HHS Public Access

Author manuscript

Ann Rheum Dis. Author manuscript; available in PMC 2021 December 22.

Published in final edited form as:

Ann Rheum Dis. 2020 July ; 79(7): 883–890. doi:10.1136/annrheumdis-2020-217200.

Improving rheumatoid arthritis comparative effectiveness research through causal inference principles: systematic review using a target trial emulation framework

Sizheng Steven Zhao[1,2], **Houchen Lyu**[2,3], **Daniel H Solomon**[2,4], **Kazuki Yoshida**[2]

[1]Musculoskeletal Biology, Institute of Lifecourse and Medical Sciences, University of Liverpool, Liverpool, UK

[2]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States

[3]Department of Orthopaedics, General Hospital of Chinese PLA, Beijing, China

[4]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States

Abstract

Objectives.—Target trial emulation is an intuitive design framework that encourages investigators to formulate their comparative effectiveness research (CER) question as a hypothetical randomised controlled trial (RCT). Our aim was to systematically review CER studies in RA to provide examples of design limitations that could be avoided using target trial emulation, and how these limitations might introduce bias.

Methods.—We searched for head-to-head CER studies of biologic DMARDs in RA. Study designs were reviewed for seven components of the target trial emulation framework: eligibility criteria, treatment strategies, assignment procedures, follow-up period, outcome, causal contrasts of interest (i.e., intention-to-treat or per-protocol effect) and analysis plan. Hypothetical trials corresponding to the reported methods were assessed to identify design limitations that would have been avoided with an explicit target trial protocol. Analysis of the primary *effectiveness outcome* was chosen where multiple analyses were performed.

Results.—We found 31 CER studies, of which 29 (94%) had at least one design limitation belonging to the seven components. The most common limitations related to: 1) eligibility criteria: 19/31 (61%) studies used post-baseline information to define baseline eligibility; 2) causal contrasts: 25 (81%) did not define whether intention-to-treat or per-protocol effects were estimated; and 3) assignment procedures: 13 (42%) studies did not account for confounding by indication or relied solely on statistical confounder selection.

Correspondence to: Dr Sizheng S Zhao, Academic Rheumatology Department, Liverpool University Hospitals, Liverpool, L9 7AL, UK, s.zhao8@liverpool.ac.uk.

Contributors: All authors contributed to study design, data interpretation, writing and review of the manuscript and approved the final version for publication.

Competing Interest: None declared.

Patient and public involvement statement: None.

Conclusions.—Design limitations were found in 94% of observational CER studies in RA. Target trial emulation is a structured approach for designing observational CER studies that helps to avoid potential sources of bias.

Keywords

target trial emulation; causal inference; observational study; comparative effectiveness; rheumatoid arthritis

Introduction

There are a growing number of pharmacological treatment options in rheumatology, particularly high-cost biologic and targeted synthetic disease modifying anti-rheumatic drugs (bDMARDs and tsDMARDs). This enlarging armamentarium begs the question of how to choose the optimal treatment for a given condition. Head-to-head randomised controlled trials (RCTs) of these new and emerging therapies – the preferred evidence – are scarce and provide only limited guidance; for example, few have directly compared treatment options for rheumatoid arthritis (RA) patients who have failed one or more bDMARDs. When RCTs are not feasible, timely, or ethical, observational data can help fill the need for comparative effectiveness information [1–3].

Observational comparative effectiveness research (CER) studies are common in rheumatology, as are their critics. The target of much criticism lies in the sheer number of methodological approaches available to the analyst [4], and the profound effect that nuanced differences in methods can have on results [5,6]. Large sample sizes in these studies may instil false confidence in the presence of critical design flaws. Improvement and standardisation of methodology guided by causal inference principles are paramount and increasingly demanded by many clinical journals [7]. One barrier, however, is that detailed theory based on "potential outcomes" [8] can seem complex and unfamiliar.

A more intuitive approach is to ask: "How would I answer my CER question as an *RCT*?" according to the recently popularized "target trial emulation" framework of observational study design and analysis [3]. This way of thinking is rooted in the principles of causal inference, which were first conceived in the context of randomized experiments [9] and extended to observational studies [2,10,11]. Principled re-analyses of existing observational studies using the target trial emulation framework have repeatedly shown to reduce bias and better align results with actual RCTs [6,12–16].

Through a systematic review of observational CER studies in RA, we provide examples of design limitations that might have been avoided by using target trial emulation, and how these limitations might introduce bias. Since the practice of explicitly writing down the target trial protocol is relatively new [3], we did not expect the reviewed study designs to incorporate its terminology; rather, we retrospectively imposed the framework components as a structured way to appraise them.

Methods

Target trial emulation framework

At the heart of the target trial emulation framework are two protocols –"target trial protocol" and "emulation protocol"– for each observational CER study (Figure 1). When prospectively *designing* an observational CER study, researchers first consider how the question can be answered and formulated as a hypothetical RCT — the "target trial" [3]. Systematically specifying this protocol helps ensure that the question is clinically meaningful for decision making (e.g., "among eligible patients, which treatment strategy maximizes benefit?" rather than simply "is exposure X associated with outcome Y?")[17,18]. The "emulation protocol" describes how available observational data might be used to obtain the best approximation of the "target trial protocol". The protocol is then reformulated when data limitations and feasibility of the ideal emulation are realised (Figure 1) [10]. This framework helps to avoid common methodologic pitfalls[3,19] and better align results with RCTs.

The framework can also be applied retrospectively for *appraising* existing studies in a structured process [3]. The description of an observational CER study can be seen as the emulation protocol for the inferred target trial. After inferring the corresponding target trial protocol, it is possible to assess the clinical question [17,18] as well as subtle design limitations that fail to emulate a sound target trial [19]. This retrospective application is our approach in this review.

Systematic literature review

We searched EMBASE, Medline and PubMed in August 2019. Search terms are shown in online supplementary materials. We only included *head-to-head effectiveness* comparisons to demonstrate the utility of target trial emulation beyond existing good design practice (Supplementary Table S1) [4,20,21]. We restricted to comparisons of *different classes* of bDMARDs, where RCTs are scarce but evidence is needed to guide clinical practice. Studies that did not report effectiveness (i.e., only reported drug retention or adverse/comorbidity events) were excluded. Independent reviewers (SSZ, HL) assessed study eligibility and performed data extraction (Table 1 and Supplementary Table S2); discrepancies were resolved through discussion moderated by a third reviewer (KY). Where multiple analyses were presented in one study, analysis for the primary *effectiveness outcome* was chosen. We appraised each study's design against each of the target trial emulation components below [3]; the specific questions we asked of each study's design (i.e., data extraction items) are listed in Table 1 (left column).

