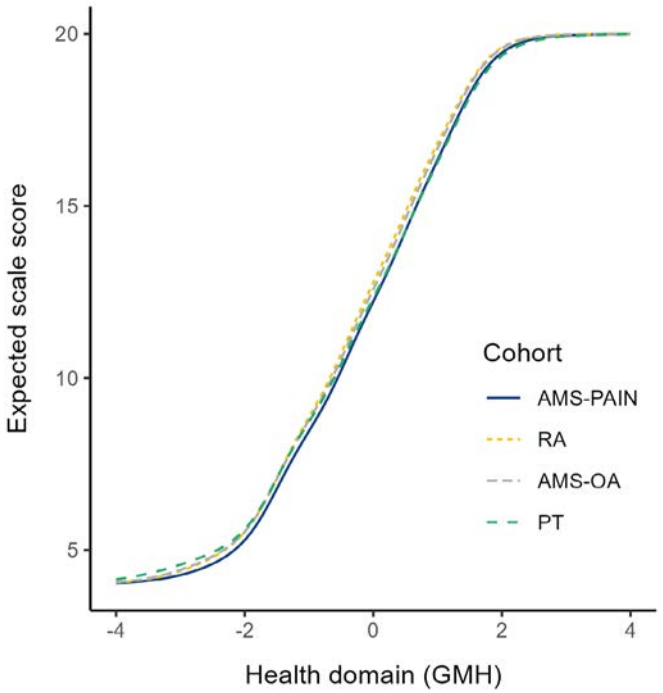

Figure 2. Test characteristic curves of the GMH subscale across the studied cohorts. Each line represents the relationship between the health domain and the expected raw subscale score for a given value of the health domain. Health domain values have a mean of 0 and an SD of 1. The minimal separation between the lines indicates that the differential item functioning caused by Global10 has little impact on the total score. AMS-OA, Amsterdam osteoarthritis cohort; AMS-PAIN, Amsterdam pain cohort; GMH, global mental health; PT, physiotherapy cohort; RA, rheumatoid arthritis cohort. Color figure can be viewed in the online issue, which is available at http://onlinelibrary.wiley.com/doi/10.1002/acr.25674/abstract.

DISCUSSION

This study showed that the PROMIS-GH can be used to measure GMH and GPH in individuals with chronic musculoskeletal pain, those with RA, those with hip/knee OA, or in those undergoing physiotherapy. First, we confirmed that it is legitimate and necessary to calculate separate GMH and GPH scores. Second, we showed that the GMH and GPH scores can be converted to interval scores. Third, we demonstrated that the GMH and GPH subscales have adequate coverage and can be used to discriminate between individuals with slightly different levels of GMH and GPH. Fourth, we demonstrated that these subscales can be used to measure physical and mental health precisely, including in people with low levels of global mental or physical health. Fifth, we found that the two subscale scores can be used to compare GMH and GPH across musculoskeletal pain diagnoses, sex, age, education, administration modality, and across the clinical and general populations unbiased. Finally, we found room for improvement in the GMH subscales, as Global10 exhibited the highest misfit to an IRT model, poor precision and, albeit with minimal impact on the GMH score, DIF across cohorts.

Generic HR-QoL measures such as the PROMIS-GH are widely used in the field of MSDs for research (eg, clinical trials) and economic and quality assessments and are recommended for use in clinical practice.[10,35,36] These measures are useful throughout the course of care as they provide benchmarks and allow the tracking of clinical progress, thus facilitating shared decision-making during initial consultations and improving clinician-patient communication during treatment.[35] Compared with domain- and disease-specific instruments, generic measures are valuable as they efficiently capture the overall health status, reflecting the extent to which patients feel and function "normally" and can be harmonized across patient populations. This aligns with the fact that patients with MSDs rank enhancing participation in day-to-day living and overall functioning among the most important treatment goals[37,38] and that different conditions often affect overlapping health domains.[36] In addition, generic measures overcome limitations of disease-specific instruments, such as the cost and time demands of maintaining different measures for different patient groups in electronic health records, the complexity of interpreting different scales and cut-offs, and the burden on patients with multiple conditions who would be required to complete overlapping questionnaires.[39] Disease- and domain-specific questionnaires may be preferred when a more detailed assessment of a particular health aspect is needed, for example when evaluating targeted interventions or when specific information is required for decision-making.

Compared with other generic HR-QoL measures, the PROMIS-GH has multiple advantages. A prior review examined the psychometric properties and suitability for clinical and research use of various instruments and promoted using the PROMIS-GH or the 12-item Short Form (SF-12) as brief measures of physical and mental health.[40] However, unlike the PROMIS-GH, the SF-12 has not yet undergone a thorough IRT analysis in patients with MSDs and requires purchasing licenses, manuals, and scoring software. Other brief tools, such as the EuroQol 5 Dimension (five items), showed worse psychometric characteristics or limited research.[40] Longer instruments, such as the 29-item PROMIS, the 36-item Short Form (completion time range: 7–10 minutes), or the World Health Organization QoL short version (26 items; completion time range: 4–20 minutes) have not been tested with IRT analyses in individuals with chronic pain and are more time-consuming.[40] To obtain a general measure of GMH and GPH, the PROMIS-GH should be preferred, also because of its brevity (reported completion time of 2 minutes).[7]

Our current results partly align with, and partly differ from, those of our previous study addressing the psychometric properties of the PROMIS-GH in a general population.[24] Both studies found that the GMH and GPH subscales meet the IRT assumptions and provide a precise measure of the relevant trait, and in both studies Global10 showed the highest misfit to the IRT model and low precision. Our previous research

identified some issues that we did not observe in this study, such as a slight deviation from monotonicity in Global06 and a slight age DIF for Global08. Taking all these results together, there are no major differences in the psychometric properties of the PROMIS-GH in these clinical populations versus a general population. In this study, we found no evidence of DIF between the clinical and general populations, and in our previous study we observed no DIF between the US and Dutch general populations.[24] Based on these considerations, it can be argued that the results of this study can be generalized to the English version of the PROMIS-GH and US populations.

The analyses raised some issues that turned out to have little practical implications. We found that RMSEA values exceeded the prespecified cutoff for the analysis of the scale unidimensionality. It has been argued, however, that this fit index is not suitable for assessing this property, as it is weakly associated with indices suggesting the strength of a general health domain.[39] Additionally, the exploratory bifactor analysis showed that the two subscales can be considered "unidimensional enough."[28,41] In the analysis of the fit to the IRT models, tests for item-level misfit yielded significant results. However, significant tests do not necessarily imply that the misfit has practical significance. Indeed, the magnitude of misfit was negligible. It is likely that the large sample size increased the risk of Type I errors.[42] Finally, the GMH subscale lacks an item with difficulty higher than two, indicating that no item is specifically targeted at individuals with GMH 2 SDs above the mean. However, this limitation is minor, as in clinical populations it is more important to differentiate individuals at the lower end of the health spectrum.

Global10 exhibited the highest misfit to the IRT model and low precision, similarly to the previous study conducted in the Dutch general population,[24] which might have biased its IRT parameters.[43] This could be attributed to content issues, specifically its phrasing, which combines experiencing emotional problems, but also being bothered by them. Despite its shortcomings, there are some arguments for continuing to use the subscale with this item included. Global10 exhibited the lowest difficulty and provided the highest information at very low levels of mental health. Consequently, removing it may reduce the coverage of GMH. Therefore, our current recommendation is to retain Global10. However, future studies should consider replacing this item with a better item or comparing the current and reworded item.

The DIF analysis revealed that the PROMIS-GH can be used to compare groups based on sex, age, education, cohort, disease duration, and administration modality. We found a negligible DIF across cohorts in Global10, which had few practical implications as it did not impact the GMH score. The fact that the PROMIS-GH functions similarly in the general population and clinical populations implies that the manual and the online scoring services, calibrated on individuals from the general population, can also be used for individuals with MSDs.[25,44] The subgroup analyses in each cohort align with those of the total sample, with some minor issues potentially attributable to multiple testing.

This study has several strengths, including its large sample size and the comparison of the parameters of the populations with MSDs with those of the general population. Nonetheless, it is to be noted that the large sample size might have influenced statistical testing, increasing Type I errors in the assessment of misfits to the IRT model. This limitation was addressed by assessing the impact of misfits using graphical methods. The main limitation of this study is that generalizing these results to other clinical populations (eg, individuals with stroke or other MSDs or shorter disease duration) should be done with caution.

Our findings support scoring the PROMIS-GH using its GMH and GPH subscale scores, which can be converted to interval scores using the online scoring service[25] or the conversion tables provided in the manual.[44] The PROMIS-GH should be preferred over other instruments when a brief instrument is required for the assessment of GPH and GMH in individuals with chronic musculoskeletal pain, those with RA, those with hip/knee OA, or those undergoing physiotherapy.

ACKNOWLEDGMENT

The authors thank Hannah Forest for her valuable help with English-language editing.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Pellicciari confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Liu S, Wang B, Fan S, et al. Global burden of musculoskeletal disorders and attributable factors in 204 countries and territories: a secondary analysis of the Global Burden of Disease 2019 study. BMJ Open 2022;12(6):e062183.
2. Beaudart C, Biver E, Bruyère O, et al. Assessment of quality of life in musculo-skeletal health. Aging Clin Exp Res 2018;30(5):413–418.
3. Pogatzki-Zahn E, Schnabel K, Kaiser U. Patient-reported outcome measures for acute and chronic pain: current knowledge and future directions. Curr Opin Anaesthesiol 2019;32(5):616–622.
4. Meregaglia M, Malandrini F, Angelini S, et al. The assessment of patient-reported outcomes for the authorisation of medicines in Europe: a review of European public assessment reports from 2017 to 2022. Appl Health Econ Health Policy 2023;21(6):925–935.

5. US Food & Drug Administration. FDA patient-focused drug development guidance series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making. Published February 14, 2024. Accessed January 29, 2025. https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical
6. Cella D, Yount S, Rothrock N, et al; PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45(5)(suppl 1):S3–S11.
7. Hays RD, Bjorner JB, Revicki DA, et al. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009;18(7):873–880.
8. Terwee CB, Roorda LD, de Vet HCW, et al. Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). Qual Life Res 2014;23(6):1733–1741.
9. Gangannagaripalli J, Albagli A, Myers SN, et al. A standard set of value-based patient-centered outcomes and measures of overall health in adults. Patient 2022;15(3):341–351.
10. Chiarotto A, Terwee CB, Kamper SJ, et al. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. J Clin Epidemiol 2018;102:23–37.
11. Salinas J, Sprinkhuizen SM, Ackerson T, et al. An international standard set of patient-centered outcome measures after stroke. Stroke 2016;47(1):180–186.
12. Khalil LS, Darrith B, Franovic S, et al. Patient-Reported Outcomes Measurement Information System (PROMIS) global health short forms demonstrate responsiveness in patients undergoing knee arthroplasty. J Arthroplasty 2020;35(6):1540–1544.
13. Nicholson AD, Kassam HF, Pan SD, et al. Performance of PROMIS Global-10 compared with legacy instruments for rotator cuff disease. Am J Sports Med 2019;47(1):181–188.
14. Suriani RJ, Kassam HF, Passarelli NR, et al. Validation of PROMIS Global-10 compared with legacy instruments in patients with shoulder instability. Shoulder Elbow 2020;12(4):243–252.
15. Lapin B, Davin S, Stilphen M, et al. Validation of PROMIS CATs and PROMIS global health in an interdisciplinary pain program for patients with chronic low back pain. Spine 2020;45(4):E227–E235.
16. Katzan IL, Lapin B. PROMIS-GH (Patient-Reported Outcomes Measurement Information System Global Health) scale in stroke: a validation study. Stroke 2018;49(1):147–154.
17. Anderson M, van Kessel R, Wood E, et al. Understanding factors impacting patient-reported outcome measures integration in routine clinical practice: an umbrella review. Qual Life Res 2024;33(10):2611–2629.
18. Crins MHP, Terwee CB, Klausch T, et al. The Dutch-Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol 2017;87:47–58.
19. Crins MHP, Terwee CB, Westhovens R, et al. First validation of the full PROMIS pain interference and pain behavior item banks in patients with rheumatoid arthritis. Arthritis Care Res (Hoboken) 2020;72(11):1550–1559.
20. 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;62:2569–2581.
21. Knoop J, van der Leeden M, van der Esch M, et al. Association of lower muscle strength with self-reported knee instability in osteoarthritis of the knee: results from the Amsterdam Osteoarthritis cohort. Arthritis Care Res (Hoboken) 2012;64(1):38–45.
22. Altman R, Alarcon G, Appelrouth D, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. Arthritis Rheum 1991;34:505–514.
23. Crins MHP, van der Wees PJ, Klausch T, et al. Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. PLoS One 2018;13(2):e0192187.
24. Pellicciari L, Chiarotto A, Giusti E, et al. Psychometric properties of the Patient-Reported Outcomes Measurement Information System Scale v1.2: global health (PROMIS-GH) in a Dutch general population. Health Qual Life Outcomes 2021;19(1):226.
25. HealthMeasures. HealthMeasures scoring service. Accessed March 10, 2021. www.assessmentcenter.net/ac_scoringservice
26. Elsman EBM, Roorda LD, Crins MHP, et al. Dutch reference values for the Patient-Reported Outcomes Measurement Information System Scale v1.2 - Global Health (PROMIS-GH). J Patient Rep Outcomes 2021;5(1):38.
27. Reeve BB, Hays RD, Bjorner JB, et al; PROMIS Cooperative Group. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5)(suppl 1):S22–S31.
28. Rodriguez A, Reise SP, Haviland MG. Applying bifactor statistical indices in the evaluation of psychological measures. J Pers Assess 2016;98(3):223–237.
29. Mokken RJ. A Theory and Procedure of Scale Analysis: With Applications in Political Research. Vol 1. Walter de Gruyter; 2011.
30. Kang T, Chen TT. Performance of the generalized S-X2 item fit index for polytomous IRT models. J Educ Meas 2008;45(4):391–406.
31. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60(1):34–42.
32. Holland PW, Wainer H. Differential Item Functioning. Routledge; 2012.
33. Zumbo B. A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores. Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.
34. Chalmers RP, Counsell A, Flora DB. It might not make a big DIF: improved differential test functioning statistics that account for sampling variability. Educ Psychol Meas 2016;76(1):114–140.
35. Holmes MM, Lewith G, Newell D, et al. The impact of patient-reported outcome measures in clinical practice for pain: a systematic review. Qual Life Res 2017;26(2):245–257.
36. Terwee CB, Zuidgeest M, Vonkeman HE, et al. Common patient-reported outcomes across ICHOM Standard Sets: the potential contribution of PROMIS®. BMC Med Inform Decis Mak 2021;21(1):259.
37. Goudman L, De Smedt A, Linderoth B, et al. Identifying goals in patients with chronic pain: a European survey. Eur J Pain 2021;25(9):1959–1970.
38. Slater H, Jordan JE, O'Sullivan PB, et al. "Listen to me, learn from me": a priority setting partnership for shaping interdisciplinary pain training to strengthen chronic pain care. Pain 2022;163(11):e1145–e1163.
39. Terwee CB, Roorda LD. Country-specific reference values for PROMIS® pain, physical function and participation measures compared to US reference values. Ann Med 2023;55(1):1–11.
40. Busija L, Ackerman IN, Haas R, et al. Adult measures of general health and health-related quality of life. Arthritis Care Res (Hoboken) 2020;72(Suppl 10):522–564.
41. Reise SP, Scheines R, Widaman KF, et al. Multidimensionality and structural coefficient bias in structural equation modeling: a bifactor perspective. Educ Psychol Meas 2013;73(1):5–26.

42. Sinharay S, Haberman SJ. How often is the misfit of item response theory models practically significant? Educ Meas 2014;33(1):23–35.
43. Crişan DR, Tendeiro JN, Meijer RR. Investigating the practical consequences of model misfit in unidimensional IRT models. Appl Psychol Meas 2017;41(6):439–455.
44. HealthMeasures. PROMIS Global Health scoring manual. Accessed April 10, 2021. https://www.healthmeasures.net/images/PROMIS/manuals/Scoring_Manuals_/PROMIS_Global_Health_Scoring_Manual.pdf
45. Svetina D, Rutkowski L. Multidimensional measurement invariance in an international context: fit measure performance with many groups. J Cross Cult Psychol 2017;48(7):991–1008.
46. DeMars C. Item Response Theory. Oxford University Press; 2010.
47. Kalinowski ST. A graphical method for displaying the model fit of item response theory trace lines. Educ Psychol Meas 2019;79(6):1064–1074.
48. Baker FB. The Basics of Item Response Theory. 2nd ed. ERIC; 2001.
49. Petrillo J, Cano SJ, McLeod LD, et al. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. Value Health 2015;18(1):25–34.
50. Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw 2011;39(8):1–30.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 662–669
DOI 10.1002/acr.25682

American College of Rheumatology
Empowering Rheumatology Professionals

Clinical Conditions Associated With a High Antinuclear Antibody Titer in Individuals Without Autoimmune Disease

Matthew Chung,[1,2] John P. Shelley,[1,2] Gul Karakoc,[2] John Still,[2] Xiaodi Ruan,[2] Jonathan Mosley,[2,3] C. Michael Stein,[2] and Vivian K. Kawai[2,4]

Objective. Antinuclear antibodies (ANAs) are present at high titers in 2% of the general population, but their clinical significance in individuals without an autoimmune (AI) disease is not known. We tested the hypothesis that the presence of a high ANA titer in non-AI conditions is associated with disease.

Methods. We conducted a retrospective case-control study in the Vanderbilt University Medical Center's de-identified electronic medical record system. Individuals without AI disease who had an ANA test were classified into three groups: high titer (HT; ANA ≥ 1:640), low titer (LT; ANA ≤ 1:80), and negative (NG) ANA results. The prevalence of diagnoses recorded within 90 days of the ANA test were compared among groups in a phenome-wide association study adjusting for age at ANA testing, sex, median body mass index (BMI), and reported race. A P value $<5 \times 10^{-5}$ was considered significant.

Results. A total of 28,781 individuals qualified for the study: 3.1% in the HT, 12.3% in the LT, and 84.6% in the NG groups. BMI was similar among groups (P value = 0.345), but individuals in the HT group were older ($P = 3.9 \times 10^{-73}$). A high ANA titer increased risk of 46 and 67 clinical diagnoses when comparing the HT group with the LT and the NG groups, respectively. The most significant associations in both comparisons included liver disorders and complications and risk factors for liver disease.

Conclusion. A high ANA titer in the absence of an AI disease was associated with increased risk of liver disorders and related risk factors and complications.

