Dear IVI,

Thank you for the opportunity to provide input into the IVI-RA cost-effectiveness model and user interface.

Model Structure

1.) The model is very similar to other models that have been developed previously (ie, based on HAQ), however, the user interface allowing flexibility in model structure and inputs is novel. The number of possible model structures in the absence of guidance up-front as to why one would choose any given structure is likely to be a deterrent for an average user. Would like to see a new innovative approach within this model structure to address limitations of prior research. For example, a key criticism for the ICER approach was lack of a patient-centric approach. Would like to see how the patient research conducted by IVI can be incorporated. On page 33 of the model description document under Limitations there is mention of patient preference being captured by HAQ. We believe that patient preference should be defined by attributes of therapies (eg, mode of administration, efficacy/safety, number of patients treated, etc) as opposed to HAQ.
2.) The ability to choose one of three possible model structures allows the user a good deal of freedom, however, it begs the question of which model structure is the most appropriate. If all models use HAQ to map to utility, it could be argued that the simplest model structure which directly links the treatment to the HAQ would be preferable.
3.) Similarly, the six different switching rules allow a lot of flexibility. However, given the importance of HAQ in determining utility, it is surprising that it wouldn't be considered in a treatment switch decision.
4.) For mapping HAQ to EQ-5D, is there the ability to use EQ-5D utilities directly from RA patients? For example, several RA registries are now routinely collecting EQ-5D (eg, Corrona).
5.) Another important option to consider is the incorporation of radiographic progression which was not allowed as an option in either the model structure or in the treatment switching criterion. We have conducted a network metanalysis on this topic which has been presented at a medical conference and a subsequent manuscript is forthcoming.
6.) For model structures involving sequential biologic treatment, it is unclear whether the short period of HAQ rebound adequately captures the disutility associated with treatment failure.
7.) The model could be compared to the body of evidence using treat to target strategies, and should consider incorporation of T2T strategic studies .

Source Data

Network Meta-Analysis:

1.) Without displaying the RCTs that contributed to the network meta-analysis it is difficult to assess the quality and completeness of the NMA and its results. Additionally, on page 45 there is a statement that the estimates for the new NMA were in the same range as NICE, however, there are 2 treatments where there is a significant difference (tocilizumab and certolizumab).
2.) The change in HAQ estimated from the NMA (Table 3) was -0.443 for ABT SC + MTX, -0.552 for ADA + MTX, and -0.555 for CZP + MTX, yet in Table 8 when evaluating the simulated change in HAQ based on H3 (which links treatment directly to HAQ) is -0.31 for ABT SC + MTX, -0.31 for ADA + MTX and -0.38 for CZP + MTX. It is strange that a 0.109 difference between ABT SC + MTX and ADA + MTX did not lead to a simulated difference in HAQ, but a 0.003 difference between ADA + MTX and CZP + MTX led to a 0.07 difference in HAQ. Additionally, can you post the supporting literature used to map ACR scores to estimate HAQ, and send a copy to the group to review.
3.) Table 3 reports the results for different treatments in combination with MTX, however, in the online model these treatments are also available as monotherapy. It is unclear where the estimates for monotherapy originate from.
4.) Is there a more representative source for mapping ACR response to EULAR response? Although the VARA registry does capture both ACR and EULAR response, the data are from veterans are not generalizable to the general RA population (eg, increased males, smoking).

Economic inputs:

1. Drug pricing is all based on WAC with a generic infusion cost not accounting for differences in infusion time. Consideration should be taken for differences in infusion time.

2. Also, all drug costs appear to be based on initial dosing and do not incorporate increased cost for products that allow increased dosing (eg, infliximab, tocilizumab). For example, the average dose for infliximab that is used in regular practice is much higher than the 3mg/kg and the dosing interval is frequently decreased.
3. There are subtle differences in monitoring frequency associated with these therapies and would like to see these differences noted.

Online Model:

1. The one-thousand patient cap on the number of patients to simulate cuts down on computing time, however it is smaller than typical microsimulations which often simulate between ten-thousand and one hundred-thousand patients. This may lead to an increase in variance for the estimates.
2. In RA especially, there is an abundance of possible treatment options and sequences. It would be very helpful in this situation to be able to simultaneously compare more than five arms allowed in the treatment sequences tab.
3. Many US payors would require flexibility to evaluate treatments over a shorter time horizon.
4. We had some difficulties with the run time being excessively slow and the user would give up.

TO: darius.lakdawalla@thevalueinitiative.org; Jason.shafrin@thevalueinitiative.org
Subject: Beta version of IVI-RA Model – Strategic Advisory Panel Comments

Dear Darius and Jason:

As a leader in immunology research, Bristol-Myers Squibb (BMS) acknowledges the importance of understanding and fully characterizing the value that innovative therapies provide to patients, and appreciates the opportunity to comment on the Innovation and Value Initiative (IVI) Rheumatoid Arthritis (RA) Model as part of IVI's Open Source Value Project (OSVP). BMS is dedicated to advancing the science of immunology and to sharing and disseminating the results of our research to ensure that our work can benefit the widest range of patients.

We have reviewed the IVI-RA model and believe it could benefit from the following revisions:

Emphasize All Elements of Value Rather than a Focus on Quality-Adjusted Life-Years (QALYs)
The beta version of IVI-RA proposes the primary health outcome is the QALY. Experts have identified other value elements and BMS believes rather than promulgating the use of this limited measure, value frameworks should aim to incorporate as many elements of value as possible. For example, ISPOR's Special Task Force (STF) on Value Assessment Frameworks has identified twelve different components of value in their draft whitepaper – including QALY, net costs, productivity, adherence improving factors, reduction in uncertainty, fear of contagion, insurance value, severity of disease, value of hope, real option value, equity, and scientific spillovers. This catalogue of elements provides one of the first, truly forward-looking steps in comprehensively characterizing all of the multidimensional facets of value. We understand that in the coming weeks IVI will aim to incorporate 'insurance value' as an additional element to the IVI-RA model, which is a promising first step as the elements identified by ISPOR's STF are critical to assessing value. To ensure there is not an undue focus on QALYs, we encourage IVI to consider incorporating other value elements as well.

