ACR Open Rheumatology
Vol. 4, No. 10, October 2022, pp 897–902
DOI 10.1002/acr2.11484

AMERICAN COLLEGE of RHEUMATOLOGY
Empowering Rheumatology Professionals

BRIEF REPORT

Comparative Efficacy Randomized Controlled Trials in Rheumatology Guidelines

Katie Henry,[1] Desh Nepal,[1] Erin Valley,[1] Connor Pedersen,[1] Alí Duarte-García,[2] and Michael Putman[1]

Background. Comparative efficacy randomized controlled trials (RCTs) compare two active interventions in a head-to-head design. They are useful for informing clinical practice guidelines, but the degree to which such trials inform clinical practice guidelines in rheumatology is unknown.

Methods. The American College of Rheumatology (ACR) and European Alliance of Associations for Rheumatology (EULAR) websites were searched from January 1, 2017, to June 12, 2021, for clinical practice guidelines. RCTs referenced by each guideline were identified, and information regarding design and outcomes were extracted. Clinical practice recommendations from each guideline were also analyzed.

Results. Fifteen ACR- and nine EULAR-endorsed guidelines were included, which cited 609 RCTs and provided 481 recommendations. Referenced RCTs enrolled an average of 418 patients (SD 985), most commonly evaluated biologic/targeted synthetic disease-modifying antirheumatic drugs (70.1%), and infrequently used a head-to-head design (28%). A minority of recommendations received a high level of evidence (LOE) by the Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) methodology (2.9%) or an "A" grade by the Oxford Centre for Evidence based Medicine Standards (OCEBM) methodology (28.9%). LOE was higher for recommendations informed by RCTs ($P < 0.001$) or head-to-head RCTs ($P = 0.008$). Many recommendations received a strong recommendation despite low (8 [2.6%]) or very low (25 [8.3%]) LOE.

Conclusion. Less than one in six rheumatology guideline recommendations are informed by head-to-head RCTs. Recommendations that were informed by head-to-head RCTs were more likely to have a high LOE by both GRADE and OCEBM. Efforts to introduce more comparative efficacy RCTs should be undertaken.

INTRODUCTION

Randomized controlled trials (RCTs) are considered the "gold standard" for evaluating the efficacy of novel interventions (1). As opposed to placebo-controlled RCTs, comparative efficacy RCTs evaluate active therapies in a head-to-head design. Because they provide high-quality data about the relative efficacy and safety of two or more medical interventions, they are especially useful for informing clinical practice (2). Comparative efficacy RCTs also require large sample sizes to be adequately powered and are not required for Food and Drug Administration (FDA) approval (3). For these reasons, placebo-controlled trials appear to be more frequently performed, both in general medicine (4) and in the field of rheumatology (3). The degree to which pragmatic RCTs have informed rheumatology practice guidelines or impacted clinical practice is unknown.

Clinical practice guidelines, which combine evidence-based synthesis and consensus-based expert opinion to form recommendations about therapeutic interventions, provide an ideal surrogate for assessing the impact of comparative efficacy RCTs.

This work was supported by the Medical College of Wisconsin Medical Student Summer Research Program (MSSRP) and the Center for Immunology Summer Research Program Fellowship (Katie Henry); by the Medical College of Wisconsin MSSRP and the National Institute on Aging Training Grant T35AG029793 (Meurer) (Erin Valley); by the Medical College of Wisconsin MSSRP (to Connor Pedersen); by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, the Rheumatology Research Foundation Scientist Development Grant, and the Centers for Disease Control and Prevention (Ali Duarte); and by a Rheumatology Research Foundation Scientist Development Grant (Michael Putman).

[1]Katie Henry, BS, Desh Nepal, MD, Erin Valley, BA, Connor Pedersen, BA, Michael Putman, MD, MSci: Medical College of Wisconsin, Milwaukee; [2]Alí Duarte-García, MD, MSc: Mayo Clinic, Rochester, Minnesota.

Data will be made available upon reasonable request.

Author disclosures are available at https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Facr2.11484&file=acr211484-sup-0001-Disclosureform.pdf.

Address correspondence to Michael Putman, MD, MSci, Medical College of Wisconsin Hub for Collaborative Medicine, 8701 Watertown Plank Road, Rheumatology 8th Floor, Milwaukee, WI 53226. Email: mputman@mcw.edu.