INTRODUCTION

Antinuclear antibodies (ANAs) represent a diverse group of autoantibodies that bind to specific molecular components within the nucleus of a cell. Their presence in blood serves as an important diagnostic tool in clinical care[1] because it provides important diagnostic information through the resulting patterns and titers.[2]

ANAs are biomarkers commonly used in the diagnosis of several autoimmune (AI) diseases,[1] with a positive ANA test being almost mandatory for the diagnosis of systemic lupus erythematosus (SLE)[3] because more than 95% of patients with SLE test positive and the classification criteria for SLE require an ANA titer of at least 1:80.[3] A positive ANA test is also associated with an increased risk of AI conditions other than SLE, such as Sjögren disease, systemic sclerosis, mixed connective tissue disease, AI liver disease, and idiopathic inflammatory myopathies among others.[1]

Although ANAs are well known for their association with various AI diseases, their occurrence in the general population suggests broader biologic implications.[4] A small study reported that healthy individuals with ANA$^+$ (≥1:120) showed suppression of T cell signatures and lower cytokine levels compared to their counterpart ANA$^-$ individuals[4,5]; but it is unclear whether this altered immunologic landscape might contribute to the development or prevention of various clinical conditions.

Some clinical studies have indicated that individuals with a positive ANA test, particularly at higher titers,[6] have an increased

Supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH (grant R01-AR-076516); the Arthritis National Research Foundation, All Arthritis Grant Program Award; and the Lupus Research Alliance–Bristol Myers Squibb Accelerator Award. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU, which is supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH-funded Shared Instrumentation Grants S10-RR-025141, S10-OD-017985, and S10-OD-025092 and Clinical and Translational Science Award grants UL1-TR-002243, UL1-TR-000445, and UL1-RR-024975. The funding sources had no role in the study design, execution, analyses, data interpretation, or decision to submit the results.

[1]Vanderbilt University, Nashville, Tennessee; [2]Vanderbilt University Medical Center, Nashville, Tennessee; [3]University of Texas Southwestern Medical Center, Dallas; [4]University of Alabama at Birmingham.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25682).

Author disclosures and graphical abstract are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25682.

Address correspondence via email to Vivian K. Kawai, MD, MPH, at vkawai@uabmc.edu.

Submitted for publication May 13, 2025; accepted in revised form October 7, 2025.

SIGNIFICANCE & INNOVATIONS

- This is the first large study that systematically evaluates the impact of a high antinuclear antibody (ANA) titer (≥1:640) in individuals without an autoimmune disease using a high-throughput approach.
- A high ANA titer is strongly associated with liver-related disorders including nonalcoholic and alcoholic liver disease in individuals without autoimmune disease.
- Nonautoimmune liver disease should be included in the differential diagnosis in patients who have an ANA titer ≥1:640.

risk of cardiovascular events,[6] cancer,[7] infections,[8] and all-cause mortality[6,9]; but these findings have not been consistent.[10,11] Recently, we showed that a positive ANA test at a titer ≥1:80 in individuals without an AI disease was significantly associated with increased risk for Raynaud phenomenon and idiopathic fibrosing alveolitis[12]; however, the clinical implications of a high ANA titer in people without AI disease is unclear.

We hypothesized that individuals without AI disease who have a high ANA titer are more likely to have immune dysregulation that would manifest as differential disease risk. To test this hypothesis, we used a phenome-wide association study (PheWAS) approach to identify clinical diagnoses that are enriched in patients without AI disease who have a high ANA titer compared to individuals with a low titer and negative ANA results.

PATIENTS AND METHODS

Study population. The study was approved by the Vanderbilt University Medical Center (VUMC) Institutional Review Board. We selected individuals from the de-identified copy of the electronic medical record (EMR). Participants were 18 years or older and had at least one clinician-ordered ANA test. Because titers were reported in ≥95% of participants with a positive ANA test since 2000, only ANA results recorded between January 2000 and October 2022 were included. Only ANA tests performed using indirect immunofluorescence of human epithelial type 2 (HEp-2) cells were selected. The hospital clinical laboratory used protocols recommended by the American College of Rheumatology Task Force[13] using anti-human IgG conjugated in HEp-2 cells from Immuno Concepts and Inova Diagnostics laboratory. Testing was performed following manufacturer's recommendations, and while manual interpretation was performed in both assays, the Inova Diagnostics assay also provided automated interpretation.[14]

For individuals with more than one eligible ANA result, we selected the test with the highest titer. We excluded any individual with a diagnosis of an AI disorder commonly considered to be associated with a positive ANA test (Supplementary Table S1).[1]

To define the clinical impact of a high ANA titer, eligible participants were divided into three mutually exclusive groups: (a) high titer (HT) included individuals with an ANA titer ≥1:640,[15] (b) low titer (LT) included individuals with ANA titer ≤1:80, and (c) negative (NG) included individuals who only had negative ANA results in their EMR. Individuals with a titer between 1:80 and 1:640 were excluded from the analyses.

Covariates. Demographic and clinical data including age at test, sex, reported race, and body mass index (BMI; kg/m^2) were extracted from the EMR and used as covariates. The median BMI was calculated using plausible BMIs recorded in the EMR, which we defined as being between 15 and 60 as previously used.[16] Only participants with data available for all covariates were included in the analyses.

Statistical analysis. To define the clinical impact of a high ANA titer, the prevalence of different clinical diagnoses was compared between HT and LT groups and between the HT and NG groups using a PheWAS approach,[17] which is a method for high-throughput analyses using EMR data. To define which clinical diagnoses were more likely associated with a high ANA titer, only diagnostic codes first recorded within 90 days of the selected ANA results were included. International Classification of Diseases (ICD) codes (versions 9 and 10) were extracted from the EMR and transformed into Phecodes, which aggregate one or more related ICD codes into distinct diseases or traits.[18] The ICD-to-Phecode map (version 1.2) was used to create Phecodes. For each clinical diagnosis, cases were defined as individuals with two or more instances of the specific Phecode in the EMR. Controls were defined as individuals without the Phecode or any closely related Phecode. Analyses were adjusted for age at ANA testing, sex, median BMI, and race. Only Phecodes with 100 cases or more were analyzed to assure power. PheWAS was performed using the PheWAS R package,[17] and associations with a P value $<5 \times 10^{-5}$ were considered significant. Continuous and categorical variables are shown as median (interquartile range [IQR]) and frequency (percentage), respectively, and compared using a Kruskal–Wallis rank test and a chi-squared test.

Case validation. Because the quality of phenotype definition varies when ICD codes are aggregated, the medical record of a random sample of 50 individuals for the 10 most significant associations was manually reviewed to confirm the diagnosis and estimate the positive predictive value (PPV) for the selected Phecodes. If related Phecodes were among top associations, the Phecode with the highest number of cases was selected for chart review.

Data availability. The data sets generated and/or analyzed during the current study are available from the

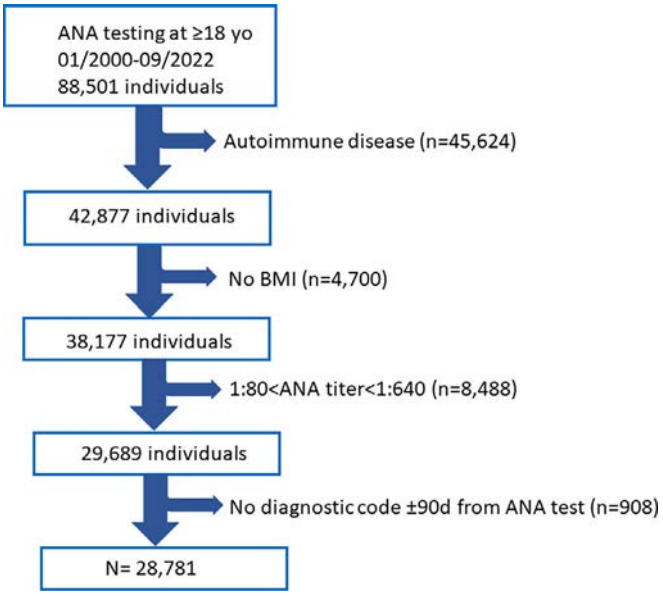

Figure 1. Flowchart of the study showing selected individuals with an ANA test result at 18 years or older in the Vanderbilt University Medical Center EMR system. ANA, antinuclear antibody; BMI, body mass index; EMR, electronic medical record; yo, years old. Color figure can be viewed in the online issue, which is available at http://onlinelibrary.wiley.com/doi/10.1002/acr.25682/abstract.

corresponding author on reasonable request. Phecode maps and R code for the PheWAS are openly available at https://phewascatalog.org/phewas.

Ethics approval and consent to participate. The study was approved by the VUMC Institutional Review Board; the study met Exemption 4 criteria and is not considered human participant research under the 2018 Revised Common Rule requirement of the Office for Human Research Protections. Thus, consent form was not required. Data used for the study come from the de-identified EMR system at VUMC. No experiments in humans were performed as part of the study, de-identified data that were already collected were used for in silico analysis, and all methods were performed in accordance with relevant guidelines and regulations. The authors did not have direct contact or knowledge of the individuals' identities.

RESULTS

Between January 2000 and September 2022, there were 88,501 participants who were 18 years or older and had at least one clinician-ordered ANA test available. We excluded individuals with a diagnosis of an ANA-related AI disease in their EMR (n = 45,624), those without BMI information (n = 4,700), and those with an ANA titer higher than 1:80 but lower than 1:640 (n = 8,488). There were 29,689 individuals with an eligible ANA result, and 28,781 of them had at least one diagnosis code first recorded within ±90 days of the eligible ANA result (Figure 1 shows the flowchart of the study design).

Table 1 shows the demographic and clinical characteristics of the 28,781 eligible individuals: 3.1% (n = 883) were in the HT, 12.3% (n = 3,544) in the LT, and 84.6% (n = 24,354) in the NG groups, respectively. Overall, the study population was largely of European ancestry, and there were more women. Individuals in the HT groups were older (median age 55 [IQR 42, 66] years) than those in the LT (median 48 [IQR 37, 60] years) and NG (median 45 [IQR 33, 56] years) groups, and the proportion of White individuals was smaller. BMI did not differ significantly among the three groups (median 28.7 [IQR 24.5, 33.3] kg/m^2, median 28.1 [IQR 24.2, 33.3] kg/m^2, and median 28.2 [IQR 24.1, 33.4] kg/m^2 for the HT, LT, and NG groups, respectively). An ANA pattern was available in 23.4% and 95.8% of the individuals in the LT and HT groups, respectively, and the distribution of patterns was very similar in both groups (Table 1).

Table 1. Characteristics of study population categorized by their ANA result*

Characteristics	Negative (n = 24,354)	Low titer (n = 3,544)	High titer (n = 883)
Age at testing, y[a]	45.0 (33.0, 56.0)	48.0 (37.0, 60.0)	55.0 (42.0, 66.0)
Female, n (%)	15,912 (65.3)	2,577 (72.7)	613 (69.4)
Body mass index,[a] kg/m^2	28.2 (24.1, 33.4)	28.1 (24.2, 33.3)	28.7 (24.5, 33.3)
Reported race, n (%)			
White	17,779 (73.0)	2,643 (74.6)	596 (67.5)
Black	2,633 (10.8)	326 (9.2)	85 (9.6)
Asian	442 (1.8)	52 (1.4)	11 (1.3)
Hispanic	617 (2.5)	77 (2.2)	15 (1.7)
Other/unknown	2,883 (11.8)	446 (12.6)	176 (19.9)
ANA pattern,[b] n (%)		(n = 830)	(n = 846)
Homogeneous	NA	476 (57.3)	478 (56.5)
Speckled	NA	309 (37.2)	329 (38.9)
Nucleolar	NA	42 (5.1)	28 (3.3)
Others	NA	3 (0.4)	11 (1.3)

* ANA, antinuclear antibody; NA, not applicable.
[a] Age at testing and body mass index are shown as median (interquartile range).
[b] ANA patterns were present in 23% and 95.8% of patients in the low (ANA titer ≤ 1:80) and high titer (ANA titer ≥ 1:640) groups, respectively.

Clinical associations comparing HT and LT groups. Individuals in the HT group had increased risk of 46 clinical Phecode diagnoses (Figure 2). Top associations were chronic liver diseases (Phecode 571), nonalcoholic fatty liver disease or steatohepatitis (NAFLD/NASH; Phecode 571.5), and alcohol-related liver disease (Phecode 317). Other significant associations were related to other liver disorders and their complications, disorders known to increase liver disease (alcoholism, metabolic and biliary disorders), and respiratory, gastrointestinal, and mood disorders (Supplementary Table S2). A Phecode (338) for pain was more common in the HT group compared to the LT group (odds ratio [OR] = 4.1, P value = 1.9×10^{-13}), but myalgia (Phecode 770) was less common (OR = 0.27, P value = 1.3×10^{-5}).

Clinical associations comparing HT and NG groups. When compared to individuals with a negative ANA test, 67 clinical diagnoses were more prevalent in the HT group (Figure 3). Fifty-nine of these corresponded to diagnoses or closely related diagnoses—morbid obesity, bariatric surgery, substance disorders for alcoholism, and insulin pump usage—that were also observed in the comparison between the HT and LT groups. Top associations included the same liver disorders observed when comparing HT with LT groups. New associations included increased risk for gastrointestinal disorders, biliary tract disorders, poisoning by different agents (such as antibiotics, other anti-infectives, and analgesics/antipyretics/antirheumatics), and screening for skin cancers (Supplementary Table S3).

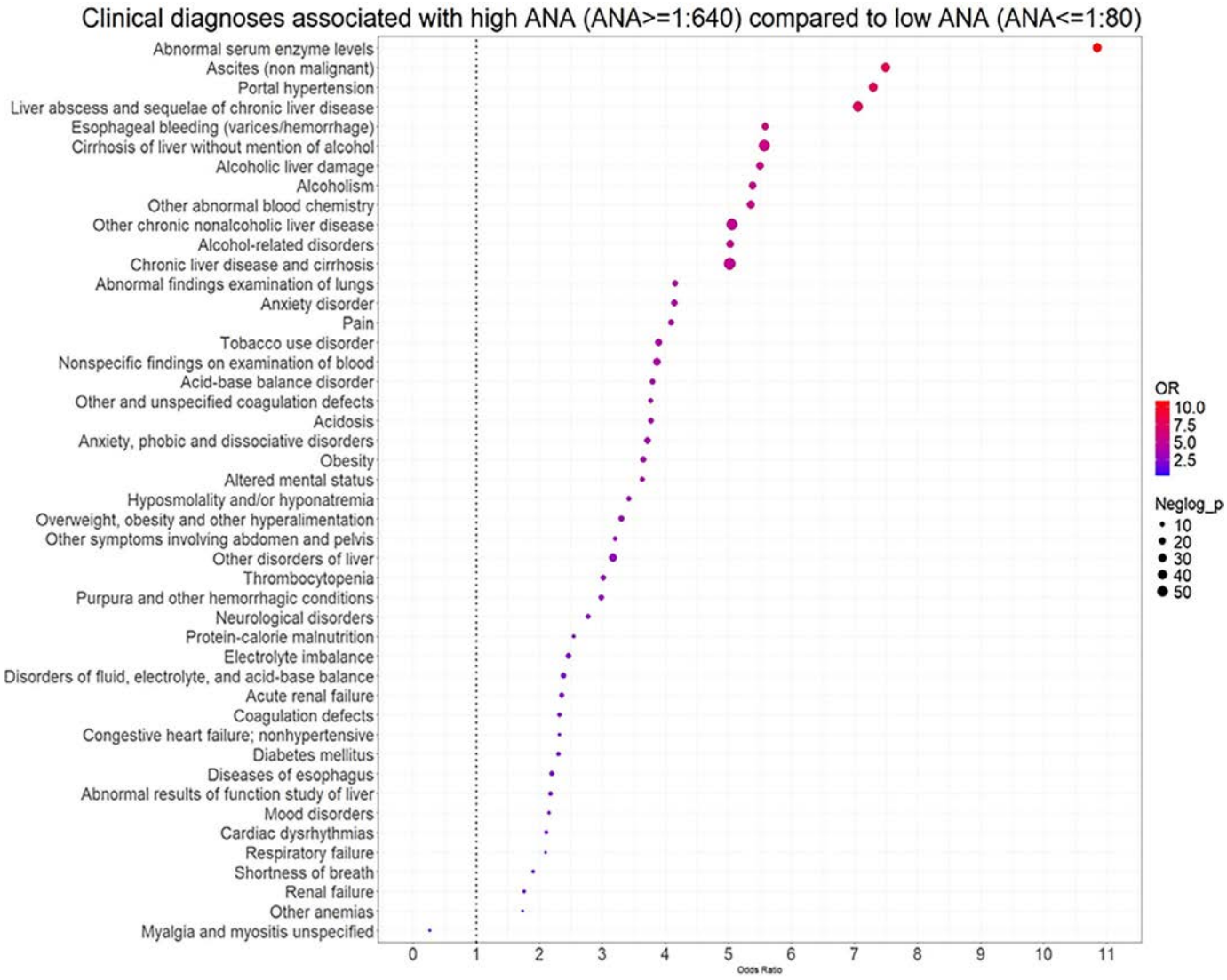

Figure 2. Significant clinical associations when individuals with high titers (ANA ≥ 1:640) are compared with individuals with low ANA titers (ANA ≤ 1:80). The magnitude (OR) and strength (P value) of the association are represented by the circle size and color, respectively. The dotted vertical line represents no association, and circles on the left and right from the dotted line represent decreased and increased risk, respectively. ANA, antinuclear antibody; OR, odds ratio.