Expand the Role of Safety in the Model and Differentiate Safety Outcomes
In addition to inadequate efficacy, lack of safety in the form of adverse events is also a common reason for RA patients switching medications. The model's primary focus is on patients who switch or discontinue their medication due to efficacy only. This presents a bias toward products with a better efficacy profile and worse safety profile. The model allows switch due to adverse events, but only in the case of serious infection, as these "have a significant cost impact and increased risk over background rates to be meaningful to include." Additional adverse events are important to patients and result in treatment switching. Therefore, the model should expand beyond serious infections and include other adverse events.

The model also assumes that the infection rate is equal among targeted disease-modifying antirheumatic drugs (tDMARDs) because "the published results for specific tDMARDs are estimated with very little precision." Evidence shows that specific tDMARDs have decreased infection risk compared to other tDMARDs. In RA patients with prior exposure to a biologic agent, exposure to etanercept, infliximab, or rituximab was associated with a greater 1-year risk of hospitalized infection compared with the risk associated with exposure to abatacept.[1] Among RA patients who experienced a hospitalized infection while on anti-TNF therapy, abatacept and etanercept were associated with the lowest risk of subsequent infection compared to other biologic therapies.[2] The expected costs of serious infections were lower for IV and SC abatacept than for infliximab and adalimumab in hypothetical analyses based on two large clinical trials.[3] Therefore, the assumption of equal infection rates negatively impacts the performance of these tDMARDs in the model. The model should differentiate rates of serious infection among tDMARDs.

TO: darius.lakdawalla@thevalueinitiative.org; Jason.shafrin@thevalueinitiative.org
Subject: Beta version of IVI-RA Model – Strategic Advisory Panel Comments

Patient Centricity

With regards to patient centricity, we recommend IVI better report their efforts to make the model patient centric. For example, the patient focus groups conducted in preparation are a key resource in developing the models, but stakeholders have no way of knowing how IVI used these patient insights to inform methodology. Wherever patient insights were used to inform the approach and methodology, we recommend noting that in IVI-RA's explanation of the methodology. We are also encouraged from the briefing sessions offered that IVI aims to better capture variation in patient circumstances and preferences in future IVI-RA iterations and suggest IVI continue to make this a priority as it revises its approach.

Allow for Analysis Based on Poor Prognostic Factors

Although this model allows more patient heterogeneity than previous cost-effectiveness models, there is still no way to analyze subgroups, such as those with poor prognostic factors and more severe disease. Even if this data is not available across every clinical trial in the network meta-analysis (NMA), the final tool should allow the user to conduct a scenario analysis for different subgroups.

Select the Most Appropriate Outcomes

The model relies on each agent's impact on Heath Assessment Questionnaire (HAQ) Disability Index score during the initial treatment phase and incorporates parameters such as progression defined by HAQ over time, mortality rates as a function of HAQ-based progression, quality of life based on HAQ, and cost based on HAQ-progression. Focusing on this efficacy measure during the initial treatment phase, and no other outcomes, is a limitation. The model should seek out ways to incorporate other measures such as Disease Activity Score (DAS) and American College of Rheumatology (ACR) response (ACR20, ACR50, ACR70, ACR90) to define model outcomes, even if evidence is not as robust as that for HAQ.

Additional Considerations

- Willingness to pay will not be well received by policymakers, patient groups, and patients. This terminology will be concerning to many relevant stakeholders.
- The limitations of using wholesale acquisition cost (WAC) and how that impacts the results should be stated to the model user, in case they do not have access to the appropriate discount and rebate data.
- The model should define what is included in the drug monitoring cost input.
- The utility score in the model depends on serious infections in addition to HAQ. The disutility of an infection in the model can vary by 20% in either direction. The methodology should explain why the disutility data on serious infections is not robust.

Bristol-Myers Squibb appreciates IVI's efforts to engage stakeholders in the development of the IVI-RA model and we look forward to providing continued input to IVI as it continues work on the OSVP. We hope to continue this dialogue and welcome the opportunity to meet to further discuss the concerns identified in our comments. If you have any questions, please do not hesitate to contact:

- Michael Ryan, PharmD, Head of U.S. Value, Access & Payment (michael.ryan1@bms.com)
- Mitch Higashi, PhD, Head of U.S. Health Economics Outcomes Research (mitch.higashi@bms.com).

References:

1. Yun H, Xie F, Delzell E, Levitan EB, Chen L, Lewis JD, Saag KG, Beukelman T, Winthrop KL, Baddley JW, Curtis JR. Comparative Risk of Hospitalized Infection Associated With Biologic Agents in Rheumatoid Arthritis Patients Enrolled in Medicare. Arthritis Rheumatol 2016, 68(1):56-66.
2. Yun H, Xie F, Delzell E, Chen L, Levitan EB, Lewis JD, Saag KG, Beukelman T, Winthrop K, Baddley JW, Curtis JR. Risk of hospitalized infection in rheumatoid arthritis patients receiving biologics following a previous infection while on treatment with anti-TNF therapy. Ann Rheum Dis 2015, 74(6):1065-71.
3. S Johnston, S Kelly, A Nadkarni, KL Wilson, B Limone, M Hochberg. Healthcare Costs Associated with Serious Infections Among Biologic-Naïve Rheumatoid Arthritis Patients Initiating First-Line Biologic Treatment. ACR/ARHP Annual Scientific Meeting, November 14-19, 2014, Boston, MA.