1. **Eligibility criteria.—**All RCTs are expected to have clear, predefined eligibility criteria before the study begins. Those with contraindications to any of the treatment strategies must be excluded. Obviously, RCT eligibility criteria can only consist of *baseline information* that are available to investigators at the time of prospective enrolment. To emulate such a target trial, observational cohorts first need to be defined using information up to the baseline (often called "time zero"), but not beyond [3]. For example, some observational studies require at least one follow-up in the eligibility criteria. This practice does not have an RCT equivalent, since trial investigators cannot see into the future at the

time of each patient's enrolment. Cheating the baseline criteria via such an oracle can bias results in either direction [22].

Including key confounders in the emulation eligibility criteria may help comparability. One example is the number/type of prior bDMARD failure; prior bDMARD failure is an important predictor of response. This is intuitive when conceptualised as equivalent RCTs: a trial comparing bDMARD naïve patients, or a trial comparing switching of therapy after one or more TNFi failures. We examined each study's eligibility criteria for use of post-baseline information and specification of the prior bDMARD treatment history (Table 1).

2. Treatment strategies.—An RCT protocol specifies detailed treatment strategies beyond the drug name. The protocol also defines criteria for discontinuation or modification, and relevant concomitant care that are allowed or prohibited during the follow up [23]. Each major treatment change should be defined as complying with or violating the protocol. For example, change of concomitant csDMARDs or discontinuation of bDMARD due to remission may be protocol-compliant, whereas switch to another bDMARD due to insufficient response may be considered a protocol violation. How treatment strategies are defined will have implications for the definition and analysis of the per-protocol effect (components 6 and 7). Observational CER studies also have to specify treatment strategies. Not all datasets have enough details to define granular treatment strategies, which may require the emulated target trial to be simplified (Figure 1). We reviewed how treatment changes were defined as part of the treatment strategy.

3. Assignment procedures.—In the simplest RCT design, participants have equal chance of being assigned to each treatment strategy, and each treatment group will have comparable distribution of prognostic factors. To emulate random treatment assignment in observational CER, all theoretical confounding factors - measured and unmeasured - need to be adjusted for. Inability to completely account for confounding is the commonest criticism for observational studies. While this may be the case, there are many ways to improve emulation of random assignment. Two considerations are the use of active-comparators [20] (which should be present by default in head-to-head comparisons) and methods for selecting confounders [24]. Confounding factors should not be selected *solely* based on their statistical association with the exposure or outcome. Selection should instead be based on subject knowledge and/or literature review, preferably supported by directed acyclic graphs (DAGs) that make assumptions transparent [24]. For the review, we focused on how confounding factors were selected.

4. Follow-up period.—RCTs have a well-defined start, schedule and end of follow-up. In rheumatology trials, efficacy typically focus either on percentage of participants achieving a response definition at a certain time-point or multiple repeated assessments over the study period. Each RCT participant is consented for a schedule of follow-ups and a finite study period (e.g., 12 months after treatment initiation). By contrast, observational data often come from *ad hoc* clinic visits, where frequency may be associated with patient and disease characteristics. The data to be used in an observational analysis should ideally reflect the same (or similar) data that would have been collected in the target trial [25]. This reduces

issues from differential health utilization and provides greater structure for missing data assessment. We assessed whether duration to end of follow-up was defined.

5. Outcome.—RCT outcomes typically include a response definition at a certain time point, e.g., DAS28<2.6 at week 12. In observational CER, an appropriate outcome definition also needs to be accompanied by assessment time(s), which is why follow-up needs to be clearly prespecified. A common practice for binary response outcomes is to include assessments within a certain window (e.g., 12 ±3 months), but this is less commonly considered for repeated continuous outcomes (e.g., use all available follow-up DAS28 scores).

The choice of RCT outcome is often linked to power (sample size) calculations. This is not a component in the target trial emulation framework [3], but an underpowered observational study can only emulate an underpowered RCT at best. In the majority of observational CER studies, sample size is not a decision. Investigators should estimate whether there is sufficient power to pursue their analysis at the design stage. Underpowered endeavours should be avoided or at least highlighted as a major limitation. Note that post hoc power calculations, whether in trials or emulations, are not informative [26,27]. We reviewed whether outcomes were clearly defined and whether statistical power limitations were discussed.

6. Causal contrast of interest.—Researchers should clearly define the answer they want – the intention-to-treat (ITT) or per-protocol effect – before thinking about the data or analysis [28], in observational CER studies as is the standard for RCTs. The two estimands require different analyses, have different interpretations and often have different effect sizes. In RCTs, ITT analyses estimate the effect of being assigned to treatment strategies, regardless of what happens thereafter (even if treatment is not initiated). The observational analogue of the ITT effect is the effect of *initiating* the treatment strategies; i.e., ITT analysis will include outcomes from patients who remained on and those who discontinued the drug (or deviated from the protocol in any other manner). The per-protocol effect is the effect of the treatment strategy when fully adhered to, hence the importance of clearly defining it. Discontinuing treatment (for whatever reason) may be the only specified “protocol deviation” in observational CER studies, in which case per-protocol analysis will include only the “on-treatment” population. We assessed whether the authors defined their causal contrast of interest and what they were.

7. Analysis plan.—Analyses for the ITT effect in RCTs do not require confounding-adjustment and is the direct comparison of the average outcomes of treatment arms. However, in the presence of differential loss-to-follow-up, missing data handling that preserves the original randomized cohort are required for valid ITT analysis (e.g., by imputing non-response). By contrast, analysis for the per-protocol effect occurs in a subset of data that artificially censor individuals at the time they deviated from the treatment strategies. Such non-random censoring likely introduces selection bias. Advanced statistical methods using post-baseline time-varying covariates are required to adjust for this bias [3]. We reviewed analysis plans together with the declared or implied causal contrast as above.