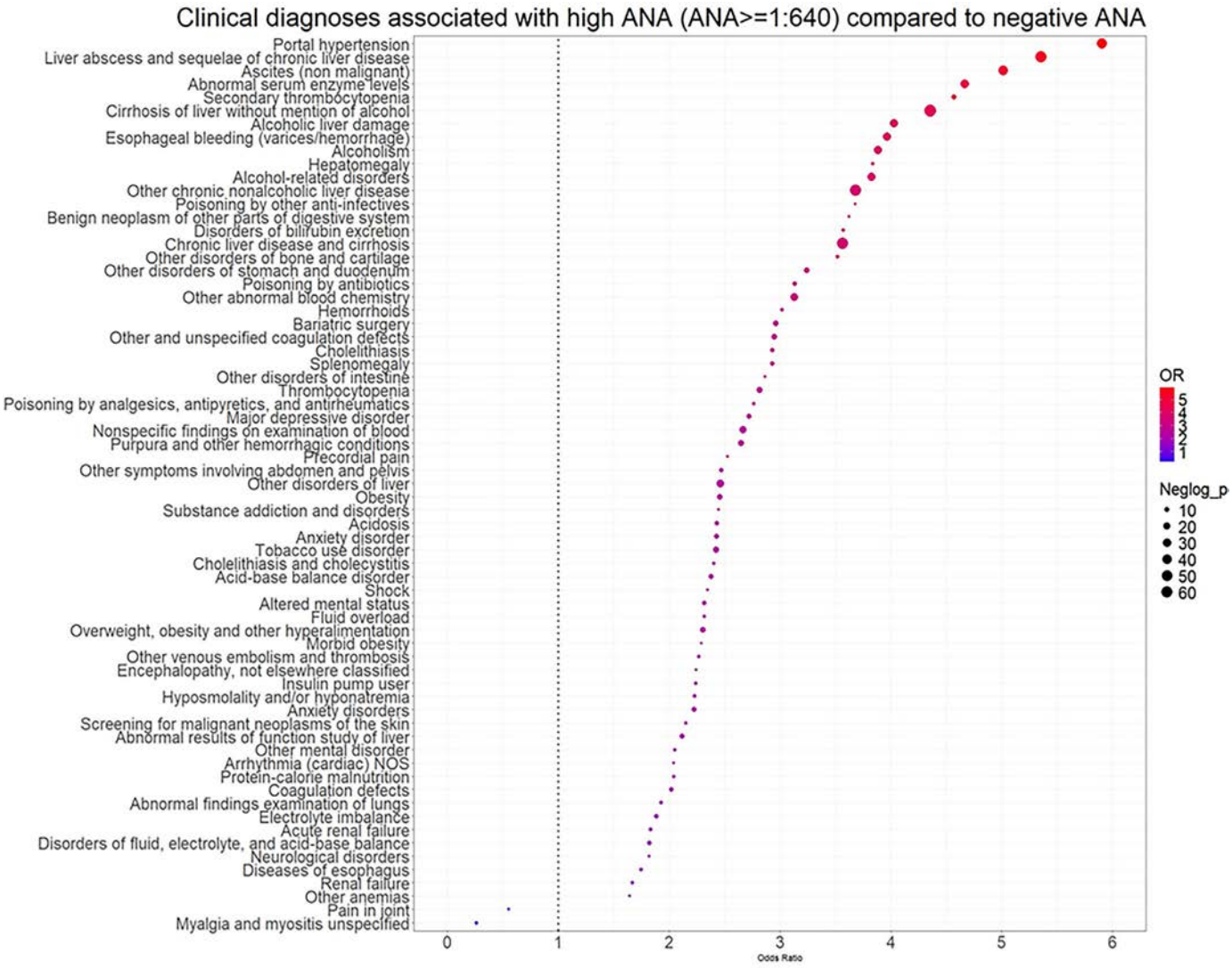

Figure 3. Significant clinical associations when individuals with high ANA titers (ANA ≥ 1:640) are compared with individuals with a negative ANA test. The magnitude (OR) and strength (*P* value) of the association are represented by the circle size and color, respectively. The dotted vertical line represents no association, with circles on the left and right from the dotted line representing decreased and increased risk, respectively. ANA, antinuclear antibody; NOS, not otherwise specified; OR, odds ratio.

A sensitivity analysis excluding BMI as a covariate showed similar results when comparing the HT with the LT group (Supplementary Table S4) and the NG group (Supplementary Table S5). Among individuals in the HT group, ANA patterns were not different between those with and without liver disease (Supplementary Table S6).

Case validation. Because the most significant clinical diagnoses associated with high ANA titers were liver- and alcohol-related disorders, we selected 50 random cases with a high ANA titer and Phecode 573 ("Other disorders of the liver") to perform chart review. We confirmed the presence of non-AI liver disease in 44 individuals (PPV of 88%) with 3 individuals having possible primary biliary cholangitis, primary biliary cirrhosis, or AI hepatitis. Fatty liver-related diseases were the most common source for liver disease followed by alcohol consumption and hepatitis C infection. Another top association was related to alcohol disorder (alcoholism—Phecode 317.1 and alcohol liver damage—Phecode 317.11), with 46 of the 50 random charts reviewed for these categories indicating diagnoses attributed to alcohol consumption (PPV = 92%).

DISCUSSION

A striking finding was that a high ANA titer is strongly associated with liver-related disorders such as NAFLD and NASH as well as alcoholic liver disorders. High titers of ANAs are

uncommon in people without AI disease[10,19]; however, few studies have examined the clinical implications of these high ANA titers.[15,20] Although ANA titers (>1 SD above the mean in healthy controls) are associated with upregulation of type I tumor necrosis factor signatures,[20] little is known about their associations with disease.

In a laboratory-based study, 5% of individuals tested had ANA titers ≥1:640, and 44% of them did not have a diagnosis of an AI disorder.[15] Among these individuals with high ANA titer without AI disease, 43% and 24% of them returned for a 5-year and 10-year follow-up, respectively.[21,22] Although ANA results remained positive in 91% of those who returned at a 5-year follow-up, only 3 of 62 participants had a new diagnosis of an AI disease.[21] As for those who returned for a 10-year follow-up, 75% continued to have positive ANA results, and 2 of 34 participants were diagnosed with an AI disease at 10-year follow-up.[22] These findings suggest that the majority of high titer positive ANA results are not explained by the future development of an AI disease.

Using the results of ANA testing performed as part of routine clinical care, we defined the associations between a high titer positive ANA test and disease in a large population that did not have AI disease. The most striking finding was the increased prevalence of liver disease and diagnoses related to liver disease, particularly NAFLD and NASH. An association between a positive ANA test and AI disease is well-recognized,[1] and we had therefore excluded people with liver diseases such as AI hepatitis and primary biliary cholangitis from the study cohort; review of a sample of medical records showed that undiagnosed AI liver disease did not account for our findings.

Also, a high ANA titer was associated with increased risk of alcoholic liver disease (ALD) and alcohol consumption, suggesting that liver injury of differing etiology can predispose to a high titer ANA. The finding that high ANA titers were associated with increased risk of alcohol- and tobacco-related disorders is in agreement with previous clinical and animal studies that suggested increased formation of autoantibodies,[23] including ANAs,[24] under conditions of increased oxidative stress,[25] such as tobacco use[26] and chronic alcohol use.[27]

Although non-AI liver disease is not currently recognized as a frequent cause of a high titer ANA, a positive ANA test occurs in approximately a third of individuals with NAFLD[28] in the absence of AI hepatitis.[29] The relationship between the presence of a positive ANA test and histologic features in the liver of patients with NAFLD is uncertain.[30–32] A small clinical study suggested that a positive ANA test, even at high titer (≥1:640), was not associated with disease activity and severity of histologic changes,[30] whereas other studies reported an association between positive ANA results and higher levels of hepatic fibrosis and inflammation.[31,32] We found that, in addition to associations with liver disease, a high titer ANA was associated with increased risk of several diagnoses characteristic of more severe liver disease, including cirrhosis, liver failure, portal hypertension, esophageal bleeding, coagulopathy, ascites, and protein–calorie malnutrition among others. Whether these associations reflect accelerated progression of NAFLD, as suggested by previous studies,[32,33] cannot be determined in our study.

In addition to liver disease, a high ANA titer was significantly associated with increased risk of several metabolic disorders (overweight, obesity, and diabetes mellitus) or their surrogates (bariatric surgery, insulin pump user). Although the median BMI among the three ANA groups was similar, the percentage of individuals in the HT group that fell in the overweight, obese, and morbidly obese categories was almost twice than that of those assigned to the same categories in the LT and NG groups. A relationship between high ANA titers and metabolic disorders has not been established previously. Cross-sectional epidemiologic studies suggest an inverse association between BMI and a positive ANA test,[19,34] which is counterintuitive because obesity is a state of chronic systemic inflammation that could lead to immune dysregulation and obesity is a risk factor for several AI disorders.[35,36] Although some evidence suggests a positive association between positive ANA results and BMI only at high levels of inflammation,[35] we cannot exclude the possibility that the association with obesity is driven by NAFLD or NASH.

We found that a high ANA titer was inversely associated with myalgia and pain in joints. This is not surprising considering that individuals with musculoskeletal pain are more likely to be screened for AI disorders with an ANA test, and those with an AI disease are more likely to have a high ANA titer.[36–38]

Our study has some limitations: first, some liver disorders associated with a high ANA titer may have been incorrectly assigned to NAFLD or NASH or ALD; however, review of a random selection of medical records showed that these were the most common conditions mentioned in the physician clinical notes. Second, it is possible that some individuals with a high ANA titer will develop an AI disease in the future; to decrease this possibility, we excluded any individuals who had a diagnosis of an AI disease in their EMR at any time, not only before the ANA test. Third, although we tried to capture only incident diagnoses, we cannot exclude the possibility that new patients with chronic conditions diagnosed elsewhere had an ANA test done at the first medical encounter in VUMC, and thus, some chronic conditions were captured. Fourth, because we studied a large cohort of patients from a tertiary care hospital, our findings may not be generalizable to other settings. Fifth, although the International Consensus on Antinuclear Antibody Patterns (ICAP) has released several recommendations for ANA testing,[39] most of the ANA results used in the analyses were performed before any of these recommendations were released. Sixth, we did not study antibodies associated with DFS70, which is present as a dense fine speckled pattern. These antibodies are difficult to identify using HEp-2 indirect fluorescent antibody[40] and require solid phase assays to be detected accurately. Whether the presence of these

antibodies may explain some of the observed associations requires further investigation. Nevertheless, our study includes clinical and laboratory results from real-world patient care encounters from a large university teaching hospital, which is a strength of the study. Other strengths included the use of the EMR system to define a large cohort of individuals tested and limiting the associated diagnoses to those occurring within ±90 days of the ANA, which reduces spurious associations due to prevalent disease. We also adjusted our analyses by sex, age, race, and BMI because these are factors that influence ANA results.

Our study is the first to address the clinical relevance of a high ANA titer in individuals without an AI disease using real-world data and high-throughput analyses. Although previous small clinical studies suggested an association between presence of ANA and non-AI liver disease, our results show that a high ANA titer is associated not only with liver disease but also with complications associated with advanced liver disease. Thus, this suggests that non-AI liver disease should be considered in differential diagnosis of a high ANA titer. In summary, we found that a high ANA titer (≥1:640) in individuals without an AI disease is more common in patients with liver disease and related sequelae.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Kawai confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Solomon DH, Kavanaugh AJ, Schur PH; American College of Rheumatology Ad Hoc Committee on Immunologic Testing Guidelines. Evidence-based guidelines for the use of immunologic tests: antinuclear antibody testing. Arthritis Rheum 2002;47(4):434–444.
2. Mariz HA, Sato EI, Barbosa SH, et al. Pattern on the antinuclear antibody-HEp-2 test is a critical parameter for discriminating antinuclear antibody-positive healthy individuals and patients with autoimmune rheumatic diseases. Arthritis Rheum 2011;63(1):191–200.
3. Aringer M, Costenbader K, Daikh D, et al. 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. Arthritis Rheumatol 2019;71(9):1400–1412.
4. Slight-Webb S, Lu R, Ritterhouse LL, et al. Autoantibody-positive healthy individuals display unique immune profiles that may regulate autoimmunity. Arthritis Rheumatol 2016;68(10):2492–2502.
5. Slight-Webb S, Smith M, Bylinska A, et al. Autoantibody-positive healthy individuals with lower lupus risk display a unique immune endotype. J Allergy Clin Immunol 2020;146(6):1419–1433.
6. Liang KP, Maradit Kremers H, Crowson CS, et al. Autoantibodies and the risk of cardiovascular events. J Rheumatol 2009;36(11):2462–2469.
7. Vlagea A, Falagan S, Gutiérrez-Gutiérrez G, et al. Antinuclear antibodies and cancer: a literature review. Crit Rev Oncol Hematol 2018;127:42–49.
8. Cainelli F, Betterle C, Vento S. Antinuclear antibodies are common in an infectious environment but do not predict systemic lupus erythematosus. Ann Rheum Dis 2004;63(12):1707–1708.
9. Dinse GE, Parks CG, Weinberg CR, et al. Antinuclear antibodies and mortality in the National Health and Nutrition Examination Survey (1999-2004). PLoS One 2017;12(10):e0185977.
10. Akmatov MK, Röber N, Ahrens W, et al. Anti-nuclear autoantibodies in the general German population: prevalence and lack of association with selected cardiovascular and metabolic disorders-findings of a multicenter population-based study. Arthritis Res Ther 2017;19(1):127.
11. Selmi C, Ceribelli A, Generali E, et al. Serum antinuclear and extractable nuclear antigen antibody prevalence and associated morbidity and mortality in the general population over 15 years. Autoimmun Rev 2016;15(2):162–166.
12. Zanussi JT, Zhao J, Wei WQ, et al. Clinical diagnoses associated with a positive antinuclear antibody test in patients with and without autoimmune disease. BMC Rheumatol 2023;7(1):24.
13. Meroni PL, Schur PH. ANA screening: an old test with new recommendations. Ann Rheum Dis 2010;69(8):1420–1422.
14. Buchner C, Bryant C, Eslami A, et al. Anti-nuclear antibody screening using HEp-2 cells. J Vis Exp 2014;(88):e51211.
15. Vaile JH, Dyke L, Kherani R, et al. Is high titre ANA specific for connective tissue disease? Clin Exp Rheumatol 2000;18(4):433–438.
16. Cheng FW, Gao X, Mitchell DC, et al. Body mass index and all-cause mortality among older adults. Obesity (Silver Spring) 2016;24(10):2232–2239.
17. Carroll RJ, Bastarache L, Denny JCR. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 2014;30(16):2375–2376.
18. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med Inform 2019;7(4):e14325.
19. Satoh M, Chan EK, Ho LA, et al. Prevalence and sociodemographic correlates of antinuclear antibodies in the United States. Arthritis Rheum 2012;64(7):2319–2327.
20. Li QZ, Karp DR, Quan J, et al. Risk factors for ANA positivity in healthy persons. Arthritis Res Ther 2011;13(2):R38.
21. Myckatyn SO, Russell AS. Outcome of positive antinuclear antibodies in individuals without connective tissue disease. J Rheumatol 2003;30(4):736–739.
22. Wijeyesinghe U, Russell AS. Outcome of high titer antinuclear antibody positivity in individuals without connective tissue disease: a 10-year follow-up. Clin Rheumatol 2008;27(11):1399–1402.
23. Lian M, Hua J, Sheng L, et al. Prevalence and significance of autoantibodies in patients with alcoholic liver disease. J Dig Dis 2013;14(7):396–401.
24. Laskin CA, Vidins E, Blendis LM, et al. Autoantibodies in alcoholic liver disease. Am J Med 1990;89(2):129–133.
25. Krzemień P, Kasperczyk S, Banach M, et al. Serum antinuclear autoantibodies are associated with measures of oxidative stress and lifestyle factors: analysis of LIPIDOGRAM2015 and LIPIDOGEN2015 studies. Arch Med Sci 2021;19(5):1214–1227.
26. Morissette MC, Gao Y, Shen P, et al. Role of BAFF in pulmonary autoantibody responses induced by chronic cigarette smoke exposure in mice. Physiol Rep 2016;4(24):e13057.
27. Tuma DJ. Role of malondialdehyde-acetaldehyde adducts in liver injury. Free Radic Biol Med 2002;32(4):303–308.
28. Yatsuji S, Hashimoto E, Kaneda H, et al. Diagnosing autoimmune hepatitis in nonalcoholic fatty liver disease: is the International

Autoimmune Hepatitis Group scoring system useful? J Gastroenterol 2005;40(12):1130–1138.

29. Kohut T, Shah A, Russo P, et al. Autoimmune antibodies in children and adolescents with nonalcoholic fatty liver disease. J Pediatr Gastroenterol Nutr 2022;75(3):264–268.
30. Vuppalanchi R, Gould RJ, Wilson LA, et al; Nonalcoholic Steatohepatitis Clinical Research Network (NASH CRN). Clinical significance of serum autoantibodies in patients with NAFLD: results from the nonalcoholic steatohepatitis clinical research network. Hepatol Int 2012; 6(1):379–385.
31. Adams LA, Lindor KD, Angulo P. The prevalence of autoantibodies and autoimmune hepatitis in patients with nonalcoholic fatty liver disease. Am J Gastroenterol 2004;99(7):1316–1320.
32. Niwa H, Sasaki M, Haratake J, et al. Clinicopathological significance of antinuclear antibodies in non-alcoholic steatohepatitis. Hepatol Res 2007;37(11):923–931.
33. Zhou YJ, Zheng KI, Ma HL, et al. Association between positivity of serum autoantibodies and liver disease severity in patients with biopsy-proven NAFLD. Nutr Metab Cardiovasc Dis 2021;31(2):552–560.4
34. Dinse GE, Parks CG, Weinberg CR, et al. Increasing prevalence of antinuclear antibodies in the United States. Arthritis Rheumatol 2022;74(12):2032–2041.
35. Blanco I, Labitigan M, Abramowitz M. The inverse association between obesity and anti-nuclear antibodies is modified by systemic inflammation and maybe associated with body composition [abstract]. Arthritis Rheum 2012;64(S10): S411–S412.
36. Menor Almagro R, Rodríguez Gutiérrez JF, Martín-Martínez MA, et al. Association between antinuclear antibody titers and connective tissue diseases in a Rheumatology Department. Reumatol Clin 2017;13(3): 150–155.
37. Li H, Zheng Y, Chen L, et al. High titers of antinuclear antibody and the presence of multiple autoantibodies are highly suggestive of systemic lupus erythematosus. Sci Rep 2022;12(1):1687.
38. Wang KY, Yang YH, Chuang YH, et al. The initial manifestations and final diagnosis of patients with high and low titers of antinuclear antibodies after 6 months of follow-up. J Microbiol Immunol Infect 2011; 44(3):222–228.
39. Damoiseaux J, Andrade LEC, Carballo OG, et al. Clinical relevance of HEp-2 indirect immunofluorescent patterns: the International Consensus on ANA patterns (ICAP) perspective. Ann Rheum Dis 2019;78(7): 879–889.
40. Bossuyt X. DFS70 autoantibodies: clinical utility in antinuclear antibody testing. Clin Chem 2024;70(2):374–381.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 670–682
DOI 10.1002/acr.25662

American College of Rheumatology
Empowering Rheumatology Professionals

REVIEW ARTICLE

Accuracy of Diagnostic Codes and Algorithms Used to Identify Rheumatoid Arthritis and Juvenile Idiopathic Arthritis in Administrative Claims and Electronic Health Records: Systematic Review and Meta-Analysis

Constanza Saka-Herrán,[1] Jessica Bennett,[1] Yara Alkabti,[2] Muhammad Fatir,[2] Barbara Clyne,[3] Caroline McCarthy,[4] Gráinne Tynan,[1] Nikki Dunne,[1] Michelle Flood,[1] Eoghan McCarthy,[5] and Frank Moriarty[1]

Objective. This systematic review aimed to assess the diagnostic accuracy of algorithms used to identify rheumatoid arthritis and juvenile idiopathic arthritis in electronic health records.

Methods. We searched Medline, Embase, and Cochrane Central Register for Controlled Trials databases and included studies that validated case definitions against a reference standard, such as rheumatologist-confirmed diagnosis or American College of Rheumatology/EULAR classification criteria. Title and abstract screening, full-text review, data extraction, and quality assessment were all completed in duplicate. Results were synthesized narratively and using a bivariate random-effects meta-analysis of sensitivity and specificity.

