2016

# Bayesian Network Meta-Analysis for Biologic Therapies in Rheumatoid Arthritis

Yizhou Ye
*University of Rhode Island*, james0602@gmail.com

# BAYESIAN NETWORK META-ANALYSIS FOR BIOLOGIC THERAPIES IN RHEUMATOID ARTHRITIS

BY

YIZHOU YE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2016

MASTER OF SCIENCE THESIS

OF

YIZHOU YE

APPROVED:

Thesis Committee:

Major Professor Gavino Puggioni

Natallia Katenka

Stephen Kogut

Liliana Gonzalez

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2016

## ABSTRACT

Meta-analysis can only compare studies with the same interventions, while a network meta-analysis can analyze studies with different interventions. Without medical data for direct comparisons, network meta-analysis can utilize existing trials to assess the relative efficacy of competing treatments. Three classes of statistical models are proposed to perform a network meta-analysis: fixed-effects, random-effects, and mixed-effects (meta-regression). The most appropriate model should be selected with the aid of a series of statistical tests including I2 statistic for heterogeneity and DIC for model fitness. Bayesian network meta-analysis provides pooled effect sizes (odds ratio for dichotomous outcome) for each treatment and their 95% probability credible intervals. After a systematic literature review, 20 randomized clinical trials of biologic anti-rheumatic therapies in combination with methotrexate in rheumatoid arthritis patients were identified. Random-effects model was used for ACR20 and ACR70 criteria treatment outcome whereas the mixed-effects model was used for ACR50 treatment outcome. Based on the analysis, we found that all biologics DMARDs were superior to placebo except for ANA in all datasets and RTX in ACR70 dataset. ETN was had the highest probability to be the best treatment in all three datasets. CTZ had the highest probability to be the second best option in ACR20 and ACR50 datasets, and TCZ held the second place in the ACR70 dataset. The rest of the rank probabilities vary by dataset but placebo was the lowest ranked option in all datasets. Therefore, despite the limitations of this study, the results are consistent with current knowledge that biologic DMARDs are superior to placebo and although more research remains to be done, ETN may be the most effective option for rheumatoid arthritis.

## ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my major advisor Dr. Gavino Puggioni, who has been tremendously knowledgeable and supportive for this entire thesis project. Every conversation I had with him was joyful, whether it was about statistics, basketball, traveling, or gastronomy. I truly appreciate him taking over to supervise my work while already being on dozens of committees. It would not been possible for me to complete this project without his inspiration and encouragement.

I would also like to thank my thesis committee, Drs. Liliana Gonzalez, Stephen Kogut and Natallia Katenka for their comments and inputs on the final work. Special thanks to Dr. Gonzalez for encouraging me to pursuit a degree in statistics and advising me on research topic selection.

In addition, I would also like to thank my friends Ke Bian, Yi-Tzai Chen, Aseel Eid, Marek Marczak, Robert McConeghy, Ajinkya Pawar, Yuanjun Shen, Qi Tang, Zhengxi Wei, Zidan Wu, Jing Yang, Si Yang, and Zonghao Zhu for their friendship and support all these years. I greatly enjoyed your company and I learned immensely from each one of you.

Lastly, I would like to thank my family for their unconditional love and unlimited support.

# PREFACE

Rheumatoid arthritis is a systemic chronic autoimmune disease that brings immense pain and burden to patients and society. Several biologic disease-modifying anti-rheumatic drugs (DMARDs) are currently available to control and alleviate the symptoms. Treatments for rheumatoid arthritis have undergone considerable clinical trials, yet there still remain gaps in our knowledge. Only a few head-to-head trials have compared the relative efficacy between treatment options.

Following the manuscript format, this thesis sought:

To estimate the relative efficacy of biologic DMARDs from randomized clinical trials that measured treatment outcome with American College of Rheumatology improvement criteria at 24 weeks in both frequentist and Bayesian network meta-analysis. The results of the Bayesian models were included in the manuscript as fifth chapter of the thesis.

## TABLE OF CONTENTS

ABSTRACT ..... ii

ACKNOWLEDGEMENTS ..... iii

PREFACE ..... iv

TABLE OF CONTENTS ..... v

LIST OF FIGURES ..... vii

LIST OF TABLES ..... viii

1 INTRODUCTION ..... 1

1.1 Randomized Clinical Trials and Meta-Analysis ..... 1

1.2 Network Meta-Analysis ..... 2

1.3 Rheumatoid Arthritis ..... 4

2 METHODOLOGY ..... 6

2.1 Literature Review and Data Collection ..... 6

2.2 Data Cleaning ..... 7

2.3 Model Selection ..... 9

3 STUDY DATASET ..... 20

3.1 Literature Search Results ..... 20

3.2 Evidence Network ..... 25

4 FITTING FREQUENTIST MODELS ..... 26

4.1 ACR20 Dataset ..... 26

4.2 ACR50 and ACR70 Dataset ..... 33

MANUSCRIPT INTRODUCTORY PAGE ..... 36

MANUSCRIPT ..... 37

APPENDIX A ..... 54

APPENDIX B ..... 55

BIBLIOGRAPHY ..... 56

## LIST OF FIGURES

**Figure** **Page**

Figure 1. Examples of Different Types of Evidence Networks ..... 3

Figure 2. Structures of Variance Components in Random-effects Model ..... 12

Figure 3. Flow Chart of Literature Review and Study Selection Process for Eligible Studies ..... 21

Figure 4. Evidence Network of Included Randomized Clinical Trials ..... 25

Figure 5. Forest Plot of Treatment Effects, Frequentist FE Models, ACR20 Dataset ..... 27

Figure 6. Forest Plot of Treatment Effects, Frequentist RE models, ACR20 Dataset, $\rho$=0.5 ..... 30

Figure 7. Forest Plot of Treatment Effects, Bayesian RE Model, ACR20 Dataset ..... 49

Figure 8. Rank Probabilities Plot, Bayesian RE Model, ACR20 Dataset ..... 49

Figure 9. Forest Plot of Treatment Effects, Bayesian ME Model with Duration of Disease, ACR50 Dataset ..... 50

Figure 10. Rank Probabilities Plot, Bayesian ME Model with Duration of Disease, ACR50 Dataset ..... 50

Figure 11. Forest Plot of Treatment Effects, Bayesian RE Model, ACR70 Dataset ..... 51

Figure 12. Rank Probabilities Plot, Bayesian RE Model, ACR70 Dataset ..... 51

## LIST OF TABLES

**Table** **Page**

Table 1. Study Characteristics and Baseline Patient Demographics ..... 23

Table 2. Treatment Response in ACR Criteria ..... 24

Table 3. Relative Treatment Effects to Placebo, Frequentist FE Model, ACR20 Dataset 27

Table 4. Relative Treatment Effects to Placebo, Frequentist RE Model, ACR20 Dataset ($\rho = 0.5$) ..... 29

Table 5. Relative Treatment Effects to Placebo, Frequentist RE Model, ACR20 Dataset ($\rho$ estimated) ..... 29

Table 6. Covariate Assessment and Ranked Treatments, Frequentist ME Model, ACR20 Dataset ..... 32

Table 7. Heterogeneity Comparison, Frequentist Models, ACR20 Dataset ..... 32

Table 8. Model Fitness Comparison, Frequentist ME Models, ACR20 Dataset ..... 32

Table 9. Covariate Assessment and Ranked Treatments, Frequentist Models, ACR50 Dataset ..... 35

Table 10. Heterogeneity and Fitness Comparison, Frequentist Models, ACR50 Dataset 35

Table 11. Covariate Assessment and Ranked Treatments, Frequentist Models, ACR70 Dataset ..... 35

Table 12. Heterogeneity and Fitness Comparison, Frequentist Models, ACR70 Dataset 35

Table 13. Rank Probabilities of Treatment Groups, ACR20 Dataset ..... 48

Table 14. Rank Probabilities of Treatment Groups, ACR50 Dataset ..... 48

Table 15. Rank Probabilities of Treatment Groups, ACR70 Dataset ..... 48

Table 16. Model Fitness and Heterogeneity Comparison, Bayesian Models ..... 55

# 1 INTRODUCTION

## 1.1 Randomized Clinical Trials and Meta-Analysis

In the history of thousands of years of medicine, evidence-based health care decision making has never played a bigger role in clinical practice as it is today.[1] Evidence-based health care aims at using the best up-to-date evidence to make decisions about individual patient care. Randomized clinical trial, the gold standard for evidence-based health decision-making, commonly compares one therapeutic intervention with placebo or standard care. Randomization is used to make sure that the different treatment groups will have the same baseline characteristics to remove confounding. When conducting clinical trials, both the patients and the practitioners are blinded to treatment group assignment (double-blind) to minimize potential bias. Sometimes, even the researchers are masked about group assignment during data analysis phase in order to observe the true treatment effects (triple-blind). It is quite common that multiple clinical trials are conducted for the same intervention in different scenarios and this is when meta-analysis, a statistical analytical method that allows one to combine evidence from a series of studies can increase the precision of measurement. It allows us to evaluate information across studies in a common framework to reduce uncertainty of efficacy evaluation .[2,3] Meta-analysis requires clinical trials of the same comparators. Clinical trials and standard meta-analysis can only demonstrate efficacy over placebo or standard care. They do not reveal evidence about relative head-to-head comparison between multiple treatments or therapies. On the other hand, pharmaceutical companies have

minimal incentives to conduct gratuitous head-to-head clinical trials that are both time-consuming and expensive.

## 1.2 Network Meta-Analysis

Even though a few clinical trials have involved more than one intervention, few studies have looked at more than two interventions. Driven by the need for relative treatment efficacy assessment, advanced methods have been developed to incorporate multiple interventions. Unlike standard meta-analysis, which can only incorporate trials of one intervention, network meta-analysis can incorporate clinical trials of multiple interventions and can estimate the relative efficacy between any pair of interventions. With a series of direct comparisons that compare treatment A and B (AB trials) and treatment A and C (AC trials), an indirect comparison can be made between treatment B and C via the common comparator A. When available evidence includes both direct and indirect comparisons of multiple interventions, sometimes even to different comparators, a network of evidence is formed hence the name network meta-analysis. We describe the comparison between treatment B and C using referent A as a comparison anchored on A. As shown in Figure 1, an evidence network can be anchored on one or multiple treatment groups and can have both open and close loops.[4,5]

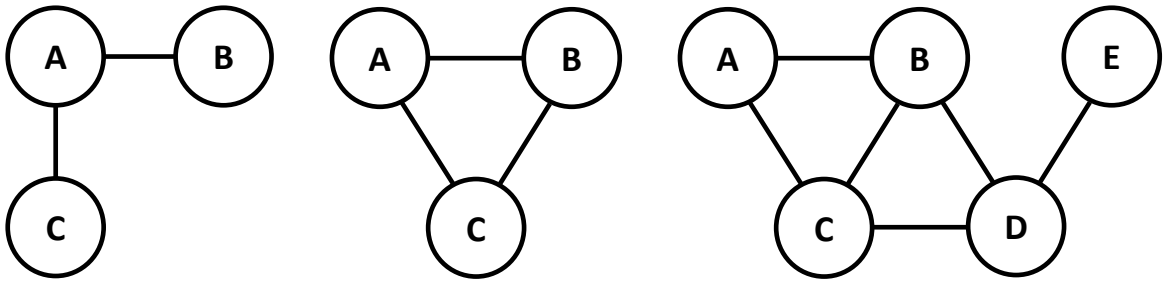

Figure 1. Examples of Different Types of Evidence Networks

For network meta-analysis, randomization stands within but not across included clinical trials . Study characteristics and patient demographics between clinical trials might differ and could also contribute to the variance of relative effect size estimation.[4,6,7] Potential treatment effect modifiers ("modifiers") may exist in patient characteristics, inclusion and exclusion criteria, study protocol development, study outcome measurement, and administrative agency requirement. Modifiers could bias the results of indirect comparisons and alter the magnitude and direction of treatment effect estimation.[7,8] Modifiers can also be referred as study-level covariates or covariates in network meta-analysis.

There are two assumptions for network meta-analysis. The first assumption is the similarity assumption, which assumes included studies are similar and the modifiers of treatment effects are balanced across included studies. The other assumption is consistency assumptions, which requires included studies to be consistent with the direct and indirect estimation based on $d_{AC}$ and $d_{AB}$: $d_{BC} = d_{AC} - d_{AB}$ for direct and indirect comparisons. Discrepancy between the direct and indirect estimations compromises the similarity assumption and introduces bias into treatment effect estimation.[7,9] When these assumptions stand, the relative effect between BC ($d_{BC}$) can be estimated indirectly given relative effect between AC ($d_{AC}$) and AB ($d_{AB}$) without a direct BC trials: $d_{BC} = d_{AC} - d_{AB}$.

For statistical analysis, network meta-analysis can be conducted in the frequentist framework and the Bayesian framework. The frequentist approach provides point

estimation and its 95% confidence intervals (CIs) for the true treatment effect. The Bayesian network meta-analysis has the advantage of the ability to incorporate previous knowledge about the treatment effects in a prior and provide a posterior probability distribution for the effect size and associated 95% credible intervals (CrIs). Compliance with the similarity and consistency assumptions is difficult due to small sample size (number of studies in a meta-analysis, not overall number of patients as only summary data is available in meta-analyses) and potential sampling error of each included study. Furthermore, unknown confounding is difficult remove completely. It is crucial to evaluate the heterogeneity in an evidence network. Heterogeneity assessments helps researchers to understand the impact of assumption violation.[4,10,11]

## 1.3 Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a systemic chronic autoimmune disease that is responsible for approximately 9 million physician visits and 250,000 hospitalizations in the United States every year.[12] The resulting swollen and painful joints often lead to disabilities, diminished quality of life, and premature mortality.[13] Cost-effective conventional diseases modifying anti-rheumatic drugs (DMARDs) have served as first-line therapy to relieve the symptoms. However, patients tend to develop tolerance and have diminished response over time. Switching to or additional biologic agents is recommended in the event of inadequate response to conventional DMARDs. Biologic DMARDs further divide into several classes: (1) tumor necrosis factor inhibitors (TNFi): adalimumab (ADA), certolizumab pegol (CTZ), etanercept (ETN), golimumab (GLB), and infliximab (IFX); (2) T-cell costimulatory modulato: abatacept (ABT); (3)

recombinant human interleukin-1 (IL-1) receptor antagonist: anakinra (ANA); (4) B-cell-depleting monoclonal antibody: rituximab (RTX); (5) monoclonal antibody to the IL-6 receptor: tocilizumab (TCZ); and (6) janus kinase inhibitor: tofacitinib (TOF). All these drugs have undergone randomized clinical trials against methotrexate for regulatory approval but there have been few head-to-head trials to compare the relative efficacy among them.

