Accepted author's manuscript. Published in final edited form as: Rheumatology 2018; 57(2): 354-369. Publisher DOI: 10.1093/rheumatology/kex394

Rheumatoid arthritis patients treated in trial and real world settings: comparison of randomized trials with registries

Gablu Kilcher[1], Noemi Hummel[1], Eva M. Didden[1], Matthias Egger[1], Stephan Reichenbach[1, 2] on behalf of the GetReal Work Package 4

[1] Institute of Social and Preventive Medicine, University of Bern, Switzerland

[2] Department of Rheumatology, Immunology and Allergology, University Hospital and University of Bern, Switzerland

Key words: Baseline characteristics, randomized controlled trials, observational studies, rituximab, tocilizumab, etanercept , rheumatoid arthritis

Correspondence to:

Stephan Reichenbach MD

Institute of Social and Preventive Medicine (ISPM)

University of Bern

CH- 3012 Bern

Switzerland

E-mail: stephan.reichenbach@ispm.unibe.ch

ABSTRACT

Objective: To investigate whether patients with rheumatoid arthritis (RA) enrolled in randomized controlled trials (RCTs) and observational studies may differ in terms of characteristics that could modify treatment effects, leading to an efficacy-effectiveness gap.

Methods: We conducted systematic literature reviews to identify RCTs and observational studies with RA, treated with rituximab, tocilizumab or etanercept. We extracted baseline characteristics and compared the data of RCTs and observational studies using fixed-effects meta-analyses for the RCTs and random-effects meta-analyses for the observational studies. We also assessed whether the baseline characteristics changed over time.

Results: Compared to patients enrolled in RCTs, those from observational studies were on average 3.0 years older (p<0.001), suffered from RA for 3.1 years longer (p<0.001), had 1.6 more prior disease modifying drugs (p=0.001), and had a lower Disease Activity Score 28 (DAS 28) (difference -0.6, p<0.001). C-reactive protein and erythrocyte sedimentation rate levels were slightly higher in RCTs. The Health Assessment Questionnaire-Disability Index (HAQ-DI) score was slightly lower in the RCT group. No differences were found in the percentages of included females or rheumatoid factor positivity. Over time, we found a significant decrease of -0.08 in DAS-28 and a decrease of -0.04 in HAQ-DI both in patients in RCTs and in patients from registries. Furthermore, ESR and CRP declined over time in RCT patients, but not in patients participating in observational studies.

Conclusion: There are substantial systematic differences in patient characteristics between randomized controlled trials and registries in RA. The efficacy seen in RCTs may not reflect real-world effectiveness.

Key messages

- There is no systematic review of the efficacy-effectiveness gap in rheumatoid arthritis.
- There are important differences in rheumatoid arthritis patients regarding the efficacy-effectiveness gap.
- Randomized controlled trials enrolled patients with better prognostic factors, potentially overestimating the treatment effect.

INTRODUCTION

The randomized controlled trial (RCT) is the gold standard for assessing the efficacy of pharmacologic treatments and other interventions (1). The main advantage of random treatment allocation is the high internal validity of estimates of treatment effects. Estimates from RCTs may, however, lack external validity (2) due to their highly standardized design, strict inclusion and exclusion criteria and fixed treatment regimens that may often be at odds with real world conditions (3,4).

In health technology assessment it is essential to gauge the effectiveness of drugs in the real world settings where they will be used (5). Several authors have recommended using non-randomized studies, clinical databases and registry data (i.e., observational studies) to assess whether RCT-based estimates apply to a target population (5–8). Patient characteristics may differ between RCTs and observational studies, and may modify treatment effects (7). A treatment may be less effective or more effective depending on age, stage of disease, or comorbidities (8–11). For example, studies comparing treatment effects between RCTs and observational studies in cardiovascular disease showed that patients with acute coronary syndrome included in clinical trials were younger, more likely to be men, and had fewer co-morbidities and risk factors when compared to registry patients (12,13). Similar results were found by Ezekowitz and colleagues, who compared characteristics of patients with heart failure between RCTs and observational studies (14) . In this context Eichler and colleagues (8) coined the term efficacy-effectiveness gap to describe the gap between treatment effects observed in RCTs and those observed in real world settings.