The analysis plan in the emulation protocol should look identical to the target trial protocol except for the need to adjust for baseline confounding. Treatment strategy, causal contrast and analysis plan are dependent upon each other and were reviewed together. Where causal contrasts were not declared, we assessed how censoring was defined in each study (i.e., treatment strategies) to infer the authors' implied causal contrast. The chosen statistical method affects how missing data and censoring are handled; we therefore reviewed the statistical model and missing data handling in this section.

Results

A total of 31 studies met our inclusion criteria. The selection flowchart is shown in Supplementary Figure S1. 21 studies compared bDMARDs from two classes [29–40,43,45,47,52–54,56,57,59] and 10 compared three or more classes [41,42,44,46,48–51,55,58]; there were no studies of tsDMARDs. Information extracted from each study are detailed in supplementary Table S2. Only one study explicitly emulated a target trial [53].

1. Eligibility criteria.

19 out of 31 (61%) studies explicitly included post-baseline information in their eligibility criteria by requiring at least one follow-up [29–45] and/or a minimal duration of follow-up [46,47,59]. The proportions of participants excluded without follow-up were typically large (up to 69% [34]), but frequently unreported. An additional seven studies [48–50,52,55,56,58] implicitly made these exclusions (up to 82% [33]) by using complete-case analyses.

10 studies specified an exact number of prior bDMARDs, while 21 studies included varying numbers of prior bDMARD failure. 12 studies combined bDMARD experienced patients (i.e., ≥1 prior bDMARD failures) or stratified analyses to that effect [30–37,44,46,54,58]. Eligibility criteria of nine studies included both bDMARD naïve and experienced patients, among which five did not stratify to separate treatment effect for these two groups [40,44,48–50].

2. Treatment strategies.

Treatments under comparison were well defined, but not treatment *strategies*. Only one study [51] defined whether treatment changes were protocol compliant (e.g., biosimilar switch and discontinuation due to remission were permitted). Studies comparing bDMARD monotherapy [37,49,50,54] were unclear whether initiation of csDMARD (although rare in clinical practice) would be artificially censored. Due to limited descriptions of treatment strategies, we instead examined what investigators artificially censored in the analysis to infer what they considered protocol compliant (i.e., whether discontinuation was censored; see Supplementary Table S2). However, what the authors censored was not always clearly described.

3. Assignment procedures.

13 of the 31 (42 %) studies used only pre-defined confounders [31–34,36–38,41,44,46,52–54]; none explicitly cited literature review or DAGs. Five (16%) studies used only statistical variable selection [29,30,43,45,47], such as univariate or stepwise P-value-based selection, or

change-in-outcome selection. One study included post-baseline variables in the selection process [29]. Nine (29%) studies performed no adjustments for confounding [39,40,42,47–50,55,56] (one deliberately excluded baseline values of the outcome [30]). Active-comparator, new-user design was used in all except one study that included both new and prevalent users [50].

4. Follow-up period.

14 of 31 (45 %) studies used binary outcomes with clear assessment time points akin to RCTs, thereby defining their end to follow-up [35,41,42,45,46,48,50,51,53–57,59]. 18 (58 %) studies included continuous outcomes (e.g., DAS28 over time), among which six did not specify an end to the study period [30–32,40,41,54]. This was typically when linear mixed models were used with all available data. One extreme example used the last available follow-up before therapy switch, which could be any time-point beyond 1 year [41].

5. Outcome.

Outcomes were clearly defined in all studies. Only four studies (two of which adopted RCT design) explicitly performed sample size calculations [29,31,43,52]. There was typically no discussion about power limitations, even when sample sizes were as small as <50 in each arm [58]. Four studies reported a joint outcome [42,50,51,54], i.e., the proportion achieving response *and* remaining on drug. When this was achieved using LUNDEX ("fraction of starters still in the study multiplied by the fraction responding" [60]), statistical comparisons and confidence intervals were not provided.

6. Causal contrast of interest.

Only six studies defined their causal contrasts prior to describing analyses [31,47,49–51,53]. Inferring from analysis methods, 20 studies included some version of the ITT effect [29,31,33–37,39,41,43,44,46–48,51,53,54,57–59]; all except one [53] excluded patients without follow-up (through eligibility criteria or complete-case analysis) which is not compatible with the traditional ITT definition. 12 studies declared or implied per-protocol effects in part of their analysis [30–32,38,40,42,45,49–52,55], but none subsequently adjusted for post-baseline time-varying confounding. Causal contrasts could not be clearly determined in two studies, due to inclusion of prevalent users [50] and lack of clarity on whether discontinuation was defined as non-response [56].

7. Analysis plan.

Most studies either used (generalised) linear models for outcomes at a fixed time-point [35,36,41,43,44,51,53,57,59], or linear mixed models for repeated continuous outcome measures [30–34,37,45,47,54]. One study used generalised estimating equations [46]. Eight studies used pairwise comparisons (e.g., t-test, chi squared test) or ANCOVA [29,39,48,49,52,55,56,58]. Three studies did not perform any statistical comparison, two of which due to the use of LUNDEX [42,50].

17 out of 31 studies used complete-case analyses [29–33,39,41,43,44,47–49,52,55,56,58,59]. Linear mixed models can handle missing (at random) data by default. Only three studies used multiple imputation for missing outcome data [51,52,57], while nine studies used single imputation (e.g. last observation carried forward, or non-response imputation) to obtain ITT

effects [34–38,45,46,53,54]. It was often unclear whether outcomes of those who discontinued treatment were included or excluded from analysis. Reasons for discontinuation (e.g., switch to another bDMARD or biosimilar or remission) were rarely differentiated in treatment strategies which impacts definition of causal contrast and its analysis. The analyses of 10 studies artificially censored individuals discontinuing the initial bDMARD ("on-treatment" analysis) [31,37,38,40,42,45,49,50,52,55].

Discussion

We use the target trial emulation framework to identify design limitations in CER studies in RA. There was significant methodologic variation despite restricting to a relatively narrow topic with simple designs. One study described the target trial[53] although not to the extent recommended[3]. 94% had at least one design limitation with the potential to introduce bias; the most common were: 1) including post-baseline information in eligibility criteria, 2) not defining the causal contrasts (i.e., ITT or per-protocol effects), and 3) inadequate emulation of random assignment with unadjusted comparisons and reliance on statistical confounder selection.