# 2 METHODOLOGY

## 2.1 Literature Review and Data Collection

A systematic literature review was conducted to collect data from peer-reviewed journal articles. A pre-determined search strategy (see Appendix A) was used to search for qualifying clinical trials in PubMed electronic databases through January 1999 to December 2015. Identified studies were screened based on titles and abstracts followed by a full-text review on selected studies. The references of full-text reviewed studies were also examined for potential additions.

Studies included had to meet the following inclusion criteria: (1) the study was a multisite randomized clinical trial for biologic therapies of interest; (2) the study was conducted in adult human rheumatoid arthritis patients; (3) the study outcome was measured with American College of Rheumatology (ACR) criteria at week 24. The following exclusion criteria were also applied: (1) the study population included patients of other inflammatory diseases such as juvenile arthritis, Crohn's disease, or psoriatic arthritis; (2) the study patients did not receive background MTX; (3) the study did not report number of patients with a 50% or 70% improvement in ACR criteria (ACR50 or ACR70).

Study outcome data extracted was the number and percentage of responders to treatment based on the ACR criteria at week 24. Treatment outcomes measured in ACR20/50/70 were considered as separate datasets and evaluated respectively. Study characteristics including authorship, year of publication, and biologic intervention

received were extracted from selected papers. Patient demographics including mean age, percentage of females, mean disease duration in years, swollen joint count (SJC), tender joint count (TJC), C-reactive protein (CRP) level, erythrocyte sedimentation rate (ESR), and percentage of rheumatoid factor positive (RF+) were also extracted from selected papers. Means and standard errors were also extracted for continuous patient demographics if available. Data extracted for ESR and CRP were converted to the unit of mm/h and mg/L respectively if reported in other units. Study covariates were centered at their weighted means to make the results more interpretable.

## 2.2 Data Cleaning

In meta-analysis, study-level characteristics, specifically those continuous characteristics such as age, duration of disease, and some laboratory measurement, are often reported in the same format among included studies but sometimes various measurements are provided due to preference of the authors, different practice patterns, and even the year of the publication which makes it difficult to pool the data for analysis. First, the same measurement may appear in divergent units of researcher's choice. For example, an important laboratory test for RA patient is the C-reactive protein (CRP) level test that has been commonly reported in both mg/L and mg/dL. To pool study data into a single model, the data need to be in a universal format and scale. In this scenario, we have manually inspected the units of all study-level characteristics and adjusted them based on a common unit.

It is a common practice to use the sample mean and standard deviation to describe patient demographics in clinical trials. However, some studies reported other forms of

summary data using the median, the range, or the interquartile. Formulas have been developed to estimate the mean ($\bar{x}$) from the median ($m$), low and high end of the range ($a$ and $b$), and sample size ($n$). When the median and range was reported instead of the mean, the median served as the best estimator of the mean when the number of patients was no smaller than 25 in each treatment arm ($n \geq 25$). When there was less than 25 patients in each arm ($n < 25$), the mean was estimated by formula:[14,15]

$$\bar{x} \approx \frac{a + 2m + b}{4}$$

Similarly, for studies that reported median ($m$), first and third interquartile ($q_1$ and $q_3$), and sample size ($n$). The following equation was used to estimate the mean regardless of the sample size:[14]

$$\bar{x} \approx \frac{q_1 + m + q_3}{3}$$

Missing data among covariates was defined as not reported due to the authors' choice therefore missing data in this study was unrelated to the actual values of the missing data and considered missing at random. Predictive mean matching approach was used to perform multiple imputation on missing values in covariates. Three imputed datasets were created for missing values and the results of the three imputed datasets were pooled for final treatment effect estimation.

## 2.3 Model Selection

A series of models have been proposed and utilized for network meta-analysis. The models have advanced from the simpler fixed-effects and random-effects models to the more complicated mixed-effects models (meta-regression) that allow treatment-by-covariate interactions. Analyses based on these models have been conducted in both frequentist and Bayesian frameworks.

### *Fixed-Effects Model*

The basic assumption of a fixed-effect (FE) network meta-analysis model is that there is a common effect size shared by all included studies. In other words, the FE model assumes homogeneity of effect sizes among selected studies meaning each individual study used the same research design, methodology, and measurement. The effect size is considered unknown but fixed in this model. This model estimates the true effect size μ from a sample of selected studies that only differ from the true value by sampling errors. Therefore, the mean of individual study effect sizes serves as an unbiased estimator for μ, and the confidence interval of μ could be calculated from standard error of the mean. The true effect size μ is shared by all the included studies and each observed treatment effect sizes is sampled from a normal distribution with the mean and the variance. In the treatment network, the placebo group is the most common comparison group and hence selected as the primary reference (treatment A). The treatment outcome investigated in this thesis follows a binary distribution (reached specific treatment outcome or not). Therefore, with the probability of treatment success $p_{jk}$ for each patient receiving treatment $k$ in the $j$th study, the number of observing $r_{jk}$ treatment success in $n_{jk}$ patients

follows a binomial distribution.[16] The fixed-effects network meta-analysis model can be described as:[16,17]

$$r_{ik} \sim \text{Binomial}(p_{jk}, n_{jk}); \qquad \text{logit}(p_{jk}) = \eta_{jk}$$

$$\eta_{jk} = \begin{cases} \mu_{jb} \\ \mu_{jb} + d_{bk} = \mu_{jb} + (d_{Ak} - d_{Ab}) \end{cases} \qquad \begin{matrix} b\text{=A,B,C,...} & \text{if } k = b \\ k\text{=B,C,D,...} & \text{if } k \text{ is after } b \end{matrix}$$

$$d_{AA} = 0$$

where $\mu_{jb}$ and $\eta_{jk}$ is the study specific treatment effect (log-odds) for treatment *b* and *k* in the *j*th study respectively, $d_{bk}$ is the fixed relative treatment effect (log-odds ratio) of treatment *k* to treatment *b* among all treatment *b*-*k* comparison studies. Under the consistency assumption, the $d_{bk}$ can be estimated indirectly through a common comparator treatment *A* ($d_{bk}$=$d_{Ak} - d_{Ab}$).

*Random-Effects Model*

While fixed-effects model assumes a single shared effect size for each comparison pair all studies, the random-effect (RE) model assumes that the effect sizes of underlying studies follow some statistical distribution.[18] With the RE model we aim at inference about the distribution of effect sizes of selected studies from a random sample of studies. The model assumes that studies selected for were not exact replications and vary beyond sampling error . In other words, included studies were not homogeneous. Effect sizes are treated as if they were a random sample of a population effect μ and model parameters, usually the mean and the variance, are estimated in order to describe μ. In a RE network meta-analysis model, the treatment effects $\delta_{jbk}$ are drawn from a

selected distribution (commonly normal distribution) and assigned to the included studies. This model assumes that the studies included are different due to study characteristics but share the same random-effect variance $\sigma^2$. With the same logit function as the fixed-effects model, the random-effects model can be described as:[16,17]

$$r_{ik} \sim \text{Binomial}(p_{jk}, n_{jk}); \qquad \text{logit}(p_{jk}) = \eta_{jk}$$

$$\eta_{jk} = \begin{cases} \mu_{jb} & b=\text{A,B,C,...} \quad \text{if } k = b \\ \mu_{jb} + \delta_{jbk} & k=\text{B,C,D,...} \quad \text{if } k \text{ is after } b \end{cases}$$

$$\delta_{jbk} \sim \text{Normal}\ (d_{bk}, \sigma^2_{bk}) = \text{Normal}\ (d_{Ak} - d_{Ab}, \sigma^2_{bk})$$

$$d_{AA} = 0$$

where $\mu_{jb}$ and $\eta_{jk}$ is the study specific treatment effect (log-odds) for treatment *b* and *k* in the jth study respectively, $\delta_{jbk}$ is the study specific relative treatment effect (log-odds ratio) of treatment *k* to treatment *b* in the *j*th study, $d_{bk}$ is the pooled relative treatment effect of treatment *k* to treatment *b*, and $\sigma^2_{bk}$ is between-study variance for treatment *b*-*k* comparison studies. The pooled relative effect $d_{bk}$ can also be estimated indirectly through a common comparator treatment *A* ($d_{bk} = d_{Ak} - d_{Ab}$).

Correlation is expected in multi-level and multi-arm datasets. In network meta-analysis, treatment effects estimated by the same study, treatment, research group, and clinical trial site are more likely to share some similarity compared with estimation by other sources. The random-effects network meta-analysis model assesses clustering quantitatively with two variance components $\tau^2$ and . The variance component$\tau^2$ represents the amount of heterogeneity in the levels of the inner factor (for example,

study ID number). The other variance component $\rho$ represents the corresponding correlation coefficient between levels of the inner factor (for example, the three treatment arms of one clinical trial). There are several options for the variance structure to model correlation in the random-effects model. The compound symmetry (CS) structure assumes a single variance component $\tau^2$ corresponding to all levels of the inner factor and a single correlation coefficient $\rho$ for the correlation between different levels. The heteroscedastic compound symmetry (HCS) structure assumes a symmetric structure of variance components of $\tau_1^2, \tau_2^2, \ldots$ and a single correlation $\rho$ for the correlation between different levels. The unstructured variance-covariance matrix (UN) structure assumes variance components of $\tau_1^2, \tau_2^2, \tau_3^2, \ldots$ and $\rho_{12}, \rho_{13}, \rho_{23}, \ldots$ for the combinations of levels of inner factor.

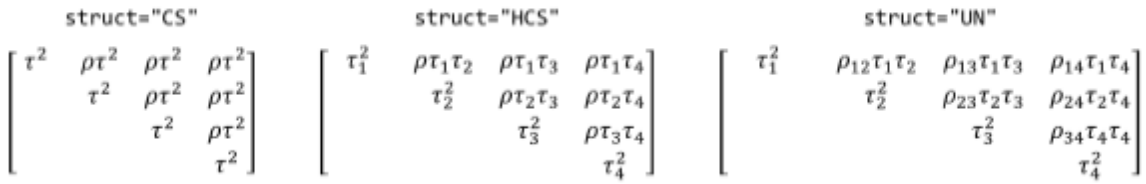


Figure 2. Structures of Variance Components in Random-effects Model

*Mixed-Effects Model*

Comparing treatment from different studies directly breaks randomization, which makes it difficult to distinguish the effect of study characteristics from treatment effects. The mixed-effects (ME) models, also known as meta-regression, include study-level treatment effect modifiers (referred as covariates often) and can further evaluate modifiers' contribution to heterogeneity. The modifiers are usually centered at its overall

mean to make the results more interpretable. The treatment effect modifier $X_j$ is incorporated in the model via the mean of the distribution of random-effects. With the same notations used as above, the mixed-effects model can be described as:[16,17]

$$r_{ik} \sim \text{Binomial}(p_{jk}, n_{jk}); \qquad \text{logit}(p_{jk}) = \eta_{jk}$$

$$\eta_{jk} = \begin{cases} \mu_{jb} \\ \mu_{jb} + \delta_{jbk} \end{cases} \qquad \begin{array}{ll} b\text{=A,B,C,...} & \text{if } k = b \\ k\text{=B,C,D,...} & \text{if } k \text{ is after } b \end{array}$$

$$\delta_{jbk} \sim \text{Normal}\left(d_{bk} + \beta_{bk} X_j, \sigma_{bk}^2\right) = \text{Normal}\left(d_{Ak} - d_{Ab} + (\beta_{Ak} - \beta_{Ab}) X_j, \sigma_{bk}^2\right)$$

$$d_{AA} = 0$$

$$\beta_{AA} = 0$$

where $\mu_{jb}$ and $\eta_{jk}$ is the study specific treatment effect (log-odds) for treatment *b* and *k* in the jth study respectively, $\delta_{jbk}$ is the study specific relative treatment effect (log-odds ratio) of treatment *k* to treatment *b* in the *j*th study, $d_{bk}$ is the pooled relative treatment effect of treatment *k* to treatment *b*, $\beta_{bk}$ is the pooled effect of the covariate, $X_j$is the matrix of the covariate in the *j*th study, and $\sigma_{bk}^2$ is between-study variance for treatment *b*-*k* comparison studies. The pooled relative treatment effect $d_{bk}$ and covariate effect $\beta_{bk}$ can be estimated indirectly through a common comparator treatment *A* ($d_{bk} = d_{Ak} - d_{Ab}$; $\beta_{bk} = \beta_{Ak} - \beta_{Ab}$).

Using study-level covariate information may affect the magnitude and the direction of the relation between the study outcome and the covariate. Therefore, even

though ME models provide information about heterogeneity, one should always be cautious when interpreting the results.

*Bayesian Framework*

In frequentist meta-analysis, treatment effects are estimated by weighing individual estimates by the inverse of their variance. In complicated situations such as network meta-analysis and meta-regression, treatment effect estimation and inference are based on maximum likelihood. In the Bayesian framework, model parameters are considered as random variables whose uncertainty is quantified with probability. The prior probability distribution reflects prior belief about possible values of the parameters. Additional information from previous clinical trials, observational studies, and reviews can be integrated in models to inform researchers in the form of prior distributions. Meanwhile non-informative prior distributions minimize influence of parameters on the posterior results when no information is given about model parameters. The likelihood represents the probably probability of the data given the value of the parameters. A posterior probability distribution of the parameter is obtained by combining the prior probability distribution and the likelihood as Bayes' theorem explains:[19,20]

$$p\,(\text{parameters} \mid \text{data}) = \frac{P\,(\text{data} \mid \text{parameters})\,P\,(\text{parameters})}{P\,(\text{data})}$$

$$\propto p\,(\text{data} \mid \text{parameters})\,p\,(\text{parameters})$$

The likelihood $p$ (data | parameters) is determined by the study dataset collected. The prior $p$ (parameters) is based on current understanding and beliefs about the model parameters. In network meta-analysis of multiple treatments, the treatment effect

functions have more than one parameter. This makes it almost impossible to obtain exact posterior quantities for all parameters. The Monte Carlo method generates random sample values of the parameters from their posterior distributions and approximates all of these posterior quantities of interest to an arbitrary degree of precision. For example, in order to study parameter $\theta$, we let $y_1, \dots, y_n$ to be the numerical values of a sample from a distribution $p(y_1, \dots, y_n| \theta)$. A number *S* of independent and identically distributed samples of $\theta$ could be acquired from the posterior distribution $p(\theta|y_1, \dots, y_n)$:

$$\theta^{(1)}, \dots, \theta^{(S)} \sim (\theta|y_1, \dots, y_n)$$

The empirical distribution of the samples $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ is known as a Monte Carlo approximation to $p(\theta|y_1, \dots, y_n)$. The more $\theta$ sampled, the empirical distribution provides a closer approximation to the true density therefore the better the approximation.