A comparison of baseline characteristics of patients with rheumatoid arthritis (RA) in RCTs and observational studies is lacking. We performed a systematic review extracting baseline characteristics from available RCTs and observational studies in RA. This review was a deliverable of Workpackage 4 of the GetReal project (incorporating real-life data into drug development), a consortium of academia, pharmaceutical companies, health technology assessment agencies, regulators and patient organizations (15). Using case studies, WP4 developed best practices in evidence synthesis and predictive modelling, with the goal of improving estimates of the real world effectiveness of drugs by incorporating the results of RCTs with other sources of clinical data, including observational data. WP4 obtained access to individual participant data from clinical trials of three widely used biologics, namely etanercept (ETN), rituximab (RTX) and tocilizumab (TCZ), as well as patient registries in RA. Our systematic review thus also focused on ETN, RTX and TCZ.

METHODS

Search strategy

We performed two systematic reviews; one literature search was done for observational studies, the other for RCTs. We applied study design search filters from the BMJ Evidence Centre Information Specialists to the Embase and Medline databases using Ovid (16). We performed the search for observational studies on March 4, 2015, and the search for RCTs on April 24, 2015. The detailed search strategies can be found in Supplementary Tables S1-S4, available at *Rheumatology* online. In addition, we manually searched known registries and screened reference lists of all papers.

Inclusion criteria and study selection

We included studies of adult patients diagnosed with RA who were treated with RTX, TCZ or ETN. Studies were required to have reported the following outcomes: Disease Activity Score 28 (DAS-28), including C-reactive protein (DAS-28-CRP) or erythrocyte sedimentation rate (DAS-28-ESR), or Health Assessment Questionnaire-Disability Index (HAQ-DI) scores. The studies had to include at least 30 patients per study arm.

The retrieved titles and abstracts of the identified articles were imported into EppiReviewer 4 (17). Duplicates across databases were removed, and for each treatment, the latest publication fulfilling the inclusion criteria was used. Each paper was independently assessed by two reviewers (G.K. and N.H. or G.K. and E.D.), based on title and abstract and, if the study was potentially eligible, on the full text of the article. Disagreements were resolved by consensus, after discussion with M.E. or S.R. whenever necessary.

Data extraction

Data from each included paper were extracted using a standardized form developed for this project. Extracted data covered three areas; first, general data included author, publication year, study design, country, overall number of patients in the study, follow-up time and the main objective of the study; second, treatment data included drug, dose, frequency and route of administration; and third, data on patient characteristics at baseline included number of patients receiving each drug, age, gender, current smoking, disease duration, comorbidities, ESR, CRP, seropositivity for rheumatoid factor (RF) or anti-citrullinated protein antibody (ACPA), DAS-28 and HAQ-DI, switching from another biologic agent to the current drug, number of prior disease-modifying anti-rheumatic drugs (DMARDs), and use of corticosteroids and other drugs. We extracted dichotomous data as numbers and percentages. For continuous data, we extracted the mean or median, together with the standard deviation or range (minimum/maximum or interquartile range).

Statistical Analysis

We converted medians and ranges to means with standard deviations using the methods described by Wan et al. (18). For binary data we used the variance estimator (v) for proportions (p) to derive the standard error: v= p (1 – p). We performed meta-analyses of patient characteristics separately for RCTs and observational studies, overall and by drug. If necessary, we first combined the data from the study arm into a single mean or proportion using fixed-effect meta-analyses. Secondly, we combined the data separately for RCTs and observational studies using random-effects meta-analyses with Knapp-Hartung adjustment (19). We used mixed-effects meta-regression analyses to assess the differences in patient characteristics between RCTs and observational studies by including the study design as a dichotomous covariate. We used restricted maximum-likelihood estimation to assess between-study variance (tau-squared) and applied the Knapp-Hartung adjustment. We stratified our main analysis by study type to explore whether the approaches differed in terms of baseline characteristics. In further meta-regression analyses, we included the year of publication of the studies to examine whether patient characteristics of patients included in RCTs or observational studies changed over calendar time. In a sensitivity analysis, we excluded phase IV and pragmatic trials, as reported by the trialists. All analyses were done with the R package metaphor (20).