Submitted for publication March 18, 2022; accepted in revised form June 10, 2022.

High-quality clinical practice guidelines in rheumatology are often generated by the American College of Rheumatology (ACR) and European Alliance of Associations for Rheumatology (EULAR). Despite extensive literature reviews, over half of ACR recommendations are based on low levels of evidence (LOEs) ("C") (5). The objective of this project was to characterize the degree to which rheumatology clinical practice guideline recommendations are supported by comparative efficacy RCTs. Associations between the presence of comparative efficacy RCTs and the strength and LOE were assessed.

METHODS

Search strategy. The ACR and EULAR websites (https://www.rheumatology.org/Practice-Quality/Clinical-Support/Clinical-Practice-Guidelines and https://www.eular.org/recommendations_management.cfm, respectively), were searched on June 12, 2021, for ACR- or EULAR-endorsed clinical practice guidelines. Guidelines were included if they were published between January 1, 2017, and June 12, 2021. Guidelines were excluded if a more recent version was published or if they did not use either the Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) (6) system for ACR guidelines or Oxford Centre for Evidence based Medicine Standards (OCEBM) (7) system for EULAR guidelines to rate recommendation LOE. To identify guidelines not posted to the ACR or EULAR websites, we also performed a literature search of the NCBI database PubMed using the following terms: ((EULAR OR "European Alliance of Associations for Rheumatology" OR ACR OR "American College of Rheumatology") AND "Practice Guideline"[pt]).

Guideline references. The following guideline characteristics were extracted by one reviewer (KH): PubMed Identifier, title, disease state, endorsement, and publication year. For each guideline, all cited references in the guideline and supplementary materials were searched for RCTs. If a systematic literature review was performed to inform the clinical practice guideline, references from the associated literature review were also searched. For references defined as an RCT, the following information was extracted: year, pharmaceutical industry funding, number of participants, study arm(s), primary outcome measure $P < 0.05$, use of head-to-head comparative evidence, use of non-inferiority or superiority trial design, use of single or multiple centers, and blinding.

We classified an RCT as head-to-head if it included at least two distinct active treatments that were compared against each other. Study arms were categorized as conventional synthetic disease-modifying antirheumatic drug (csDMARD) (i.e., methotrexate, mycophenolate, hydroxychloroquine), biologic and targeted synthetic DMARDs (b/tsDMARD) (i.e., adalimumab, secukinumab, tofacitinib), nonsteroidal anti-inflammatory drug (NSAID), steroid, non-pharmacologic, urate lowering therapy (ULT), placebo, or other if the intervention did not fit within an aforementioned category. Study designs were defined by b/tsDMARD, csDMARD, NSAID, ULT, and Other versus placebo or head-to-head.

Guideline recommendations. The recommendations from each guideline were extracted by one reviewer (KH). A recommendation was counted if it had a corresponding LOE. Points to consider and overarching principles were not extracted. The following information was extracted directly from the recommendation: LOE, strength/grade of recommendation, and whether the recommendations involved a pharmaceutical treatment (further classified by csDMARD, b/tsDMARD, NSAID, ULT, steroid, or other). Two reviewers (KH, DN) working independently identified whether recommendations were informed by any RCT or by any head-to-head RCT using the following methodology: First, direct references within the recommendation rationale were analyzed and cross-referenced with data extracted from references. Next, all corresponding Patient/Population, Intervention, Comparison, and Outcomes questions within the supplementary materials were analyzed for any RCT or any head-to-head RCT. Last, any additional relevant supplementary materials were evaluated for recommendations informed by any RCT or any head-to-head RCT. Disagreements were resolved by consensus and adjudicated by a third reviewer (EV) as necessary.

Statistical analysis. Descriptive statistics were used to describe the extraction variables. Categorical variables were compared using a Fisher exact test. All *P* values were two-sided and considered statistically significant if less than 0.05 with no adjustment for multiple comparisons. All analyses were performed on R version 4.04.