For complicated multi-parametric models such as network meta-analysis models, it is difficult to sample from the joint posterior distribution directly. In such cases, Gibbs sampler provides a computationally feasible alternative in which samples are drawn from the full conditional distributions of parameters. The Gibbs sampler is an iterative algorithm in which all parameters start at an initial value. The algorithm needs to run for another number *T* of iterations. Each parameter is sampled from full conditional and has its current value updated at each iteration. The current value of each parameter $\theta^{(T)}$ only depends on the previous value $\theta^{(T-1)}$ and is independent from any other past values $\theta^{(T-2)}$, $\theta^{(T-3)}$,…, $\theta^{(1)}$. In this so-called Markov chain, the algorithm will converge to the true distributions as the number *T* that goes to up towards infinity, By applying the same algorithm to the regular Monte Carlo approximation, the joint posterior distributions of

the parameters could be approximated by this Markov Chain Monte Carlo (MCMC) method.

No previous studies have assessed study-level covariates in network meta-analysis for biologic anti-rheumatic therapies. No prior distributions were used in Bayesian network meta-analysis models. No assumptions were made about the treatment effects of the biologic therapies from external sources other than the studies included in the systematic review. We estimated the posterior distribution with the following non-informative priors to limit inference to clinical trials data only.

*Model Assessment*

Checking the homogeneity and consistency assumptions are as important as analyzing the data. Heterogeneity test checks whether the studies are evaluating the same treatment effect. In network meta-analysis, the Cochran's Q test is still the usual statistic for testing heterogeneity which uses the weighted sum of squared differences between individual study effect and the overall fixed-effects mean estimation.[21] P-values are obtained by comparing the statistic with a $\chi^2$ distribution with *k*-(*n*-1) degrees of freedom (where *k* is the number of studies, and *n* is the number of treatments compared). The Q-value can be calculated with the following equations:[17,22]

$$Q = \sum w_i \left(y_i - \overline{y}_w\right)^2$$

$$\overline{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

$$w_i = \frac{1}{s_i^2}$$

Where $y_i$ and $s_i^2$ ($i$ = 1,2,3,…,$k$) are the estimated treatment effects and corresponding variances of each included studies, $w_i$ are weights for each study, and $\overline{y}_w$ is the weighted fixed-effects mean estimation. The results of Cochrane's Q test needs to be interpreted carefully though, since it has low power when studies have small number of studies included. A statistically significant result may indicate heterogeneity, but a non-significant result does not indicate homogeneity. A *p*-value of 0.10 is often used instead of 0.05.[23] On the other hand, Q also has too much power as a test of heterogeneity if the number of studies is large. It is important to consider to what extent the results of studies are consistent. If confidence intervals for the results of individual studies (generally depicted graphically using horizontal lines) have poor overlap, this generally indicates the presence of statistical heterogeneity.

Due to the limitation of the *Q*-test, the $I^2$ statistic which describes the percentage of variation across studies that is due to heterogeneity was also used in model assessment:[22]

$$I^2 = 100\% \times [Q - (k - 1)] / Q$$

In random-effects models, $I^2$ estimates the percentage of how much of the total variability in the effect size estimates (heterogeneity plus sampling variability) attributed to heterogeneity among the true effects. In mixed-effects models, $I^2$ estimates the percentage of the unaccounted variability (residual heterogeneity plus sampling variability) can be attributed to residual heterogeneity. $I^2$ is an intuitive and simple expression of the heterogeneity of studies. The Good Research Practice guideline of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR)

suggested the following thresholds: $I^2 < 30\%$ is considered mild heterogeneity, and $I^2 > 50\%$ is considered notable heterogeneity.[17]

The assessment of consistency is a currently popular but yet accomplished field of research. The current common practice is to compare the calculated effect size estimations from both direct and indirect comparisons respectively for inconsistency. Another approach is to use the forest plot to examine the consistency graphically. In the Bayesian framework, the posterior density of the direct, indirect, and the network overall estimations are plotted together to assess consistency.

*Model Fitness*

Discrepancy assessment of model fitness was conducted for goodness of fit with in both frequentist and Bayesian frameworks. The goodness of fit summarizes and describes the difference between values observed in datasets and values expected in a model in order to help researchers to select the best fitting model. In the frequentist framework, the maximum-likelihood approach was used. Akaike information criterion (AIC) and Bayesian information criterion (BIC) values were used to assess model fitness in the frequentist framework. The model with smaller values of AIC and BIC is a better fitting model. In the Bayesian framework, the deviance information criterion (DIC) was used to assess goodness of fit. The DIC is hierarchical model selection criteria and is particularly useful in Bayesian framework when the posterior distributions of parameters were obtained Markov Chain Monte Carlo simulation. Similarly, a smaller value of DIC indicates a better model fitness.

*Analytical Packages*

The analysis was carried out in a freely available open source statistical program R. Statistical methods in both frequentist and Bayesian framework are provided via R packages.[24] R package *mice* (version 2.24, released November 9, 2015) was used for multiple imputation. R package *metafor* (version 1.9-8, released September 28, 2015) was used for the analysis in frequentist framework. In the *metaphor* package, fixed-effects and random-effects models could be fit on both arm-based and contrast-based data. A number of study-level covariates could be added to the random-effects model for meta-regression. R package *gemtc* (version 0.8, released March 1, 2016) was used for the Bayesian analysis. The package *gemtc* could only include one study-level covariate (regressor) in the mixed-effects model. Free Bayesian software JAGS (Just Another Gibbs Sampler, version 4.2.0, released February 19, 2016)) was used to form Gibbs sampling.

# 3 STUDY DATASET

## 3.1 Literature Search Results

The search strategy identified 1,334 articles (Figure 3). After title and abstract review, 1,298 studies were excluded due to the following reasons: (1) the study outcome was not measured in ACR criteria (n = 1,104); (2) no study outcome reported at week 24 (n = 180); (3) the study was conducted for psoriatic arthritis (n=8); (4) other reasons (n = 6). After full-text review, 16 articles were further excluded due to the following reasons: (1) the patients were receiving other concomitant RA treatments other (n=10); (2) the study did not have background MTX in combination with the biologics (n=5); (3) the clinical trial for toficitanib, an oral biologic agent with different indication from other biologics (n=1).

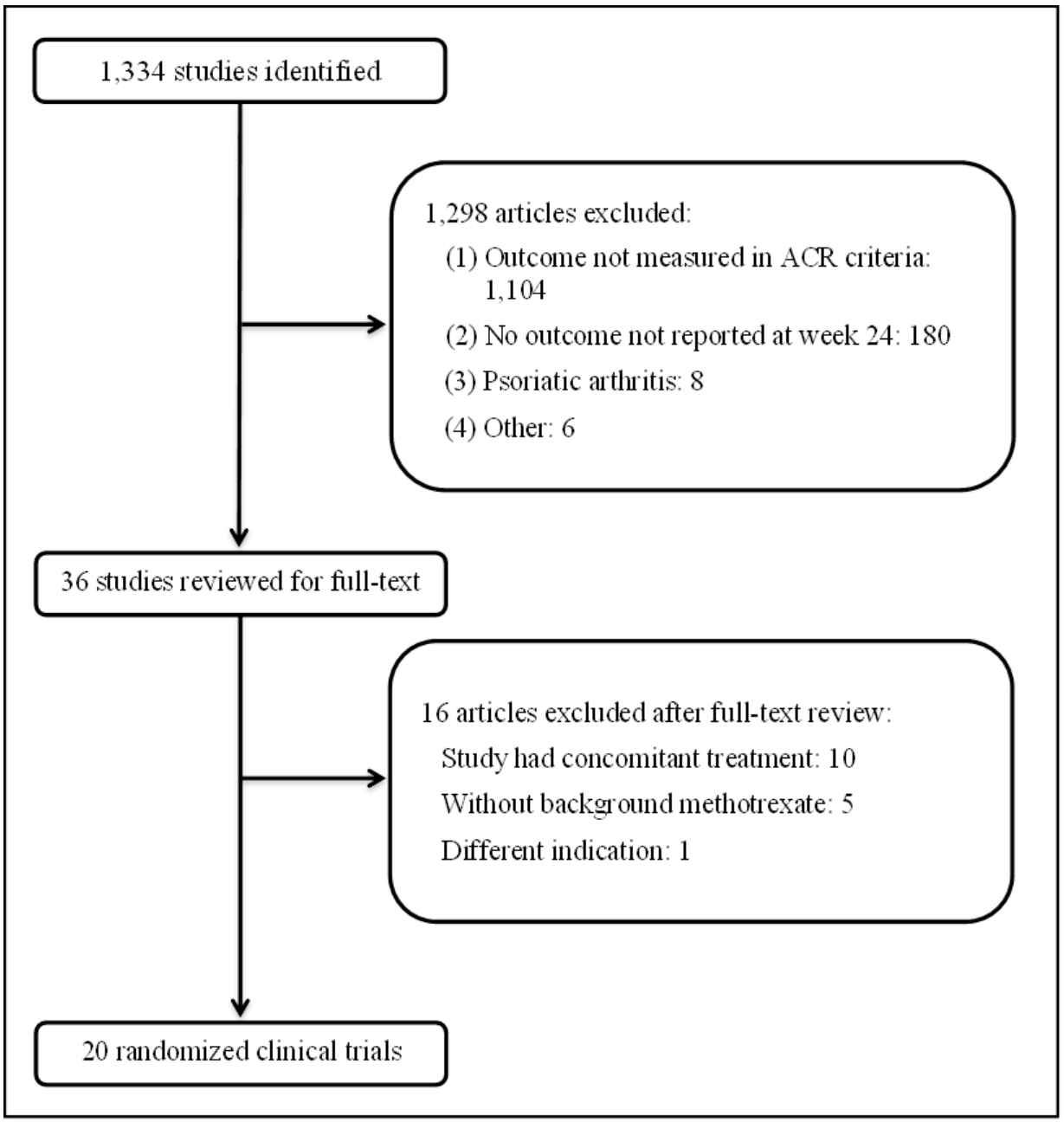


Figure 3. Flow Chart of Literature Review and Study Selection Process for Eligible Studies

Abbreviation: ACR = American College of Rheumatology

The 20 selected studies included a total of 7,666 patients that received biologic DMARDs for rheumatoid arthritis. Treatment arms were balanced within each study indicating sufficient randomization. Table 1 shows the study characteristics and patient demographics of each study, whereas Table 2 shows the numbers of responders and response rates for ACR20/50/70 criteria.[25-44] The evidence network is demonstrated in Figure 2. Each node represents a biologic therapy. Each edge between nodes represents a direct comparison and the width of the edges represents the number of clinical trials contributing to the comparison pair. Placebo was the common comparator among all included studies. Nineteen studies (95%) were two-armed clinical trials that compared one biologic therapy with placebo while one study (5%) was a three-armed trial that compared two biologics and placebo. For study covariates, three (15%) studies reported the median for all patient demographic covariates; two (10%) studies reported the median for ESR and CRP, and two (10%) other studies report the geometric mean for also ESR and CRP. Baseline ESR, CRP, and RF+ were missing in the ten (50%), one (5%), and two (10%) studies, respectively. Imputed values were analyzed separately before pooling model estimations for interpretation.

Table 1. Study Characteristics and Baseline Patient Demographics

| Study characteristics | | | | Baseline patient demographics | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Study (Trial name)** | **Publication Year** | **Intervention** | **Number of patients** | **Age (years)** | **Female (%)** | **Duration (years)** | **SJC** | **TJC** | **ESR (mm/h)** | **CRP (mg/L)** | **RF+ (%)** |
| Keystone | 2004 | ADA | 207 | 56.1 | 76.3 | 11 | 19.3 | 27.3 | NR | 18 | 81.6 |
| | | Placebo | 200 | 56.1 | 73 | 10.9 | 19 | 28.1 | NR | 18 | 89.5 |
| Weinblatt (ARMADA) | 2003 | ADA | 67 | 57.2 | 74.6 | 12.2 | 17.3 | 28 | NR | 21 | NR |
| | | Placebo | 62 | 56 | 82.3 | 11.1 | 16.9 | 28.7 | NR | 31 | NR |
| van Vollenhoven (AUGUST-2) | 2011 | ADA | 79 | 53 | 81 | 8.8 | 16.2 | 27.8 | 41.7 | 16.6 | 81 |
| | | Placebo | 76 | 54 | 84 | 8.4 | 16.4 | 24.3 | 39.3 | 16.5 | 83 |
| van Vollenhoven | 2012 | ADA | 204 | 52.5 | 79.4 | 8.1 | 16.4 | 26.7 | 48.5 | 17.5 | 68.2 |
| | | Placebo | 108 | 53.8 | 75.9 | 7.9 | 16.7 | 27.3 | 48 | 16.1 | 66.3 |
| Keystone (RAPID 1) | 2008 | CTZ | 393 | 51.4 | 82.4 | 6.1 | 21.7 | 30.8 | 43.5 | 16 | 79.6 |
| | | Placebo | 199 | 52.2 | 83.9 | 6.2 | 21.2 | 29.8 | 45 | 16 | 82.8 |
| Smolen (RAPID 2) | 2009 | CTZ | 246 | 52.2 | 83.7 | 6.1 | 20.5 | 30.1 | NR | NR | 77.5 |
| | | Placebo | 127 | 51.5 | 84.3 | 5.6 | 21.9 | 30.4 | NR | NR | 78.2 |
| Choy | 2012 | CTZ | 126 | 53 | 72.2 | 9.4 | 22.8 | 29 | NR | NR | 73.8 |
| | | Placebo | 121 | 55.6 | 66.1 | 9.9 | 22.2 | 31 | NR | NR | 78.5 |
| Weinblatt | 1999 | ETN | 59 | 48 | 90 | 13 | 20 | 28 | 25 | 22 | 84 |
| | | Placebo | 30 | 53 | 73 | 13 | 17 | 28 | 36 | 26 | 90 |
| Kay | 2008 | GLB | 35 | 57 | 85.7 | 8.2 | 14 | 28 | NR | 21 | NR |
| | | Placebo | 35 | 52 | 74.3 | 5.6 | 13 | 22 | NR | 20 | NR |
| Keystone (GO-FORWARD) | 2009 | GLB | 89 | 52 | 80.9 | 4.5 | 13 | 26 | NR | 10 | 86.5 |
| | | Placebo | 133 | 52 | 82 | 6.5 | 12 | 21 | NR | 8 | 81.2 |
| Maini (ATTRACT) | 1999 | IFX | 86 | 56 | 81 | 8.4 | 19 | 32 | NR | 31 | 84 |
| | | Placebo | 88 | 51 | 80 | 8.9 | 19 | 24 | NR | 30 | 77 |
| Westhovens (START) | 2006 | IFX | 360 | 53 | 80 | 7.8 | 15 | 22 | NR | 16 | 82.8 |
| | | Placebo | 363 | 52 | 83.2 | 8.4 | 15 | 22 | NR | 12 | 80.7 |
| Schiff (ATTEST) | 2008 | ABT | 156 | 49 | 83.3 | 7.9 | 21.3 | 31.6 | 49.4 | 31 | 87.2 |
| | | IFX | 165 | 49.1 | 82.4 | 7.3 | 20.3 | 31.7 | 47.8 | 33 | 84.8 |
| | | Placebo | 110 | 49.4 | 87.3 | 8.4 | 20.1 | 30.3 | 47 | 27 | 77.3 |
| Weinblatt (AMPLE) | 2013 | ABT | 318 | 51.4 | 81.4 | 1.9 | 15.8 | 25.4 | NR | 16 | 75.5 |
| | | ADA | 328 | 51 | 82.3 | 1.7 | 15.9 | 26.3 | NR | 15 | 77.4 |
| Kremer | 2003 | ABT | 115 | 55.8 | 75 | 9.7 | 21.3 | 30.8 | NR | 29 | 99 |
| | | Placebo | 119 | 54.7 | 66 | 8.9 | 21.8 | 29.2 | NR | 32 | 90 |
| Kremer | 2006 | ABT | 433 | 51.5 | 77.8 | 8.5 | 21.4 | 31 | NR | 33 | 81.8 |
| | | Placebo | 219 | 50.4 | 81.7 | 8.9 | 22.1 | 32.3 | NR | 28 | 78.5 |
| Cohen | 2004 | ANA | 250 | 56 | 79 | 11 | 20.1 | 26.8 | 41.5 | 27 | 76 |
| | | Placebo | 251 | 57 | 75 | 10 | 20 | 24.5 | 42.9 | 26 | 78 |
| Emery (SERENE) | 2010 | RTX | 337 | 51.6 | 80.4 | 6.85 | 19 | 27.9 | NR | NR | 74.4 |
| | | Placebo | 172 | 52.16 | 85.5 | 7.48 | 20.9 | 30.2 | NR | NR | 75 |
| Smolen (OPTION) | 2008 | TCZ | 205 | 50.8 | 85 | 7.5 | 19.5 | 31.9 | 51.2 | 26 | 83 |
| | | Placebo | 204 | 50.6 | 78 | 7.8 | 20.7 | 32.8 | 49.7 | 24 | 71 |
| Kremer (LITHE) | 2011 | TCZ | 398 | 53.4 | 82 | 9.3 | 17.3 | 29.3 | 46.4 | 23 | 83 |
| | | Placebo | 393 | 51.3 | 83 | 9 | 16.6 | 27.9 | 46.5 | 22 | 82 |