Excluding those without future follow-up in eligibility criteria cannot emulate an RCT. Beyond the conceptual conundrum, this practice can bias results in either direction when loss to follow-up differ across treatment arms and are associated with outcomes [22]. Many complete case analyses also implicitly exclude those without follow-up. The underlying motive is to deal with missing data. Naïve methods of missing data handling, such as complete case analysis or single imputation, are not recommended for RCTs [61,62]. This is still more relevant for observational studies where the proportions missing are much higher (the potential for bias increases as the proportion of missing data increases). Larger sample sizes seen in observational CER may instil greater confidence in the wrong result. Common alternative approaches include multiple imputation or likelihood-based methods depending on the pattern of missingness. Defining a follow-up duration and desired outcome assessment times can help assess missing data patterns and mechanisms. Note that the proportion of missing data does not dictate the validity of multiple imputation, rather it is the mechanisms of missingness and amount of information held by auxiliary variables (that inform generation of imputations) [63]. Investigators might also attempt to reduce missing data by selecting participants with higher likelihood of follow-up without looking at post-baseline data (analogous to pre-randomisation run-in periods of RCTs [64]). A full discussion on how to handle missing data is beyond the scope of this review. We instead refer readers to reference [62] for an introduction and [61,65] for comprehensive overview. If analysis restricted to those with follow-up is unavoidable, comparison of included and excluded individuals should be clearly presented as a minimum. However, this would render causal contrasts difficult to define.

Choice of causal contrasts relates to a similar underlying missing data issue. ITT is appealing for its (perceived) simplicity in both trials and observational CER: analyse outcomes in all those assigned treatment strategies regardless of what happens thereafter (i.e., adherence or protocol deviations). Unlike trials, observational data often have significant loss to follow-up. Imputing non-response to all missing cases may result in

null ITT effect, even if one treatment is superior. ITT analyses also have limitations when studying harms of treatment and in non-inferiority comparisons [66]. Per-protocol effects have many advantages over ITT, but valid estimation can be challenging. One case-study against per-protocol effects was the survival difference between adherers and non-adherers to placebo in a cardiology trial, which could not be removed with simple statistical adjustment [67]. Successful adherence adjustment is possible with modern methods that adjust for post-baseline time-varying prognostic factors [67]. As is recommended for pragmatic trials, observational CER studies should present both ITT and per-protocol effects [66] or, as a minimum, declare the causal contrast before analysis. This was rarely done in the studies reviewed, partly because clear definitions were impossible when post-baseline information was used for eligibility criteria and/or because treatment strategies were not described. Further discussion of causal contrasts and related methods can be found in references [10,19,68].

Another common design issue was the emulation of random assignment in RCTs. Inclusion of prevalent users was uncommon because we reviewed head-to-head (i.e., active-comparator) studies, but this practice and consequent time-dependent biases are common in the rheumatology literature (e.g., immortal time bias resulting from a period of follow-up during which the study outcome cannot occur [69]). Applying intuition from the target trial protocol can prevent such “self-inflected injuries” [19]. Selecting an active-comparator with similar indications also has several advantages for confounding adjustment [20,21]. Valid comparison requires individuals to have non-zero probability of receiving either treatment (i.e., no absolute contra-indications). Statistical adjustment for confounding also requires sufficient overlap in participant characteristics across treatment groups, which will be greater when treatments have similar indications; this is also true for unmeasured confounding (e.g., frailty) [4].

Approaches for choosing confounders varied. Statistical approaches to covariate selection (e.g., based on statistical associations that are widely used in studies of predictors) should generally be avoided for CER [7,24]. Selection should be based more on subject knowledge and/or literature review, preferably supported by directed acyclic graphs (DAGs) that make assumptions transparent [7]. This helps avoid including variables that can introduce bias when adjusted, such as mediators (causal intermediate variables that are part of the treatment effect) and colliders (variable causally influenced by ≥2 variables that induces spurious associations if adjusted) [7]. Adjusting for covariates that are not prognostic but strongly associate with treatment assignment (possible when selection is purely based on statistical associations with treatment group) does little to reduce bias but increase variance [70]. Declaring *a priori* confounders also improves transparency, which is another criticism of observational studies [5]. Propensity score methods - although not necessarily superior to traditional multivariable regression - provide an intuitive emulation of randomization since it separates confounding adjustment from outcome analysis. They also help assess utility of the comparison; for example, analyses should be avoided or cautiously approached when propensity score overlap is poor (i.e., violation of the positivity assumption required for valid causal inference) [4].

This review was limited to comparisons of effectiveness, but target trial emulation applies equally to safety and other types of time-to-event outcomes (existing target trial emulation examples are typically applications for the latter). Additional design considerations for studies of safety outcomes, such as infections, are discussed in reference [71]. A central concept in the emulation framework - clear definition of time-zero – may seem obvious in the examples reviewed, but the principle is equally relevant for other observational designs such as case-control studies [72,73]. Our restriction in scope left out secondary analyses, which were also fraught with design issues, e.g., comparisons of a bDMARD as first vs second line treatment [45,57] (try considering if this can be implemented as an RCT). Post hoc adherence-adjusted response using the LUNDEX [60] was found in several papers, with the primary aim of accounting for missing outcome data. Assuming non-informative censoring, the LUNDEX estimates the "proportion of patients who not only remain on a particular therapeutic regimen but also fulfil certain response criteria" [60]. Adherence issues are avoided at the cost of changing the outcome definition. Future work in this area should incorporate more modern approaches to adherence adjustment [67], in emulation of more rigorous definitions and handling of missing data in RCT literature [62].

In conclusion, target trial emulation builds on existing good design practices to make robust observational designs intuitive. It ensures that clear and clinically relevant questions are asked, incorporates causal inference principles, and has been shown to better align observational results with actual RCTs. Future CER studies should avoid using post-baseline information to define eligibility, clearly define the causal contrast pursued (ideally presenting both ITT and per protocol effects), and consider confounding using prior knowledge (preferably using causal diagrams that make assumptions transparent). This framework is beginning to be adopted in the rheumatology literature [53,74], but further improvement and standardisation of CER methodology is essential as more drugs become available, often without (timely) head-to-head RCTs to compare their effectiveness.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

We thank Dr Anders Huitfeldt for helpful suggestions.