RESULTS

We identified 15 ACR (n = 6) and EULAR (n = 9) endorsed guidelines published between January 1, 2017, and June 12, 2021. The guidelines assessed 10 different diseases (Supplementary Table 1) and provided 481 recommendations. ACR guidelines had an average of 51.3 (SD 26.3) recommendations, and EULAR guidelines had an average of 19.4 (SD 15.4) recommendations. The majority of recommendations involved a pharmaceutical agent (78.2%), roughly half were informed by at least one RCT (46.2%), and a minority were informed by at least one head-to-head RCT (15.0%). Diseases with the largest proportion of recommendations informed by any head-to-head RCT included gout (34.0%), rheumatoid arthritis (32.8%), and juvenile idiopathic arthritis (19.6%), whereas recommendations for vasculitis and Sjogren's syndrome guidelines had no recommendations informed by any head-to-head RCTs (Figure 1A).

The included guidelines provided 682 references to RCTs, 609 of which were unique. Referenced RCTs enrolled an average of 413 patients (SD 940) and most commonly evaluated

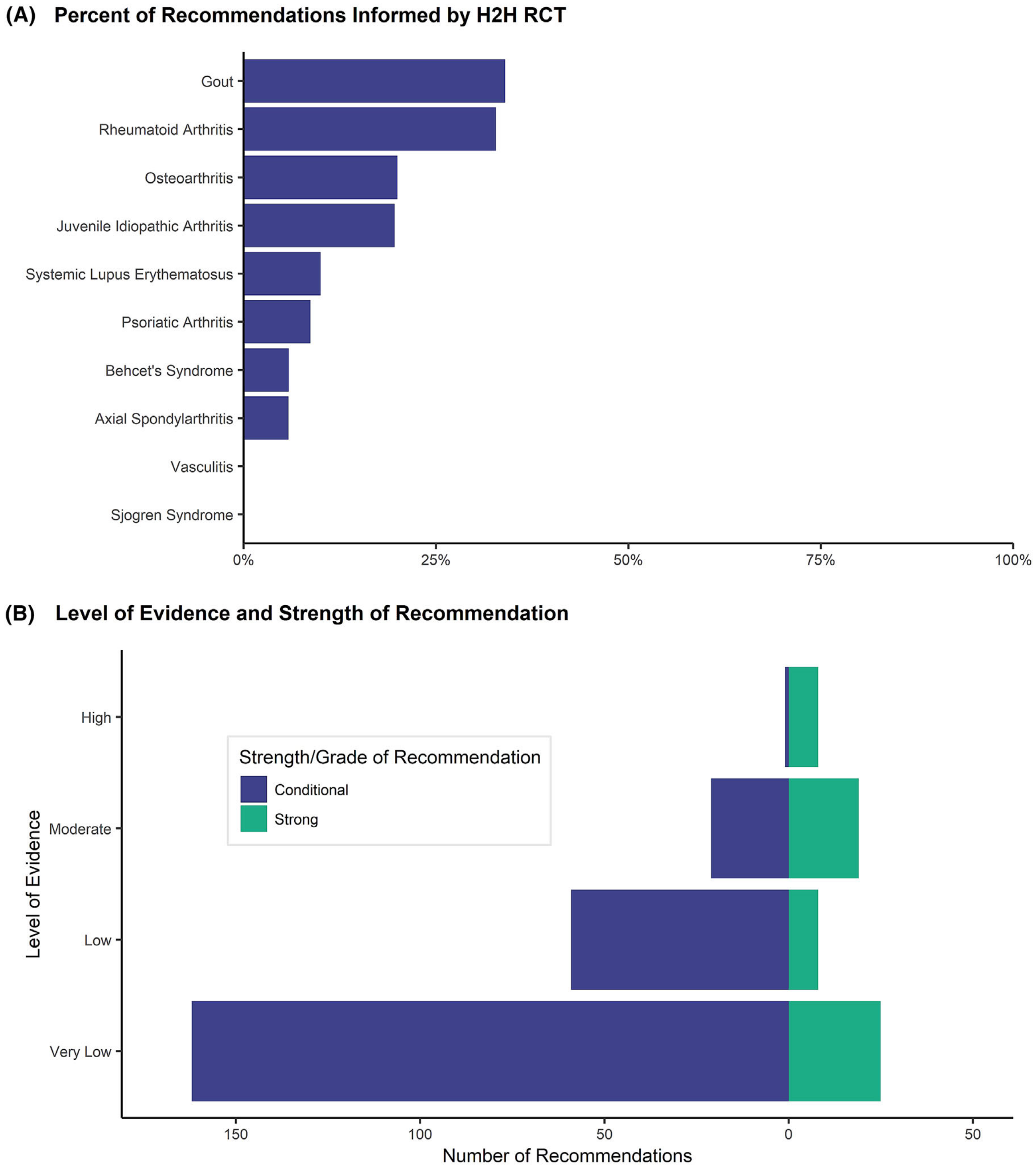

Figure 1. (**A**) Percent of recommendations informed by head-to-head RCTs for each disease state's clinical practice guideline(s). (**B**) Level of evidence and corresponding strength of recommendation for GRADE rated recommendations.