A Description of the IVI-RA Model v0.1

Devin Incerti* Jeroen P. Jansen*

July 18, 2017

Contents

*Innovation and Value Initiative

List of Figures

List of Tables

1 Overview

This document describes version 0.1 of the Innovation and Value Initiative's (IVIs) rheumatoid arthritis (RA) cost-effectiveness model. The IVI-RA model is an individual patient simulation (IPS) that simulates patients one at a time. The model can be run with multiple perspectives (e.g., health care sector, societal) and accounts for both parameter and structural uncertainty. Since the range of defendable scientific approaches is large, the IVI-RA model consists of 336 possible model structures. Structural uncertainty can be quantified by estimating cost-effectiveness across these different model structures and parameter uncertainty is quantified using probabilistic sensitivity analysis (PSA).

The model will be made available as an R package and the source code will be available at our GitHub repository. The IPS was primarily written in C++ so that PSA and analyses of structural uncertainty can be run in a reasonable amount of time. The model can either be run using R or online with our user-friendly R Shiny web application.

This document is structured as follows. We begin by discussing treatment strategies that can be modeled in Section 2. Section 3 outlines the competing model structures. Section 4 examines the data needed to define a population and run an analysis. Section 5 describes the statistical techniques used to estimate the model parameters and the data sources used. Section 6 details the simulation techniques used to implement the IVI-RA model and quantify uncertainty. Section 7 focuses on methods for validating the model. Lastly, Section 8 discusses limitations and areas for improvement.

2 Treatment strategies

The primary purpose of the model is to evaluate the cost-effectiveness of treatments for rheumatoid arthritis (RA). Since patients typically use multiple treatments over a lifetime, the model is capable of simulating a treatment sequence of any arbitrary length. Treatments that can be included in a sequence include conventional disease-modifying anti-rheumatic drugs (cDMARDs) such as methotrexate as well as the following biologic DMARDs (bDMARDs):

- **Tumor necrosis factor (TNF) inhibitors**: etanercept, adalimumab, infliximab, certolizumab, golimumab
- **non-TNF inhibitors**: abatecept, tocilizumab, rituximab
- **Janus kinase/signal transducers and activators of transcription (JAK/STAT) inhibitors**: tofacitinib

At the end of a sequence, patient switch to non-biologic therapy (NBT), which encompasses a range of therapies that clinicians may feel is appropriate for all patients such as methotrexate and sulfasalazine (Stevenson et al. 2016, 2017).

3 Competing model structures

The IVI-RA model is a discrete-time IPS with 6 month cycles that can be run using a number of different model structures. Like most RA cost-effectiveness models, the model measures changes in disease severity using the Health Assessment Questionnaire (HAQ) Disability Index score (Tosh et al. 2011; Carlson et al. 2015; Stephens et al. 2015; Stevenson et al. 2016; Institute for Clinical and Economic Review 2017; Stevenson et al. 2017). In particular, at the start of the simulation, each patient is assigned a baseline HAQ score. Subsequently, the impact of the disease measured by the HAQ trajectory over time is modeled as a function of a sequence of treatments (Figure 1). In the absence of treatment, HAQ deteriorates at a certain rate as depicted by the dashed line in the figure. Treatment is separated into two distinct phases: an initial phase of up to 6 months, consistent with data reported from randomized controlled trials (RCTs), and a maintenance phase thereafter until discontinuation.

During the initial treatment phase HAQ is modeled as a change from baseline. Three possible model structures labeled **H1-H3** are possible.

- **H1**: Treatment → ACR → HAQ
- **H2**: Treatment → ACR → EULAR → HAQ
- **H3**: Treatment → HAQ

In **H1**, treatment influences HAQ through its effect on the American College of Rheumatology (ACR) response criteria, which is similar to the structure used in other US based cost-effectiveness models (e.g. Carlson et al. 2015; Institute for Clinical and Economic Review 2017). ACR 20/50/70 response is defined as at least a 20/50/70% improvement. In the simulation, we convert these overlapping ACR categories to four mutually exclusive categories: no response (defined as less than 20% improvement), ACR 20% to <50% improvement, ACR 50% to <70% improvement, and ACR

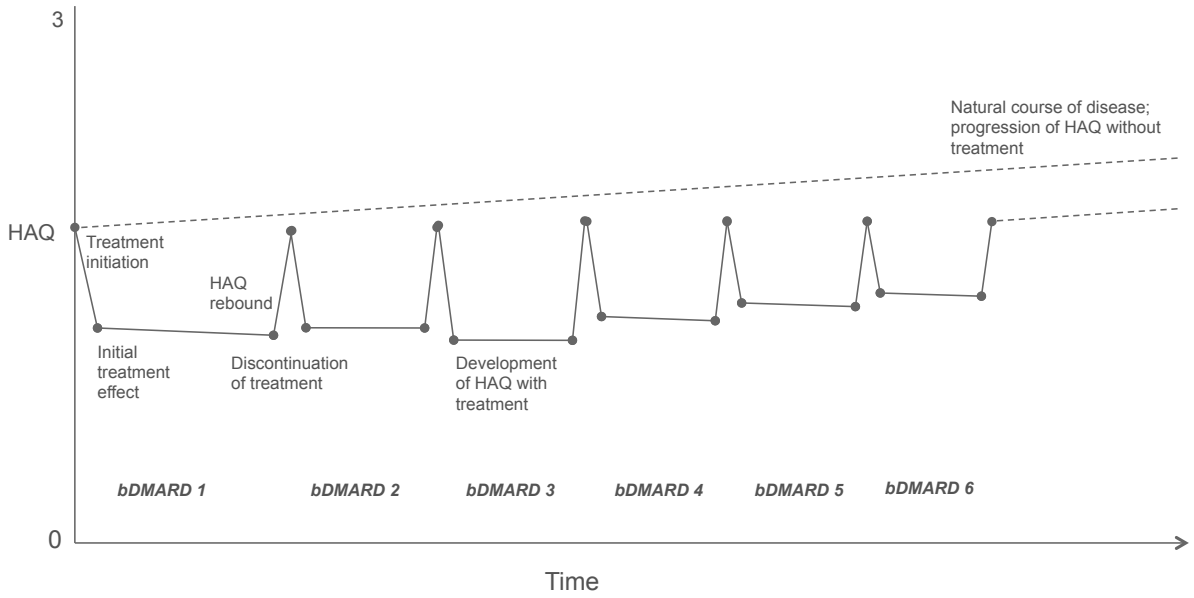


Figure 1: Model structure regarding development of HAQ with sequential biologic treatment