ABT = abatacept; ADA = adalimumab; ANA = anakinra; CRP = C-reactive protein; CTZ = certolizumab pegol; ESR = erythrocyte sedimenation rate; ETN = etanercept; GLB = golimumab; IFX = infliximab; mg/L = milligram per liter; mm/h = millimeters per hour; NR = not reported; RF+ = tested positive for rheumatoid factor; RTX = rituximab; SJC = swollen joint count; TCZ = tocilizumab; TJC = tender joint count

Table 2. Treatment Response in ACR Criteria

| Study and year | Intervention | Sample size | Treatment responses | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | ACR20 | | ACR50 | | ACR70 | |
| | | | Rate | Count | Rate | Count | Rate | Count |
| Keystone 2004 | ADA | 207 | 0.633 | 131 | 0.391 | 81 | 0.208 | 43 |
| | Placebo | 200 | 0.295 | 59 | 0.095 | 19 | 0.025 | 5 |
| Weinblatt 2003 | ADA | 67 | 0.672 | 45 | 0.552 | 37 | 0.269 | 18 |
| | Placebo | 62 | 0.145 | 9 | 0.081 | 5 | 0.048 | 3 |
| van Vollenhoven 2011 | ADA | 79 | 0.710 | 56 | 0.380 | 30 | 0.180 | 14 |
| | Placebo | 76 | 0.460 | 35 | 0.150 | 11 | 0.050 | 4 |
| van Vollenhoven 2012 | ADA | 204 | 0.470 | 96 | 0.280 | 57 | 0.090 | 18 |
| | Placebo | 108 | 0.280 | 30 | 0.120 | 13 | 0.020 | 2 |
| Keystone 2008 | CTZ | 393 | 0.588 | 231 | 0.371 | 146 | 0.214 | 84 |
| | Placebo | 199 | 0.136 | 27 | 0.076 | 15 | 0.030 | 6 |
| Smolen 2009 | CTZ | 246 | 0.573 | 141 | 0.325 | 80 | 0.159 | 39 |
| | Placebo | 127 | 0.087 | 11 | 0.031 | 4 | 0.008 | 1 |
| Choy 2012 | CTZ | 126 | 0.459 | 58 | 0.180 | 23 | 0.000 | 0 |
| | Placebo | 121 | 0.229 | 28 | 0.059 | 7 | 0.017 | 2 |
| Weinblatt 1999 | ETN | 59 | 0.710 | 42 | 0.390 | 23 | 0.150 | 9 |
| | Placebo | 30 | 0.270 | 8 | 0.030 | 1 | 0.000 | 0 |
| Kay 2008 | GLB | 35 | 0.600 | 21 | 0.371 | 13 | 0.086 | 3 |
| | Placebo | 35 | 0.371 | 13 | 0.057 | 2 | 0.000 | 0 |
| Keystone 2009 | GLB | 89 | 0.596 | 53 | 0.371 | 33 | 0.202 | 18 |
| | Placebo | 133 | 0.278 | 37 | 0.135 | 18 | 0.053 | 7 |
| Maini 1999 | IFX | 86 | 0.500 | 43 | 0.270 | 23 | 0.080 | 7 |
| | Placebo | 88 | 0.200 | 18 | 0.050 | 4 | 0.000 | 0 |
| Westhovens 2006 | IFX | 360 | 0.580 | 209 | 0.321 | 116 | 0.140 | 50 |
| | Placebo | 363 | 0.255 | 93 | 0.097 | 35 | 0.047 | 17 |
| Schiff 2008 | ABT | 156 | 0.667 | 104 | 0.404 | 63 | 0.205 | 32 |
| | IFX | 165 | 0.594 | 98 | 0.370 | 61 | 0.242 | 40 |
| | Placebo | 110 | 0.418 | 46 | 0.200 | 22 | 0.091 | 10 |
| Weinblatt 2013 | ABT | 318 | 0.660 | 210 | 0.460 | 146 | 0.240 | 76 |
| | ADA | 328 | 0.650 | 213 | 0.430 | 141 | 0.220 | 72 |
| Kremer 2003 | ABT | 115 | 0.600 | 69 | 0.365 | 42 | 0.165 | 19 |
| | Placebo | 119 | 0.350 | 42 | 0.118 | 14 | 0.017 | 2 |
| Kremer 2006 | ABT | 433 | 0.679 | 294 | 0.399 | 173 | 0.198 | 86 |
| | Placebo | 219 | 0.397 | 87 | 0.168 | 37 | 0.065 | 14 |
| Cohen 2004 | ANA | 250 | 0.380 | 95 | 0.170 | 43 | 0.060 | 15 |
| | Placebo | 251 | 0.220 | 55 | 0.080 | 20 | 0.020 | 5 |
| Emery 2010 | RTX | 337 | 0.525 | 177 | 0.261 | 88 | 0.095 | 32 |
| | Placebo | 172 | 0.233 | 40 | 0.093 | 16 | 0.052 | 9 |
| Smolen 2008 | TCZ | 205 | 0.590 | 121 | 0.440 | 90 | 0.220 | 45 |
| | Placebo | 204 | 0.260 | 53 | 0.110 | 22 | 0.020 | 4 |
| Kremer 2011 | TCZ | 398 | 0.570 | 227 | 0.320 | 127 | 0.120 | 48 |
| | Placebo | 393 | 0.270 | 106 | 0.100 | 39 | 0.020 | 8 |

## 3.2 Evidence Network

To improve transparency of the analyses, Figure 5 illustrated the evidence network of the frequentist network meta-analysis. As demonstrated in the figure, placebo was the common comparator among all included studies and one study compared two medications with placebo in the same trial. Each of the nodes with label represented a treatment option and the drug's name. Each of the edges represented a treatment comparison and the widths of the edges represented the number of clinical trials contributing to the comparison pairs.

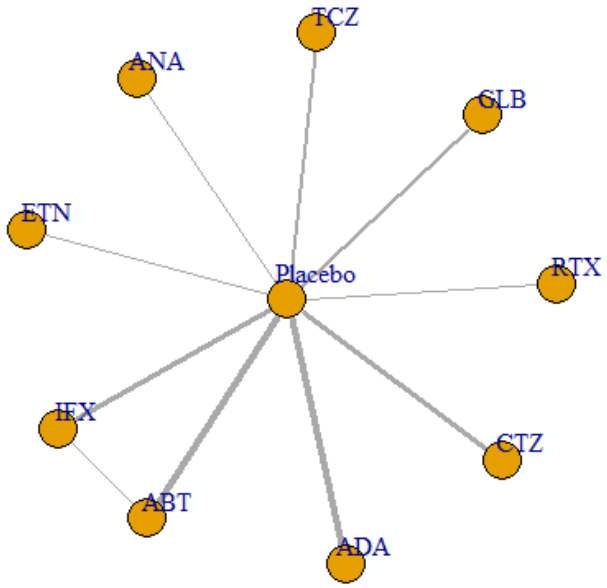


Figure 4. Evidence Network of Included Randomized Clinical Trials

# 4 FITTING FREQUENTIST MODELS

Results of frequentist models are demonstrated in this chapter. Dataset for ACR20/50/70 criteria were analyzed respectively. For each dataset, fixed-effects, random-effects, and mixed-effects model with selected effect modifiers were fit and compared.

## 4.1 ACR20 Dataset

This section summarizes and reports frequentist model results for ACR20 dataset. Fixed-effect and random-effects models were fit to the ACR20 dataset with no study covariates. Mixed-effects models with covariates reported in Table 1 were fit respectively to assess their individual contribution to heterogeneity. Model estimations were reported for each model and the most suitable model for this dataset was used to draw conclusions.

*Fixed-Effects Model*

Fixed-effects model was fit to the ACR20 dataset first. Based on model estimation (Table 3), all medications demonstrated a significant superiority over placebo (degrees of freedom [df] = 12). CTZ showed the best treatment effect among all medications under ACR20 criteria, followed by ETN, TCZ, RTX, ADA, GLB, IFX, and ABT (in this order). ANA showed the least, but significant superiority over placebo among all medications.

Table 3. Relative Treatment Effects to Placebo, Frequentist FE Model, ACR20 Dataset

| Treatment | Estimated Coefficient | OR | p-value |
|---|---|---|---|
| CTZ | 1.931 | 6.899 | < 0.001 |
| ETN | 1.916 | 6.794 | < 0.001 |
| TCZ | 1.324 | 3.758 | < 0.001 |
| RTX | 1.295 | 3.651 | < 0.001 |
| ADA | 1.288 | 3.624 | < 0.001 |
| GLB | 1.235 | 3.437 | < 0.001 |
| IFX | 1.111 | 3.038 | < 0.001 |
| ABT | 0.785 | 2.192 | < 0.001 |
| ANA | 0.781 | 2.184 | 0.002 |

Model degrees of freedome = 12

Forest plot of odds ratios (ORs) and associated confidence intervals (CIs) for each treatment is displayed in figure 5.

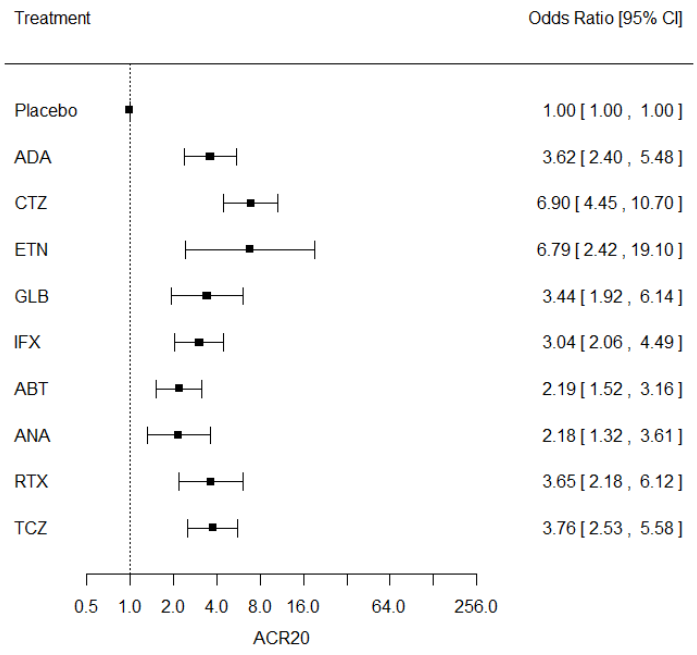


Figure 5. Forest Plot of Treatment Effects, Frequentist FE Models, ACR20 Dataset

For the test of heterogeneity, the null hypothesis which is there is no significant heterogeneity among included studies was rejected (Q [df = 12] = 64.686, *p*-value < 0.001). From which $I^2$ was calculated as 100% × ([64.686 – 12] / 64.686) = 81.4%. The $I^2$

statistic indicated that substantial heterogeneity existed in the dataset and 81.4% of the total variability in the estimated treatment effects was due to heterogeneity. With substantial heterogeneity observed among included studies, the assumption of fixed-effect model was violated. Therefore, appropriate models should be explored and used to interpret the data.

*Random-Effects Model*

Random-effects model was fit to the dataset given its more reasonable assumptions. In our dataset, one study was a three-armed study that compared two different biologic RA medications with placebo. Some of the medications have been studied in multiple clinical trials. Therefore, two levels of grouping variable were considered in the random-effects model. Treatment effects estimated with different values of the outer grouping variable (study ID) were assumed to be independent, while those with the same value of the outer grouping variable shared correlated random effects based on the inner grouping variable (treatment group). Compound symmetry (CS) structure was used as it produced the best model fit among all structures. Two different random-effects models were used to explore the impact of correlation in depth. One model estimated $\rho$ based on the dataset while the other preset $\rho$ a value of 0.5.

Table 4 below shows the relative effects compared to placebo for each treatment with a fixed $\rho$ of 0.5 and the forest plot of this model can be found in figure 6. As seen in Table 5, the $\rho$ estimated model yield similar results to the other model with a slightly better model fitness (discussed later in this thesis). Based on the model estimation, all medications except ANA demonstrated a significant superiority over placebo. CTZ was

still the most efficacious treatment option among all medications, followed by ETN, ADA, TCZ, RTX, GLB, IFX, and ABT (in this order). ANA also showed superiority over placebo but the effect was not statistically significant. Compared with the results of the fixed-effects model, the random-effects model reported the same directions for treatment effects and minor differences in magnitudes. However, it affected the significance level of some of the treatment options. Once significant, ANA now has non-significant superiority over placebo in this model.