b/tsDMARDs (70.1%), used a double blind design (79.6%), and were funded by the pharmaceutical industry (66.3%). Less than a third (28%) of referenced RCTs used a head-to-head design. As compared with placebo-controlled trials, head-to-head trials were more likely to be non-inferiority (15.9% vs. 3.7%, $P < 0.001$) and open label (23.4% vs. 10.5%, $P < 0.001$) and less likely to evaluate b/tsDMARDs (62.5% vs. 55.0%, $P < 0.05$). The most common trial design was b/tsDMARD versus placebo (35.1%) followed by b/tsDMARD versus b/tsDMARD (14.4%) and b/tsDMARD versus csDMARD (6.1%) (Supplementary Table 2). There were no significant differences with respect to decade of study and the proportion of trials that used a head-to-head design ($P = 0.056$) (Table 1).

Among EULAR-endorsed recommendations, which used the OCEBM grading system, 28.9% received an "A" grade, 22.0% received a "B" grade, 33.5% received a "C" grade, and 15.6% received a "D" grade. Among ACR-endorsed recommendations, which used the GRADE system, 2.9% recommendations had a high LOE, 13.6% a moderate LOE, 21.8% a low LOE, and 61.7% a very low LOE. The recommendations LOEs were

Table 1. RCT references informing ACR and EULAR clinical practice guidelines published in 2017–2021, n = 609

Characteristic		Overall N (%)	Not head to head n (%)	Head to head n (%)	*P* value
Intervention	b/tsDMARD	367 (60.3%)	273 (62.5%)	94 (55.0%)	<0.05
	csDMARD	77 (12.6%)	46 (10.5%)	31 (18.1%)	
	ULT	43 (7.1%)	32 (7.3%)	10 (5.8%)	
	Steroid	22 (3.6%)	11 (2.5%)	11 (6.4%)	
	NSAID	21 (3.4%)	12 (2.7%)	9 (5.3%)	
	Other	79 (13.0%)	63 (14.4%)	16 (9.4%)	
Blinding	Double	452 (79.6%)	345 (84.4%)	106 (67.1%)	<0.001
	Single	36 (6.3%)	21 (5.1%)	15 (9.5%)	
	Open	80 (14.1%)	43 (10.5%)	37 (23.4%)	
Multicenter vs. single center	Multicenter	490 (84.8%)	356 (85.4%)	134 (83.2%)	0.52
	Single center	88 (15.2%)	61 (14.6%)	27 (16.8%)	
Superiority vs. non inferiority	Superiority	526 (88.3%)	390 (90.3%)	136 (82.9%)	<0.001
	Non-inferiority	42 (7.0%)	16 (3.7%)	26 (15.9%)	
	Both	1 (0.2%)	0	1 (0.6%)	
	Equivalence	27 (4.5%)	26 (6.0%)	1 (0.6%)	
Pharmaceutical funding	Yes	403 (66.3%)	295 (67.7%)	108 (63.2%)	0.56
	No	153 (25.2%)	106 (24.3%)	47 (27.5%)	
	Unknown	52 (8.6%)	35 (8.0%)	16 (9.4%)	
Primary outcome measure significant	Yes	462 (76.7%)	333 (77.1%)	129 (75.9%)	0.35
	No	139 (23.1%)	99 (22.9%)	40 (23.5%)	
	Unknown	1 (0.2%)	0	1 (0.6%)	

Abbreviations: b/tsDMARD, biological and targeted synthetic disease-modifying antirheumatic drug; csDMARD, conventional synthetic disease-modifying antirheumatic drug; NSAID, nonsteroidal anti-inflammatory drug; ULT, urate lowering therapy.

significantly higher for recommendations that were informed by any RCT (*P* < 0.001 for OCEBM and *P* < 0.001 for GRADE) or by head-to-head RCTs (*P* = 0.008 for OCEBM and *P* < 0.001 for GRADE) (Figure 2, Supplementary Table 3). With regard to the strength of recommendation as assessed by the ACR-endorsed GRADE recommendations, 80.0% were conditional and 20.0% were strong. The strength of recommendation was more likely to be conditional if the LOE was very low (53.5% vs. 8.3%, *P* < 0.001) or low (19.5% vs. 2.6%, *P* < 0.001). However, many recommendations still received a strong recommendation despite low (8 [2.6%]) or very low (25 [8.3%]) LOE (Figure 1B).