RESULTS

We identified 308 references in our literature search for RCTs and 594 for observational studies, and considered 89 RCTs and 194 observational studies to be potentially eligible (Supplementary Figure S1 and S2). Fifty-one RCTs and 76 observational studies met our inclusion criteria and were included in the meta-analysis.

Study characteristics

The eligible studies were published between 1999 and 2015 for RCTs and between 2003 and 2015 for observational studies. Among RCTs, we included 5 Phase II studies, 23 Phase III studies, 10 Phase IV studies and one pragmatic trial. For the remaining 12 RCTs we could not retrieve any information on the phase. Observational studies comprised 17 cohort, 28 registry and to 31 case series studies. Most observational studies (71; 93.4%) were conducted in a single country whereas almost half of RCTs (25; 49.0%) were multi-country trials, mostly involving European countries. The number of study participants ranged from 70 to 1262 patients for RCTs and from 30 to 8908 for observational studies. Among RCTs, we included 17 TCZ, 10 RTX, and 24 ETN trials, and among observational studies, we included 16 TCZ, 28 RTX 28 and 32 ETN studies. Tables 1 and 2 summarize characteristics of RCTs and observational studies.

Comparison of patient characteristics

Compared to patients participating in RCTs, those from observational studies were on average 3.0 years older (p<0.001), suffered from RA for 3.1 years longer (p<0.001) and had 1.6 more prior DMARDs (p=0.001, Figure 1). Patients in RCTs had higher disease activity: the DAS-28 was 0.6 points higher in RCT than in observational studies (p<0.001, Figure 2). CRP and ESR levels were also slightly higher in RCTs, but differences failed to reach conventional levels of statistical significance (Figure 2).