70% improvement or greater. The rationale for using ACR response rather than HAQ directly is that the evidence base relating treatment to ACR response is larger than the evidence based relating treatment to HAQ. **H2** follows the National Institute for Health and Care Excellence (NICE) cost-effectiveness model (Stevenson et al. 2016, 2017) and models the effect of treatment on HAQ indirectly through its effect on ACR response and, in turn, the three categories of the European League Against Rheumatism (EULAR) response (no response, moderate response, or good response). Finally, since modeling the effect of treatment on HAQ through intermediary variables may mediate treatment response, in **H3**, treatment impacts HAQ directly.

Treatment switching during the initial treatment phase is modeled using 6 possible structures labeled **S1-S6**.

- **S1:** Treatment → ACR → Switch
- **S2:** Treatment → ACR → ΔDAS28 → DAS28 → Switch
- **S3:** Treatment → ACR → ΔSDAI → SDAI → Switch
- **S4:** Treatment → ACR → ΔCDAI → CDAI → Switch
- **S5:** Treatment → ΔDAS28 → DAS28 → Switch
- **S6:** Treatment → ACR → EULAR → Switch

S1 follows a common approach where ACR non-responders discontinue treatment (e.g. Carlson et al. 2015; Institute for Clinical and Economic Review 2017). One drawback of this approach is that it is not consistent with current treat-to-target guidelines in the United States (Singh et al. 2016). In **S2-S5**, treatment switching consequently depends on disease activity (remission, low, moderate, high) (Anderson et al. 2012). In **S2-S4**, ACR response predicts the change in disease

activity from baseline, which along with baseline disease activity, predicts absolute disease activity. Patients with moderate or high disease switch treatment while patients with low disease activity or in remission continue treatment. Disease activity is measured using either the Disease Activity Score with 28-joint counts (DAS28) (Prevoo et al. 1995), Simplified Disease Activity Index (SDAI) (Smolen et al. 2003; Aletaha and Smolen 2005), or the Clinical Disease Activity Index (CDAI) (Aletaha et al. 2005).

S5 is similar to **S2-S4**, but models the effect of treatment on changes in DAS28 directly, rather than indirectly through ACR response. We also aimed to model the direct effect of treatment on SDAI and CDAI, but sufficient clinical trial data are not available. Finally, since in the UK, the British Society for Rheumatology and the British Health Professionals in Rheumatology recommends using the EULAR response (Deighton et al. 2010), treatment switching in **S6** depends on EULAR response. In particular, following the NICE model, we assume that EULAR non-responders discontinue treatment while moderate and good responders continue treatment (Stevenson et al. 2016). The reasoning is that rules stipulated by NICE require a DAS28 improvement of more than 1.2 to continue treatment which is associated with moderate or good EULAR response.

Not all model structures **S1-S6** can be used with each of **H1-H3**. If **H1** is used, then **S1-S5** are available, but **S6** is not because EULAR response is not simulated. In **H2**, **S1-S6** are all available while in **H3** only **S5** can be used since ACR response is not simulated. The 12 possible model structures and the number of each structure are outlined in Table 1.

Table 1: Model structures for initial treatment phase

	S1	S2	S3	S4	S5	S6
H1	1	2	3	4	5	-
H2	6	7	8	9	10	11
H3	-	-	-	-	12	-

Notes: Rows denote the model structure used to relate treatment to HAQ and columns denote the model structure used to determine treatment switching. Each number denotes a unique model structure (i.e., 1 corresponds to H1 and S1, and 8 corresponds to H2 and S3) and the "-" denotes a model structure combination that is not possible. There are 12 possible model structures for the initial treatment phase.

In the maintenance phase, two model structures can be used to simulate the long-term progression of HAQ. First, as is common in cost-effectiveness analyses (CEAs) of therapies for RA, HAQ is assumed to progress at a constant linear rate over time (see Tosh et al. 2011; Wailoo et al. 2008). However, since emerging evidence suggests that the rate of HAQ progression is non-linear and varies across patients (Gibson et al. 2015), our second scenario simulates HAQ progression using a latent class growth model (LCGM) (Norton et al. 2014) with 4 distinct HAQ trajectories and a rate of HAQ progression that decreases over time within each trajectory. Upon discontinuation of treatment, the HAQ score rebounds by a proportion of the improvement experienced at the end of the initial 6-month period with that treatment.

The duration of the maintenance phase (i.e., time to discontinuation of maintenance treatment) is simulated using parametric time-to-event distributions. When **S1** is used, time to treatment discontinuation is simulated using a single time-to-event curve because we have been unable to obtain curves stratified by ACR response categories. In contrast, when **S2-S5** are selected, the time-to-event curves are stratified by disease activity level so patients with lower disease activity at the end of the initial treatment phase stay on treatment longer, on average. Likewise, when structure **S6** is used, the time-to-event distributions are stratified by EULAR response category and patients

with good response at the end of the initial treatment phase tend to stay on treatment longer than patients with a moderate response. In each case, time to discontinuation can be simulated using one of 7 possible distributions (exponential, Weibull, Gompertz, normal, gamma, log-logistic, generalized gamma).