Results. A total of 35 studies were included. Algorithms varied widely in complexity, ranging from single *International Classification of Diseases* (ICD) codes to combinations including disease-modifying antirheumatic drugs (DMARDs), hospitalization records, and specialist diagnosis. Algorithms combining ICD codes with DMARD prescriptions (pooled sensitivity 0.79 [95% confidence interval (CI) 0.61–0.90], specificity 0.96 [95% CI 0.72–1.00], positive predictive value [PPV] 0.78 [95% CI 0.63–0.88]) or requiring an ICD code assigned by a rheumatologist (pooled sensitivity 0.91 [95% CI 0.70–0.98], specificity 0.94 [95% CI 0.49–1.00], PPV 0.70 [95% CI 0.64–0.75]) showed the highest accuracy, with balanced sensitivity, specificity, and PPV. Less restrictive algorithms demonstrated high sensitivity but lower PPV. Substantial heterogeneity was observed across studies, likely due to differences in algorithm structure, data sources, and validation methods. Despite this variability, we used conceptually coherent categories to allow for meaningful synthesis, prioritizing clinical interpretability.

Conclusions. These findings support the use of more specific algorithms when diagnostic certainty is essential and highlight the need for further validation of high-performing algorithms across diverse health care systems.

INTRODUCTION

Rheumatoid arthritis (RA) and juvenile idiopathic arthritis (JIA) are autoimmune diseases characterized by joint inflammation and systemic complications leading to disability and reduced quality of life.[1] Despite differences in onset and classification criteria, they share key pathogenic mechanisms, clinical manifestations, and therapeutic strategies.[2,3] A challenge in RA and JIA research is obtaining robust epidemiologic data to understand the natural history of disease and the impact of therapy on disease outcomes.[4] Electronic health records (EHRs), have greatly facilitated studying these conditions in large populations, providing comprehensive data on patient demographics, symptoms, diagnoses, and treatment history[5,6] in a real-world setting.

Supported by the Health Research Board Applied Programme Grant Awards scheme (grant APRO-2023-028).

[1]School of Pharmacy and Biomolecular Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland; [2]School of Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland; [3]Department of Public Health and Epidemiology, School of Population Health, Royal College of Surgeons in Ireland, Dublin, Ireland; [4]Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland; [5]Department of Rheumatology, Beaumont Hospital/Royal College of Surgeons in Ireland, Dublin, Ireland.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25662).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25662.

Address correspondence via email to Frank Moriarty, PhD, at frankmoriarty@rcsi.com.

Submitted for publication June 19, 2025; accepted in revised form September 22, 2025.

SIGNIFICANCE & INNOVATIONS

- This review provides an updated and expanded synthesis of validated algorithms to identify rheumatoid arthritis (RA) and juvenile idiopathic arthritis (JIA) in electronic health records, addressing a growing need as real-world data become increasingly central in rheumatology research. Compared to earlier reviews, it includes more recent studies, a broader range of geographic and health care settings, both adult and pediatric populations, and structured categorization of algorithm components. It is also the first to perform meta-analysis stratified by algorithm components and to assess the risk of bias using the QUADAS-2 tool. These advances offer practical guidance for researchers for selecting or developing algorithms tailored to their setting and data source.
- This review offers evidence-based guidance for selecting RA and JIA identification algorithms based on their accuracy across diverse settings. By categorizing algorithms and summarizing accuracy measures, the findings support researchers and clinicians in choosing or adapting definitions depending on research purposes and available data sources. These findings may also inform future algorithm development and validation studies.
- The review highlights gaps in validation efforts and emphasizes the need to validate high-performing algorithms across diverse health care settings and evolving coding systems, ensuring accurate disease identification in current and future research.

However, the reliability and validity of research based on EHR data depend on the accuracy of the diagnostic codes and algorithms used to identify the condition.[7,8] Combining multiple diagnostic codes with clinical data, such as prescriptions and laboratory results, enhances RA identification accuracy. For instance, algorithms incorporating at least three physician diagnostic codes achieve specificity and positive predictive values (PPVs) of over[9] 90%. The inclusion of disease-modifying antirheumatic drug (DMARD) prescriptions with diagnostic codes, significantly improves case identification.[10] Standardized algorithms that integrate *International Classification of Diseases* (ICD) codes with other clinical parameters have also been assessed for their utility in diverse health care systems.[11]

Despite advances in identifying RA and JIA cases, limitations persist due to coding inaccuracy, population heterogeneity, and differences across health care systems and methodologic approaches. A 2013 systematic review, part of the Mini-Sentinel pilot project sponsored by the US Food and Drug Administration, reported substantial variability in the accuracy of algorithms (PPV 34%–97%) applied to EHRs.[12] The data sources of interest were limited to the United States or Canada, reflecting Mini-Sentinel's focus on US-based data for surveillance system,[13] thus limiting the global applicability of the findings.

The aim of this systematic review is to synthesize and critically assess evidence on the accuracy of diagnostic codes and algorithms used to identify RA and JIA in EHRs and other administrative databases. By expanding and updating the scope, this review provides a more comprehensive evaluation of the tools used to identify the disease across various key data sources in the context of the significant changes in RA and JIA treatment and outcomes over the past decade.

METHODS

The study protocol was registered on PROSPERO (registration ID: CRD42024619130). The systematic review is reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) statement.[14]

Eligibility criteria. We included studies assessing the accuracy of codes and algorithms identifying RA and JIA in EHRs or other administrative databases against a reference standard based on clinician-documented diagnosis, clinician-conducted chart review, another form of clinician-confirmed diagnosis, or application of the American College of Rheumatology (ACR)/EULAR criteria[15] (eg, to general practitioner records, medical charts, office records) regardless of study design. Studies relying on self-reported diagnoses as the reference standard were excluded, as previously recommended.[16] Eligible studies had to clearly report algorithm components and describe methods for accuracy assessment. We excluded case reports, editorials, commentaries, and narrative reviews. Although systematic reviews were not included, relevant ones were used for reference checking. No restrictions were placed on language or publication date.

Information sources. We searched Medline (PubMed interface), Embase, and the Cochrane Central Register for Controlled Trials. Search strategies used medical subject headings (MeSH) terms and text-words aligned with our eligibility criteria (Supplementary Material S1) and were adapted for each database. To ensure literature saturation, we manually reviewed references of eligible studies and included articles from prior systematic reviews.

Search strategy. Our strategy was based on the Mini-Sentinel project search strategies,[12,17] updated with additional terms. We reviewed the PubMed MeSH terms, as well as abstracts and full texts of primary studies from a prior systematic review, to identify relevant terminology. The search was developed collaboratively and was validated by confirming that it retrieved all articles from the Chung et al[12] review. After the Medline search was finalized, it was adapted to the other databases. Full search strategies and validation process details are in

Supplementary Material S2. Searches were conducted on November 11, 2024.

Study records. *Data management.* We used Covidence software for record management and deduplication. Screening questions for title and abstract review were developed and tested. Reviewers completed training and pilot screening (50 abstracts by 5 reviewers) to calibrate eligibility criteria.

Study selection. Of the reviewers CS-H, JB, YA, M. Fatir, and FM, two independently screened titles and abstracts against eligibility criteria. Full texts of potentially eligible studies were reviewed in duplicate. Disagreements were resolved by discussion or a third reviewer. Reasons for full-text exclusions were documented.

Data collection process and items. Two reviewers independently extracted data using a standardized template based on the Standards for Reporting of Diagnostic Accuracy Studies criteria,[18] adapted for administrative databases. Discrepancies were resolved through discussion or a third reviewer. Extracted data included study characteristics, algorithms definitions, validation methods, and accuracy measures. The template is available as a data supplement (see Availability of data section). Classifications for the target condition are provided in Supplementary Material S3.

Diagnostic accuracy measures. Primary accuracy measures were sensitivity, specificity, PPV, and negative predictive value. When not reported, these were calculated from two-by-two tables. For practical interpretation, accuracy was categorized is high (≥80%), moderate (60%–80%), or low (<60%).[19] The unit of assessment was the individual. Further details are provided in Supplementary Material S3.

Risk of bias and applicability. Quality was assessed using the QUADAS-2 tool, which examines risk of bias and applicability across four domains: patient selection, index test, reference standard, and flow/timing.[20] Domains were rated as low, unclear, or high. Overall study bias followed QUADAS-2 guidance (ie, any high-risk domain rendered the study high risk overall).[14] Two reviewers independently assessed study quality, resolving disagreements through discussion.

Synthesis of results. We categorized algorithms based on the number and type of codes used to identify RA and JIA (eg, ICD, prescriptions, diagnostic tests, procedures). Classification followed the review by Shrestha et al,[21] with modifications made as necessary:

- Less restrictive algorithms: required a single diagnostic code from an outpatient visit or unspecified source (eg, single ICD code)
- Restrictive algorithms: required multiple codes or inclusion of specific types (eg, procedures, prescriptions, tests, or hospitalization codes)

Algorithms were categorized as either restrictive or less restrictive. For further analysis, we developed subcategories among restrictive algorithms based on expected accuracy and usage frequency in included studies. These were RA diagnosis codes and DMARD prescription codes, two or more ICD codes for RA, one or more ICD codes for RA by rheumatologist, and one or more ICD hospitalization codes for RA (full descriptions are in Supplementary Material S3).

Meta-analysis. We performed a bivariate random-effects meta-analysis using a generalized linear mixed-effects model to jointly synthesize sensitivity and specificity estimates, accounting for within- and between-study variability and their correlation.[22–24] Pooled PPV estimates were obtained via meta-analysis of proportions using the *meta* package in R, applying random-effects models within subgroups defined by algorithm and reference standard categories (*metaprop* function).

For each study, sensitivity and specificity with 95% confidence intervals (CIs) were calculated. Forest plots were generated for PPV, sensitivity, and specificity, displaying individual and pooled estimates by algorithm subgroup to capture between-study heterogeneity. Heterogeneity was assessed using I^2 statistics derived from variance components of the random-effects model. Analyses were conducted using R (version 4.4.3).

Availability of data. This research is based on published literature. All data used in this research are already included in the article or the supplementary material. The data extraction template is available at the Open Science Framework (https://osf.io/xfpzg).

RESULTS

Study selection. The search yielded 4,109 records; after removing duplicates, 3,452 were screened based on eligibility criteria, and 108 underwent full-text review. Seventy-seven articles were excluded (reasons provided in Figure 1). Reference checks of the 31 eligible articles identified 4 additional studies, totaling 35 included in the review (Figure 1).

Study characteristics. Most studies were from the United States and Canada (n = 21), followed by Europe, (mainly Italy, the United Kingdom, Denmark, and Sweden; n = 10), Asia (Japan and South Korea; n = 3), and Australia (n = 1). The majority were retrospective cohort studies (n = 25), followed by cross-sectional (n = 6) and case-control studies (n = 4). Two-thirds (n = 23, 66%) were published since the 2013 review (Supplementary Table S1).

The included studies (n = 35) used diverse electronic databases, including national health insurance claims (eg, Korean National Health Insurance database,[25] Medicare,[10,26–28] and Ontario Health Insurance Plan[9,29]), hospital discharge records

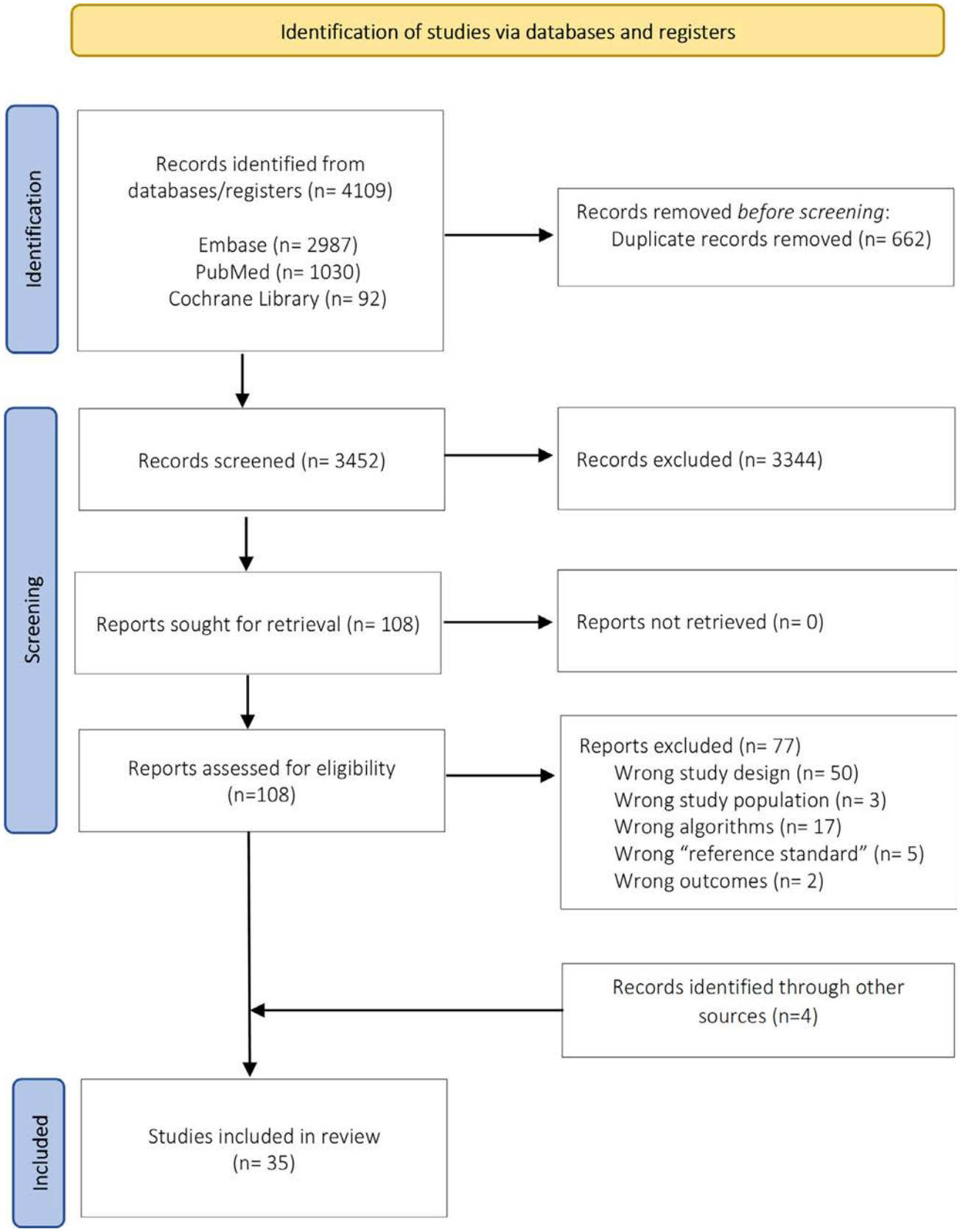

Figure 1. Flow diagram for study selection. Color figure can be viewed in the online issue, which is available at http://onlinelibrary.wiley.com/doi/10.1002/acr.25662/abstract.

(eg, Stockholm County Medical Information System,[30] Danish National Patient Registry,[31,32] Partners Healthcare database used by Brigham and Women's Hospital and Massachusetts General Hospital,[33–35] and Canadian Institute for Health Information Discharge Abstract Database[9,29]), and institutional EHRs (eg, Vanderbilt University Medical Center's Synthetic Derivate[33,36,37]; University of California, Los Angeles [UCLA] Health System[38]; Mayo Clinic Biobank[39]; and EHRs from specific rheumatologic institutions or disease registries[11,40–42]). Some also incorporated pharmacy data[31,43,44] and physician billing records.[42,45,46]

Most studies focused on patients with RA, although five examined patients with JIA.[36,40,46–48] Samples sizes ranged from small cohorts, for example, 31 and 94 participants,[47,49] to large datasets exceeding 55,000 individuals.[25] Women made up 50% to 80% of participants. Mean age varied, typically between 50 and 70 years for RA and <19 years for JIA. Full study characteristics are detailed in Supplementary Table S1.

RA case identification relied mainly on ICD codes (Table 1), with ICD-9 (62.8%) and ICD-10 (48.6%) codes being the most used (Supplementary Table S1). Some older datasets also used ICD-8 codes.[30,31] Additional criteria were applied in several studies, including medications (54.3%) and laboratory markers (22.8%) such as rheumatoid factor (RF) and anti–cyclic citrullinated peptide (anti-CCP) antibodies[35,50] (Table 1). Some studies implemented complex algorithms, including the eMERGE algorithm, which combined ICD codes with RF values and other autoimmune disease exclusions.[38,39] The algorithm developed by Liao et al,[35] incorporating ICD codes, prescriptions, and laboratory values, was validated in other settings by Carroll et al[33] and Huang et al[34] (Supplementary Table S1).

Most studies (71.4%) used a rheumatologist's clinical diagnosis as the reference standard (Table 1), whereas others applied established classification criteria, including the 1987 ACR criteria[51] and the 2010 ACR/EULAR criteria[15] and for JIA, the 2001 International League of Associations for Rheumatology criteria[52] (Supplementary Table S1). Four studies relied on medical records review by trained nurses or researchers, without clear specification of the reference standard[27,28,37,50] (Table 1).

Risk of bias and applicability. Risk of bias and applicability concerns varied across studies, with patient selection (unclear risk 54%, high risk 26%) and flow/timing being the most affected domains (unclear risk 57%, high risk 26%) (Table 2, Supplementary Figure S1). Patient selection showed the highest risk, mainly due to nonrandom inclusion, use of ICD codes without additional verification, and concerns regarding representativeness. The index test domain showed the lowest risk (63%), generally due to clearly defined and consistently applied algorithms, though some studies lacked clarity (eg, specific medications included in the algorithm) or did not report blinding to the reference standard. The reference standard domain showed low risk (57%), using established standards such as rheumatologist diagnosis or ACR/EULAR criteria,[15] though some had unclear validation methods. Flow and timing had mixed results, with high risk in studies that did not include all participants in the analysis or did not assess all diagnostic accuracy measures due to study design.

Overall, six studies were rated high risk of bias[26–28,30,41,49] due to limitations in multiple domains, potentially leading to misclassification and reduced generalizability. In contrast, studies by Carroll et al,[33] Cho et al,[25] Liao et al,[35] Singh et al,[53] and Widdifield et al[9] had low risk, with strong design across all domains (Table 2).

Applicability concerns were generally low (patient selection 51%, index test 66%, reference standard 66%), suggesting good relevance to real-world settings. However, studies using specialized databases or narrow inclusion criteria (eg, pediatric populations, single-center databases, incident RA cases) had moderate concerns because they may not be generalizable to broader RA populations (Table 2, Supplementary Figure S1).