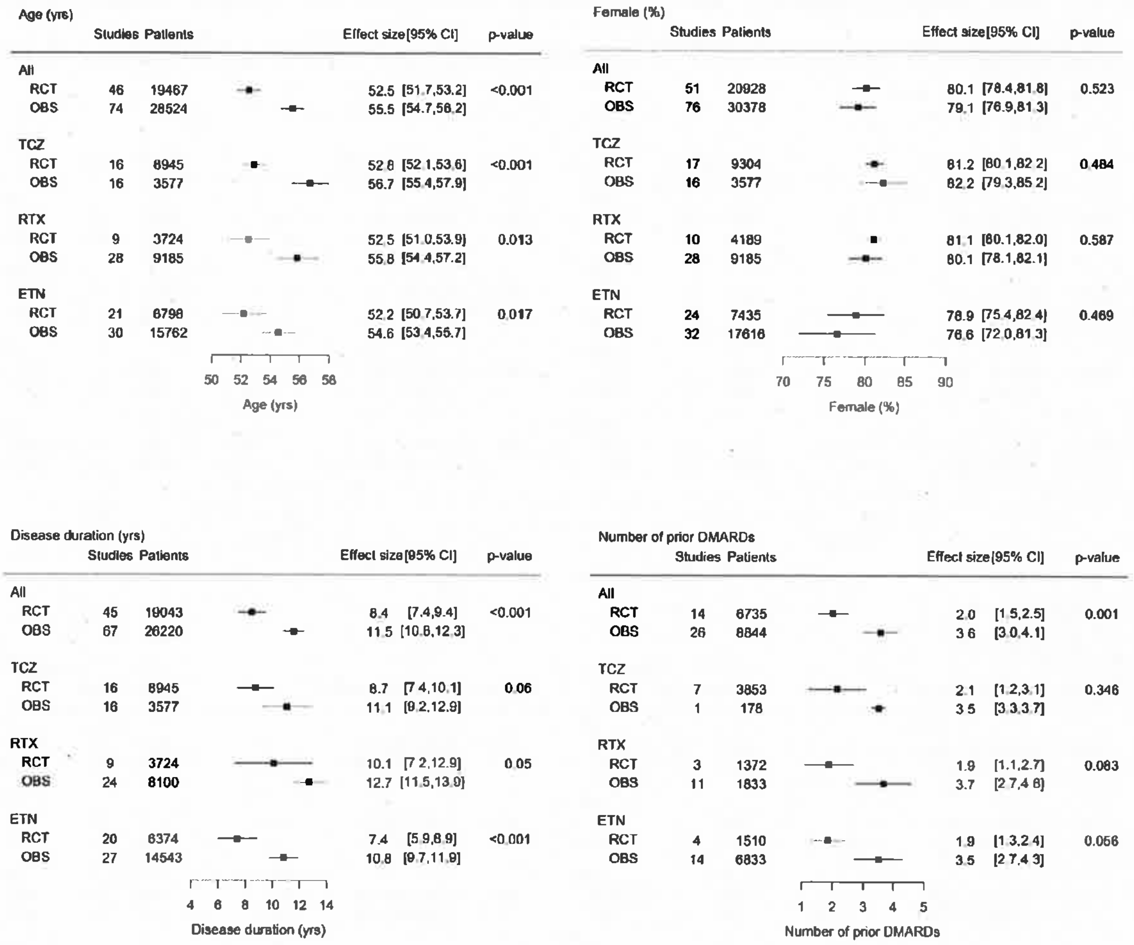


Figure 1. Comparison between randomized controlled trials and observational studies for age, gender, disease duration, and number of prior DMARDs

Similarly, there was little evidence for any difference between HAQ-DI scores, rheumatoid factor positivity or the proportion of women participating in the studies. Analyses stratified by drug showed that differences generally were in the same direction for the three drugs, but tended to be more pronounced for TCZ and RTX than for ETN (Figures 1 and 2). Patients on TCZ were 3.9 years older in observational studies than in RCTs (p<0.001), their disease duration was 2.4 years longer (p=0.06), they had been exposed on average to 1.4 additional DMARDs (p=0.346) and the DAS-28 was 1.2 lower than in RCTs (p<0.001). Similarly, patients on RTX were 3.3 years older (p=0.013), their disease duration was 2.6 years longer (p=0.05), they had been exposed to 1.8 more DMARDs (p=0.083) and the DAS-28 was 1.1 point (p<0.001) lower in observational studies than in RCTs. Patients on ETN were 2.4 years older in observational studies than in RCTs (p=0.017), their disease duration was 3.4 years longer (p<0.001) , and they had been exposed to 1.6 more DMARDs (p=0.056).