In line with Stevenson et al. (2016) the adverse events included in the model are limited to serious infections; we assume that only serious infections have a significant cost impact and increased risk over background rates to be meaningful to include (Ramiro et al. 2017). During the initial treatment phase, a patient immediately stops treatment if a serious infection occurs; during the maintenance phase, time on treatment depends on the sampled time to treatment discontinuation and a patient experiences a serious infection if the individual's sampled time to the adverse event is shorter than the sampled time to treatment discontinuation.

Baseline HAQ scores (and changes in HAQ scores from baseline) are used to determine mortality relative to age/sex specific rates for the US general population (assumed to have a HAQ score of 0). Treatment, therefore, has an indirect effect on mortality through its effect on HAQ.

Individual HAQ scores at a particular point in time were also used to simulate EuroQol five dimensions questionnaire (EQ-5D) utility scores (0-1 range), which, in turn, were used to simulate quality-adjusted life-years (QALYs). However, since a number of different methods have been used to convert HAQ into utility, our model contains two different possible mapping algorithms. Our preferred algorithm is the Alava et al. (2013) mixture model, which uses a much larger sample size than other statistical models and has been shown to have better predictive accuracy. Other algorithms are typically estimated using clinical trial data (e.g. Carlson et al. 2015; Stephens et al. 2015) and consequently have limited generalizability. The second utility algorithm available within our model is based on a linear regression analysis of real-world data by Wailoo et al. (2006) that has been used in a few previous CEAs (e.g. Wailoo et al. 2008; Institute for Clinical and Economic Review 2017).

Annual hospitalization days and productivity losses are simulated as a function of HAQ. Health sector costs considered in the models are related to drug acquisition and administration, adverse events, general management of RA, and hospitalization. Non-health sector costs are limited to work-related productivity loss.

The flow diagram in Figure 2 describes the flow of a single patient through the simulation. The simulation runs for a patient's entire lifespan beginning with treatment initiation and ending in death. The rectangles in the figure represent "processes" determining the effect of treatment on disease progression and the diamonds represent "decisions" that determine whether a patient will switch to a new treatment.

The influence diagram in Figure 3 summarizes the assumed relationships among different variables in the model. Each arrow represents the direct effect of one parameter on another. Dashed lines represent relationships that depend on the structural assumptions used. Figure 3a focuses on the effect of treatment on disease progression and adverse events while Figure 3b looks at the variables influencing the primary health and cost outcomes.

Model outcomes depend on patient characteristics, which have a direct effect on HAQ progression, mortality, and utility. The primary health outcome is the QALY, which depends on mortality and utility. Total costs consist of health care sector costs and productivity losses. The components of health sector costs include drug acquisition and administration costs, general management and monitoring costs, adverse event costs, and hospitalization costs. Analyses from a societal perspective would include productivity losses while analyses from a health care sector perspective would not.