Table 4. Relative Treatment Effects to Placebo, Frequentist RE Model, ACR20 Dataset ($\rho = 0.5$)

Table 4. Treatment Effects, Frequentist RE Model, ACR20 Dataset ($\rho = 0.5$)

| Treatment | Estimated Coefficient | OR | p-value |
|---|---|---|---|
| CTZ | 1.929 | 6.880 | < 0.001 |
| ETN | 1.861 | 6.429 | 0.008 |
| ADA | 1.361 | 3.899 | < 0.001 |
| TCZ | 1.337 | 3.809 | < 0.001 |
| RTX | 1.286 | 3.618 | 0.017 |
| GLB | 1.155 | 3.175 | 0.009 |
| IFX | 1.065 | 2.902 | < 0.001 |
| ABT | 0.865 | 2.375 | 0.001 |
| ANA | 0.777 | 2.174 | 0.148 |

Model degrees of freedome = 12

Table 5. Relative Treatment Effects to Placebo, Frequentist RE Model, ACR20 Dataset ($\rho$ estimated)

| Treatment | Estimated Coefficient | OR | p-value |
|---|---|---|---|
| CTZ | 1.929 | 6.881 | < 0.001 |
| ETN | 1.861 | 6.429 | 0.008 |
| ADA | 1.362 | 3.902 | < 0.001 |
| TCZ | 1.337 | 3.809 | < 0.001 |
| RTX | 1.286 | 3.618 | 0.019 |
| GLB | 1.155 | 3.173 | 0.010 |
| IFX | 1.002 | 2.725 | < 0.001 |
| ABT | 0.845 | 2.328 | 0.002 |
| ANA | 0.777 | 2.174 | 0.152 |

Model degrees of freedome = 12

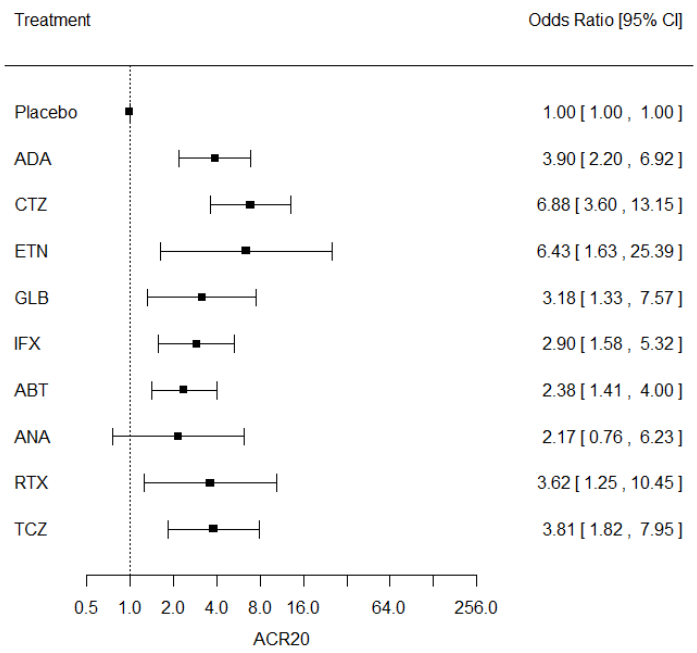


Figure 6. Forest Plot of Treatment Effects, Frequentist RE models, ACR20 Dataset, ρ=0.5

Since the same dataset was used as the fixed-effects model and the random-effects model acknowledges the existence of heterogeneity, the test of heterogeneity remained the same conclusion (Q [df = 12] = 64.195, p-value < 0.001), as expected. $I^2$ calculated equals to 81.3%. The $I^2$ statistic indicated that substantial heterogeneity existed in the dataset and 81.3% of the total variability in the estimated treatment effects was due to heterogeneity. Random-effects models showed a profound improvement in model fitness (ρ = 0.5 model: AIC = 40.60, BIC = 45.45; ρ estimated model: AIC = 42.55, BIC = 47.88) compared with fixed-effects model (AIC = 106.26, BIC = 120.81)). Assigning a value of 0.5 to the variance components $\rho$ showed a small further improvement. The model fitness results also favored the random-effects model over the fixed-effects model.

*Mixed-Effects Model*

Meta-regression was conducted to explore covariates' contribution to heterogeneity among results of studies using mixed-effects model. Covariates were centered at their overall means to make the results more interpretable. To achieve this, the difference between the weighted mean of a covariate and the covariate was fit in the model. Each covariate was assessed respectively to assess individual effect. Given the same multi-level structure, the same variance components were reported in this model. Variance component $\rho$ was fixed to 0.5 in the mixed-effects models.

Table 6 below shows the results of single-covariate mixed-effects models along with the ranked treatment effects based on odds ratios. For the study covariates, none of them demonstrated a significant effect on treatment success. Duration of disease had the biggest effect (OR = 1.105) and age had the slightest effect (OR = 1.002). Each of the odds ratios was close to one suggesting that these covariates had no effects on treatment success. For the treatments, similar superiority and significance was observed as well with different ranks of treatments. It is worth mentioning that the model with covariate duration of disease reported the same for the most and the least efficacious treatment as the random-effects model but different ranks in the middle. The models with covariates SJC or TJC reported ETN the best treatment while CTZ was constantly the best treatment in other models.

Table 6. Covariate Assessment and Ranked Treatments, Frequentist ME Model, ACR20 Dataset

| Covariate | Ranked Relative Treatments to Placebo and ORs | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **OR** | **1st** | **2nd** | **3rd** | **4th** | **5th** | **6th** | **7th** | **8th** | **9th** |
| Age | CTZ | ETN | ADA | TCZ | RTX | GLB | IFX | ABT | ANA |
| 1.002 | 6.885*** | 6.464* | 3.898*** | 3.817*** | 3.623* | 3.163* | 2.899** | 2.383** | 2.157 |
| Female Percentage | CTZ | ETN | ADA | TCZ | RTX | GLB | IFX | ANA | ABT |
| 1.039 | 7.366*** | 5.547* | 4.239*** | 3.581*** | 3.389* | 3.115** | 2.996*** | 2.477 | 2.463*** |
| Duration of Disease | CTZ | RTX | ETN | GLB | TCZ | ADA | IFX | ABT | ANA |
| 1.105 | 7.272*** | 3.900** | 3.840 | 3.749** | 3.581*** | 3.206*** | 2.909*** | 2.608*** | 1.665 |
| SJC | ETN | CTZ | GLB | ADA | TCZ | RTX | IFX | ABT | ANA |
| 1.074 | 6.308** | 5.564*** | 4.833* | 4.306*** | 3.881*** | 3.389* | 3.242*** | 2.228** | 1.978 |
| TJC | ETN | CTZ | GLB | ADA | RTX | IFX | TCZ | ANA | ABT |
| 1.074 | 6.484*** | 5.928*** | 4.307** | 4.16*** | 3.477* | 3.457*** | 3.239** | 2.591 | 2.189** |

Significance level: *** p < 0.001, ** p < 0.01, * p < 0.05 (same below), df = 11 for all ME models

Table 7. Heterogeneity Comparison, Frequentist Models, ACR20 Dataset

| Model | Test of (residual) heterogeneity | I^2 |
|---|---|---|
| RE model | Q [df = 12] = 64.195, p-value < 0.001 | 81.3% |
| ME model with covariate age | Q [df = 11] = 63.922, p-value < 0.001 | 82.8% |
| ME model with covariate female percentage | Q [df = 11] = 60.648, p-value < 0.001 | 81.9% |
| ME model with covariate duration of disease | Q [df = 11] = 43.510, p-value < 0.001 | 74.7% |
| ME model with covariate SJC | Q [df = 11] = 53.636, p-value < 0.001 | 79.5% |
| ME model with covariate TJC | Q [df = 11] = 57.689, p-value < 0.001 | 80.9% |

Table 8. Model Fitness Comparison, Frequentist ME Models, ACR20 Dataset

| Model | AIC | BIC |
|---|---|---|
| Random-effects model, ρ fixed to 0.5 | 40.60 | 45.45 |
| Mixed-effects model with covariate age | 41.78 | 46.16 |
| Mixed-effects model with covariate female percentage | 40.53 | 44.91 |
| Mixed-effects model with covariate duration of disease | 40.04 | 44.42 |
| Mixed-effects model with covariate SJC | 41.25 | 45.63 |
| Mixed-effects model with covariate TJC | 41.11 | 45.49 |

For the test of residual heterogeneity, all the models still demonstrated significant residual heterogeneity (Table 7). Statistical heterogeneity is inevitable due to clinical and methodological differences in network meta-analysis.[45] $I^2$ statistic assesses the impact of heterogeneity on the analysis. The mixed-effects models reported similar $I^2$ values to random-effects model. The inclusion of covariates of duration of disease had the largest impact on $I^2$, which was a decrease of about 6% while other covariates had unnoticeable effects on $I^2$. The $I^2$ statistic indicates that over 70% of the variability in treatment effect estimation is due to heterogeneity rather than sampling error. Including covariates did not improve model fitness as all models reported similar AICs and BICs (Table 8).

## 4.2 ACR50 and ACR70 Dataset

This section summarizes and reports frequentist model results for ACR50 dataset. The relative treatment effects and impact of study covariates in the ACR50 and ACR70 dataset is presented in table 9 and Table 11 respectively. Table 10 and Table 12 present the model comparison results for ACR50 and ACR70 dataset.

In the ACR50 dataset, duration of disease demonstrated significant impact on heterogeneity and model fitness. Therefore, inference regarding relative treatment effect will be based on the mixed-effects model with covariate duration of disease. CTZ appeared to be the most efficacious treatment. ANA again was the least efficacious treatment option.

In the ACR70 dataset, duration of disease and SJC both demonstrated significant impact on heterogeneity and model fitness. Therefore, inference regarding relative treatment effect will be based on the mixed-effects model with covariate duration of

disease. When adjusted for duration of disease, CTZ appeared to be the most efficacious treatment, whereas when adjusted for SJC, GLB was the most efficacious treatment. However, due to the small number of patients that reached the treatment success in ACR70 treatment criteria, the effects of study covariates observed in the ACR70 dataset could be by chance.

Table 9. Covariate Assessment and Ranked Treatments, Frequentist Models, ACR50 Dataset

| Model | Covariate OR | Ranked Treatments and ORs | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th |
| FE Model (df = 12) | N/A | ETN | CTZ | TCZ | ADA | GLB | RTX | IFX | ANA | ABT |
| | | 18.53** | 6.979*** | 4.951*** | 4.875*** | 4.330*** | 3.446*** | 3.172*** | 2.399* | 2.148*** |
| RE Model (df=12) | N/A | ETN | CTZ | TCZ | ADA | GLB | IFX | RTX | ABT | ANA |
| | | 12.66* | 6.729*** | 5.118*** | 4.925*** | 4.674** | 3.567*** | 3.364* | 2.619*** | 2.367 |
| ME Model (df = 11) | Age | ETN | CTZ | TCZ | GLB | ADA | RTX | IFX | ABT | ANA |
| | 1.096 | 16.33** | 6.682*** | 5.556*** | 4.563** | 4.013*** | 3.580* | 3.329*** | 2.872*** | 1.637 |
| | Female Percentage | ETN | CTZ | TCZ | ADA | GLB | IFX | RTX | ABT | ANA |
| | 1.004 | 12.48* | 6.772*** | 5.093*** | 4.981*** | 4.698** | 3.602*** | 3.343* | 2.638*** | 2.397 |
| | Duration of Disease | CTZ | ETN | GLB | TCZ | RTX | ADA | IFX | ABT | ANA |
| | 1.161** | 7.490*** | 5.833 | 5.811*** | 4.565*** | 3.767*** | 3.618*** | 3.544*** | 3.050*** | 1.585 |
| | SJC | ETN | GLB | ADA | TCZ | IFX | CTZ | RTX | ABT | ANA |
| | 1.148* | 12.20** | 10.50*** | 5.856*** | 5.325*** | 4.504*** | 4.459*** | 2.964* | 2.343*** | 1.971 |
| | TJC | ETN | GLB | ADA | CTZ | IFX | TCZ | RTX | ANA | ABT |
| | 1.113 | 12.83** | 7.557*** | 5.411*** | 5.349*** | 4.760*** | 4.016*** | 3.168* | 3.089* | 2.338*** |

Table 10. Heterogeneity and Fitness Comparison, Frequentist Models, ACR50 Dataset

| Model | Test of (Residual) Heterogeneity | I^2 | AIC | BIC |
|---|---|---|---|---|
| FE model | Q [df = 12] = 50.215, p-value < 0.001 | 76.1% | 98.06 | 112.6 |
| RE model | Q [df = 12] = 49.156, p-value < 0.001 | 75.6% | 41.11 | 45.96 |
| ME model with covariate age | Q [df = 11] = 48.277, p-value < 0.001 | 77.2% | 41.35 | 45.73 |
| ME model with covariate female percentage | Q [df = 11] = 48.803, p-value < 0.001 | 77.5% | 42.22 | 46.60 |
| ME model with covariate duration of disease | Q [df = 11] = 20.311, p-value = 0.041 | 45.8% | 37.14 | 41.51 |
| ME model with covariate SJC | Q [df = 11] = 29.133, p-value = 0.002 | 62.2% | 39.39 | 43.77 |
| ME model with covariate TJC | Q [df = 11] = 37.131, p-value < 0.001 | 70.4% | 40.32 | 44.70 |

Table 11. Covariate Assessment and Ranked Treatments, Frequentist Models, ACR70 Dataset

| Model | Covariate OR | Ranked Treatments and ORs | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th |
| FE Model (df = 12) | N/A | ETN | TCZ | CTZ | ADA | GLB | ANA | IFX | RTX | ABT |
| | | 10.98 | 8.587*** | 8.146*** | 6.671*** | 4.711* | 3.140 | 2.890* | 1.900 | 1.856 |
| RE Model (df=12) | N/A | ETN | TCZ | CTZ | ADA | GLB | IFX | ANA | ABT | RTX |
| | | 11.48 | 8.496*** | 6.844*** | 5.722*** | 4.728* | 3.443*** | 2.950 | 2.530** | 1.831 |
| ME Model (df = 11) | Age | ETN | TCZ | CTZ | GLB | ADA | IFX | ABT | RTX | ANA |
| | 1.127 | 16.00 | 9.388*** | 7.222*** | 4.793* | 4.261** | 3.177** | 2.985** | 1.986 | 1.820 |
| | Female Percentage | ETN | TCZ | CTZ | ADA | GLB | IFX | ANA | ABT | RTX |
| | 0.961 | 13.39 | 9.048*** | 7.357*** | 5.249*** | 4.889** | 3.267*** | 2.574 | 2.518*** | 1.960 |
| | Duration of Disease | CTZ | TCZ | GLB | ETN | ADA | IFX | ABT | RTX | ANA |
| | 1.171*** | 9.577*** | 7.121*** | 6.286*** | 5.057 | 4.187*** | 3.493*** | 2.966*** | 2.063 | 1.930 |
| | SJC | GLB | ETN | TCZ | ADA | IFX | CTZ | ANA | ABT | RTX |
| | 1.228*** | 16.49*** | 10.87 | 9.210*** | 7.543*** | 5.110*** | 4.263*** | 2.245 | 2.188*** | 1.516 |
| | TJC | ETN | GLB | ADA | TCZ | IFX | CTZ | ANA | ABT | RTX |
| | 1.160* | 11.68 | 9.629*** | 6.575*** | 6.191*** | 5.394*** | 5.150*** | 4.260* | 2.177*** | 1.685 |

Table 12. Heterogeneity and Fitness Comparison, Frequentist Models, ACR70 Dataset

| Model | Test of (Residual) Heterogeneity | I^2 | AIC | BIC |
|---|---|---|---|---|
| FE model | Q [df = 12] = 29.980, p-value = 0.003 | 60.0% | 91.92 | 106.5 |
| RE model | Q [df = 12] = 30.169, p-value = 0.003 | 60.2% | 50.51 | 55.36 |
| ME model with covariate age | Q [df = 11] = 29.690, p-value = 0.002 | 63.0% | 50.21 | 54.59 |
| ME model with covariate female percentage | Q [df = 11] = 27.628, p-value = 0.004 | 60.2% | 50.33 | 54.71 |
| ME model with covariate duration of disease | Q [df = 11] = 14.659, p-value = 0.199 | 25.0% | 46.04 | 50.42 |
| ME model with covariate SJC | Q [df = 11] = 12.815, p-value = 0.306 | 14.2% | 44.08 | 48.46 |
| ME model with covariate TJC | Q [df = 11] = 18.418, p-value = 0.072 | 40.3% | 47.88 | 52.26 |

## MANUSCRIPT INTRODUCTORY PAGE

The manuscript “Bayesian Network Meta-analysis for Biologic Therapies in Rheumatoid Arthritis” is prepared for submission to the journal of *Modern Rheumatology*.