DISCUSSION

Less than one in six of the recommendations from ACR and EULAR clinical practice guidelines were informed by any head-to-head RCT. Recommendations that were informed by head-to-head RCTs were more likely to be "strong" recommendations by the GRADE approach and were more likely to have a high LOE by both GRADE and OCEBM. Similar to other recent investigations, the majority of recommendations had a low or very low LOE.

Over 80% of rheumatology guideline recommendations are not informed by head-to-head RCTs. Without comparative efficacy RCTs to inform such recommendations, they must be made using indirect comparisons of randomized data or lower-quality evidence, such as observational data or expert opinion. Likely reflecting this, most recommendations were graded as having a low quality of evidence, and the majority of recommendations using the GRADE system were scored as conditional. Multiple reasons for the lack of head-to-head trials in rheumatology should be considered, including the high cost of head-to-head RCTs, lack of governmental funding for clinical trials, and a reliance on industry funding for clinical trials that has prioritized industry incentives (8). These forces are not unique to rheumatology, and similar observations have been made in other fields (5,9).

These factors may also be visible in the remarkable heterogeneity in RCTs and head-to-head RCTs informing recommendations for different rheumatological diseases. The introduction of highly effective and lucrative biologic disease-modifying antirheumatic drugs in the late 1990s revolutionized the care of rheumatoid arthritis and psoriatic arthritis (10). Other diseases, such as Sjogren's syndrome and systemic lupus erythematosus, have experienced high-profile trials that failed to show a benefit (11), possibly fueling a lesser degree of interest in advancing therapies. Notably, the guidelines for the management of gout had the highest percentage of head-to-head references and recommendations informed by head-to-head RCTs. This observation highlights two critical areas in which changes to public policy could be beneficial. First, the FDA required head-to-head trials against the established standard of care, allopurinol, as part of the drug development program for a novel agent, febuxostat (3). Second, the FDA required multiple post-marketing trials to clarify potential safety signals (3), which inadvertently provided high-quality data about efficacy as well (12). This combination of upfront requirements for efficacy as compared with the standard of care as well as requirements for large post-marketing surveillance studies could substantially improve the quality of information generation in rheumatology.