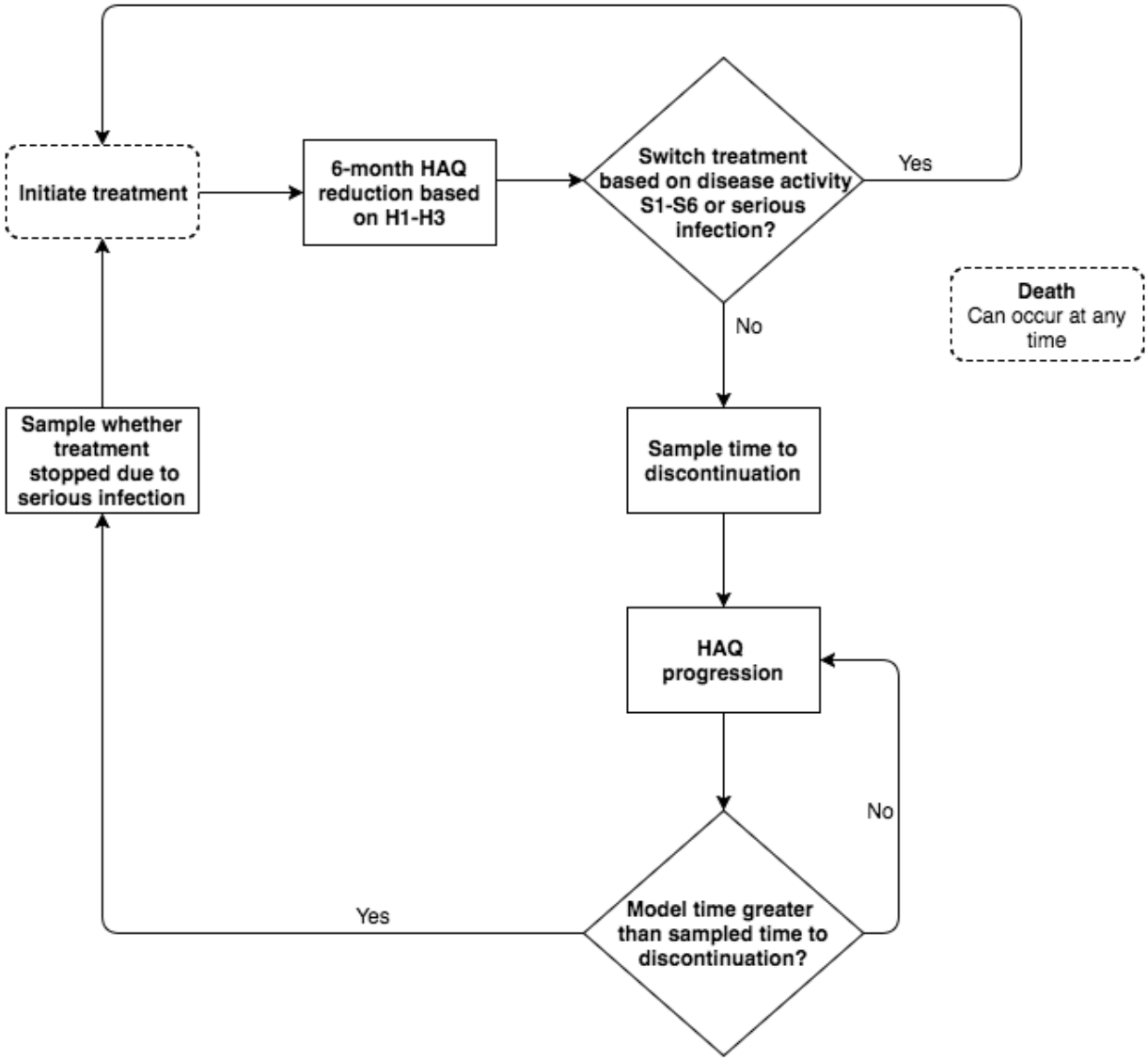
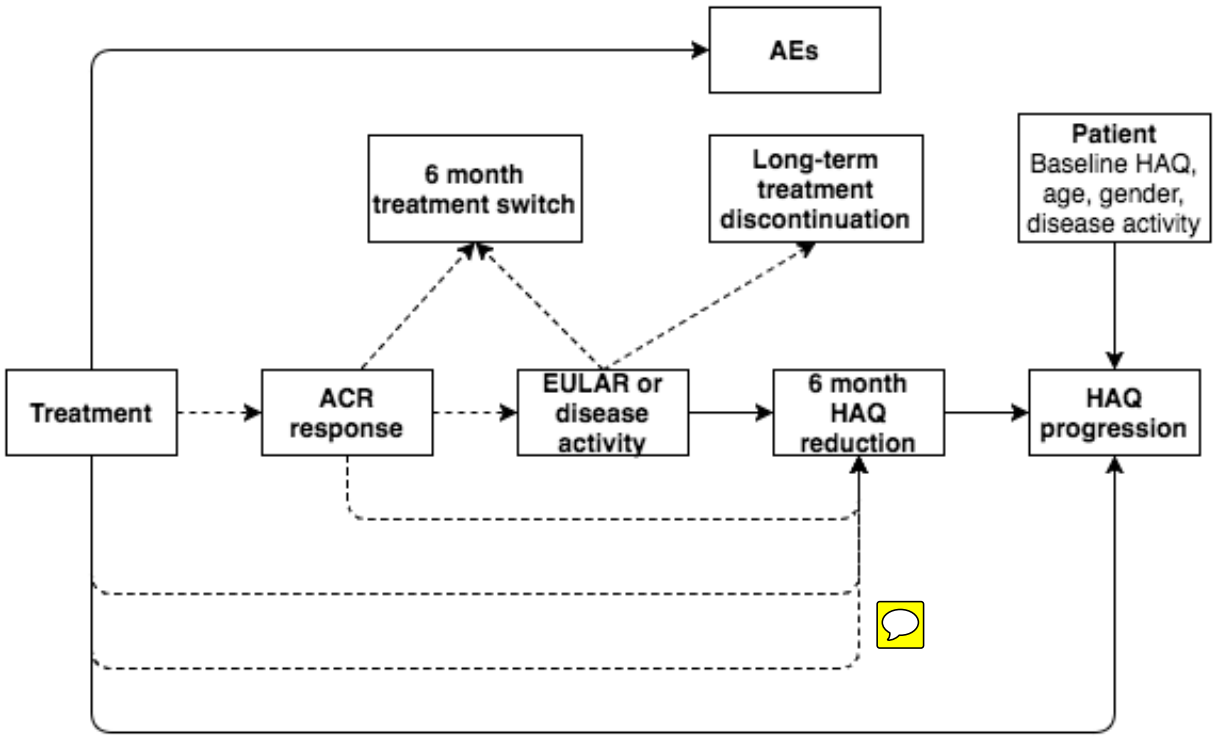


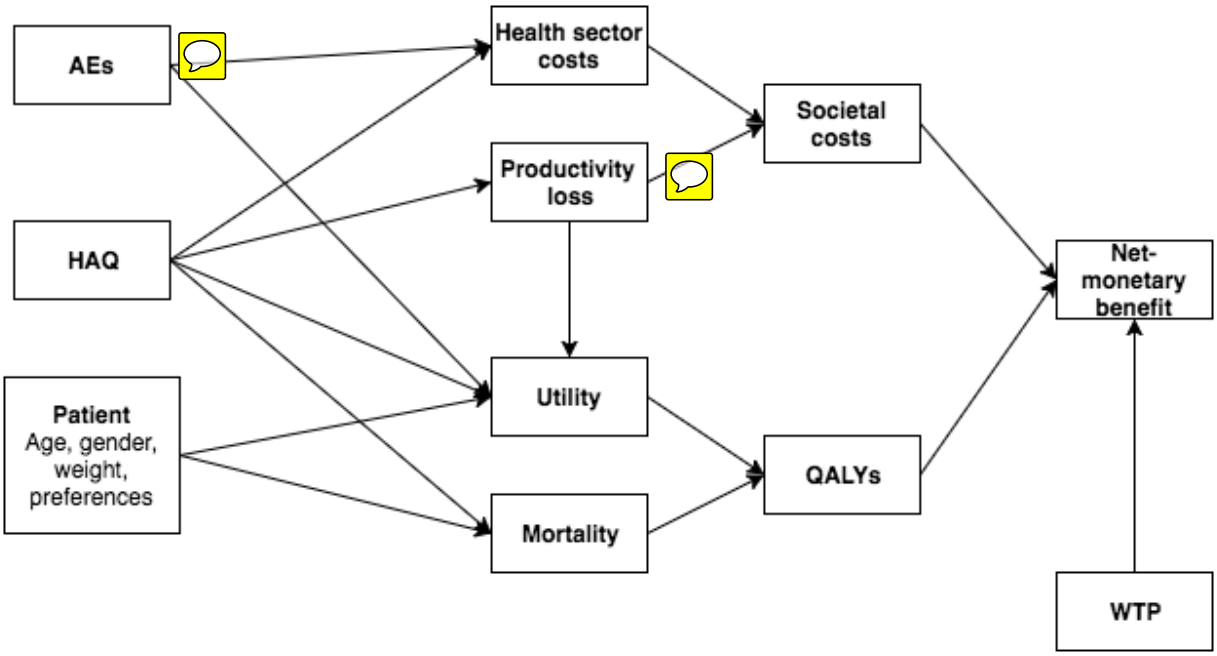
Figure 2: Flow diagram of the simulation for a single patient