**MANUSCRIPT**

## Bayesian Network Meta-analysis for Biologic Therapies in Rheumatoid Arthritis

### Introduction

Rheumatoid arthritis (RA) is a systemic chronic autoimmune disease with substantial burden. The 2013 Nationwide Inpatient Survey estimated that RA was listed as the principal diagnosis for 8,115 hospitalizations with total hospital charges of $359 million.[46] In 2007, there were 2.6 physician office visits including 1.9 million visits to a specialist.[47] The 2005 United States Medical Expenditure Panel Survey (MEPS) showed the medical costs associated with RA were 22.3 billion ($2,085 per person).[48] The resulting swollen and painful joints often lead to disabilities, diminished quality of life, and premature mortality.[13] Cost-effective conventional diseases modifying anti-rheumatic drugs (DMARDs) have served as first-line therapy to relieve the symptoms. However, patients tend to develop tolerance and have diminished response over time. Switching to or additional biologic agents is recommended in the event of inadequate response to conventional DMARDs. Biologic DMARDs further divide into several classes: (1) tumor necrosis factor inhibitors (TNFi): adalimumab (ADA), certolizumab pegol (CTZ), etanercept (ETN), golimumab (GLB), and infliximab (IFX); (2) T-cell costimulatory modulato: abatacept (ABT); (3) recombinant human interleukin-1 (IL-1) receptor antagonist: anakinra (ANA); (4) B-cell-depleting monoclonal antibody: rituximab (RTX); (5) monoclonal antibody to the IL-6 receptor: tocilizumab (TCZ); and (6) janus kinase inhibitor: tofacitinib (TOF). All these drugs have undergone randomized clinical trials

against methotrexate for regulatory approval but there have been few head-to-head trials to compare the relative efficacy among them.

Randomized clinical trial, the gold standard for evidence-based health decision-making, commonly compares one therapeutic intervention with placebo or standard care. Even though a few clinical trials have involved more than one intervention, few studies have looked at more than two interventions. Driven by the need for relative treatment efficacy assessment, advanced methods have been developed to incorporate multiple interventions. Unlike standard meta-analysis, which can only incorporate trials of one intervention, network meta-analysis can incorporate clinical trials of multiple interventions and can estimate the relative efficacy between any pair of interventions. With a series of direct comparisons between a common comparator (placebo) and one or two interventions, biologic therapies can be indirectly compared. Bayesian network meta-analysis can incorporate previous knowledge about the treatment effects in form of prior information of effect sizes and provide a posterior probability distribution for the effect size and its 95% credible interval (CrI).[4,17]

The difference between clinical trials could also contribute to the effect size estimation. Study characteristics and patient demographics between clinical trials might differ and randomization does not hold across these studies.[4,6,7] Potential treatment effect modifiers (study-level covariates) may be caused by patient characteristics, inclusion and exclusion criteria, study protocol development, study outcome measurement, and administrative agency requirement. Treatment effect modifiers could bias the results of indirect comparisons and alter the magnitude and direction of treatment effect estimation.[7,8] There are several published network meta-analyses of biologics DMARDs

but few have examined the effects of study level covariates. Therefore, the present network meta-analysis was conducted in the Bayesian framework to compare the relative efficacy of biological DMARDs when used in combination with methotrexate (MTX).

## Materials and Methods

### *Materials*

A systematic literature review was conducted to collect data from peer-reviewed journal articles. A pre-determined search strategy (see Appendix A) was used to search for qualifying clinical trials in PubMed electronic databases through January 1999 to December 2015. Identified studies were screened based on titles and abstracts followed by a full-text review on selected studies. The references of full-text reviewed studies were also examined for potential additions.

Studies included had to meet the following inclusion criteria: (1) the study was a multisite randomized clinical trial for biologic therapies of interest; (2) the study was conducted in adult human rheumatoid arthritis patients; (3) the study outcome was measured with American College of Rheumatology (ACR) criteria at week 24. The following exclusion criteria were also applied: (1) the study population included patients of other inflammatory diseases such as juvenile arthritis, Crohn's disease, or psoriatic arthritis; (2) the study patients did not receive background MTX; (3) the study did not report number of patients with a 50% or 70% improvement in ACR criteria (ACR50 or ACR70).

Study outcome data extracted was the number and percentage of responders to treatment based on the ACR criteria at week 24. Treatment outcomes measured in ACR20/50/70 were considered as separate datasets and evaluated respectively. Study characteristics including authorship, year of publication, and biologic intervention received were extracted from selected papers. Patient demographics including mean age, percentage of females, mean disease duration in years, swollen joint count (SJC), tender joint count (TJC), C-reactive protein (CRP) level, erythrocyte sedimentation rate (ESR), and percentage of rheumatoid factor positive (RF+) were also extracted from selected papers. Means and standard errors were also extracted for continuous patient demographics if available. Data extracted for ESR and CRP were converted to the unit of mm/h and mg/L respectively if reported in other units. Study covariates were centered at their weighted means to make the results more interpretable.

When other forms of summary data such as the geometric mean, median, range, or interquartile were reported, the arithmetic mean was calculated. Study covariates reported in geometric mean were considered missing. When a study reported the median ($m$), low and high end of the range ($a$ and $b$), and sample size ($n$), the median was used as the the mean when $n \geq 25$ and the mean was calculated by $\frac{a+2m+b}{4}$ when $n < 25$.[14,15] When a study reported $m$, first and third interquartile ($q_1$ and $q_3$), and $n$, the mean was calculated by the average of $m$, $q_1$, and $q_3$ regardless of the sample size.[14] Missing data in epidemiological and clinical research undermines the power and validity of clinical researches.[49] The predictive mean matching approach was used to perform multiple imputation on missing values in covariates. Three imputed values were predicted for each

missing value for analysis and the results of the three imputed datasets were pooled for final treatment effect estimation.

*Model Selection*

There are several choices of models for network meta-analysis. The fixed-effects network meta-analysis model (FE model) which assumes homogeneity of effect sizes among selected studies was fit to the datasets first.

The random-effects model (RE model) assumes that included studies are heterogeneous and the difference observed in effect sizes is not due to sampling error. RE model treats measured effect sizes and estimated model parameters as the random sample of the true population effect size and true model parameters.

Then mixed-effects models (ME models) were used to investigate the sources of heterogeneity between included studies and the contribution to heterogeneity of included covariates. The ME model, also known as meta-regression, includes study-level treatment effect modifiers or covariates and can further evaluate their contribution to heterogeneity. The treatment effect modifier $X_j$ is incorporated in the model via the mean of the distribution of random-effects.

*Statistical Analysis*

Bayesian network meta-analysis models were used to estimate relative efficacy between biologic DMARDs. In the Bayesian framework, model parameters are considered random variables. The prior probability distribution reflects prior beliefs about possible values of the parameters. The likelihood represents the probably

probability of the data as a function of the parameters. A posterior probability distribution of the parameter is obtained by combining the prior probability distribution and the likelihood. Network meta-analysis models usually have multiple parameters for effect size functions. The Monte Carlo method generates random sample values of the parameters from their posterior distributions and approximates all of these posterior quantities of interest to an arbitrary degree of precision. The empirical distribution provides a closer approximation to the true density as the number of repeated sample increases. For complicated multi-parametric models such as network meta-analysis, it is difficult to sample from the joint posterior distribution directly. In such cases, Gibbs sampler provides a computationally feasible alternative in which samples are drawn from the full conditional distributions of parameters. In this so-called Markov Chain, the algorithm will converge to the true distributions as the number of sampling goes to up towards infinity, By applying the same algorithm to the regular Monte Carlo approximation, the joint posterior distributions of the parameters could be approximated by this Markov Chain Monte Carlo (MCMC) method.

No assumption was made for the treatment effect parameters, the heterogeneity parameter, and the covariate regression coefficients. The posterior distribution was estimated with non-informative priors to limit inference to collected data only. A uniform prior distribution (*uniform* [0, 2]) was used for the heterogeneity parameter σ in the random- and mixed-effects models. No prior distributions were assigned for the treatment effect parameters $d_{Ak}$ and the regression coefficients $\beta_{Ak}$ in the mixed-effect models.

Rank probabilities were calculated for the treatment options for ACR20/50/70. The rank probability indicates the probabilities for each treatment to be the most, second,

third… the least efficacy treatment option. The treatments were ranked by their efficacy relative to placebo for each of the MCMC iterations. A frequency table based on these rankings was normalized by the number of iterations to calculate the rank probabilities.

Heterogeneity was tested with the $I^2$ statistics, which describes the percentage of variation across studies. $I^2$ is an intuitive and simple expression of the heterogeneity of studies. The random-effects model was used if a significant variation in measured effect sizes among included studies was observed. In the mixed-effects model, $I^2$ estimates the percentage of the unaccounted variability (residual heterogeneity plus sampling variability) attributed to residual heterogeneity. The Good Research Practice guideline of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) suggested the following thresholds: $I^2 < 30\%$ is considered mild heterogeneity, and $I^2 > 50\%$ is considered notable heterogeneity.[17] Discrepancy assessment of model fitness was conducted for goodness of fit. The deviance information criterion (DIC) is hierarchical model selection criteria and is particularly useful in Bayesian analysis. A smaller value of DIC indicates a better model fitness.

All analyses were conducted using the statistical program R. R package *gemtc* (version 0.8, released March 1, 2016) was used for the Bayesian analysis. R package *mice* (version 2.24, released November 9, 2015) was used for multiple imputation. Bayesian software JAGS (Just Another Gibbs Sampler, version 4.2.0, released February 19, 2016)) was used for Gibbs sampling.

## Results

*Systematic Review*

The search strategy identified 1,337 articles. After title and abstract review, 1,301 studies were excluded for the following reasons: (1) the study was not conducted for the biologics of interest (n = 962); (2) the study was not a multisite randomized clinical trial (n = 242); (3) the study utilized different treatment outcome measurement (n = 63); (4) the study was not conducted in other population (n = 34). After full-text review, 16 articles were further excluded due to the following reasons: (1) the patients were receiving other concomitant RA treatments other (n=10); (2) the study did not have background MTX in combination with the biologics (n=5); (3) the clinical trial for toficitanib, an oral biologic agent with different indication from other biologics (n=1).

The 20 selected studies included a total of 7,666 patients that received biologic DMARDs for rheumatoid arthritis. Treatment arms were balanced within each study indicating sufficient randomization. Table 1 shows the study characteristics and patient demographics of each study, whereas Table 2 shows the numbers of responders and response rates for ACR20/50/70 criteria[25-44]. The evidence network is demonstrated in Figure 2. Each node represents a biologic therapy. Each edge between nodes represents a direct comparison and the width of the edges represents the number of clinical trials contributing to the comparison pair. Placebo was the common comparator among all included studies. Nineteen studies (95%) were two-armed clinical trials that compared one biologic therapy with placebo while one study (5%) was a three-armed trial that compared two biologics and placebo. For study covariates, three (15%) studies reported

the median for all patient demographic covariates, two (10%) studies reported the median for ESR and CRP and mean for other covariates, and two (10%) other studies report the geometric mean for also ESR and CRP. Baseline ESR, CRP, and RF+ were missing in the ten (50%), one (5%), and two (10%) studies, respectively. Imputed values were analyzed separately before pooling model estimations for interpretation.

*Network Meta-analysis Results*

Model results for estimated relative efficacy of biologic DMARDs are demonstrated in this section. ACR20/50/70 outcomes were analyzed respectively. This section summarizes and reports model results for ACR20 dataset. The ACR20 response rate ranged from 38.0% to 71.0% (Table 2). Notable heterogeneity was observed in the dataset (Fixed-effects model: $I^2 = 58\%$). The Random-effects model demonstrated an improvement in model fitness from the fixed-effects model but had similar DIC values when covariates were included (Appendix B). Therefore, the random-effects network meta-analysis model was used for the relative treatment effect estimation. All biologics except ANA demonstrated a statistically significant superiority over placebo, whereas ANA showed a statistically non-significant superiority over placebo (Figure 7). The rank probability (Figure 8) demonstrated the probabilities of each treatment to be the best to the least efficacious among the treatment options. ETN had the highest probability (42.6%) to be the most efficacious treatment. Followed by CTZ, which had similar probabilities for being the best and the second best (36.7% vs 35.5). Rank probabilities of each treatment are found in Table 13. Effect of study covariates was estimated with mixed-effects models for each covariate. Duration of disease had the largest but not statistically significant impact among all covariates (OR=1.59, 95% CrI=0.71-3.48).

Female percentage (OR=1.27, 95% CrI=0.63-2.53), mean age (OR=1.27, 95% CrI=0.54-3.11), TJC (OR=1.21, 95% CrI=0.52-2.91, and SJC (OR=1.20, 95% CrI=0.42-3.44) also had no significant impact on treatment effects.