Funding:

DHS was supported by grants from the National Institute of Health (NIH-P30-AR072577 (VERITY)). KY was supported by the Rheumatology Research Foundation Career Development Bridge Funding Award.

Data availability statement:

All data relevant to the study are included in the article or uploaded as online supplementary information.

References

1. Schneeweiss S Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther. 2007 8;82(2):143–56. [PubMed: 17554243]
2. Rubin DB. Estimating causal effects of treatment in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
3. Hernan MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 4 15;183(8):758–64. [PubMed: 26994063]
4. Stürmer T, Wang T, Golightly YM, Keil A, Lund JL, Jonsson Funk M. Methodological considerations when analysing and interpreting real-world data. Rheumatology. 2020 1 1;59(1):14–25. [PubMed: 31834408]
5. Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015 9;68(9):1046–58. [PubMed: 26279400]
6. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. Nat Med. 2019 10;25(10):1601–6. [PubMed: 31591592]
7. Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, et al. Control of Confounding and Reporting of Results in Causal Inference Studies. Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. Ann Am Thorac Soc. 2018 9 19;16(1):22–8.
8. Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. J Am Stat Assoc. 2005 3;100(469):322–31.
9. Rubin D Comment on Randomization Analysis of Experimental Data: The Fisher Randomization Test. J Am Stat Assoc 198075371591–593.
10. Hernán M, Robins J. Causal Inference: What If. [Internet]. Boca Raton: Chapman & Hall/CRC; 2020 [cited 2020 Jan 13]. Available from: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
11. Robins J A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model. 1986;7(9–12):1393–512.
12. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Stampfer MJ, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiol Camb Mass. 2008 11;19(6):766–79.
13. García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol. 2017;32(6):495–500. [PubMed: 28748498]
14. Danaei G, Rodríguez LAG, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. Stat Methods Med Res. 2013 2;22(1):70–96. [PubMed: 22016461]
15. Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining Bias in Studies of Statin Treatment and Survival in Patients With Cancer. JAMA Oncol. 2018 1;4(1):63–70. [PubMed: 28822996]
16. Dickerman BA, Giovannucci E, Pernar CH, Mucci LA, Hernán MA. Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019 01;188(3):579–86. [PubMed: 30496346]
17. Hernán MA. Counterpoint: Epidemiology to Guide Decision-Making: Moving Away From Practice-Free Research. Am J Epidemiol. 2015 11 15;182(10):834–9. [PubMed: 26507306]
18. Didelez V Commentary: Should the analysis of observational data always be preceded by specifying a target experimental trial? Int J Epidemiol. 2016 01;45(6):2049–51. [PubMed: 27063602]
19. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016 11;79:70–5. [PubMed: 27237061]

20. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. Nat Rev Rheumatol. 2015 7;11(7):437–41. [PubMed: 25800216]
21. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. Curr Epidemiol Rep. 2015 12;2(4):221–8. [PubMed: 26954351]
22. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiol Camb Mass. 2004 9;15(5):615–25.
23. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med. 2013 2 5;158(3):200–7. [PubMed: 23295957]
24. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol. 2019 3;34(3):211–9. [PubMed: 30840181]
25. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? EGEMS Wash DC. 2016;4(1):1203. [PubMed: 27668265]
26. The CONSORT 2010 Statement. 7a. Sample size. [Internet]. [cited 2020 Mar 31]. Available from: http://www.consort-statement.org/checklists/view/32—consort-2010/83-sample-size
27. Hoenig JM, Heisey DM. The Abuse of Power. Am Stat. 2001 2 1;55(1):19–24.
28. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal inference and estimands in clinical trials. Stat Biopharm Res. 2019 12 2;1–30.
29. Emery P, Gottenberg JE, Rubbert-Roth A, Sarzi-Puttini P, Choquette D, Taboada VMM, et al. Rituximab versus an alternative TNF inhibitor in patients with rheumatoid arthritis who failed to respond to a single previous TNF inhibitor: SWITCH-RA, a global, observational, comparative effectiveness study. Ann Rheum Dis. 2015;74(6):979–84. [PubMed: 24442884]
30. Finckh A, Ciurea A, Brulhart L, Kyburz D, Moller B, Dehler S, et al. B cell depletion may be more effective than switching to an alternative anti-tumor necrosis factor agent in rheumatoid arthritis patients with inadequate response to anti-tumor necrosis factor agents. Arthritis Rheum. 2007 5;56(5):1417–23. [PubMed: 17469098]
31. Finckh A, Moller B, Dudler J, Walker UA, Kyburz D, Gabay C, et al. Evolution of radiographic joint damage in rituximab-treated versus TNF-treated rheumatoid arthritis cases with inadequate response to TNF antagonists. Ann Rheum Dis. 2012;71(10):1680–5. [PubMed: 22419773]
32. Finckh A, Ciurea A, Brulhart L, Moller B, Walker UA, Courvoisier D, et al. Which subgroup of patients with rheumatoid arthritis benefits from switching to rituximab versus alternative anti-tumour necrosis factor (TNF) agents after previous failure of an anti-TNF agent? Ann Rheum Dis. 2010 2;69(2):387–93. [PubMed: 19416802]
33. Gomez-Reino JJ, Maneiro JR, Ruiz J, Roselló R, Sanmarti R, Romero AB. Comparative effectiveness of switching to alternative tumour necrosis factor (TNF) antagonists versus switching to rituximab in patients with rheumatoid arthritis who failed previous TNF antagonists: the MIRAR Study. Ann Rheum Dis. 2012 11;71(11):1861–4. [PubMed: 22736086]
34. Harrold LR, Reed GW, Kremer JM, Curtis JR, Solomon DH, Hochberg MC, et al. The comparative effectiveness of abatacept versus anti-tumour necrosis factor switching for rheumatoid arthritis patients previously treated with an anti-tumour necrosis factor. Ann Rheum Dis. 2015;74(2):430–6. [PubMed: 24297378]
35. Harrold LR, Reed GW, Magner R, Shewade A, John A, Greenberg JD, et al. Comparative effectiveness and safety of rituximab versus subsequent anti-tumor necrosis factor therapy in patients with rheumatoid arthritis with prior exposure to anti-tumor necrosis factor therapies in the United States Corrona registry. Arthritis Res Ther. 2015;17(101154438):256. [PubMed: 26382589]
36. Harrold LR, Reed GW, Solomon DH, Curtis JR, Liu M, Greenberg JD, et al. Comparative effectiveness of abatacept versus tocilizumab in rheumatoid arthritis patients with prior TNFi exposure in the US Corrona registry. Arthritis Res Ther [Internet]. 2016 12 [cited 2019 Jun 6];18(1). Available from: http://arthritis-research.biomedcentral.com/articles/10.1186/s13075-016-1179-7