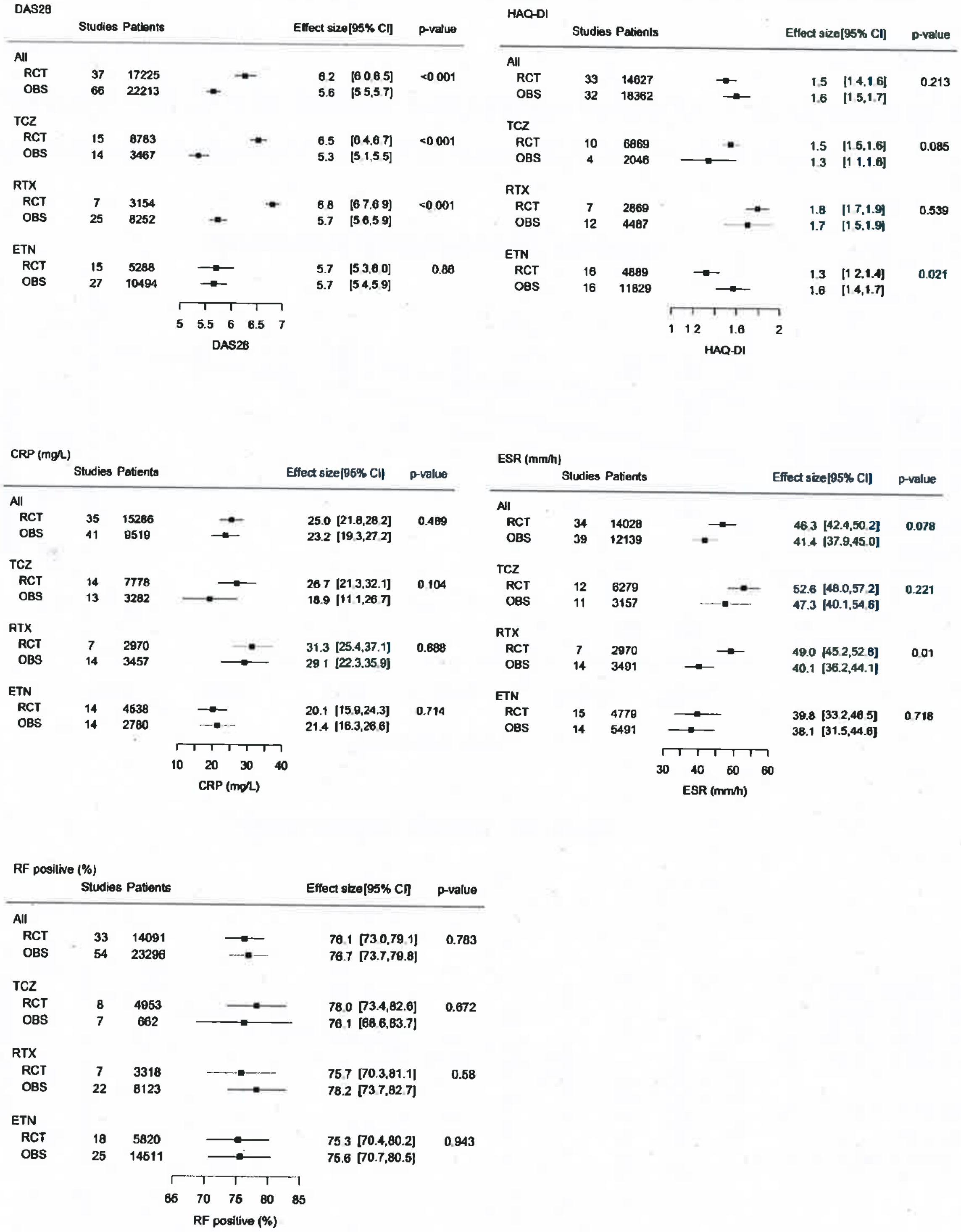


Figure 2. Comparison between randomized controlled trials and observational studies for DAS28, HAQ disability index, CRP, ESR, and RF positivity

There was no difference in DAS28 (0 points, p=0.86). Analyses stratified by study type gave similar results compared to the main analyses (Supplementary Figure S3 – S11, available at *Rheumatology* online, forest plots for all baseline characteristics).

Trends over calendar time

We found that DAS-28 declined over calendar time both in RCTs (slope of -0.08, p=0.026) and in observational studies (slope of -0.08, p=0.002) (Figure 3).