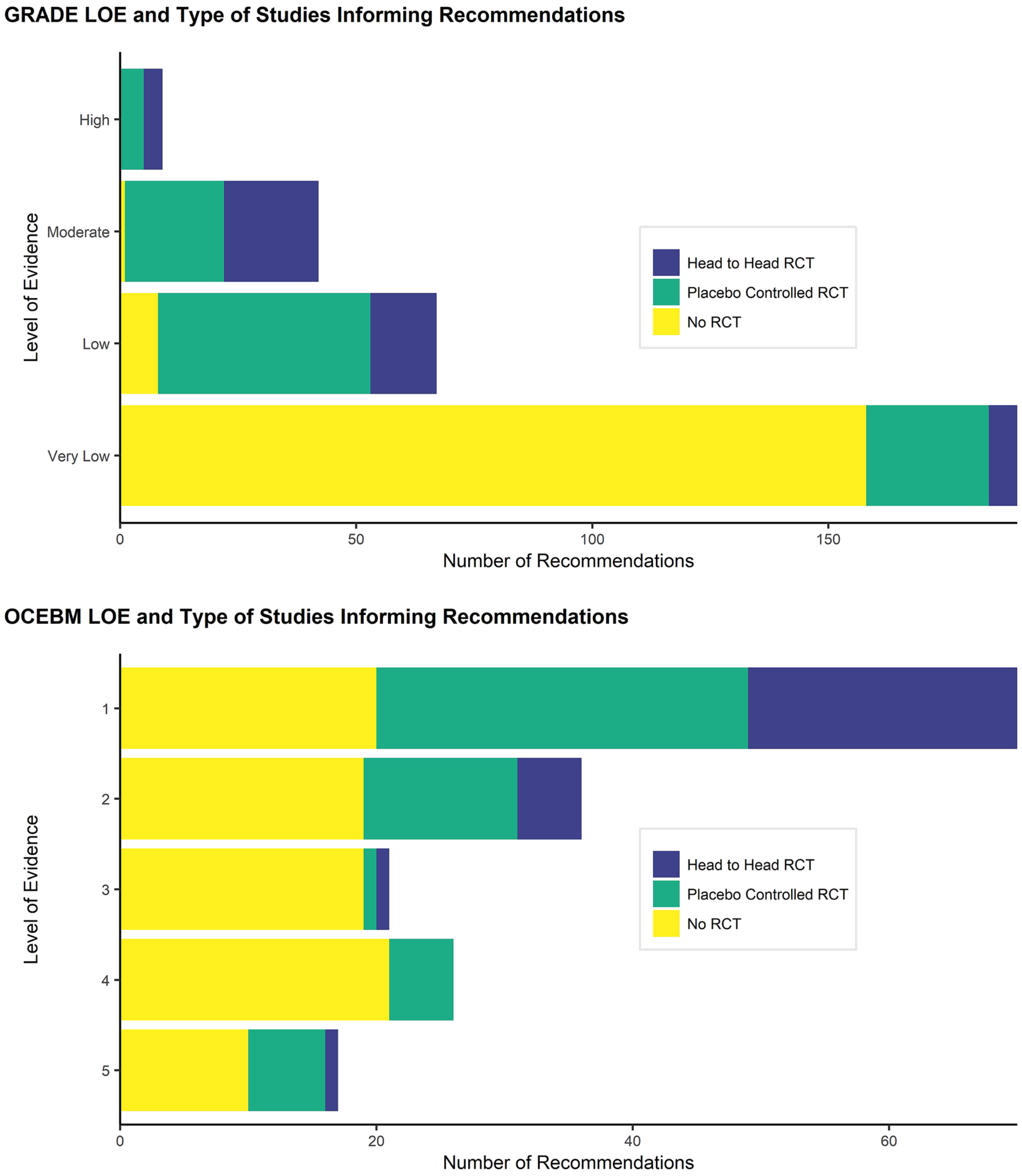

Figure 2. Type of study informing recommendations for each LOE in GRADE and OCEBM rated recommendations.

Refocusing the rheumatology research agenda in this manner could address another finding of this study—the presence of "strong" recommendations by the GRADE methodology that were based on low or very low quality of evidence. Such "discordance" has been observed in other specialties (13) and is allowed within the GRADE methodology when "most well-informed people will make the same choice" (14). In an effort to make this less subjective, the GRADE recommendations describe five merited paradigmatic scenarios, all of which pertain to either potential benefit in a life-threatening situation or potential harm (14). Observationally, many of the low or very low LOE ACR recommendations that were rated as "strong" did not appear to fit within one of the paradigmatic scenarios, instead focusing on general utility of interventions, such as NSAIDs, despite strong evidence. One potential solution, instead of providing discordant recommendations, is to follow the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) approach to define the domains in which therapies have evidence of efficacy. This would allow physicians to have more freedom to individualize choice. Nevertheless, while guidelines exist to provide recommendations, and such dissociation may be reasonable, providing a "strong" recommendation in the absence of strong data may eliminate perceived equipoise, thereby precluding RCTs and preventing the creation of confirmatory high-quality data.

Several avenues could be considered to improve the quality of data informing rheumatology practice guidelines. Perhaps most effectively, the FDA could take a leadership role. Simply informing an industry sponsor during the pre-submission process that a placebo-controlled design alone would be insufficient, which is already within their purview, could direct development programs toward comparative efficacy research (3). As discussed with the example of febuxostat, requiring more post-marketing trials may ultimately result in stronger evidence. Despite absent FDA leadership, funding agencies should prioritize the creation of high-quality comparative efficacy research. The current emphasis on the generation of basic and translational research has merit, but addressing this void by funding large-scale comparative efficacy trials may be of greater impact. Finally, guideline committees may consider requiring a higher LOE and head-to-head trials before recommending novel therapies over established standards of care, as happens in other specialties such as dermatology. Such therapies may "push out" the prior standard of care, but many such "advances" are ultimately reversed and often come at a high cost to both patients and health care systems (15).