Results of individual studies. *RA codes and algorithms.* Individual study results are detailed in Supplementary Table S2. Across studies, over 180 definitions/algorithms were evaluated. Percentage of studies that used less restrictive or restrictive algorithms and their categories are presented in Figure 2.

Less restrictive algorithms typically had high sensitivity but low specificity, increasing false-positives. For example, Almutairi et al[54] found 90% sensitivity but only 28.5% specificity using one or more RA primary codes in hospital discharge records, highlighting the risk of overidentification. Other studies[29,45,53] showed similar trends, with sensitivities >90% and specificities <65%.

Adding DMARD prescriptions to ICD codes improved specificity and PPV while maintaining good sensitivity. The study by Cho et al[25] achieved 85% sensitivity and 70% specificity using ICD-10 codes with biologic, DMARD, or nonsteroidal anti-inflammatory drug prescriptions, maintaining >85% PPV values in all developed algorithms. Similarly, Singh et al[53] reported 85% sensitivity, 83% specificity, and 81% PPV when using a similar approach. The study by Linauskas et al[31] showed PPV increased from 61.9% to 87.7% when requiring one or more DMARD prescriptions.

Requiring two or more ICD RA codes, separated over time, improved specificity while maintaining good sensitivity, though decreasing PPV. Hanly et al[45] reported 83% sensitivity, 81% specificity, and 52% PPV using two RA physician codes two or more months apart. Similar trends were seen in the studies by Kubota et al[55] (84% sensitivity, 99% specificity, and 59% PPV) and Widdifield et al[9] (84% sensitivity, 99% specificity, and 46% PPV).

Requiring one or more hospitalization records improved specificity and PPV but greatly reduced sensitivity. Convertino et al[56] found 95% specificity and 81% PPV but only 37%

Table 1. Summary characteristics and accuracy measures of included studies (N = 35)*

	Codes/algorithms					Reference standard			Measures of accuracy/results, %			
	Restrictive			Less restrictive								
Study (year)	D	M	L	D	M	CC	DR	MRR	Se	Sp	PPV	NPV
Rheumatoid arthritis												
Allebeck et al[30] (1983)	–	–	–	X	–	X	X	–	–	–	56.5–65.2	–
Almutairi et al[54] (2021)	X	X	–	–	–	X	X	–	25–90	16.6–85.7	63.6–97.9	7.6–30
Carrara et al[11] (2015)	X	X	–	X	X	X	–	–	82.4–96.3	90.3–99.8	72.5–85	97.1–99.9
Carroll et al[33] (2012)	X	X	X	–	–	–	X	–	57–79	–	87–95	–
Cho et al[25] (2013)	X	X	–	–	–	X	–	–	84.7–96.5	30.8–70.0	87.3–94.0	32.9–76.2
Convertino et al[56] (2021)	X	X	–	–	–	–	X	–	37–93	84–95	74–82	72–95
Curtis et al[26] (2018)	X	–	–	–	–	–	X	–	–	–	68–75	–
Fowles et al[27] (1995)	–	–	–	X	–	–	–	X	44.4	99.7	44.4	99.7
Hanly et al[45] (2015)	X	–	–	X	–	–	X	–	20.7–94.8	62.5–98.5	38.7–77.1	83.2–98.0
Huang et al[34] (2020)	X	X	X	–	–	X	X	–	47–89	50–95	54–93	50–90
Ibfelt et al[41] (2017)	X	–	–	–	–	–	X	–	–	–	79–92	–
Katz et al[42] (1997)	–	–	–	X	–	–	X	–	90	–	95	–
Kim et al[10] (2011)	X	X	–	–	–	X	X	–	–	–	33.6–88.9	–
Kronzer et al[39] (2020)	X	–	X	–	–	X	–	–	13–71	97–100	90–100	61–80
Kubota et al[55] (2021)	X	X	–	–	–	–	X	–	26–86.3	99.6–100	55.3–89.6	99.6–99.9
Liao et al[35] (2010)	X	X	X	–	–	X	X	–	51–63	–	88–94	–
Linauskas et al[31] (2018)	X	X	–	X	–	X	X	–	–	–	25–92.6	–
Losina et al[28] (2003)	X	–	–	–	–	–	–	X	65	–	86	–
Nanji et al[49] (2012)	–	–	–	X	–	–	X	–	–	–	70.9	–
Ng et al[50] (2012)	X	X	–	–	–	X	–	X	38.1–88.1	68.3–98.4	30.9–91.4	–
Paltta et al[43] (2021)	X	X	X	–	–	–	X	–	–	–	71–98	100–100
Pedersen et al[32] (2004)	X	–	–	–	–	X	X	–	–	–	46–58.9	–
Singh et al[53] (2004)	X	X	X	X	–	–	X	–	76.5–100	55–97.1	66.2–97	77–86.2
Sugiyama et al[67] (2024)	X	X	–	–	–	X	X	–	–	–	22.6–88.6	–
Thomas et al[48] (2008)	X	X	–	–	–	X	–	–	78–93	40–96	64.8–96	69.5–84.5
Waldenlind et al[44] (2014)	X	–	–	X	–	X	–	–	–	–	83.3–91	–
Wei et al[37] (2016)	X	X	–	X	X	–	–	X	0–60	–	0–88	–
Widdifield et al[29] (2013)	X	X	–	X	–	–	X	–	23–100	60–96	55–93	72–100
Widdifield et al[9] (2014)	X	–	–	X	–	–	X	–	22–90	97–100	20–88	99–100
Zheng et al[38] (2022)	X	–	X	–	–	X	X	–	71.9–75.5	95.2–99.5	86.4–93.8	74.4–77.2
Zhou et al[68] (2016)	X	X	–	–	–	–	X	–	83–94	94.6–99	30.9–91.3	–
Juvenile idiopathic arthritis												
Harrold et al[40] (2013)	X	–	X	–	–	–	X	–	81–100	–	45–91	–
Peterson et al[36] (2023)	X	–	–	–	–	–	X	–	63–100	–	48–97	–
Shiff et al[46] (2017)	X	–	–	–	–	–	X	–	81.9–90.8	84.8–92.6	91.1–94.9	–
Stringer and Bernatsky[47] (2015)	X	–	–	X	–	X	X	–	53–86	–	52–65	–
Thomas et al[48] (2008)	X	X	–	–	–	X	–	–	40–100	0–100	86.2–100	21–80

* CC, clinical criteria (American College of Rheumatology, EULAR, International League of Associations for Rheumatology)[15,51,52]; D, diagnosis codes; DR, diagnosis or chart review by rheumatologist; L, laboratory test codes; M, medication codes; MRR, medical records review by trained nurses or authors with no clear specification of the reference standard used; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity.

sensitivity when requiring one RA hospitalization or emergency admission code. Similarly, the study by Hanly et al[45] achieved 98.5% specificity, 77% PPV, and 20.7% sensitivity when using one or more hospitalization codes. Similar patterns were noted by Widdifield et al.[9,29]

Algorithms requiring at least one rheumatologist diagnosis code showed overall moderate to high sensitivity and specificity, with moderate PPV. The studies by Widdifield et al[9,29] achieved 81% and 99% sensitivity, 99% and 77% specificity, and 51% and 68% PPV, respectively. Similarly, Hanly et al[45] reported 88% sensitivity, 75.5% specificity, and 47% PPV when using one or more RA codes by a rheumatologist. In contrast, Carrara et al[11] reported 99.8% specificity but only 13% sensitivity when combining one RA rheumatologist certification with other ICD-9 codes.

The eMERGE algorithm, combining ICD codes, RF values, and other autoimmune disease exclusions, demonstrated high specificity (95%–99%) and PPV (94%–97%) in both studies evaluating it (studies by Zheng et al[38] and Kronzer et al,[39] respectively). However, sensitivity varied likely due to differences in sample size and study population. Zheng et al[38] (n = 4,766, UCLA Health System) reported a higher sensitivity (72%), whereas the study by Kronzer et al[39] (n = 497, Mayo Clinic Biobank included in the Rochester Epidemiology Project) achieved a sensitivity of 53% in a smaller and more geographically limited sample. However, Kronzer et al[39] found that sensitivity increased with RA

Table 2. Risk of bias and applicability concerns of included studies (QUADAS-2 tool)

	Domain 1: patient selection		Domain 2: index test		Domain 3: reference standard		Domain 4: flow and timing
Author, year, country	Risk	Concern	Risk	Concern	Risk	Concern	Risk
Allebeck et al,[30] 1983, Sweden	High	Unclear	Unclear	Unclear	High	High	High
Almutairi et al,[54] 2021, Australia	High	Unclear	Low	Low	Low	Low	High
Carrara et al,[11] 2015, Italy	Unclear	Low	Low	Low	Low	Low	Unclear
Carroll et al,[33] 2012, USA	Low	Low	Low	Low	Unclear	Low	Unclear
Cho et al,[25] 2013, South Korea	Unclear	Unclear	Low	Low	Low	Low	Low
Convertino et al,[56] 2021, Italy	Unclear	Low	Low	Low	Low	Low	Low
Curtis et al,[26] 2018, USA	High	High	High	High	Low	Low	High
Fowles et al,[27] 1995, USA	High	High	High	Unclear	High	High	Unclear
Hanly et al,[45] 2015, Canada	Unclear	Unclear	Low	Low	Low	Low	Low
Harrold et al,[40] 2013, USA	Unclear	Low	Unclear	Unclear	Unclear	Unclear	Unclear
Huang et al,[34] 2020, USA	Unclear	Low	Low	Low	Low	Low	Unclear
Ibfelt et al,[41] 2017, Denmark	High	High	Unclear	Unclear	Low	Low	High
Katz et al,[42] 1997, USA	High	High	Unclear	Unclear	Low	Low	Unclear
Kim et al,[10] 2011, USA	High	High	Low	Low	Low	Low	Unclear
Kronzer et al,[39] 2020, USA	Unclear	Unclear	Low	Low	Unclear	Unclear	Low
Kubota et al,[55] 2021, Japan	Unclear	Low	Low	Low	Low	Low	Unclear
Liao et al,[35] 2010, USA	Low	Low	Low	Low	Low	Low	Unclear
Linauskas et al,[31] 2018, Denmark	Unclear	Low	Low	Low	Unclear	Unclear	Unclear
Losina et al,[28] 2003, USA	High	High	Unclear	Low	Unclear	Low	High
Nanji et al,[49] 2012, Canada	High	High	Unclear	Unclear	Low	Low	High
Ng et al,[50] 2012, USA	Unclear	Unclear	Low	Low	Unclear	Low	Unclear
Paltta et al,[43] 2021, Finland	Unclear	Low	Unclear	Unclear	Unclear	Unclear	Unclear
Pedersen et al,[32] 2004, Denmark	Unclear	Low	Low	Low	Low	Low	High
Peterson et al,[36] 2023, USA	Unclear	Low	Low	Low	Unclear	Unclear	Unclear
Singh et al,[53] 2004, USA	Low	Low	Low	Low	Low	Low	Unclear
Shiff et al,[46] 2017, Canada	Unclear	Low	Low	Low	Low	Low	Unclear
Stringer and Bernatsky,[47] 2015, Canada	Unclear	Low	Unclear	Unclear	Unclear	Unclear	Unclear
Sugiyama et al,[67] 2024, Japan	Unclear	Unclear	Low	Low	Low	Low	Unclear
Thomas et al,[48] 2008, UK	Low	Low	Unclear	Unclear	Low	Low	High
Waldenlind et al,[44] 2014, Sweden	Unclear	Unclear	Unclear	Unclear	Low	Low	High
Wei et al,[37] 2016, USA	Unclear	Unclear	Low	Low	Unclear	Unclear	Unclear
Widdifield et al,[29] 2013, Canada	Unclear	Unclear	Low	Low	Low	Low	Low
Widdifield et al,[9] 2014, Canada	Low	Low	Low	Low	Unclear	Unclear	Low
Zheng et al,[38] 2022, USA	Low	Low	Low	Low	Unclear	Unclear	Unclear
Zhou et al,[68] 2016, UK	Low	Low	Unclear	Unclear	Unclear	Unclear	Unclear

duration and EHR history, reaching 71% for patients with ≥10 years' disease duration. Overall, the eMERGE algorithm performed well, with sensitivity possibly influenced by dataset scope.

The RA algorithm initially developed by Liao et al,[35] validated by Carroll et al[33] and Huang et al,[34] consistently showed moderate sensitivity (51%–79%) and high PPV (>87%) across different settings, showing a good balance between measures for identifying RA cases. Huang et al[34] enhanced the model by incorporating ICD-10 codes and newer biologic treatments, achieving 77% sensitivity, 95% specificity, and 91% PPV. Overall, the algorithm maintained high accuracy, with sensitivity improving slightly as case definitions and coding systems evolved.

Both the eMERGE and Liao algorithms used laboratory data (RF, anti-CCP), which improved specificity and PPV, and had limited effect on sensitivity. Paltta et al[43] reported 98% PPV when incorporating anti–citrullinated protein antibody positivity in their algorithm, compared with ICD codes alone (82%) or ICD codes and DMARDs (89%). Singh et al[53] also showed improved sensitivity (88%), specificity (91%), and PPV (93%) when adding positive RF to their algorithm, compared to ICD codes alone (sensitivity 100%, specificity 55%, PPV 66%) or ICD codes and DMARDs (sensitivity 85%, specificity 83%, PPV 81%), suggesting that laboratory data helps reduce false-positives while maintaining a balanced improvement in the accuracy measures.

JIA codes and algorithms. JIA studies evaluated similar EHR-based algorithms (Supplementary Table S3). Less restrictive algorithms, such as one or more JIA diagnosis codes, showed high sensitivity but moderate to low PPV. Harrold et al[40] and Peterson et al[36] both reported 100% sensitivity, with PPVs of 45% and 48%, respectively.

Only one study assessed the accuracy of DMARD prescription alone as a definition. Thomas et al[48] found 40% sensitivity but 100% specificity and PPV when requiring one or more DMARD prescriptions without prior alternative indication.

Requiring two or more ICD codes for JIA improved PPV while maintaining good sensitivity. Reported sensitivity ranged from 72% to 93% and PPV ranged from 60% to 100% across studies.[36,40,46–48] Peterson et al[36] noted that increasing the

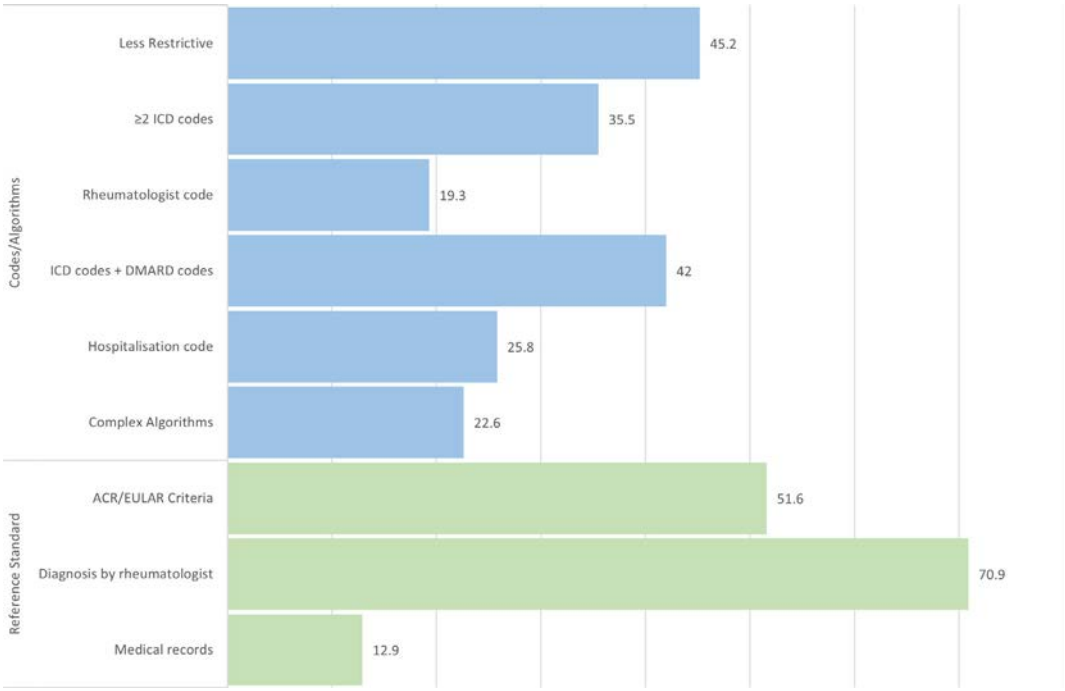

Figure 2. Codes/algorithms and reference standards used by the included studies assessing rheumatoid arthritis in electronic health records (n = 31). ACR, American College of Rheumatology; DMARD, disease-modifying antirheumatic drug; ICD, *International Classification of Diseases*.

number of JIA codes (four or more) raised PPV (83%) without substantial decrease in sensitivity (87%), compared to using one or more codes (PPV 48%). Shiff et al[46] found that specificity improved (85%–92%) with more ICD codes over longer intervals.

Two studies assessed algorithms requiring one or more JIA codes from a rheumatologist. The study by Harrold et al[40] showed that requiring a rheumatologist visit improved PPV from 69% to 90% while maintaining good sensitivity (81%). The study by Stringer et al[47] showed similar sensitivity (86%) but decreased PPV (58%), which increased to 65% when using two JIA codes eight or more weeks apart within two years, maintaining good sensitivity (81%).

Meta-analysis. Meta-analyses showed substantial variability in the accuracy of RA identification algorithms in EHRs, depending on algorithm components and reference standards. Using rheumatologist diagnosis as a reference, algorithms requiring two or more ICD codes had the highest pooled sensitivity (0.89 [95% CI 0.75–0.95]; n = 4 studies) and specificity (0.96 [95% CI 0.74–1.00]; n = 4), though with high heterogeneity (I^2 = 91.3% and I^2 = 99.8%, respectively). Algorithms combining ICD codes with DMARD prescriptions also showed good pooled sensitivity (0.79 [95% CI 0.61–0.90], I^2 = 90.6%; n = 5) and high specificity (0.96 [95% CI 0.72–1.00], I^2 = 98.9%; n = 5) (Table 3, Supplementary Figure S2). Algorithms using one or more ICD codes by a rheumatologist also showed high performance (sensitivity 0.91 [95% CI 0.70–0.98], specificity 0.94 [95% CI 0.49–1.00]; n = 3) but were based on only three studies, warranting cautious interpretation. A meta-analysis of sensitivity and specificity using ACR/EULAR classification criteria[15] as the reference standard was not feasible due to insufficient data.