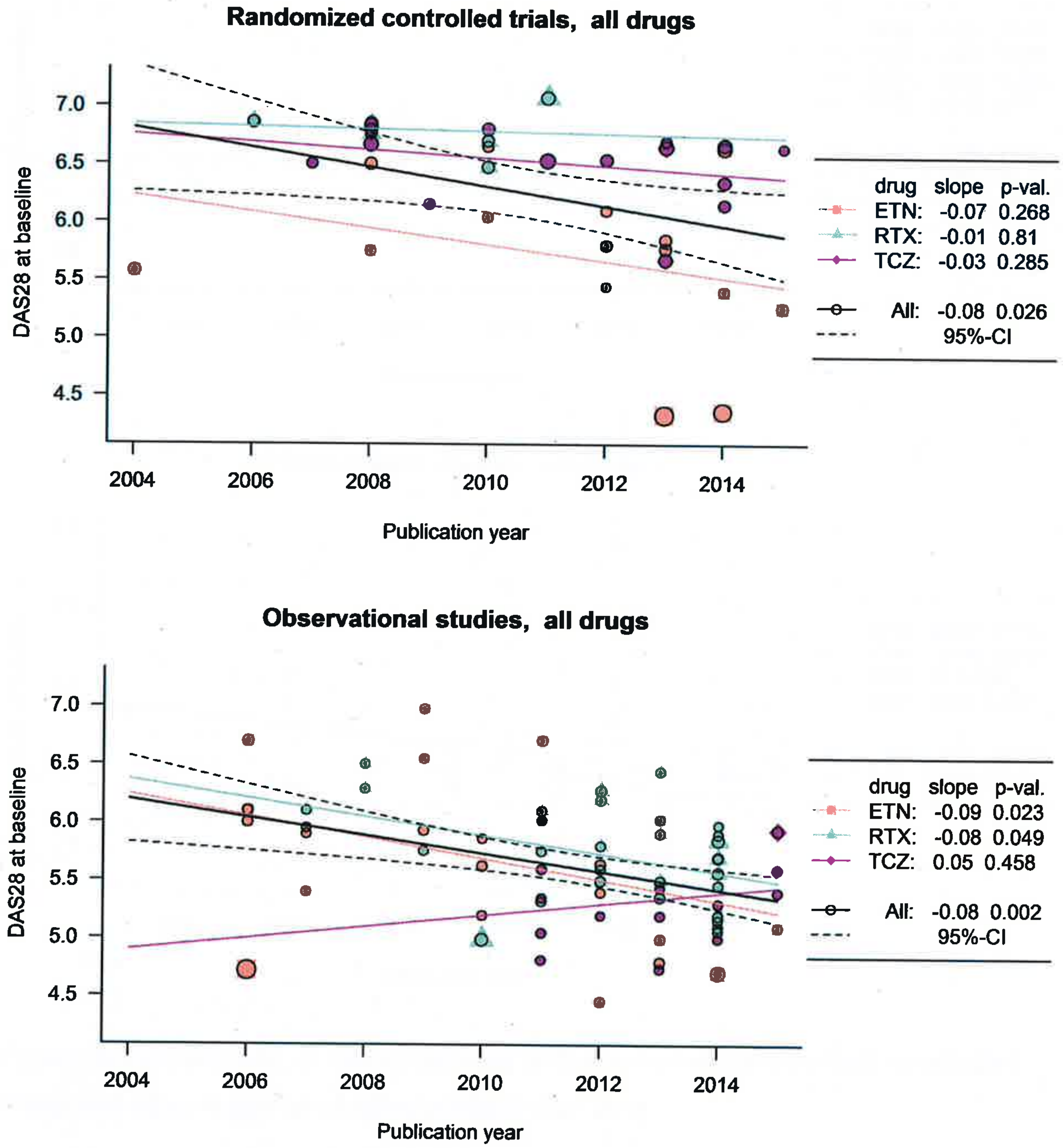
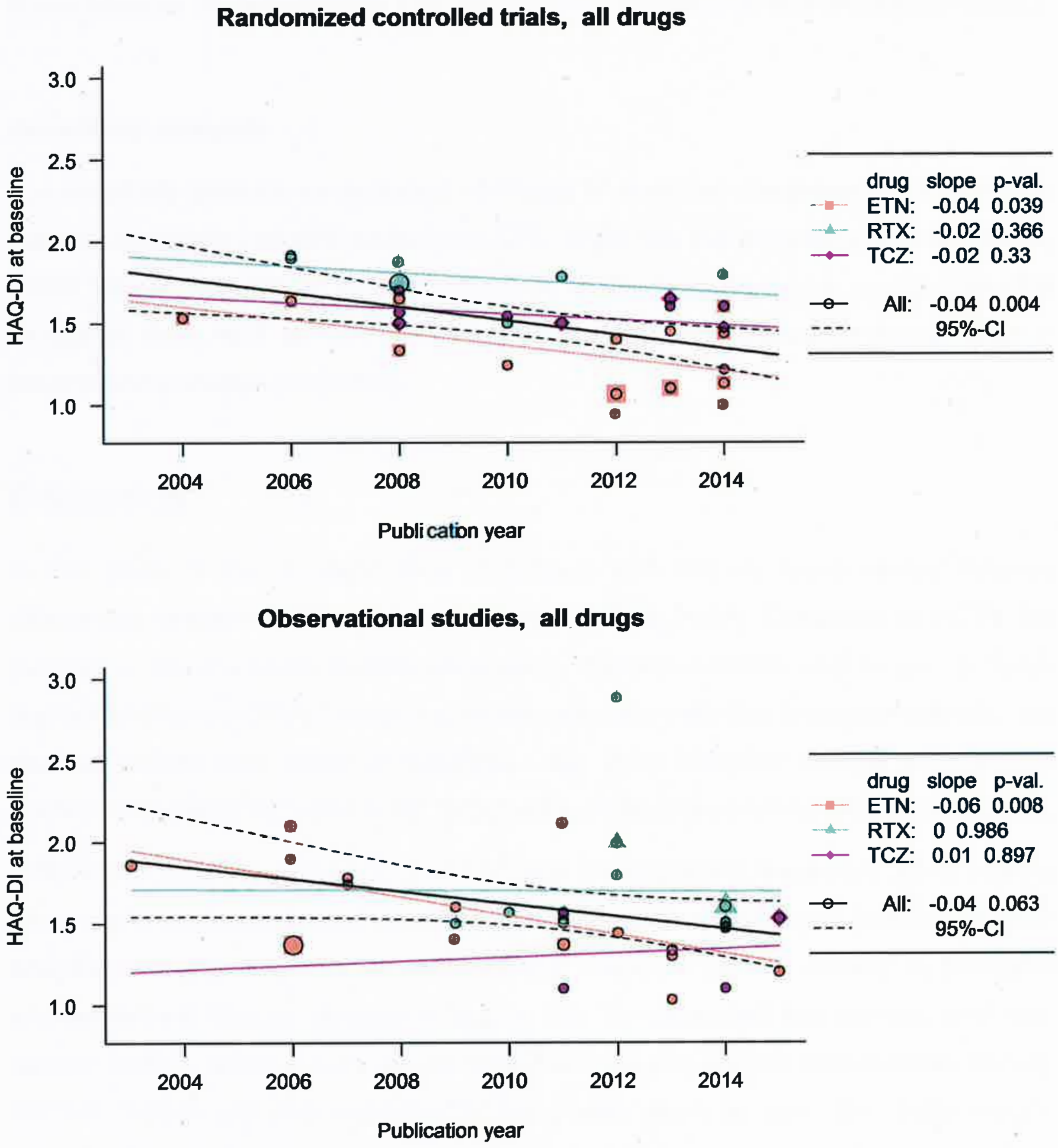