For the ACR50 criteria data, the response rate ranged from 17.0% to 55.2%. Notable heterogeneity was also observed in the dataset (Fixed-effects model: $I^2$ = 51%). The same improvement in model fitness and reduction in $I^2$ was observed in random-effects model. The covariate duration of disease had a significant effect on treatment effect (OR=2.05. 95% CrI=1.05-3.71). Covariates mean age (OR=1.87, 95% CrI=0.79-4.50), SJC (OR=1.61, 95% CrI=0.57-4.40), TJC (OR=1.40, 95% CrI=0.58-3.51), and female percentage (OR=0.92, 95% CrI=0.44-1.98) had no significant impact on treatment effects. Hence, the mixed-effects model with covariate duration of disease was used for the relative treatment effect estimation. The same treatment superiority over placebo was observed for all biologics except ANA (Figure 9). The rank probability for ACR50 criteria (Figure 10) showed that ETN had the highest probability (57.6%) to be the best treatment option. CTZ had the highest probability for being the second best treatment (45.4%), and ANA (66.9%) had the highest probability to be the least efficacious biologic. Detailed rank probabilities of each treatment can be found in Table 14.

This section summarizes and reports model results for ACR70 dataset. The ACR70 response rate ranged from 0% to 26.9%. Moderate heterogeneity was observed in the dataset ($I^2$ = 43%). None of the study covariates had a significant impact on treatment effects. The random-effects model had the smallest DIC. Therefore, the random-effects model was used for treatment effect inference. In the ACR70 criteria data, all the biologics except for ANA and RTX demonstrated a statistically significant benefit over

placebo (Figure 11). The rank probability (Figure 12) showed that ETN had the highest probability (98.2%) to be the most superior treatment. The second best treatment for ACR70 dataset was TCZ (34.7%). RTX had the highest probability to be the least efficacious treatment (29.1%). Specific rank probabilities of each treatment can be found in Table 15.

Table 13. Rank Probabilities of Treatment Groups, ACR20 Dataset

| Treatment | Probability being the nth ranked treatment | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
| ABT | 0.12% | 0.55% | 1.58% | 4.00% | 7.94% | 14.43% | 22.37% | 28.78% | 19.98% | 0.26% |
| ADA | 2.63% | 10.43% | 20.27% | 21.07% | 19.42% | 12.79% | 7.87% | 4.21% | 1.31% | 0.02% |
| ANA | 1.97% | 3.28% | 4.85% | 6.07% | 7.28% | 8.96% | 11.27% | 16.31% | 30.15% | 9.87% |
| CTZ | 36.77% | 35.48% | 14.47% | 6.79% | 3.44% | 1.83% | 0.81% | 0.34% | 0.08% | 0.00% |
| ETN | 42.67% | 18.66% | 10.87% | 7.78% | 5.38% | 4.31% | 3.52% | 3.27% | 2.78% | 0.79% |
| GLB | 3.28% | 7.30% | 11.59% | 13.21% | 13.80% | 14.13% | 13.63% | 12.19% | 9.83% | 1.05% |
| IFX | 0.62% | 2.54% | 6.17% | 10.49% | 15.28% | 19.99% | 20.18% | 16.02% | 8.46% | 0.26% |
| Placebo | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.07% | 1.07% | 13.41% | 85.46% |
| RTX | 8.03% | 11.49% | 13.80% | 12.11% | 11.69% | 10.97% | 10.28% | 10.22% | 9.29% | 2.12% |
| TCZ | 3.93% | 10.27% | 16.42% | 18.49% | 15.78% | 12.60% | 10.02% | 7.60% | 4.72% | 0.18% |

Table 14. Rank Probabilities of Treatment Groups, ACR50 Dataset

| Treatment | Probability being the nth ranked treatment | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
| ABT | 0.02% | 0.27% | 1.29% | 3.93% | 9.09% | 17.88% | 28.22% | 33.12% | 6.17% | 0.02% |
| ADA | 0.38% | 2.27% | 8.12% | 16.02% | 22.19% | 20.62% | 16.27% | 12.28% | 1.88% | 0.00% |
| ANA | 0.13% | 0.35% | 0.80% | 1.07% | 1.85% | 2.73% | 4.45% | 10.03% | 66.91% | 11.69% |
| CTZ | 27.65% | 45.41% | 17.73% | 5.18% | 2.38% | 1.04% | 0.37% | 0.18% | 0.06% | 0.01% |
| ETN | 57.64% | 11.66% | 7.94% | 4.68% | 3.34% | 3.03% | 3.28% | 5.42% | 2.49% | 0.53% |
| GLB | 10.67% | 25.78% | 29.61% | 14.24% | 7.91% | 5.34% | 3.30% | 2.44% | 0.70% | 0.02% |
| IFX | 0.23% | 1.24% | 4.53% | 11.19% | 18.18% | 24.09% | 23.39% | 14.71% | 2.43% | 0.01% |
| Placebo | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.01% | 0.32% | 12.48% | 87.19% |
| RTX | 1.87% | 6.39% | 11.76% | 15.66% | 14.70% | 13.23% | 12.98% | 17.03% | 5.86% | 0.53% |
| TCZ | 1.43% | 6.64% | 18.23% | 28.03% | 20.36% | 12.04% | 7.74% | 4.49% | 1.03% | 0.02% |

Table 15. Rank Probabilities of Treatment Groups, ACR70 Dataset

| Treatment | Probability being the nth ranked treatment | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
| ABT | 0.01% | 0.61% | 1.88% | 4.36% | 8.34% | 15.54% | 26.08% | 29.21% | 13.21% | 0.76% |
| ADA | 0.16% | 12.68% | 20.20% | 22.96% | 19.02% | 12.25% | 7.61% | 3.82% | 1.23% | 0.08% |
| ANA | 0.21% | 7.10% | 7.85% | 8.05% | 9.74% | 11.67% | 12.84% | 17.18% | 15.28% | 10.08% |
| CTZ | 0.23% | 17.01% | 18.64% | 17.51% | 14.96% | 11.74% | 9.60% | 6.61% | 2.97% | 0.74% |
| ETN | 98.15% | 0.55% | 0.23% | 0.32% | 0.14% | 0.13% | 0.15% | 0.14% | 0.17% | 0.03% |
| GLB | 0.50% | 20.66% | 16.07% | 14.84% | 13.83% | 11.63% | 9.49% | 7.88% | 4.18% | 0.93% |
| IFX | 0.10% | 4.40% | 8.11% | 12.68% | 17.29% | 22.71% | 19.01% | 11.21% | 4.18% | 0.33% |
| Placebo | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.05% | 0.42% | 4.51% | 28.68% | 66.35% |
| RTX | 0.02% | 2.20% | 2.98% | 4.20% | 5.89% | 7.38% | 10.53% | 17.18% | 29.13% | 20.49% |
| TCZ | 0.63% | 34.79% | 24.04% | 15.09% | 10.79% | 6.92% | 4.27% | 2.28% | 0.99% | 0.20% |

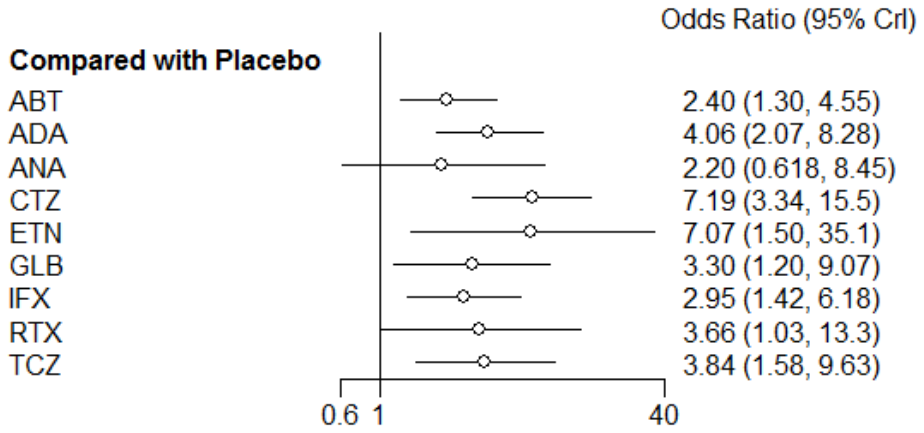


Figure 7. Forest Plot of Treatment Effects, Bayesian RE Model, ACR20 Dataset

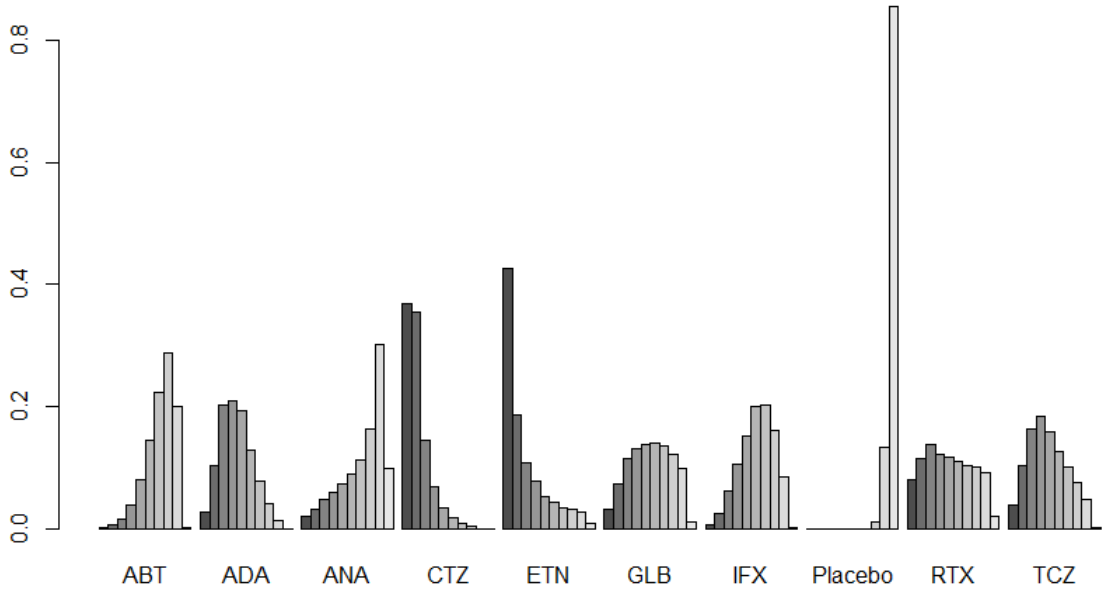


Figure 8. Rank Probabilities Plot, Bayesian RE Model, ACR20 Dataset

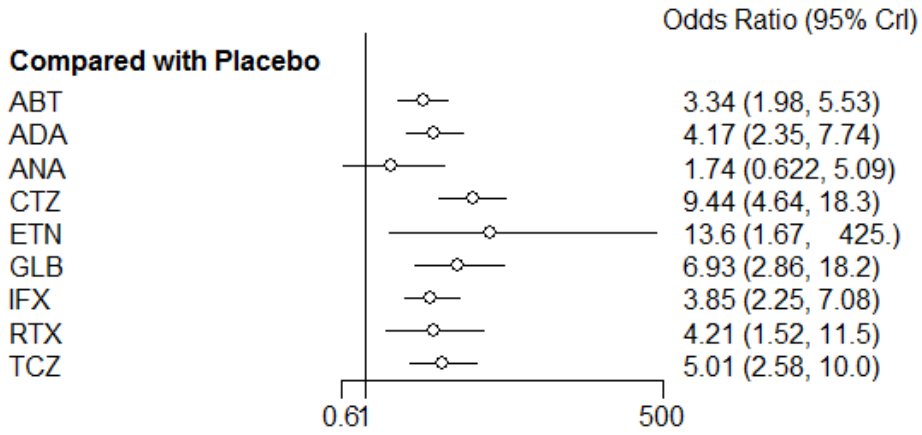


Figure 9. Forest Plot of Treatment Effects, Bayesian ME Model with Duration of Disease, ACR50 Dataset

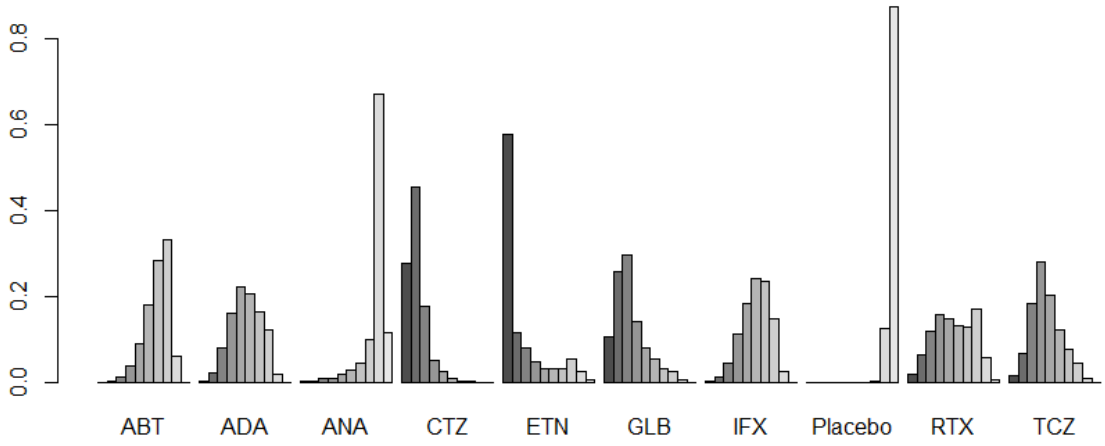


Figure 10. Rank Probabilities Plot, Bayesian ME Model with Duration of Disease, ACR50 Dataset