37. Harrold LR, Reed GW, Best J, Zlotnick S, Kremer JM. Real-world Comparative Effectiveness of Tocilizumab Monotherapy vs. Tumor Necrosis Factor Inhibitors with Methotrexate in Patients with Rheumatoid Arthritis. Rheumatol Ther. 2018 12;5(2):507–23. [PubMed: 30293218]
38. Harrold LR, Litman HJ, Connolly SE, Alemao E, Kelly S, Rebello S, et al. Comparative Effectiveness of Abatacept Versus Tumor Necrosis Factor Inhibitors in Patients with Rheumatoid Arthritis Who Are Anti-CCP Positive in the United States Corrona Registry. Rheumatol Ther. 2019 6;6(2):217–30. [PubMed: 30868550]
39. Kekow J, Muller-Ladner Schulze-Koops. Rituximab is more effective than second anti-TNF therapy in rheumatoid arthritis patients and previous TNFα blocker failure. Biol Targets Ther. 2012 7;191.
40. Leffers HC, Ostergaard M, Glintborg B, Krogh NS, Foged H, Tarp U, et al. Efficacy of abatacept and tocilizumab in patients with rheumatoid arthritis treated in clinical practice: results from the nationwide Danish DANBIO registry. Ann Rheum Dis. 2011 7 1;70(7):1216–22. [PubMed: 21551512]
41. Li N, Betts KA, Messali AJ, Skup M, Garg V. Real-world Effectiveness of Biologic Disease-modifying Antirheumatic Drugs for the Treatment of Rheumatoid Arthritis After Etanercept Discontinuation in the United Kingdom, France, and Germany. Clin Ther. 2017 8;39(8):1618–27. [PubMed: 28729087]
42. Santos-Faria D, Tavares-Costa J, Eusébio M, Leite Silva J, Ramos Rodrigues J, Sousa-Neves J, et al. Tocilizumab and rituximab have similar effectiveness and are both superior to a second tumour necrosis factor inhibitor in rheumatoid arthritis patients who discontinued a first TNF inhibitor. Acta Reumatol Port. 2019 6;44(2):103–13. [PubMed: 31243259]
43. Soliman MM, Hyrich KL, Lunt M, Watson KD, Symmons DPM, Ashcroft DM, et al. Effectiveness of rituximab in patients with rheumatoid arthritis: observational study from the British Society for Rheumatology Biologics Register. Griffiths I ID Panayi G, Scott DG, Bamji A, Bax D, Scott DL, Peters S, Taylor N, Hogg M, Tracey A, McGrother K, Silman A, editor. J Rheumatol. 2012;39(2):240–6. [PubMed: 22174201]
44. Walker UA, Jaeger VK, Chatzidionysiou K, Hetland ML, Hauge E-M, Pavelka K, et al. Rituximab done: what's next in rheumatoid arthritis? A European observational longitudinal study assessing the effectiveness of biologics after rituximab treatment in rheumatoid arthritis. Rheumatology. 2016 2;55(2):230–6. [PubMed: 26316581]
45. Yoshida K, Tokuda Y, Oshikawa H, Utsunomiya M, Kobayashi T, Kimura M, et al. An observational study of tocilizumab and TNF- inhibitor use in a Japanese community hospital: different remission rates, similar drug survival and safety. Rheumatology. 2011 11 1;50(11):2093–9. [PubMed: 21890622]
46. Gottenberg J-E, Morel J, Perrodeau E, Bardin T, Combe B, Dougados M, et al. Comparative effectiveness of rituximab, abatacept, and tocilizumab in adults with rheumatoid arthritis and inadequate response to TNF inhibitors: prospective cohort study. BMJ. 2019 1 24;l67. [PubMed: 30679233]
47. Blom M, Kievit W, Donders ART, den Broeder AA, Straten VHHP, Kuper I, et al. Effectiveness of a third tumor necrosis factor-alpha-blocking agent compared with rituximab after failure of 2 TNF-blocking agents in rheumatoid arthritis. J Rheumatol. 2011;38(11):2355–61. [PubMed: 21885487]
48. Iannone F, Ferraccioli G, Sinigaglia L, Favalli EG, Sarzi-Puttini P, Atzeni F, et al. Real-world experience of tocilizumab in rheumatoid arthritis: sub-analysis of data from the Italian biologics' register GISEA. Clin Rheumatol. 2018 2;37(2):315–21. [PubMed: 28980085]
49. Jørgensen TS, Turesson C, Kapetanovic M, Englund M, Turkiewicz A, Christensen R, et al. EQ-5D utility, response and drug survival in rheumatoid arthritis patients on biologic monotherapy: A prospective observational study of patients registered in the south Swedish SSATG registry. Kuwana M, editor. PLOS ONE. 2017 2 2;12(2):e0169946. [PubMed: 28151971]
50. Jorgensen TS, Kristensen LE, Christensen R, Bliddal H, Lorenzen T, Hansen MS, et al. Effectiveness and drug adherence of biologic monotherapy in routine care of patients with rheumatoid arthritis: a cohort study of patients registered in the Danish biologics registry. Rheumatol Oxf Engl. 2015;54(12):2156–65.