PPV results varied by reference standard and showed substantial heterogeneity. Using rheumatologist diagnosis as the reference, the highest pooled PPVs were seen in algorithms combining ICD and DMARD codes (0.78 [95% CI 0.63–0.88]; n = 10), those including one or more ICD codes by a rheumatologist (0.64 [95% CI 0.51–075]; n = 4), and those requiring one or more hospitalization codes (0.79 [95% CI 0.77–0.81]; n = 7). In contrast, when using ACR/EULAR classification criteria,[15] PPVs were slightly lower when requiring two or more ICD codes (0.62 [95% CI 0.14–0.94]; n = 2) or ICD and DMARD codes (0.73 [95% CI 0.45–0.90]; n = 6). Algorithms requiring one or more ICD codes by a rheumatologist demonstrated the highest pooled PPV (0.85 [95% CI 0.35–0.98]; n = 3), though with high heterogeneity (I^2 = 96.3%) (Table 3, Supplementary Figure S3).

DISCUSSION

This systematic review analyzed the accuracy of algorithms for identifying RA and JIA in EHRs. Overall, algorithms combining ICD codes with DMARD prescriptions or those that include one or more ICD codes assigned by a rheumatologist demonstrated the highest accuracy, balancing sensitivity, specificity, and PPV. These algorithms are more reliable for identifying confirmed RA cases, though rheumatologist-assigned codes may miss cases diagnosed in primary care, potentially leading to underestimation

Table 3. Pooled estimates of sensitivity, specificity, and PPV across algorithm categories and reference standard (diagnosis by rheumatologist, ACR/EULAR criteria[15])*

Categories by reference standard	No. of studies	Estimate[a] (95% CI)	I^2
Diagnosis by rheumatologist			
Sensitivity			
1 single ICD code	7	0.92 (0.78–0.98)	92.6
At least 2 ICD codes	4	0.89 (0.75–0.95)	91.3
At least 1 ICD code by rheumatologist	3	0.91 (0.70–0.98)	94.7
At least 1 ICD hospitalization code	6	0.44 (0.22–0.67)	97
ICD codes and DMARD codes	5	0.79 (0.61–0.90)	90.6
Specificity			
1 single ICD code	7	0.90 (0.53–0.98)	99.8
At least 2 ICD codes	4	0.96 (0.74–1.00)	99.8
At least 1 ICD code by rheumatologist	3	0.94 (0.49–1.00)	99.7
At least 1 ICD hospitalization code	6	0.99 (0.96–1.00)	98.5
ICD codes and DMARD codes	5	0.96 (0.72–1.00)	98.9
PPV			
1 single ICD code	14	0.63 (0.49–0.75)	97.8
At least 2 ICD codes	7	0.64 (0.52–0.75)	94.5
At least 1 ICD code by rheumatologist	4	0.64 (0.51–0.75)	94.8
At least 1 ICD hospitalization code	7	0.79 (0.77–0.81)	8.9
ICD codes and DMARD codes	10	0.78 (0.63–0.88)	96.5
ACR/EULAR criteria[15]			
PPV			
1 single ICD code	6	0.74 (0.54–0.87)	92.6
At least 2 ICD codes	2	0.62 (0.14–0.94)	98.2
At least 1 ICD code by rheumatologist	3	0.85 (0.35–0.98)	96.3
At least 1 ICD hospitalization code	–	–	–
ICD codes and DMARD codes	6	0.73 (0.45–0.90)	99.6

* ACR, American College of Rheumatology; CI, confidence interval; DMARD, disease-modifying antirheumatic drug; ICD, *International Classification of Diseases*; PPV, positive predictive value.
[a] Pooled estimate of sensitivity, specificity, or PPV.

of disease prevalence. Less restrictive algorithms are useful for initial case identification (highly sensitive) but may require additional criteria to enhance specificity and PPV. Requiring two or more ICD codes is a well-balanced approach, reducing false-positives (higher specificity) without significantly lowering sensitivity, though achieving moderate PPV. Hospitalization-based codes are highly specific for identifying severe RA cases but may exclude many nonhospitalized, mild/moderate cases, making it less suitable for general RA case identification.

For JIA, a single ICD code is highly sensitive but may overestimate prevalence due to false-positives. Requiring two or more spaced ICD codes improves accuracy and seems to be the best approach, providing a good balance of sensitivity, specificity, and PPV in EHRs or administrative data. However, validation evidence for JIA remains limited, highlighting the need for further research.

Our findings align with and extend previous systematic reviews on RA and connective tissue disease algorithms in EHRs. Compared to the 35 studies we included, the 2013 review by Chung et al[12] included 9 studies and found that algorithms combining diagnostic and prescription data—particularly those including specialist codes, medication data, or laboratory data—achieved higher PPVs (>80%), whereas ICD-only algorithms had more variable performance across settings, generally with low PPVs. Similar trends were seen in studies for other conditions, but there were generally lower PPVs (<65%) in general populations,[21,57] probably reflecting the challenges of identifying these conditions due to nonspecific coding and overlap with other autoimmune conditions. Similarly, a systematic review by Crossfield et al[58] on gout also highlighted variability in methods and performance, reflecting a lack of standardization in defining rheumatologic conditions using real-world data. Compared to these conditions, RA appears to be more accurately identified in EHRs, especially when algorithms include medication data or are supported by specialist confirmation.

The accuracy of hospitalization-based algorithms must be interpreted within the context of evolving health care delivery. RA hospitalizations have declined significantly over the past two decades due to earlier diagnosis, improved outpatient care, and widespread use of biologic therapies.[59–61] As a result, performance of algorithms using hospitalization codes in historic data may be less representative of current patient populations. Similarly, treatment patterns have evolved, with earlier DMARD initiation, combined therapies (eg, biologics and DMARDs),[62–64] and the development of targeted-synthetic DMARDs (eg, JAK inhibitors), which may affect the predictive value of medication-based algorithms applied to contemporary prescribing data.

The transition from ICD-9 to ICD-10 introduced more specific diagnosis codes, potentially influencing algorithm performance. In this review, 62.8% of studies examined ICD-9 codes

Table 4. Practical summary of algorithm categories and their potential uses in EHR-based research*

Categories	Sensitivity	Specificity	PPV	Potential uses	Data source
1 single ICD code	High	High	Low/moderate	Initial case identification, screening	Claims, EHRs
At least 2 ICD codes	High	High	Low/moderate	Cohort definition	Claims, EHRs
At least 1 ICD code by rheumatologist	Moderate/high	High	Moderate/high	Accurate case identification	Specialty EHRs, hospital records
At least 1 ICD hospitalization code	Low	Very high	Moderate	Severe cases, inpatient cohorts	Hospital discharge records
ICD codes and DMARDs codes	Moderate/high	High	Moderate/high	Clinical studies, treatment effects	Claims and EHRs with prescription data

* DMARD, disease-modifying antirheumatic drug; EHR, electronic health record; ICD, *International Classification of Diseases*; PPV, positive predictive value.

and 48.6% examined ICD-10 codes, reflecting variability in study periods, country-level implementation, differences in capacity to adopt new coding systems, and the use of historical data sources. Algorithms developed with ICD-9 codes may not translate directly to the ICD-10 without revalidation,[65,66] particularly for specific or rare disease subtypes. These changes underscore the need for regular update and validation of algorithms to align with current coding systems and clinical practice.

From a practical perspective, the selection of an algorithm for identifying RA or JIA should be guided by the purpose of the study and the nature of the data source. For studies aiming to evaluate treatment outcomes, assess adverse events, or conduct comparative effectiveness studies, algorithms with high specificity may be preferable to ensure the correct identification of true cases. In contrast, for surveillance or studies assessing prevalence or burden of disease, algorithms prioritizing higher sensitivity may be preferable, even at lower PPV. To further support practical application, we provide a summary table (Table 4) outlining algorithm definitions, accuracy, and considerations for their use depending on study purposes and data source, which may assist researchers and clinicians in selecting or adapting algorithms in future EHR-based studies.

Future research should focus on validating existing algorithms, such as the eMERGE[38,39] and Liao et al[35] algorithms, across diverse health care systems and populations. Although both show consistent diagnostic accuracy in diverse settings, the generalizability of the Liao algorithm beyond academic health centers remains uncertain. Validation using community-based data is needed to ensure broader applicability. Clear and transparent reporting of algorithm components is essential for replication, external validation, and broader implementation in different contexts. Additionally, the development and validation of algorithms using machine learning or natural language processing techniques hold promise for further improving case ascertainment, particularly in settings where structured coding is limited.

The studies reviewed exhibited substantial methodologic variability, making direct comparisons difficult. Differences in sample size, sampling methods, settings, population, and data sources likely contributed to the high statistical heterogeneity in several analyses. Algorithm definitions varied widely, even within similar categories, highlighting a lack of standardization in algorithm development and validation methods. Despite this, meta-analyses were feasible by grouping algorithms into broad but conceptually coherent categories, prioritizing simplicity and clinical interpretability. Importantly, algorithm performance is highly context dependent and may not be generalizable across different health care systems, data sources, or patient populations. As such, the pooled estimates should not be used to directly select algorithms for a given setting but rather to identify overall trends and guide clinicians and researchers in choosing or adapting algorithms appropriate to their specific context and data source. The results may also support future validation efforts in more targeted contexts.

Although many studies incorporated medication use, most grouped DMARDs as a single category, precluding an evaluation of the contribution of specific drugs or therapeutic classes to the accuracy. Several studies did not report full accuracy measures due to sampling designs, which involved evaluating only algorithm-positive cases and thus only the PPV. The PPV depends on disease prevalence within the population being studied. Several studies selected their populations from hospital settings or rheumatologic clinics, where the prevalence of RA is likely to be higher than in the general population, potentially leading to an overestimation of the PPVs. Another source of variability may be the data used in each study and the purpose for which the data were originally generated (ie, claims/reimbursement vs clinical care). Therefore, findings should be interpreted cautiously and, ideally, validated in general populations to better understand their performance across varying levels of disease prevalence. Additionally, sensitivity and specificity may be influenced by the characteristics of the population included in the reference standard. For example, algorithms validated using rheumatologist-confirmed diagnoses may differ in performance from those validated against classification criteria, which may be more stringent or require specific data elements. These methodologic differences highlight the need to consider context when comparing accuracy across studies. Finally, the search focused on studies explicitly reporting diagnostic validation, potentially excluding relevant studies that addressed this in broader investigations (ie, assessed

algorithm accuracy as an intermediate objective as part of an EHR-based study of RA outcomes). A protocol deviation was the omission of a gray literature search due to the nature of the review and to the expectation that relevant studies would be published primarily in peer-reviewed journals.

This systematic review incorporates comprehensive and contemporary evidence from diverse geographic areas. It comprehensively synthesizes and quantitatively analyzes RA and JIA algorithm accuracy in EHRs across diverse health care settings. It adheres to PRISMA guidelines, systematically categorizes algorithms by restrictiveness to facilitate clinically meaningful comparisons, and uses the QUADAS-2 tool for transparent quality assessment.

This review highlights the variability in accuracy of algorithms used to identify RA and JIA in EHRs. Algorithms combining diagnostic codes and prescription data, particularly DMARDs, or a rheumatologist diagnosis demonstrated the highest overall performance, with a good balance of sensitivity, specificity, and PPV. Less restrictive algorithms may aid initial case identification but require refinement for specific research purposes. The choice of algorithm should be tailored to the study objective and data source. Future efforts should prioritize external validation of high-performing algorithms across diverse settings to improve standardization and reliability in real-world data research.

ACKNOWLEDGMENT

The authors acknowledge the use of large language models to assist with English grammar and spelling during the preparation of the manuscript.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Prof. Moriarty confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

REFERENCES

1. Agarwal SK. Core management principles in rheumatoid arthritis to help guide managed care professionals. J Manag Care Pharm 2011; 17(9 suppl B):S03–S08.
2. Abramowicz S, Kim S, Prahalad S, et al. Juvenile arthritis: current concepts in terminology, etiopathogenesis, diagnosis, and management. Int J Oral Maxillofac Implants 2016;45(7):801–812.
3. Hochberg MC. Adult and juvenile rheumatoid arthritis: current epidemiologic concepts. Epidemiol Rev 1981;3(1):27–44.
4. Sangha O. Epidemiology of rheumatic diseases. Rheumatology (Oxford) 2000;39(suppl 2):3–12.
5. Vinet E, Chakravarty EF, Simard JF, et al. Use of administrative databases to assess reproductive health issues in rheumatic diseases. Rheum Dis Clin North Am 2018;44(2):327–336.
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58(4):323–337.
7. De Coster C, Quan H, Finlayson A, et al. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. BMC Health Serv Res 2006; 6(1):77.
8. Benchimol EI, Manuel DG, To T, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J Clin Epidemiol 2011;64(8):821–829.
9. Widdifield J, Bombardier C, Bernatsky S, et al. An administrative data validation study of the accuracy of algorithms for identifying rheumatoid arthritis: the influence of the reference standard on algorithm performance. BMC Musculoskelet Disord 2014;15(1):216.
10. Kim SY, Servi A, Polinski JM, et al. Validation of rheumatoid arthritis diagnoses in health care utilization data. Arthritis Res Ther 2011; 13(1):R32.
11. Carrara G, Scirè CA, Zambon A, et al. A validation study of a new classification algorithm to identify rheumatoid arthritis using administrative health databases: case-control and cohort diagnostic accuracy studies. Results from the RECord linkage On Rheumatic Diseases study of the Italian Society for Rheumatology. BMJ Open 2015;5(1):e006029.
12. Chung CP, Rohan P, Krishnaswami S, et al. A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. Vaccine 2013;31(suppl 10): K41–K61.
13. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf 2012;21(suppl 1):82–89.
14. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020;370:m2632.
15. 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.
16. Widdifield J, Labrecque J, Lix L, et al. Systematic review and critical appraisal of validation studies to identify rheumatic diseases in health administrative databases. Arthritis Care Res (Hoboken) 2013;65(9): 1490–1503.
17. McPheeters ML, Sathe NA, Jerome RN, et al. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. Vaccine 2013;31(suppl 10):K2–K6.
18. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11):e012799.
19. Nelson DE, Holtzman D, Bolen J, et al. Reliability and validity of measures from the behavioral risk factor surveillance system (BRFSS). Soz Praventivmed 2001;46(suppl 1):S3–S42.
20. Reitsma H, Rutjes A, Whiting P, et al. Assessing methodological quality. In: Deeks J, Bossuyt P, Gatsonis C, eds. Cochrane Handbook of Systematic Reviews of Diagnostic Test Accuracy. Version 1.0. The Cochrane Collaboration; 2009:1–24.
21. Shrestha S, Dave AJ, Losina E, et al. Diagnostic accuracy of administrative data algorithms in the diagnosis of osteoarthritis: a systematic review. BMC Med Inform Decis Mak 2016;16(1):82.
22. Macaskill P, Gatsonis C, Deeks J, et al. Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonics C, eds. Cochrane

Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 1.0. The Cochrane Collaboration; 2010:1–61.

23. Takwoingi Y, Dendukuri N, Schiller I, et al. Supplementary material 1 to chapter 10: code for undertaking meta-analysis. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0. The Cochrane Collaboration; 2023:1–83.
24. Barker TH, Migliavaca CB, Stein C, et al. Conducting proportional meta-analysis in different types of systematic reviews: a guide for synthesisers of evidence. BMC Med Res Methodol 2021;21(1):189.
25. Cho SK, Sung YK, Choi CB, et al. Development of an algorithm for identifying rheumatoid arthritis in the Korean National Health Insurance claims database. Rheumatol Int 2013;33(12):2985–2992.
26. Curtis JR, Xie F, Chen L, et al. Evaluation of a methodologic approach to define an inception cohort of rheumatoid arthritis patients using administrative data. Arthritis Care Res (Hoboken) 2018;70(10):1541–1545.
27. Fowles JB, Lawthers AG, Weiner JP, et al. Agreement between physicians' office records and Medicare Part B claims data. Health Care Financ Rev 1995;16(4):189–199.
28. Losina E, Barrett J, Baron JA, et al. Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. J Clin Epidemiol 2003;56(6):515–519.
29. Widdifield J, Bernatsky S, Paterson JM, et al. Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: a validation study using the medical records of rheumatologists. Arthritis Care Res (Hoboken) 2013;65(10):1582–1591.
30. Allebeck P, Ljungström K, Allander E. Rheumatoid arthritis in a medical information system: how valid is the diagnosis? Scand J Soc Med 1983;11(1):27–32.
31. Linauskas A, Overvad K, Johansen MB, et al. Positive predictive value of first-time rheumatoid arthritis diagnoses and their serological subtypes in the Danish National Patient Registry. Clin Epidemiol 2018; 10:1709–1720.
32. Pedersen M, Klarlund M, Jacobsen S, et al. Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. Eur J Epidemiol 2004;19(12):1097–1103.
33. Carroll RJ, Thompson WK, Eyler AE, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012;19(e1):e162–e169.
34. Huang S, Huang J, Cai T, et al. Impact of ICD10 and secular changes on electronic medical record rheumatoid arthritis algorithms. Rheumatology (Oxford) 2020;59(12):3759–3766.
35. 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.
36. Peterson HM, Vela KL, Barnado A, et al. Developing electronic health record algorithms that accurately identify patients with juvenile idiopathic arthritis. Semin Arthritis Rheum 2023;59:152167.
37. Wei WQ, Teixeira PL, Mo H, et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(e1): e20–e27.
38. Zheng HW, Ranganath VK, Perry LC, et al. Evaluation of an automated phenotyping algorithm for rheumatoid arthritis. J Biomed Inform 2022;135:104214.
39. Kronzer VL, Wang L, Liu H, et al. Investigating the impact of disease and health record duration on the eMERGE algorithm for rheumatoid arthritis. J Am Med Inform Assoc 2020;27(4):601–605.
40. Harrold LR, Salman C, Shoor S, et al. Incidence and prevalence of juvenile idiopathic arthritis among children in a managed care population, 1996-2009. J Rheumatol 2013;40(7):1218–1225.
41. Ibfelt EH, Sørensen J, Jensen DV, et al. Validity and completeness of rheumatoid arthritis diagnoses in the nationwide DANBIO clinical register and the Danish National Patient Registry. Clin Epidemiol 2017;9: 627–632.
42. Katz JN, Barrett J, Liang MH, et al. Sensitivity and positive predictive value of Medicare Part B physician claims for rheumatologic diagnoses and procedures. Arthritis Rheum 1997;40(9):1594–1600.
43. Paltta J, Heikkilä HK, Pirilä L, et al; FinnGen investigators. The validity of rheumatoid arthritis diagnoses in Finnish biobanks. Scand J Rheumatol 2023;52(1):1–9.
44. Waldenlind K, Eriksson JK, Grewin B, et al. Validation of the rheumatoid arthritis diagnosis in the Swedish National Patient Register: a cohort study from Stockholm County. BMC Musculoskelet Disord 2014;15(1):432.
45. Hanly JG, Thompson K, Skedgel C. The use of administrative health care databases to identify patients with rheumatoid arthritis. Open Access Rheumatol 2015;7:69–75.
46. Shiff NJ, Oen K, Rabbani R, et al. Validation of administrative case ascertainment algorithms for chronic childhood arthritis in Manitoba, Canada. Rheumatol Int 2017;37(9):1575–1584.
47. Stringer E, Bernatsky S. Validity of juvenile idiopathic arthritis diagnoses using administrative health data. Rheumatol Int 2015;35(3): 575–579.
48. Thomas SL, Edwards CJ, Smeeth L, et al. How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the general practice research database? Arthritis Rheum 2008;59(9):1314–1321.
49. Nanji JA, Choi M, Ferrari R, et al. Time to consultation and disease-modifying antirheumatic drug treatment of patients with rheumatoid arthritis–northern Alberta perspective. J Rheumatol 2012;39(4):707–711.
50. Ng B, Aslam F, Petersen NJ, et al. Identification of rheumatoid arthritis patients using an administrative database: a Veterans Affairs study. Arthritis Care Res (Hoboken) 2012;64(10):1490–1496.
51. 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.
52. Petty RE, Southwood TR, Manners P, et al. International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. J Rheumatol 2004;31(2): 390–392.
53. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. Arthritis Rheum 2004;51(6):952–957.
54. Almutairi K, Inderjeeth C, Preen DB, et al. The accuracy of administrative health data for identifying patients with rheumatoid arthritis: a retrospective validation study using medical records in Western Australia. Rheumatol Int 2021;41(4):741–750.
55. Kubota K, Yoshizawa M, Takahashi S, et al. The validity of the claims-based definition of rheumatoid arthritis evaluated in 64 hospitals in Japan. BMC Musculoskelet Disord 2021;22(1):373.
56. Convertino I, Cazzato M, Giometto S, et al. Validation of algorithms for selecting rheumatoid arthritis patients in the Tuscan healthcare administrative databases. Sci Rep 2021;11(1):20314.
57. Moores KG, Sathe NA. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. Vaccine 2013;31(suppl 10):K62–K73.
58. Crossfield SSR, Lai LYH, Kingsbury SR, et al. Variation in methods, results and reporting in electronic health record-based studies evaluating routine care in gout: a systematic review. PLoS One 2019; 14(10):e0224272.
59. Iyer P, Gao Y, Field EH, et al. Trends in hospitalization rates, major causes of hospitalization, and in-hospital mortality in rheumatoid arthritis in the United States from 2000 to 2014. ACR Open Rheumatol 2020;2(12):715–724.