This project has a number of limitations. First, we only evaluated guidelines from two societies, and our findings may not be generalizable to other countries or other professional society recommendations. Second, we focused on trials that informed clinical practice guidelines in order to identify those that were most impactful. The true rate of head-to-head RCT production in rheumatology cannot be inferred from these data. Third, changes over time may be influenced by both the rate of comparative efficacy research and also the degree to which different trial types are cited by guideline committees. Fourth, not all recommendations included an explicit statement for one treatment over another. Most recommendations imply some comparison—i.e., a recommendation "for methotrexate as first-line therapy" implies using methotrexate over hydroxychloroquine as first line therapy—and in some cases, the recommendation may be "common sense" and a comparator may not be necessary. Finally, this was a descriptive study and cannot assess causation.

Less than one in six rheumatology guideline recommendations are informed by comparative efficacy research. The majority of RCTs informing rheumatology guidelines evaluated bs/tsDMARDs and used a placebo-controlled design. Efforts to encourage more comparative efficacy RCTs in rheumatology, both through the FDA and through funding agencies, should be undertaken.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, approved the final version to be published, and agree to be accountable for all aspects of the work.

Study conception and design. Henry, Nepal, Valley, Pedersen, Duarte-García, Putman.
Acquisition of data. Henry, Nepal, Valley, Pedersen, Duarte-García, Putman.
Analysis and interpretation of data. Henry, Nepal, Valley, Pedersen, Duarte-García, Putman.

REFERENCES

1. Akobeng A. Understanding randomised controlled trials. Arch Dis Child 2005;90:840–4.
2. Estellat C, Ravaud P. Lack of head-to-head trials and fair control arms: randomized controlled trials of biologic treatment for rheumatoid arthritis. Arch Intern Med 2012;172:237–44.
3. US Food and Drug Administration. Development & approval process | Drugs. Published June 1, 2021. URL: https://www.fda.gov/drugs/development-approval-process-drugs.
4. Bourgeois FT, Murthy S, Mandl KD. Comparative effectiveness research: an empirical study of trials registered in ClinicalTrials.gov. PloS One 2012;7:e28820.
5. Duarte-García A, Zamore R, Wong JB. The evidence basis for the American College of Rheumatology Practice Guidelines. JAMA Intern Med 2018;178:146–8.
6. Brozek JL, Akl EA, Compalati E, Kreis J, Terracciano L, Fiocchi A, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. Allergy 2009;64:669–77.
7. University of Oxford Centre for Evidence-Based Medicine (CEBM). Oxford Centre for Evidence-Based Medicine: levels of evidence (March 2009). 2009. URL: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009.
8. Fanaroff AC, Califf RM, Windecker S, Smith SC, Lopes RD. Levels of evidence supporting American College of Cardiology/American Heart Association and European Society of Cardiology Guidelines, 2008-2018. JAMA 2019;321:1069–80.
9. Weissman S, Goldowsky A, Aziz M, Mehta TI, Sharma S, Lipcsey M, et al. Colorectal cancer screening guidelines are primarily based on low-moderate-quality evidence. Dig Dis Sci 2021;66:4208–19.
10. Albrecht K, Callhoff J, Zink A. Long-term trends in rheumatology care: achievements and deficits in 25 years of the German national rheumatology database. Z Rheumatol 2019;78:65–72.
11. Fisher BA, Everett CC, Rout J, O'Dwyer JL, Emery P, Pitzalis C, et al. Effect of rituximab on a salivary gland ultrasound score in primary Sjögren's syndrome: results of the TRACTISS randomised double-blind multicentre substudy. Ann Rheum Dis 2018;77:412–6.
12. Putman MS, Harrison Ragle A, Ruderman EM. The quality of randomized controlled trials in high-impact rheumatology journals, 1998-2018. J Rheumatol 2020;47:1446–9.
13. Sims CR, Warner MA, Stelfox HT, Hyder JA. Above the GRADE: evaluation of guidelines in critical care medicine. Crit Care Med 2019;47:109–13.
14. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook. 2013. URL: https://gdt.gradepro.org/app/handbook/handbook.html.
15. Cifu AS, Prasad VK. Medical debates and medical reversal. J Gen Intern Med 2015;30:1729–30.