51. Frisell T, Dehlin M, Di Giuseppe D, Feltelius N, Turesson C, Askling J, et al. Comparative effectiveness of abatacept, rituximab, tocilizumab and TNFi biologics in RA: results from the nationwide Swedish register. Rheumatology [Internet]. 2019 1 23 [cited 2019 Jan 23]; Available from: https://academic-oup-com.liverpool.idm.oclc.org/rheumatology/advance-article/doi/10.1093/rheumatology/key433/5298542
52. Choy EH, Bernasconi C, Aassi M, Molina JF, Epis OM. Treatment of Rheumatoid Arthritis With Anti-Tumor Necrosis Factor or Tocilizumab Therapy as First Biologic Agent in a Global Comparative Observational Study: Comparative Effectiveness of Tocilizumab and TNF Inhibitors in RA. Arthritis Care Res. 2017 10;69(10):1484–94.
53. Grøn KL, Glintborg B, Nørgaard M, Mehnert F, Østergaard M, Dreyer L, et al. Comparative Effectiveness of Certolizumab Pegol, Abatacept, and Biosimilar Infliximab in Patients With Rheumatoid Arthritis Treated in Routine Care: Observational Data From the Danish DANBIO Registry Emulating a Randomized Trial. Arthritis Rheumatol Hoboken NJ. 2019 12;71(12):1997–2004.
54. Lauper K, Nordström DC, Pavelka K, Hernández MV, Kvien TK, Kristianslund EK, et al. Comparative effectiveness of tocilizumab versus TNF inhibitors as monotherapy or in combination with conventional synthetic disease-modifying antirheumatic drugs in patients with rheumatoid arthritis after the use of at least one biologic disease-modifying antirheumatic drug: analyses from the pan-European TOCERRA register collaboration. Ann Rheum Dis. 2018 9;77(9):1276–82. [PubMed: 29730637]
55. Boyadzhieva V, Stoilov N, Ivanova M, Petrova G, Stoilov R. Real World Experience of Disease Activity in Patients With Rheumatoid Arthritis and Response to Treatment With Varios Biologic DMARDs. Front Pharmacol [Internet]. 2018 11 20 [cited 2020 Jan 21];9. Available from: https://www.frontiersin.org/article/10.3389/fphar.2018.01303/full
56. Torrente-Segarra V, Acosta Pereira A, Morla R, Ruiz JM, Clavaguera T, Figuls R, et al. VARIAR Study: Assessment of Short-term Efficacy and Safety of Rituximab Compared to an Tumor Necrosis Factor Alpha Antagonists as Second-line Drug Therapy in Patients With Rheumatoid Arthritis Refractory to a First Tumor Necrosis Factor Alpha Antagonist. Reumatol Clínica Engl Ed. 2016 11;12(6):319–22.
57. Kihara M, Davies R, Kearsley-Fleet L, Watson KD, Lunt M, Symmons DPM, et al. Use and effectiveness of tocilizumab among patients with rheumatoid arthritis: an observational study from the British Society for Rheumatology Biologics Register for rheumatoid arthritis. Clin Rheumatol. 2017 2;36(2):241–50. [PubMed: 27913894]
58. Pascart T, Philippe P, Drumez E, Deprez X, Cortet B, Duhamel A, et al. Comparative efficacy of tocilizumab, abatacept and rituximab after non-TNF inhibitor failure: results from a multicentre study. Int J Rheum Dis. 2016 11;19(11):1093–102. [PubMed: 27018857]
59. Romao VC, Santos MJ, Polido-Pereira J, Duarte C, Nero P, Miguel C, et al. Comparative Effectiveness of Tocilizumab and TNF Inhibitors in Rheumatoid Arthritis Patients: Data from the Rheumatic Diseases Portuguese Register, Reuma.pt. BioMed Res Int. 2015;2015(101600173):279890. [PubMed: 26000286]
60. Kristensen LE, Saxne T, Geborek P. The LUNDEX, a new index of drug efficacy in clinical practice: results of a five-year observational study of treatment with infliximab and etanercept among rheumatoid arthritis patients in southern Sweden. Arthritis Rheum. 2006 2;54(2):600–6. [PubMed: 16447237]
61. National Research Council (US) Panel on Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials. [Internet]. National Academies Press (US); 2010 [cited 2020 Jan 8]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK209899/
62. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The Prevention and Treatment of Missing Data in Clinical Trials. N Engl J Med. 2012 10 4;367(14):1355–60. [PubMed: 23034025]
63. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019 6;110:63–73. [PubMed: 30878639]

64. Laursen DRT, Paludan-Müller AS, Hróbjartsson A. Randomized clinical trials with run-in periods: frequency, characteristics and reporting. Clin Epidemiol. 2019 2 11;11:169–84. [PubMed: 30809104]

65. Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. Handbook of Missing Data Methodology [Internet]. 1st ed. Chapman and Hall/CRC; 2014 [cited 2020 Jan 7]. Available from: https://www.taylorfrancis.com/books/9781439854624

66. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. ArXiv191106030 Stat [Internet]. 2019 11 19 [cited 2020 Jan 8]; Available from: http://arxiv.org/abs/1911.06030

67. Murray EJ, Hernán MA. Improved adherence adjustment in the Coronary Drug Project. Trials. 2018 3 5;19(1):158. [PubMed: 29506561]

68. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. N Engl J Med. 2017 10 5;377(14):1391–8. [PubMed: 28976864]

69. Iudici M, Porcher R, Riveros C, Ravaud P. Time-dependent biases in observational studies of comparative effectiveness research in rheumatology. A methodological review. Ann Rheum Dis. 2019 4 1;78(4):562–9. [PubMed: 30755417]

70. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol. 2006 6 15;163(12):1149–56. [PubMed: 16624967]

71. Solomon DH, Lunt M, Schneeweiss S. The risk of infection associated with tumor necrosis factor α antagonists: Making sense of epidemiologic evidence. Arthritis Rheum. 2008;58(4):919–28. [PubMed: 18383377]

72. Schneeweiss S, Suissa S. Discussion of Schuemie et al: "A plea to stop using the case-control design in retrospective database studies". Stat Med. 2019;38(22):4209–12. [PubMed: 31489683]

73. Suissa S, Dell'Aniello S, Vahey S, Renoux C. Time-window Bias in Case-control Studies: Statins and Lung Cancer. Epidemiology. 2011 3;22(2):228–231. [PubMed: 21228697]

74. Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, et al. Opioid use, postoperative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study. Lancet Rheumatol. 2019 12 1;1(4):e229–36.

What is already known about this subject?

- Target trial emulation is an intuitive design approach that encourages researchers to formulate their question as a hypothetical randomised controlled trial (RCT). Using observational data to emulate such a target trial helps avoid common biases and has been shown to better align results with actual RCTs.