Notes: Rectangles represent "processes" determining the effect of treatment on disease progression, Diamonds represent "decisions" that determine whether a patient will switch to a new treatment. Dotted lines denote start of a new treatment or the end of the simulation.

The value of treatment is estimated using the net-monetary benefit (NMB), which is calculated by multiplying QALYs by a willingness to pay threshold and subtracting costs ($NMB = QALYs \cdot WTP - Costs$).



(a) Treatment effects



(b) Model outcomes

Figure 3: Influence diagram outlining structural relationships

Notes: ACR: American College of Rheumatology; EULAR: European League Against Rheumatism; HAQ: Health Assessment Questionnaire: AEs: adverse events; QALYs: quality-adjusted life-years; WTP: willingness to pay. Disease activity refers to the Disease Activity Score with 28-joint counts (DAS28), the Simplified Disease Activity Index (SDAI), or the Clinical Disease Activity Index (CDAI).

4 Populations

To run the IPS, a patient population must be specified. The model is designed for patients who are cDMARD experienced. The patient characteristics that must be included in the analysis are age, HAQ, gender, weight, the number of previous DMARDs, and disease activity. These variables are measured at the start of the simulation (i.e., model cycle 0).

Two default options for the patient population are available. First, a homogeneous cohort of men and women with gender-specific weights but otherwise identical characteristics can be used. Second, a heterogeneous cohort of patients with gender-specific weights but varying across all other characteristics can be specified. Other populations (i.e., for certain subgroups or based on registry data) can be used as well but are not prespecified in our R package.

Our default population consists of individuals that, on average, have high disease activity. The proportion that is female, age, the number of previous DMARDs, baseline HAQ, and DAS28 are based on the values reported in Curtis et al. (2010). Mean values for the SDAI and CDAI are from the US301 clinical trial—which had a DAS28 score similar to the value from Curtis et al. (2010)—summarized in Smolen et al. (2003). Summaries of each variable are reported in Table 2. Details on the algorithm for simulating heterogeneous patients are described in Appendix B.

Table 2: Default patient population

	Mean	Standard deviation	Minimum	Maximum
Age	55.00	13.00	18	85
Male	0.21	-	-	-
Female weight (kg)	75.00	-	-	-
Male weight (kg)	89.00	-	-	-
Previous DMARDs	3.28	1.72	0	-
DAS28	6.00	1.20	0	9.4
SDAI	43.00	13.00	0	86
CDAI	41.00	13.00	0	76
HAQ	1.50	0.70	0	3

5 Source data and parameter estimation

5.1 Comparative treatment efficacy from NMA

The effect of treatment on ACR response, DAS28, and HAQ at 6 months for bDMARD naive patients are estimated using Bayesian network meta-analyses (NMA) of published randomized controlled trials (RCTs). Primary outcomes were ACR response, change in DAS28 from baseline at 6 months, and the change in HAQ from baseline at 6 months. Results from the NMA are shown in Table 3. Details of the systematic literature review and the statistical methodology are provided in the Appendix (Section H.4).

Treatment effects for bDMARD experienced patients were reduced by applying the relative risk ratio reported in Carlson et al. (2015). More specifically, reductions in DAS28 and HAQ scores for bDMARD experienced patients are assumed to be, on average, 84% of the reduction in DAS28 and HAQ scores for bDMARD naive patients. Likewise, we assume that the mean reduction in

Table 3: NMA estimates of ACR response, change in DAS28, and change in HAQ for biologic naive patients

	ACR response				
	ACR20	ACR50	ACR70	ΔDAS28	ΔHAQ
cDMARDs	0.265 (0.248, 0.283)	0.102 (0.093, 0.113)	0.032 (0.028, 0.036)	-1.146 (-1.220, -1.078)	-0.231 (-0.253, -0.209)
ABT IV + MTX	0.559 (0.479, 0.636)	0.312 (0.241, 0.385)	0.142 (0.100, 0.192)	-2.262 (-2.656, -1.837)	-0.449 (-0.574, -0.321)
ADA + MTX	0.565 (0.485, 0.637)	0.318 (0.249, 0.385)	0.146 (0.104, 0.191)	-	-0.552 (-0.666, -0.442)
ETN + MTX	0.648 (0.527, 0.751)	0.400 (0.282, 0.516)	0.204 (0.123, 0.293)	-2.653 (-3.218, -2.052)	-0.530 (-0.652, -0.406)
GOL + MTX	0.599 (0.451, 0.733)	0.352 (0.224, 0.492)	0.171 (0.089, 0.273)	-2.609 (-3.223, -1.967)	-0.505 (-0.629, -0.386)
IFX + MTX	0.662 (0.417, 0.862)	0.425 (0.195, 0.672)	0.230 (0.074, 0.448)	-1.942 (-2.499, -1.411)	-0.437 (-0.620, -0.250)
TCZ + MTX	0.562 (0.373, 0.741)	0.320 (0.168, 0.503)	0.151 (0.062, 0.283)	-3.003 (-3.289, -2.713)	-0.500 (-0.598, -0.401)
CZP + MTX	0.740 (0.538, 0.887)	0.510 (0.292, 0.712)	0.296 (0.130, 0.491)	-3.159 (-3.585, -2.722)	-0.555 (-0.657, -0.452)
ABT SC + MTX	0.567 (0.434, 0.703)	0.322 (0.210, 0.455)	0.150 (0.082, 0.244)	-2.268 (-2.921, -1.573)	-0.443 (-0.632, -0.245)
RTX + MTX	0.570 (0.415, 0.716)	0.325 (0.196, 0.473)	0.152 (0.076, 0.257)	-2.498 (-2.957, -2.038)	-0.517 (-0.623, -0.399)
TOF + MTX	0.608 (0.447, 0.767)	0.362 (0.219, 0.535)	0.179 (0.087, 0.313)	-	-0.469 (-0.703, -0.234)