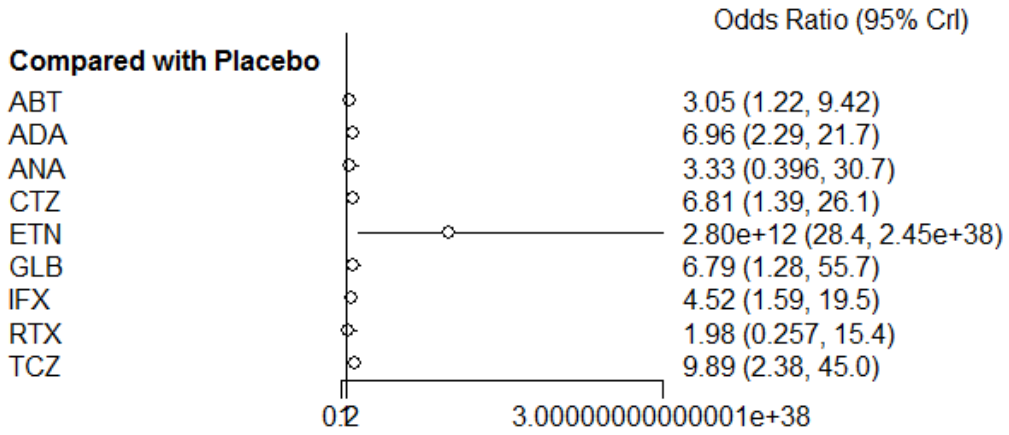


Figure 11. Forest Plot of Treatment Effects, Bayesian RE Model, ACR70 Dataset

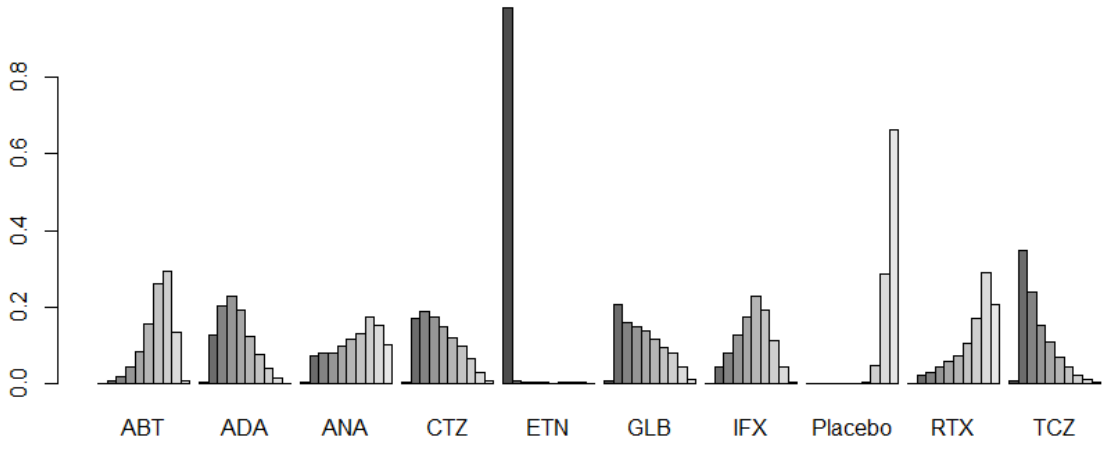


Figure 12. Rank Probabilities Plot, Bayesian RE Model, ACR70 Dataset

## Discussion

The advantage of network meta-analysis is that it can compare multiple treatments via an evidence network and increase the precision of measurement by pooling effect sizes. However, in this study, the ACR70 dataset had a small number of responders, which led to less accurate treatment effect estimation. The mean ACR70 response rate was 10.0% among all study arms whereas the mean ACR50 and ACR20 response rate were 23.6% and 44.3% respectively. The small sample sizes contributed to the less accurate treatment effect estimation and wide credible intervals in ACR70 results.

In addition, the ACR70 criteria were more stringent than those of ACR20 and ACR50. It required the patients to have at least a 70% improvement in the number of tender and swollen joints, and a 70% improvement in at least 3 of the following: (1) the patient's global assessment of disease status; (2) the patient's assessment of pain; (3) the patient's assessment of function; (4) the physician's global assessment of disease status; (5) serum C-reactive protein levels. These are the highest ACR criteria for effective treatment, which is difficult to achieve in 6 month of treatment. Four studies arms had no patients that reached this outcome.

Despite rigorous testing for effect size modifying covariate, duration of disease was the only study covariate to have statistically significant effect on treatment effect. This effect was not observed in the ACR20 or ACR70 data. The inclusion of study covariates did not have notable improvement on either DIC for model fitness or $I^2$

statistic for heterogeneity assessment. Therefore, the effect of duration of disease on treatment outcome in ACR50 data is most likely to be caused by chance alone.

The study is limited in several ways. First, the number of study and number of patients included was small. We included studies using placebo as the common comparator to minimize the bias inconsistency that additional studies might introduce. As a result, only 20 studies were included for the nine biologic products of interest. Second, pooling randomized clinical trials in a network meta-analysis is not the same as conducting a randomized controlled trial of the nine different drugs because randomization is compromised. Effects of a study-level covariate estimated in the mixed-effects model might not represent the true effect of the study covariate at the patient level. In addition, this study only assessed ACR criteria at week 24. Information from clinical trials used different outcome measurement or measured treatment outcomes at different endpoint were not included in the analysis.

Based on the network meta-analysis, we found that all biologics DMARDs were superior to placebo except for ANA in all datasets and RTX in ACR70 dataset. ETN was had the highest probability to be the best treatment in all three datasets. CTZ had the highest probability to be the second best option in ACR20 and ACR50 datasets, and TCZ held the second place in the ACR70 dataset. The rest of the rank probabilities vary by dataset but placebo was the lowest ranked option in all datasets. Therefore, despite the limitations of this study, the results are consistent with current knowledge that biologic DMARDs are superior to placebo and although more research remains to be done, ETN may be the most effective option for rheumatoid arthritis.

## APPENDIX A

## SEARCH STRATEGY FOR SYSTEMATIC LITERATURE REVIEW

The following search terms and MeSH (Medical Subject Headings) terms were used to identify eligible studies for this network meta-analysis:

("Certolizumab Pegol" OR "Abatacept" OR "Infliximab" OR "Rituximab" OR "Adalimumab" OR "Etanercept" OR "golimumab" OR "tocilizumab" OR "tofacitinib")

AND

"Randomized Controlled Trial" [Publication Type]

AND

(ACR OR ACR20 OR “ACR 20” OR ACR50 OR “ACR50” OR ACR70 OR “ACR 70” OR "American College of Rheumatology" OR "American College of Rheumatology 20% improvement criteria" OR "American College of Rheumatology 50% improvement criteria" OR "American College of Rheumatology 70% improvement criteria")

AND

("24 weeks" OR “week 24” OR “week-24”)

NOT

"psoriatic arthritis"

## APPENDIX B

## MODEL FITNESS AND HETEROGENEITY

This appendix presents model fitness and heterogeneity assessment of all models. The use of random-effects model (RE model) increased model fitness and alleviated unaccounted for heterogeneity compared with fixed-effects model (FE model) in all three datasets. The inclusion of study covariates did not have notable impact on either model fitness or heterogeneity among mixed-effects models (ME models).

Table 16. Model Fitness and Heterogeneity Comparison, Bayesian Models

| Dataset | Model | DIC | I^2 |
|---|---|---|---|
| ACR20 | FE model | 123.66 | 58% |
| | ME model with covariate age | 80.96 | 4% |
| | ME model with covariate duration of disease | 80.87 | 4% |
| | ME model with covariate SJC | 80.68 | 4% |
| | ME model with covariate RF+ | 80.66 | 4% |
| | ME model with covariate ESR | 80.46 | 3% |
| | RE model | 80.45 | 3% |
| | ME model with covariate CRP | 80.42 | 4% |
| | ME model with covariate female percentage | 80.41 | 3% |
| | ME model with covariate TJC | 80.36 | 3% |
| ACR50 | FE model | 110.48 | 51% |
| | ME model with covariate duration of disease | 80.18 | 9% |
| | ME model with covariate female percentage | 80.00 | 4% |
| | ME model with covariate SJC | 79.54 | 4% |
| | RE model | 79.33 | 3% |
| | ME model with covariate TJC | 79.20 | 3% |
| | ME model with covariate age | 78.99 | 3% |
| ACR70 | FE model | 98.61 | 43% |
| | ME model with covariate duration of disease | 83.57 | 19% |
| | ME model with covariate female percentage | 83.14 | 14% |
| | ME model with covariate SJC | 82.55 | 16% |
| | ME model with covariate age | 81.58 | 12% |
| | ME model with covariate TJC | 81.35 | 12% |
| | RE model | 80.80 | 12% |

## BIBLIOGRAPHY


1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71-2.

2. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. 1st ed: Academic Press; 1985.

3. Haidich AB. Meta-Analysis in Medical Research. Hippokratia 2010;14:29-37.

4. Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health 2011;14:417-28.

5. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005;331:897-900.

6. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;9:1-134, iii-iv.

7. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med 2009;28:1861-81.

8. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003;326:472.

9. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ 2011;343:d4909.

10. Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. Evaluation of inconsistency in networks of interventions. Int J Epidemiol 2013;42:332-45.

11. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. PLoS One 2014;9:e99682.

12. Boonen A, Severens JL. The burden of illness of rheumatoid arthritis. Clinical rheumatology 2011;30 Suppl 1:S3-8.

13. Sihvonen S, Korpela M, Laippala P, Mustonen J, Pasternack A. Death rates and causes of death in patients with rheumatoid arthritis: a population-based study. Scandinavian journal of rheumatology 2004;33:221-7.

14. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 2014;14:135.

15. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005;5:13.

16. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105-24.

17. Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value Health 2011;14:429-37.

18. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc 2009;172:137-59.

19. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res 2001;10:277-303.

20. Hoff PD. A First Course in Bayesian Statistical Methods. 1st ed: Springer; 2009.

21. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557-60.

22. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539-58.

23. Gavaghan DJ, Moore RA, McQuay HJ. An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. Pain 2000;85:415-24.

24. R Development Core Team. The R Project for Statistical Computing. 3.2.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2016.

25. Keystone EC, Kavanaugh AF, Sharp JT, et al. Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: a randomized, placebo-controlled, 52-week trial. Arthritis Rheum 2004;50:1400-11.

26. Keystone E, Heijde D, Mason D, Jr., et al. Certolizumab pegol plus methotrexate is significantly more effective than placebo plus methotrexate in active rheumatoid

arthritis: findings of a fifty-two-week, phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group study. Arthritis Rheum 2008;58:3319-29.

27. Keystone EC, Genovese MC, Klareskog L, et al. Golimumab, a human antibody to tumour necrosis factor {alpha} given by monthly subcutaneous injections, in active rheumatoid arthritis despite methotrexate therapy: the GO-FORWARD Study. Ann Rheum Dis 2009;68:789-96.

28. Weinblatt ME, Kremer JM, Bankhurst AD, et al. A trial of etanercept, a recombinant tumor necrosis factor receptor:Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. N Engl J Med 1999;340:253-9.

29. Weinblatt ME, Keystone EC, Furst DE, et al. Adalimumab, a fully human anti-tumor necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. Arthritis Rheum 2003;48:35-45.

30. Weinblatt ME, Schiff M, Valente R, et al. Head-to-head comparison of subcutaneous abatacept versus adalimumab for rheumatoid arthritis: findings of a phase IIIb, multinational, prospective, randomized study. Arthritis Rheum 2013;65:28-38.

31. Smolen J, Landewe RB, Mease P, et al. Efficacy and safety of certolizumab pegol plus methotrexate in active rheumatoid arthritis: the RAPID 2 study. A randomised controlled trial. Ann Rheum Dis 2009;68:797-804.

32. van Vollenhoven RF, Kinnman N, Vincent E, Wax S, Bathon J. Atacicept in patients with rheumatoid arthritis and an inadequate response to methotrexate: results of a phase II, randomized, placebo-controlled trial. Arthritis Rheum 2011;63:1782-92.

33. van Vollenhoven RF, Fleischmann R, Cohen S, et al. Tofacitinib or adalimumab versus placebo in rheumatoid arthritis. N Engl J Med 2012;367:508-19.

34. Choy E, McKenna F, Vencovsky J, et al. Certolizumab pegol plus MTX administered every 4 weeks is effective in patients with RA who are partial responders to MTX. Rheumatology (Oxford) 2012;51:1226-34.

35. Kay J, Matteson EL, Dasgupta B, et al. Golimumab in patients with active rheumatoid arthritis despite treatment with methotrexate: a randomized, double-blind, placebo-controlled, dose-ranging study. Arthritis Rheum 2008;58:964-75.

36. Maini R, St Clair EW, Breedveld F, et al. Infliximab (chimeric anti-tumour necrosis factor alpha monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial. ATTRACT Study Group. Lancet 1999;354:1932-9.

37. Westhovens R, Yocum D, Han J, et al. The safety of infliximab, combined with background treatments, among patients with rheumatoid arthritis and various

comorbidities: a large, randomized, placebo-controlled trial. Arthritis Rheum 2006;54:1075-86.

38. Schiff M, Keiserman M, Codding C, et al. Efficacy and safety of abatacept or infliximab vs placebo in ATTEST: a phase III, multi-centre, randomised, double-blind, placebo-controlled study in patients with rheumatoid arthritis and an inadequate response to methotrexate. Ann Rheum Dis 2008;67:1096-103.

39. Smolen JS, Beaulieu A, Rubbert-Roth A, et al. Effect of interleukin-6 receptor inhibition with tocilizumab in patients with rheumatoid arthritis (OPTION study): a double-blind, placebo-controlled, randomised trial. Lancet 2008;371:987-97.

40. Emery P, Deodhar A, Rigby WF, et al. Efficacy and safety of different doses and retreatment of rituximab: a randomised, placebo-controlled trial in patients who are biological naive with active rheumatoid arthritis and an inadequate response to methotrexate (Study Evaluating Rituximab's Efficacy in MTX iNadequate rEsponders (SERENE)). Ann Rheum Dis 2010;69:1629-35.

41. Cohen SB, Moreland LW, Cush JJ, et al. A multicentre, double blind, randomised, placebo controlled trial of anakinra (Kineret), a recombinant interleukin 1 receptor antagonist, in patients with rheumatoid arthritis treated with background methotrexate. Ann Rheum Dis 2004;63:1062-8.

42. Kremer JM, Westhovens R, Leon M, et al. Treatment of rheumatoid arthritis by selective inhibition of T-cell activation with fusion protein CTLA4Ig. N Engl J Med 2003;349:1907-15.

43. Kremer JM, Genant HK, Moreland LW, et al. Effects of abatacept in patients with methotrexate-resistant active rheumatoid arthritis: a randomized trial. Ann Intern Med 2006;144:865-76.

44. Kremer JM, Blanco R, Brzosko M, et al. Tocilizumab inhibits structural joint damage in rheumatoid arthritis patients with inadequate responses to methotrexate: results from the double-blind treatment phase of a randomized placebo-controlled trial of tocilizumab safety and prevention of structural joint damage at one year. Arthritis Rheum 2011;63:609-21.

45. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons, Ltd.; 2011.

46. Agency for Healthcare Quality and Research. HCUPnet. National and regional 2013 estimates on hospital use for all patients from the HCUP Nationwide Inpatient Sample (NIS). National statistics - principal procedure only. ICD-9-CM 714.0. (Accessed July 16, 2016, at http://hcupnet.ahrq.gov/HCUPnet.jsp.)

47. Schappert SM, Rechtsteiner EA. Ambulatory medical care utilization estimates for 2007. Vital Health Stat 13 2011:1-38.

48. Kawatkar AA, Jacobsen SJ, Levy GD, Medhekar SS, Venkatasubramaniam KV, Herrinton LJ. Direct medical expenditure associated with rheumatoid arthritis in a nationally representative sample from the medical expenditure panel survey. Arthritis Care Res (Hoboken) 2012;64:1649-56.

49. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.