What does this study add?

- Most CER studies in rheumatoid arthritis had at least one design limitation, such as using post-baseline information to define eligibility, not specifying whether interest is in intention-to-treat or per-protocol effects, and using statistical selection of confounders. Each of these issues can introduce bias and affect data analysis and interpretation.

How might this impact on clinical practice or future developments?

- The target trial emulation framework unifies and builds upon existing good design practices to make robust observational designs intuitive. Improvement and standardisation of CER methodology is essential in rheumatology as more drugs become available, often without (timely) head-to-head RCTs to compare their effectiveness.
- Future studies should avoid using post-baseline information to define eligibility, clearly define the causal contrast pursued (ideally presenting both ITT and per protocol effects), and consider confounding using prior knowledge (preferably using causal diagrams that make assumptions transparent).

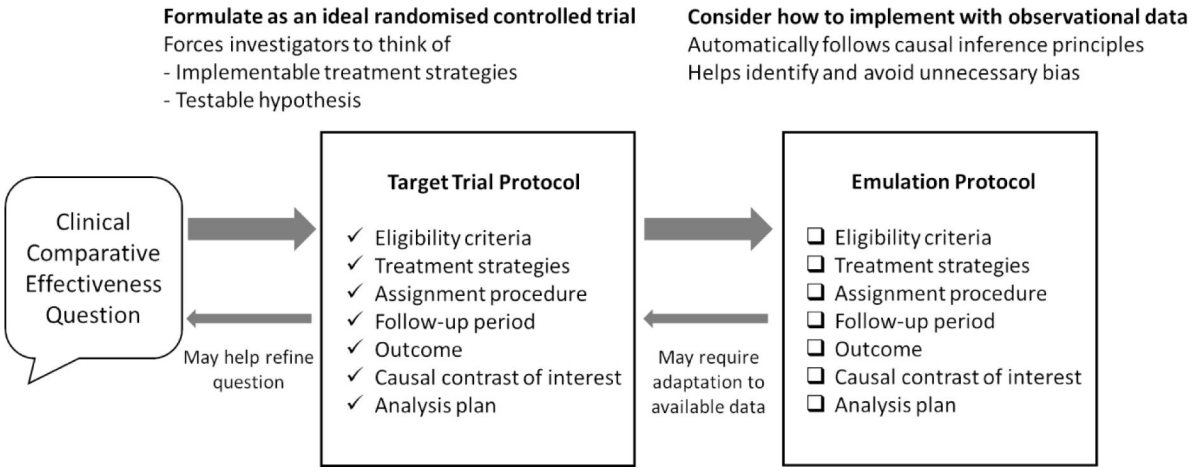

Figure 1.

The target trial protocol is used to guide observational comparative effectiveness research design. This idealized protocol may need to be reformulated once limitations of the data are realised. Divergence of the implemented observational study (emulation) from the target trial protocol should be addressed by sensitivity analyses or transparent reporting of limitations.

Table 1.

Items in the pre-specified data extraction form and a summary of findings from 31 studies.

	Data extraction questions (See Methods for rationale)	**Summary of findings from 31 studies***
1. Eligibility criteria	What were the criteria? (e.g., classification criteria RA with no prior exposure to bDMARDs)	All studies specified basic eligibility criteria such as RA definition, level of disease activity and number of prior bDMARDs.
	Was post-baseline information used to define eligibility? (e.g., requiring ≥1 follow-up)	19/31 (61%) studies explicitly required post-baseline information for eligibility [29–47].
	How many bDMARDs had been used before? (e.g., bDMARD naïve, ≥1 prior TNFi)	All studies described the number of prior bDMARDs. 5/31 (16%) studies combined response from both bDMARD-experienced and naïve patients [40,44,48–50].
2.Treatment strategies	What were the bDMARDs under comparison?	Drug name, dose and frequency were generally clearly defined, except when different TNFi were combined as one group.
	What were the treatment strategies? (e.g., discontinuation due to remission or switch to biosimilar are permitted in the protocol)	One study clearly defined treatment strategies [51].
3.Assignment procedures	How did the study use statistics to emulate random assignment? Specifically, how were confounding factors selected?	13/31 (42%) studies used only pre-defined confounders [31–34,36–38,41,44,46,52–54]. 5/31 (16%) studies used only statistical (e.g., p-value-based) variable selection [29,30,43,45,47]. 9/31 (29%) made no adjustments for confounding beyond active-comparator design [39,40,42,47–50,55,56].
4.Follow-up period	What was the specified duration of study? For studies using existing data from registries, duration of the implied trial was used.	Studies using binary response outcomes clearly defined follow-up times in all except one study [41]. 6 studies did not specify an end to follow-up at all [30–32,40,41,54].
5.Outcome	What was the primary effectiveness outcome measure and timeframe/point?	Outcomes were clearly described, but time occasionally not (see above).
	Was sample size or statistical power discussed at the design stage?	4/31 (13%) studies included sample size considerations [29,31,43,52].
6.Causal contrast of interest	Was a causal contrast of interest declared prior to analysis?	6/31 (19%) studies clearly defined causal contrast [31,47,49–51,53].
	What was the declared or inferred causal contrast?	20/31 (97%) studies examined some version of the ITT effect [29,31,33–37,39,41,43,44,46–48,51,53,54,57–59]; analyses were compatible with traditional ITT effect definition in only 1 study [53]. 12/31 (42%) studies include per-protocol analysis [30–32,38,40,42,45,49–52,55] but did not apply any post-baseline adjustments.
7.Analysis plan	What statistical model was used? (e.g., linear regression)	18/31 (42%) studies used regression-based methods. 8/31 used pairwise comparisons and 3/31 did not perform statistical comparison.
	How were missing data handled? (e.g., complete-case analysis, imputation)	17/31 (55%) studies used complete-case analysis [29–33,39,41,43,44,47–49,52,55,56,58,59]. 3/31 used multiple imputation for missing outcome data[51,52,57]. 9/31 use single imputation[34–38,45,46,53,54]. Reasons for discontinuation were rarely differentiated in analyses

* Components of the target trial emulation framework are discussed in detail in references 3 and 6. See Results for details; information extracted from individual studies are shown in online Supplementary Table S2.