60. Rai SK, Aviña-Zubieta JA, McCormick N, et al. Trends in gout and rheumatoid arthritis hospitalizations in Canada from 2000 to 2011. Arthritis Care Res (Hoboken) 2017;69(5):758–762.
61. Kiadaliri AA, Englund M. Temporal trends and regional disparity in rheumatoid arthritis and gout hospitalizations in Sweden, 1998-2015. Clin Rheumatol 2018;37(3):825–830.
62. Aletaha D, Smolen JS. The rheumatoid arthritis patient in the clinic: comparing more than 1,300 consecutive DMARD courses. Rheumatology (Oxford) 2002;41(12):1367–1374.
63. Betts KA, Griffith J, Ganguli A, et al. Economic burden and treatment patterns of cycling between conventional synthetic disease-modifying antirheumatic drugs among biologic-treated patients with rheumatoid arthritis. Clin Ther 2016;38(5):1205–1216.
64. Kim G, Barner JC, Rascati K, et al. Examining time to initiation of biologic disease-modifying antirheumatic drugs and medication adherence and persistence among Texas Medicaid recipients with rheumatoid arthritis. Clin Ther 2016;38(3):646–654.
65. Simeone JC, Liu X, Bhagnani T, et al. Comparison of ICD-9-CM to ICD-10-CM crosswalks derived by physician and clinical coder vs. automated methods. Perspect Health Inf Manag 2021;18 (Spring):1e.
66. Nam YH, Mendelsohn AB, Panozzo CA, et al. Health outcomes coding trends in the US Food and Drug Administration's Sentinel System during transition to International Classification of Diseases-10 coding system: a brief review. Pharmacoepidemiol Drug Saf 2021;30(7): 838–842.
67. Sugiyama N, Kinjo M, Jinno S, et al. Validation of claims-based algorithms for rheumatoid arthritis in Japan: results from the VALIDATE-J study. Int J Rheum Dis 2024;27(1):e15001.
68. Zhou SM, Fernandez-Gutierrez F, Kennedy J, et al; UK Biobank Follow-up and Outcomes Group. Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. PLoS One 2016; 11(5):e0154515.

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 683–687
DOI 10.1002/acr.25669

American College of Rheumatology
Empowering Rheumatology Professionals

BRIEF REPORT

Building Consensus on the Essential Elements of the Musculoskeletal Physical Examination During Rheumatology Telehealth Encounters

Lisa Zickuhr,[1] Alberto Sobrero,[2] Daniel Albert,[3] Amanda S. Alexander,[4] Tami Bonnett-Admi,[5] Sarah Dill,[6] Sharon Dowell,[7] Elizabeth D. Ferucci,[8] Connie Herndon,[9] Bharat Kumar,[10] David Leverenz,[11] Jennifer Mandal,[12] Irene J. Tan,[13] Swamy Venuturupalli,[14] Tiffany Westrich-Robertson,[15] Marcy B. Bolster,[2] and Jason Kolfenbach,[16] for the American College of Rheumatology Workforce Solutions Committee

Objective. Patients and providers encounter challenges when conducting virtual musculoskeletal physical examinations (PEs) during rheumatology telehealth encounters. Guidance for a structured virtual PE could enhance the quality of clinical information gleaned and management decisions made during rheumatology telehealth visits. This study aims to build expert consensus and identify the most essential elements as the first step in defining the virtual rheumatology musculoskeletal PE.

Methods. A team with expertise in rheumatology telehealth, consisting of rheumatology attending physicians, educators, and a patient with rheumatic disease, conducted a modified Delphi to achieve consensus on the items determined to be most essential to the virtual rheumatology musculoskeletal PE. The modified Delphi consisted of two online surveys and a virtual meeting.

Results. The team identified seven items essential to the rheumatology musculoskeletal telehealth PE. These items describe elements in a focused joint examination as well as the assessment for level of activity of inflammatory arthritis. The modified Delphi method excluded maneuvers related to assessment of muscle strength and widespread pain syndromes, determining that these elements were better conducted in person.

Conclusion. A list of PE items most essential to rheumatology musculoskeletal telehealth encounters, supported by expert opinion and established evidence, marks the first step toward standardizing, evaluating, and teaching the virtual rheumatology PE. These items, alongside anticipated future revisions and improvements, promise to enhance the quality of telehealth care delivered to people with rheumatic diseases.

INTRODUCTION

Telehealth, defined as the use of telecommunications to deliver health care, rose in its implementation during the COVID-19 pandemic and remains in common practice for many patients and providers.[1] The use of telehealth may enhance the care delivered to people with rheumatic diseases as well as expand access to care, especially for patients living in remote

Supported by a Pfizer education grant (grant 76770947) and the Rheumatology Research Foundation Clinician Scholar Educator Award.

[1]Washington University in St Louis School of Medicine, St Louis, Missouri; [2]Massachusetts General Hospital, Boston; [3]Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire; [4]University of Alabama at Birmingham School of Medicine, Birmingham; [5]3GEM Consulting, Washington, DC; [6]Lakeview Rheumatology, Yakima, Washington; [7]NEA Baptist Memorial Hospital, Jonesboro, Arkansas; [8]Alaska Native Tribal Health Consortium, Anchorage; [9]American College of Rheumatology, Atlanta, Georgia; [10]Carver College of Medicine, University of Iowa, Iowa City; [11]Duke University School of Medicine, Durham, North Carolina; [12]University of California San Francisco School of Medicine, San Francisco; [13]Einstein Healthcare Network, Philadelphia, Pennsylvania; [14]Cedars-Sinai Medical Center, Los Angeles, California; [15]AiArthritis, Inc, St. Louis, Missouri; [16]University of Colorado Anschutz School of Medicine, Aurora.

Drs Zickuhr and Sobrero are co-first authors and contributed equally to this work. Drs Bolster and Kolfenbach are co-principal investigators and contributed equally to this work.

Additional supplementary information cited in this article can be found online in the Supporting Information section (https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.25669).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25669.

Address correspondence via email to Lisa Zickuhr, MD, MHPE, at zickuhrl@wustl.edu.

Submitted for publication July 11, 2025; accepted in revised form September 29, 2025.

SIGNIFICANCE & INNOVATIONS

- The virtual musculoskeletal physical examination presents challenges during rheumatology telehealth encounters.
- A defined list of musculoskeletal physical examination maneuvers essential to rheumatology telehealth encounters can enhance the delivery of rheumatology telehealth care.
- A team of rheumatology providers, educators, and a patient achieved consensus on the most important elements of the virtual rheumatology musculoskeletal physical examination.
- Using these results, future research can evaluate the sensitivity and specificity of physical examination elements within the context of the virtual encounter, the efficacy of educational interventions for trainees, and the value of provider and patient preparation for musculoskeletal physical examinations during rheumatology telehealth encounters.

communities during long-term follow-up.[2] The relatively novel practice of rheumatology telehealth warrants further investigation and evidence-based implementation to ensure people with rheumatic diseases benefit from high-quality telehealth practices.

The musculoskeletal physical examination (PE) is essential for diagnosing and monitoring rheumatic disease.[3] However, the virtual environment introduces unique conditions for conducting PEs, as it precludes providers from performing hands-on maneuvers. These factors potentially limit the clinical information obtained during telehealth visits.[4] The Rheumatology Telehealth Competencies and the Canadian Rheumatology Association's consensus statement on best practices for virtual care advise rheumatology providers to adapt the PE to the virtual environment and recommend in-person encounters when disease-specific factors limit the virtual PE.[1,5] These statements set standards for high-value telehealth care without instructing providers on ways to accomplish them. On the other hand, studies describe the implementation of and provide evidence for virtual PE techniques that assess disease activity in specific rheumatic diseases but do not describe a standardized approach to the general rheumatology musculoskeletal telehealth PE (RMTPE).[6,7] Without consensus on RMTPE practices, the rheumatology workforce lacks guidance for preparing patients for and implementing virtual PEs, limiting the care delivered during rheumatology telehealth encounters.

This study aims to generate consensus on the essential elements of the RMTPE. These elements will help guide future studies on the virtual PE, facilitate broad implementation into rheumatology practices, and provide structured training for rheumatology trainees and providers to enhance the quality of telehealth care delivered to people with rheumatic diseases.

MATERIALS AND METHODS

The essential elements of the RMTPE were determined using the same methods that established the essential components of successful rheumatology telehealth encounters, described in full detail by Zickuhr and colleagues.[1] The American College of Rheumatology (ACR) Workforce Solutions Committee invited 14 people with expertise in rheumatology telehealth, including attending physicians, educators, and a patient with rheumatic disease, to form a Working Group dedicated to defining the essential elements of the RMTPE (Appendix A). Working Group members came from academic and community practice as well as the Northeastern, Southern, Midwest, and Western regions of the United States. Study protocols, including waived consent for surveys and verbal consent for virtual meeting participation, complied with the Declaration of Helsinki and were approved by the Washington University in St. Louis School of Medicine institutional review board (#202304097).

Expert consensus on content. The Working Group focused on defining the RMTPE elements most specific to rheumatology, seeking to highlight the nuances of our field rather than duplicate maneuvers conducted during general telehealth examinations. To generate preliminary consensus, the Group conducted a modified Delphi consisting of two rounds of surveys. It then applied inductive reasoning and finalized consensus during a virtual meeting.

The first survey presented 15 potential skills required for the RMTPE based on literature review, spanning the following four domains: (1) conducting focused joint examinations, (2) determining the activity of established inflammatory arthritis, (3) measuring muscle strength, and (4) assessing widespread pain syndromes.[5,6,8–11] The survey was distributed using Qualtrics' online survey platform to Working Group members (DA, ASA, S. Dill, S. Dowell, EDF, BK, DL, JM, IJT, SW, and TW-R) who ranked the importance of each item for a successful RMTPE on a four-point Likert scale ("not at all essential," "moderately essential," "essential," and "highly essential"). As part of the first survey, Working Group members could suggest additional items they considered to be essential to the RMTPE for incorporation into the second survey.

Rankings of "essential" and "highly essential" counted as votes for inclusion, whereas rankings of "not at all essential" and "moderately essential" counted as votes for exclusion. A 70% consensus marked the threshold for inclusion or exclusion, in line with standards established in other modified Delphi studies.[1] During the virtual meeting, Working Group codirectors (JK and LZ) led the discussion, presenting the list of elements and seeking input from members (DA, CH, JM, SD, SD, and TW-R) on each item. Discussion emphasized items that did not reach consensus after the second survey and reviewed all items marked for inclusion, revising them as necessary. Discussion points were documented

for record-keeping, and the Working Group finalized the wording of elements during the virtual meeting.

RESULTS

The Working Group reached consensus to include 6 of the 15 original survey items and proposed one additional element during the virtual meeting (Supplementary Table S1). The RMTPE Elements ("The Elements") summarize the seven elements essential to the RMTPE (Table 1).

Focused, symptomatic joint examination. Researchers agreed that assessing swelling, tenderness, and range of motion of symptomatic joints are essential elements in the RMTPE. Working Group members modified the wording of this item, specifying that rheumatology providers should "guide patients to clarify the precise location of pain" rather than "assess joint(s) for tenderness via patient self-palpation." When combined with providers' virtual inspection of symptomatic joints to assess for swelling, this approach negates reliance on patients palpating precise anatomic landmarks and allows for more accurate localization of pain and identification of potential synovitis. The Working Group also identified "[adapting] provocative testing that translates to the telehealth platform for regional musculoskeletal conditions" as an essential component of the RMTPE. Examples of provocative testing include Tinel's and Phalen's signs for carpal tunnel syndrome or Spurling's test for cervical radiculopathy. The Working Group recognized that clinicians currently lack data on the efficacy, sensitivity, and specificity of these maneuvers when implemented during telehealth but have established guidance on how to perform them.[10,12]

Table 1. Rheumatology musculoskeletal telehealth physical examination elements

Category	Elements
When performing a focused joint examination	• Assess joint(s) for swelling by visual inspection or review of submitted images[a] • Guide patients to clarify precise location of pain • Assess range of motion of affected joints • Modify provocative examination maneuvers that transfer to telehealth
When assessing for inflammatory arthritis	• Assess joint(s) for swelling by visual inspection or review of submitted images[a] • Assess for range of motion as part of longitudinal disease management • Inspect for joint deformity
When assessing muscle strength	• Opt for in-person evaluation
When assessing for widespread pain syndromes	• Opt for in-person evaluation

[a] "Review of submitted images" is a form of asynchronous telehealth care delivery wherein patients capture photographs of affected joints and share them with rheumatology providers through US Health Insurance Portability and Accountability Act–secure platforms, such as those embedded within the electronic health record.

Assess activity of inflammatory arthritis. The Working Group agreed that comprehensively evaluating for joint swelling and inspecting for joint deformities are essential when caring for patients with known inflammatory arthritis. Inspection may be done synchronously during a video-teleconference telehealth encounter or asynchronously through patient-submitted photographs of affected joints. The Group intentionally phrased the item, "Assesses for range of motion as part of longitudinal disease management," to allow providers to use their clinical judgment to determine the need for assessment of range of motion in any given telehealth encounter. For instance, it may be routine to measure the range of motion in the hands of a person with rheumatoid arthritis at every visit. In contrast, a modified Schöber test for spine mobility may be performed at specific intervals for patients with axial spondyloarthritis. The Working Group excluded the item, "Incorporates relevant patient-reported outcomes for disease activity and functional assessment (eg, Routine Assessment of Patient Index Data 3 or Multidimensional Health Assessment Questionnaire)." Although this datum is often critical to the assessment of disease activity during telehealth encounters in general, it was not deemed to be a specific component of the RMTPE and instead is considered a component of history-taking with other objective data.[1] Although assessment for joint swelling, deformity, and range of motion was included in the final consensus elements, patient-reported tender joint counts were not. Methods to coach patients through self-assessment of swollen and tender joint counts have been described in the literature, but uncertainty regarding the accuracy of the examination and potential time-consuming nature prevented its ultimate inclusion during the modified Delphi process.[13]

Muscle strength and widespread pain syndromes. The Working Group excluded muscle strength items from the RMTPE, as patients with myositis and persistent or new weakness were deemed to require an in-person evaluation. They agreed that the nuance of the strength examination, such as judging subtle changes in strength or distinguishing the effects of pain from weakness, imposed challenges to testing strength accurately during an RMTPE. The Working Group recommended rheumatology providers conduct in-person strength testing when accurate examination data were needed (eg, in patients with new-onset or inadequately controlled disease).

Similarly, the Working Group excluded items related to the evaluation of widespread pain syndromes, such as palpating myofascial tender points, collecting a psychosocial history, and using the 2010 ACR Fibromyalgia Diagnostic Questionnaire.

Members recognized that palpating myofascial tender points requires applying pressure to muscular areas rather than articular surfaces, and patients likely would not be able to precisely conduct this PE element, supporting that individuals with widespread pain would benefit from in-person examinations. Additionally, collecting a psychosocial history and administering the 2010 ACR Fibromyalgia Diagnostic Questionnaire represented virtual history- and data-gathering rather than PE elements.[1]

DISCUSSION

Telehealth's popularity has remained high in recent years, making it an important avenue of care delivery for many people with rheumatic diseases. Although the evidence supporting telehealth best practices is increasing, a consensus on adaptation of the rheumatology musculoskeletal PE to the virtual context has yet to be established. In this study, a diverse team of rheumatology providers, educators, and a patient experienced in telehealth developed the first published RMTPE Elements.

Our description of the most essential elements of RMTPEs meets a shortcoming in rheumatology telehealth guidance. Together, telehealth and in-person care support the health outcomes of people with rheumatic diseases. It is essential that rheumatology providers competently identify scenarios, including adapting to the nuances of the RMTPE, when either telehealth or in-person encounters will best serve the patient. Thus far, formal statements regarding rheumatology telehealth and research optimizing virtual assessments of disease activity have directed providers' RMTPEs.[2,5,6,7] The Elements enhance this guidance and outline the key components of the RMTPE, presenting seven consensus-based elements providers can routinely apply, with or without additional disease-specific maneuvers, to deliver high-quality telehealth care.

The Elements will facilitate future research in RMTPE. Most published data on the RMTPE have measured the accuracy of virtual PE maneuvers for calculating disease activity scores.[6,7] Applying methods for determining sensitivity and specificity, researchers can now focus on gathering validity evidence for the short list of specified maneuvers in The Elements that contribute to the foundation of rheumatology telehealth care delivery.