Figure 3. Comparison of DAS28 between randomized controlled trials and observational studies plotted over time

HAQ-DI declined slightly over calendar time both in RCTs (slope of -0.04, p=0.004) and in observational studies (slope of -0.04, p=0.063) (Figure 4).


Randomized controlled trials, all drugs

Observational studies, all drugs

Figure 4. Comparison of HAQ disability index between randomized controlled trials and observational studies plotted over time

Furthermore, ESR and CRP declined over calendar time in RCTs (slope of -1.69, p=0.009 for ESR and slope of -1.68, p=0.001 for CRP), but not significantly so in observational studies (Figure 5). There was little evidence for changes in baseline patient characteristics over time for any of the other socio-demographic or clinical characteristics (Supplementary Figures S12-S17, available at *Rheumatology* online).

Sensitivity analysis

In a sensitivity analysis we excluded 10 Phase IV trials and one pragmatic trial. Ten of the excluded trials included patients on ETN, while one trial included patients on TCZ. There was no substantial change compared to the main analyses, except that ESR increased from 46.3 to 49.4 mm/h in RCTs, which is significantly higher than in observational studies (p=0.004).

Discussion

In this study of the characteristics of patients with RA we found clinical relevant differences between RCTs and observational studies in RA. Compared to RCTs, RA patients in observational studies were older, disease duration was longer, a higher number of different DMARDs were administered before starting biologic treatment, and disease activity was lower at baseline. Over time, baseline DAS28 and HAQ-DI declined in patients included in RCTs but not in patients from observational databases.

Differences between real-world and trial data are important, especially when making decisions in everyday clinical practice. Eichler and colleagues argue that the efficacy-effectiveness gap is due to variability in drug response (8,148) caused by biological and behavioral factors. Biological factors can be separated into genetic and non-genetic factors, which in turn can be further divided into intrinsic and extrinsic factors. Intrinsic factors are characteristics of the person such as age, sex, body weight, comorbidities and baseline severity of disease, whereas extrinsic factors relate to lifestyle factors such as smoking (8).

Kirsch and colleagues studied all available data of clinical trials submitted to the FDA for the licensing of four new-generation antidepressants. They found a relationship between initial disease severity and antidepressant efficacy, an association that was due to decreased responsiveness to placebo among very severely depressed patients as opposed to increased responsiveness to medication (9). Similarly, in patients with RA a high DAS-28 score at baseline is a good predictor of a decline in the DAS-28 following treatment with ETN (149) and TCZ (120). Our review showed higher DAS-28 scores in patients enrolled in RCTs and we can therefore speculate that the response was better in trial patients than in observational studies. In other words, the treatment effect in everyday clinical praxis might be smaller than that in RCTs.