Similarly, The Elements will enhance future educational initiatives. To date, educators have been creating instructional materials with content meeting local and regional needs.[14] The Elements, grounded in the expertise of the national Working Group, specify the virtual examination maneuvers in which learners must achieve competence and outline the RMTPE most essential for inclusion in future educational initiatives. Providers and fellows-in-training can focus on skill development to achieve competence in delivering rheumatology telehealth care, including the fundamental aspects of the RMTPE.

The rigor of our methods highlights the strength of The Elements. First, we applied a modified Delphi approach to systematically define the most essential RMTPE maneuvers, combining medical literature with expert consensus. Second, the Working Group represented the expertise of rheumatology providers and medical educators and included patient experience in telehealth care from across the United States. This diversity enhances the trustworthiness and patient-centeredness of our results.

Limitations to our study include the paucity of data describing the sensitivity, specificity, and accuracy of many PE maneuvers in the virtual space. Therefore, the Working Group was unable to consider these diagnostic characteristics for all elements when building consensus. Members incorporated established data when possible, relied on published descriptions for adapting in-person musculoskeletal PE maneuvers to rheumatology telehealth, and supplemented with clinical experience when needed.[6–12] We hope The Elements inspire researchers, as the fields of dermatology and neurology have done for their specialties, to measure these diagnostic characteristics for the fundamental RMTPE defined in our work.[9,15] This context limits the application of The Elements to suggested guidance that lacks the requisite evidence for informing policies, including telehealth reimbursement frameworks. In addition, having only a single patient representative on the Working Group may reflect a limitation. Although the patient representative is a strong advocate who contributed their individual experience alongside reflections from the community they represent, additional input could further strengthen the patient-centeredness of our results.

As we consider ways to use telehealth to enhance the care delivery we provide to our patients, The Elements are an important starting point for improving rheumatology telehealth encounters. Providers can immediately incorporate these fundamental elements into their practice to enhance the quality of RMTPEs. Medical educators can design instructional and assessment interventions based on essential RMTPE maneuvers, and patient educators can create materials to prepare people with rheumatic diseases to participate in virtual PE. Lastly, researchers can start studying the validity of the RMTPE compared with in-person musculoskeletal PE. Ultimately, each of these steps stemming from The Elements will enhance the quality of telehealth care delivered to people with rheumatic diseases.

AUTHOR CONTRIBUTIONS

All authors contributed to at least one of the following manuscript preparation roles: conceptualization AND/OR methodology, software, investigation, formal analysis, data curation, visualization, and validation AND drafting or reviewing/editing the final draft. As corresponding author, Dr Zickuhr confirms that all authors have provided the final approval of the version to be published and takes responsibility for the affirmations regarding article submission (eg, not under consideration by another journal), the integrity of the data presented, and the statements regarding compliance with institutional review board/Declaration of Helsinki requirements.

ROLE OF THE STUDY SPONSOR

Pfizer had no role in the study design or in the collection, analysis, or interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication. Publication of this article was not contingent on approval by Pfizer.

REFERENCES

1. Zickuhr L, Sobrero A, Albert D, et al; American College of Rheumatology Workforce Solutions Committee. Development, usability, and validity evidence of a rheumatology telehealth feedback form. Arthritis Care Res (Hoboken) 2025;77(10):1246–1253.
2. de Thurah A, Bosch P, Marques A, et al. 2022 EULAR points to consider for remote care in rheumatic and musculoskeletal diseases. Ann Rheum Dis 2022;81(8):1065–1071.
3. Kolfenbach JR. History and physical examination. In: West S, Kolfenbach JR, eds. Rheumatology Secrets. 4th ed. Elsevier Health Sciences; 2021:42–50.
4. Avouac J, Marotte H, Balsa A, et al. Teleconsultation in rheumatology: a literature review and opinion paper. Semin Arthritis Rheum 2023;63:152271.
5. Barber CEH, Levy DM, Ahluwalia V, et al. Best practices for virtual care: a consensus statement from the Canadian Rheumatology Association. J Rheumatol 2022;49(4):408–418.
6. England BR, Barber CEH, Bergman M, et al. Adaptation of American College of Rheumatology rheumatoid arthritis disease activity and functional status measures for telehealth visits. Arthritis Care Res (Hoboken) 2021;73(12):1809–1814.
7. Askanase AD, Aranow C, Kim MY, et al. Protocol for virtual physical examination in an observational, longitudinal study evaluating virtual outcome measures in SLE. Lupus Sci Med 2023;10(2):e000952.
8. Yao P, Adam M, Clark S, et al. A scoping review of the unassisted physical exam conducted over synchronous audio-video telemedicine. Syst Rev 2022;11(1):219.
9. Lu AD, Veet CA, Aljundi O, et al. A systematic review of physical examination components adapted for telemedicine. Telemed J e-Health 2022;28(12):1764–1785.
10. Laskowski ER, Johnson SE, Shelerud RA, et al. The telemedicine musculoskeletal examination. Mayo Clin Proc 2020;95(8):1715–1731.
11. Wouters F, van der Giesen FJ, Matthijssen XME, et al. Difficulties making a fist in clinically suspect arthralgia: an easy applicable phenomenon predictive for RA that is related to flexor tenosynovitis. Ann Rheum Dis 2019;78(10):1438–1439.
12. Wainberg MC, Jurisson ML, Johnson SE, et al. The telemedicine hand examination. Am J Phys Med Rehabil 2020;99(10):883.
13. Grainger R, Townsley HR, Stebbings S, et al. Codevelopment of patient self-examination methods and joint count reporting for rheumatoid arthritis. ACR Open Rheumatol 2020;2(12):705–709.
14. Bolster MB, Kolfenbach J, Poeschla A, et al. Incorporating telemedicine in rheumatology fellowship training programs: needs assessment, curricular intervention, and evaluation. Arthritis Care Res (Hoboken) 2023;75(12):2428–2434.
15. O'Connor DM, Jew OS, Perman MJ, et al. Diagnostic accuracy of pediatric teledermatology using parent-submitted photographs: a randomized clinical trial. JAMA Dermatol 2017;153(12):1243–1248.

APPENDIX A: AMERICAN COLLEGE OF RHEUMATOLOGY WORKFORCE SOLUTIONS COMMITTEE AUTHORS

Members of the American College of Rheumatology Workforce Solutions Committee are as follows: Lisa Zickuhr, Daniel Albert, Amanda S. Alexander, Tami Bonnett-Admi, Sarah Dill, Sharon Dowell, Elizabeth D. Ferucci, Connie Herndon, Bharat Kumar, David Leverenz, Jennifer Mandal, Irene J Tan, Swamy Venuturupalli, Tiffany Westrich-Robertson, Marcy B. Bolster, Jason Kolfenbach.

American College of Rheumatology
Empowering Rheumatology Professionals

Arthritis Care & Research
Vol. 78, No. 5, May 2026, pp 688–690

LETTER

DOI 10.1002/acr.25678

Knee crepitus after anterior cruciate ligament reconstruction: not predictive, but still pragmatic: comment on the article by Couch et al

To the Editor:

The recent study by Couch et al, titled "Knee crepitus and osteoarthritis features in young adults following traumatic knee injury,"[1] offers valuable insight into the prognostic relevance of self-reported crepitus following anterior cruciate ligament reconstruction (ACLR). The authors report that although crepitus at one year is associated with full-thickness patellofemoral cartilage defects, it does not predict worsening of magnetic resonance imaging (MRI)-defined osteoarthritis (OA) features over the next five years. From the perspective of an orthopedic clinician, this is a timely and practically relevant finding that can help refine postoperative assessment protocols. However, based on my clinical experience, I believe that further nuance in the interpretation and management of crepitus could strengthen the translational value of this work.

First, the dichotomous definition of crepitus used in the study may oversimplify a heterogeneous clinical sign. In practice, what patients describe as "grinding" or "clicking" can originate from diverse anatomic sources—patellofemoral incongruence, synovial plicae, tendinous snapping, or residual chondral fraying—and the implications differ substantially depending on whether the sound is painful, consistent under load, or mechanically reproducible.[2] For more granular phenotyping, I suggest a two-step bedside assessment: (1) actively reproduce the sound during both unloaded knee extension and loaded squatting, and (2) categorize it as "painless-fine," "painful-coarse," or "nonarticular" based on palpation and provocation. This simple stratification could help clinicians better decide which patients warrant imaging or focused rehabilitation while avoiding unnecessary concern for those with benign, painless crepitus.

Second, the study's patellofemoral-centric findings support the use of targeted, compartment-specific rehabilitation pathways. In patients with painful, load-linked crepitus, I advocate an initial six- to eight-week protocol focused on patellofemoral unloading strategies—such as high-repetition closed-chain quadriceps exercises, hip abduction and rotation strengthening, cadence modification during ambulation, and symptom-guided use of patellar taping. These interventions have demonstrated value in patellofemoral pain syndromes and may plausibly benefit this subgroup post-ACLR.[3,4] If symptoms persist despite high adherence, follow-up MRI focused on patellofemoral cartilage integrity can be considered, ensuring that imaging is guided by functional nonresponse rather than sound alone. This tailored approach aligns with the anatomic specificity reported by Couch et al and allows clinicians to deploy resources more judiciously.

Third, I believe there is an opportunity to enhance the clinical translation of this research by establishing a pragmatic, minimal-data registry for patients with persistent crepitus post-ACLR. Such a registry could include standardized documentation of crepitus phenotype, pain linkage, basic strength symmetry (eg, hand-held dynamometry), and Knee Injury and Osteoarthritis Outcome Score subscale scores. Optional audio recordings of the crepitus during a standardized squat could further facilitate research into acoustic phenotyping. Scheduled follow-up at 8 to 12 weeks would enable outcome tracking and assist in identifying patients who may require imaging or surgical evaluation. This real-world approach would complement the findings of the KOALA cohort by enriching longitudinal data on symptom persistence and functional outcomes.

In summary, although the presence of knee crepitus following ACLR may not independently predict structural OA progression, it should not be dismissed as irrelevant. When linked to patellofemoral pain under load, crepitus can serve as a pragmatic clinical cue to initiate compartment-specific rehabilitation and monitor treatment response. By refining the phenotypic characterization of crepitus and linking escalation of care to objective functional outcomes rather than the presence of sound alone, orthopedic teams can provide reassurance without complacency and caution without overtreatment.

Supported by Yunnan Fundamental Research Kunming Medical University Projects (grant 2021AY070001-026).

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25678.

Shuai Chen, PhD
shuaichen982@outlook.com
The Affiliated Yan'an Hospital of Kunming Medical University
Kunming, Yunnan Province, China

1. Couch JL, Patterson BE, Crossley KM, et al. Knee crepitus and osteoarthritis features in young adults following traumatic knee injury. Arthritis Care Res 2026;78(3):398–406.
2. Codorean II, Codorean IB. Condition causing anterior knee pain. In: Codorean II, Codorean IB, eds. Clinical-MRI Correlations of Anterior Knee Pain: Common and Uncommon Causes. Springer Cham; 2023:21–57.
3. Willy RW, Hoglund LT, Barton CJ, et al. Patellofemoral pain. J Orthop Sports Phys Ther 2019;49(9):CPG1–CPG95.
4. Hasan S. The effect of patellar taping combined with isometric strength training on pain, muscle strength, and functional performance in patients with patellofemoral pain syndrome: a randomized comparative study. PeerJ 2025;13:e19381.

DOI 10.1002/acr.25681

Reply

To the Editor:

We thank Dr Chen for their thoughtful comments on our article evaluating knee crepitus and osteoarthritis features in young adults following traumatic knee injury.[1] We are pleased that our findings resonate with the clinical community and agree that continued discussion about the implications of crepitus following anterior cruciate ligament reconstruction (ACLR) is warranted.

We acknowledge the heterogenous nature of knee crepitus and that the underlying mechanism(s) or cause(s) of this clinical sign are largely unknown. This is highlighted in our recent systematic review evaluating the prevalence of knee crepitus, wherein most studies (76%) failed to provide an operational definition of this common clinical sign.[2] Despite this, knee crepitus is typically defined as a continuous audible crackling or grinding noise during knee joint movement,[3] highlighting that the isolated clicks and pops inherent to other anatomic sources (eg, tendinous snapping) are a distinct clinical sign.

In our investigation, self-reported crepitus of the index ACLR knee was assessed using item S2 of the Knee Injury and Osteoarthritis Outcome Score symptoms subscale (KOOS-S2; "Do you feel grinding, hear clicking or any other type of noise when your knee moves?") at one-year post-ACLR. Based on the phrasing of this item (specifically, "hear… any other type of noise when your knee moves"), it is possible that mechanical symptoms beyond crepitus may be captured. As per the format of the KOOS questionnaire, participants respond on an ordinal scale (ie, 0 = never, 1 = rarely, 2 = sometimes, 3 = often, 4 = always) in relation to symptoms experienced in the past week. By dichotomizing this scale and defining the presence of knee crepitus as a response of at least 3 (ie, often or always), we aimed to avoid capturing the isolated mechanical signs that would not be considered knee crepitus. It should also be noted that our previous evaluation of the diagnostic accuracy (and comparison to physical examination) of this item determined it appropriate for the detection of knee crepitus following ACLR.[3] Self-reported evaluation of knee crepitus has been implemented across several clinical populations,[4,5] with KOOS-S2 previously used to investigate the implications of knee crepitus in those at risk of incident symptomatic osteoarthritis within the multicenter longitudinal Osteoarthritis Initiative cohort.[6]

To our knowledge, there is no evidence supporting the notion that the specific characteristics of knee crepitus (or the methods from which it is reproduced) offer greater diagnostic or prognostic value, bringing into question the value of phenotyping this clinical sign. This is supported by our recent meta-analysis demonstrating that knee crepitus is common, irrespective of knee pain status or injury history.[2] It should also be noted that despite our current investigation being conducted within a symptomatic cohort, nearly all included participants experienced nonpainful crepitus (and thus considered "benign" under the proposed stratification) despite the association with structural and symptomatic outcomes. Previous attempts to phenotype the acoustic characteristics of knee crepitus using techniques, such as vibroarthrography, have yielded limited clinical utility and may substantially inflate the prevalence of this clinical sign (eg, recorded in 99% of "normal" knees).[7] Although a structured phenotyping approach integrating patient-reported symptoms and acoustic measures of evaluation may provide more granular insights, the existing literature is currently not in a place to support this approach.

Although the clinical opinion described by Dr Chen provides an interesting account of managing patients with knee crepitus, there is no peer-reviewed evidence to support dedicated rehabilitation strategies targeting knee crepitus or the stratification of care based on this clinical sign. Whether specific interventions can modify crepitus is yet to be determined, with much of the information available online lacking empirical support.[8] It is known that patients often report catastrophizing beliefs surrounding their knee crepitus.[9] On the weight of available evidence, patients presenting with knee crepitus should be counseled that despite the association with full-thickness patellofemoral cartilage defects, the presence of knee crepitus may not be a prognostic indicator for long-term knee joint health.

The proposal for a pragmatic registry of patients post-ACLR with persistent crepitus is an admirable one. Although systematic collection of phenotypic, symptom, and functional data could enhance understanding of prognosis and refine clinical pathways, given the labor-intensive requirements of establishing and maintaining a clinical data registry, greater investigations of the relevance of knee crepitus are likely warranted before such efforts can be justified. It is important to note that many existing ACLR registries (eg, Danish, Swedish, Norwegian ligament registries) already assess knee crepitus subjectively using KOOS-S2 described above. These registries may prove to be a critical source of untapped data to better understand the trajectory and prognosis of those with knee crepitus following ACLR and guide future research.

Author disclosures are available at https://onlinelibrary.wiley.com/doi/10.1002/acr.25681.

Jamon L. Couch, BHlthSc, MPhysioPrac, MExSci(S&C)
La Trobe Sport and Exercise Medicine Research Centre, School of Allied Health, Human Services and Sport and Australian International Olympic Committee Research Centre, La Trobe University, Melbourne, Victoria, Australia
and Arthritis Research Canada
Vancouver, British Columbia, Canada
Brooke E. Patterson, BPhysio (Hons), PhD
Kay M. Crossley, BAppSc (Physio), PhD
Danilo De Oliveira Silva, PhD
Michael A. Girdwood, BHlthSc, MPhysioPrac, PhD
Adam G. Culvenor, BPhysio (Hons), PhD
a.culvenor@latrobe.edu.au
La Trobe Sport and Exercise Medicine Research Centre, School of Allied Health, Human Services and Sport and Australian International

Olympic Committee Research Centre, La Trobe University
Melbourne, Victoria, Australia
Ali Guermazi, MD, PhD, MSc
Department of Radiology, Boston University School of Medicine
Boston, MA
Matthew G. King, BPhysio (Hons), PhD
Discipline of Physiotherapy, School of Allied Health, Human Services and Sport, La Trobe University
Melbourne, Victoria, Australia
Jackie L. Whittaker, PhD
Arthritis Research Canada
and Department of Physical Therapy, Faculty of Medicine, University of British Columbia
Vancouver, British Columbia, Canada

1. Couch JL, Patterson BE, Crossley KM, et al. Knee crepitus and osteoarthritis features in young adults following traumatic knee injury. Arthritis Care Res 2026;78(3):398–406.
2. Couch JL, King MG, De Oliveira Silva D, et al. Noisy knees - knee crepitus prevalence and association with structural pathology: a systematic review and meta-analysis. Br J Sports Med 2025;59:126–132.
3. Couch JL, King MG, De Oliveira Silva D, et al. Diagnostic performance of self-reported knee crepitus using a Knee Injury and Osteoarthritis Outcome Score item. Osteoarthritis Cartilage 2025;33:764–769.
4. Leal-Blanquet J, Alentorn-Geli E, Tuneu J, et al. Anterior cruciate ligament reconstruction: a multicenter prospective cohort study evaluating 3 different grafts using same bone drilling method. Clin J Sport Med 2011;21:294–300.
5. Tudisco C, Meo A, Blasucci C, et al. Arthroscopic treatment of lateral meniscal cysts using an outside-in technique. Am J Sports Med 2000;28:683–686.
6. Lo GH, Strayhorn MT, Driban JB, et al. Subjective crepitus as a risk factor for incident symptomatic knee osteoarthritis: data from the Osteoarthritis Initiative. Arthritis Care Res (Hoboken) 2018;70:53–60.
7. McCoy GF, McCrea JD, Beverland DE, et al. Vibration arthrography as a diagnostic aid in diseases of the knee. A preliminary report. J Bone Joint Surg Br 1987;69(2):288–293.
8. Serighelli F, Gomes DA, Pazzinatto MF, et al. "My knee is cracking" – what information is available on the internet about it? A systematic appraisal of the credibility, readability and accuracy of online information about knee crepitus. JSAMS Plus 2023;2:100041.
9. Robertson CJ, Hurley M, Jones F. People's beliefs about the meaning of crepitus in patellofemoral pain and the impact of these beliefs on their behaviour: a qualitative study. Musculoskelet Sci Pract 2017;28:59–64.