High numbers of prior DMARDs and higher age were associated with decreased response rates in patients with ETN (148). Older age was also associated with decreased response rate in age with TCZ (150); these two baseline characteristics differed significantly between RCTs and observational studies in our analysis. Predictive factors for better response to biologics were male gender (in ETN treated patients (148)), non-smokers (ETN (148)), RF positivity (RTX and TC (151)) and low HAQ-DI (TCZ (120) and RTX (148)). For all these factors, if data were available, we found no difference between RCTs and observational studies.

In our time trend analysis, we saw a decrease in baseline DAS-28 and HAQ-DI in RCTs over the last 10 years. A decrease in DAS-28 has also been shown for other biologics such as infliximab (152). These findings support the results of an inception cohort study published 10 years ago, where the trend was thought to be caused by a more aggressive treatment strategy (153).

In a sensitivity analysis we excluded ten Phase IV clinical trials and one pragmatic trial. Interestingly, we found no difference compared to the main results described above. In particular, the results in the ETN group where ten trials were excluded remained virtually the same. This may call into question the notion that Phase IV trials accurately represent real-world scenarios, and that their estimates of comparative effectiveness are closer to those of observational studies. We acknowledge that the number of Phase IV and pragmatic trials was small and that the results from our sensitivity analysis should be interpreted with caution.

Our review has several strengths and weaknesses. Strengths are that the review was based on a systematic literature search and study selection and screening were performed independently by two authors. Data extraction was performed by one person and checked by a second. Our search was comprehensive, but we included only English-language studies. Also, we did not look into reports by the European Medicines Agency (EMA) or Food and Drug Administration (FDA). Data for each biologic may have been assessed at different time points in each registry. For instance, in the Rabbit registry, we used data for ETN from 2006, whereas data for RTX was assessed from a publication in 2013. We cannot exclude the possibility that some of the included patients were counted twice, because patients might have switched their treatment from ETN to RTX. However, in the absence of individual patient data we can only speculate the percentage of patients who switched treatment regimens. Overall, 7 of the 28 included registries had more than one publication, and it is therefore possible that patients were counted twice.

Since we did not assess outcomes, we did not apply any risk of bias tool or similar instrument to examine the quality of studies. Our main interest was the characteristics of patients included in RCTs and observational studies and it is therefore unlikely that the comparison was distorted by publication or other selection bias.

We transformed median values into mean values. This might lead to bias in the aggregated mean if the data summarized by the median was clearly not normally distributed. Since we transformed only about 8% of the values provided in the RCTs and only about 17% of the values from the observational studies, any bias introduced is likely small. Several relevant variables were poorly reported: concomitant MTX use, concomitant DMARD use, percentage of smokers (reported in only one RCT), comorbidities and ACPA positivity. These variables were therefore not included in our analyses, despite their potential relevance in the context of generalizing results from RCTs to real world settings.

Clearly, more work is required on how best to narrow the efficacy-effectiveness gap. In Phase IV and pragmatic trials, inclusion and exclusion criteria need to be widened to reflect the real world. The baseline characteristics of patients included in these trials should better reflect what we found in observational studies. In addition, evidence synthesis and modeling approaches should be used to combine data from both RCT and observational studies to generate real-world evidence (15).

In summary, we found important differences between RA patients included in RCTs as compared to observational studies; in particular, patients with better prognostic factors were included in the RCTs, leading to potential overestimation of the treatment effect. More research is needed to overcome this efficacy-effectiveness gap in RA to generate real-world evidence.

Funding

The research leading to these results was conducted as part of the GetReal consortium. For further information please refer to www.imi-getreal.eu. This paper only reflects the personal views of the stated authors. The work leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° [115546], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA companies' in kind contribution. In addition as a special form of the IMI JU grant Eva-Maria Didden received a direct financial contribution from Boehringer Ingelheim.

Disclosure statement

GK is an employee of the Swiss Federal Office of Public Health. The content of this publication does not reflect in any way the views or policies of the Federal Office of Public Health. ME, NH, EMD and SR have no conflicts of interests to declare.