Notes: ACR20/50/70 categories are the probability of at least a 20/50/70% improvement. 95% credible intervals are in parentheses. Estimates are based on 1,000 random draws of the NMA parameters. ΔDAS28 and ΔHAQ are changes in the DAS28 and HAQ score from their baseline scores respectively; negative numbers denote reductions in baseline values. cDMARDs = conventional disease-modifying antirheumatic drugs; MTX = methotrexate; ABT IV = abatacept intravenous; ADA = adalimumab; ETN = etanercept; GOL = golimumab; IFX = infliximab; TCZ = tocilizumab; CZP = certolizumab pegol; ABT SC = abatacept subcutaneous; RTX = rituximab; TOF = tofacitinib. ACR = American College of Rheumatology.

overlapping ACR response categories (ACR 20/50/70) is 84% after failing the first bDMARD, implying that an ACR response of 60/40/20 would, on average, be reduced to 50/33/16. To capture uncertainty, the reduction in treatment effects is assumed to be distributed uniformly with lower and upper bounds of 74% and 94% respectively.

In the simulation, treatment efficacy depends on the line of therapy and whether a patient is bDMARD naive or bDMARD experienced at baseline. For bDMARD naive patients, first line treatment response is based on the NMA results for bDMARD naive patients while efficacy for all other treatments in a treatment sequence are reduced using the relative risk ratio. For bDMARD experienced patients, treatment response is reduced using the relative risk ratio at each line of therapy including the first line.

5.2 Treatment switching at 6 months

The data required to determine treatment switching at 6 months depends on the selected model structure. If **S1** is selected, then treatment switching depends on the simulated ACR response; likewise, if **S5** is selected, then treatment switching depends on the simulated level of DAS28 at 6 months. When **S2-S4** are used, treatment switching is determined by the relationship between ACR response and the change in disease activity, and in **S6**, switching is based on the relationship between ACR response and EULAR response. Details of the mapping between ACR response and change in disease activity and between ACR response and EULAR response are provided below.

5.2.1 ACR response and change in disease activity

There are currently no established mappings between mutually exclusive ACR response categories and DAS28, SDAI, or CDAI (Madan et al. 2015). However, Aletaha and Smolen (2005) provides evidence on the relationship between overlapping ACR response categories (ACR 20/50/70) and mean changes in each of the three disease activity measures. Results are reported for three cohorts—the Leflunomide datasets, the inception cohort, and the routine cohort—with 1,839, 91, and 279 patients, respectively. We transformed mean changes by overlapping ACR response categories to mean changes by mutually exclusive ACR response categories by using the number of patients in each mutually exclusive ACR response category as described in Appendix C. Smolen et al. (2003) provided the number of patients in each ACR response category in the Leflunomide dataset and Aletaha et al. (2005) provided the number of patients in the inception cohort. Mean changes in disease activity in each mutually exclusive ACR response category are shown in Table 4.

Table 4: Relationship between ACR response and change in disease activity measures

ACR response	Mean change at 6 months			
	Leflunomide dataset	Inception cohort		
	SDAI	SDAI	CDAI	DAS28
<20	0.000	0.000	0.000	0.000
20 to <50	-30.284	-13.700	-11.300	-1.550
50 to <70	-35.234	-14.882	-12.873	-1.543
≥ 70	-41.000	-30.100	-27.600	-3.310

We did not include estimates from the routine cohort for two reasons. First, we were unable to find information on the number of patients in each ACR response category. Second, patients in the

routine cohort had considerably lower disease activity levels (Aletaha and Smolen 2005; Aletaha et al. 2005) and our default population (see Section 4) consists of patients with high disease activity at baseline. Mean DAS28 in the inception cohort and routine cohort were 5.62 and 4.09, respectively, while the mean DAS 28 ranged from 6.3 to 7 across the clinical trials making up the Leflunomide dataset.

5.2.2 ACR response and change in EULAR response

ACR responses were translated into EULAR response probabilities based on evidence of their relationship reported in Stevenson et al. (2016) and obtained from the US Veterans Affairs Rheumatoid Arthritis (VARA) registry (Table 5).

Table 5: Relationship between ACR response and EULAR response

	EULAR response		
ACR response	None	Moderate	Good
<20	755	136	57
20 to <50	4	27	26
50 to <70	2	2	10
≥70	0	2	2

Notes: The VARA registry is a multicentre, US database of veterans age 19 and older. Each cell represents the number of patients in the database in a given category.

5.3 Change in HAQ at 6 months

As in Institute for Clinical and Economic Review (2017), ACR responses from the NMA were translated into HAQ scores based on evidence from the adalimumab monotherapy for treatment of rheumatoid arthritis (ADACTA) trial reported in Carlson et al. (2015) (Table 6).

Table 6: Relationship between ACR response and change in HAQ at 6 months

	HAQ change	
ACR response	Mean	Standard error
<20	-0.11	0.06765
20 to <50	-0.44	0.05657
50 to <70	-0.76	0.09059
≥70	-1.07	0.07489

Source: Carlson et al. (2015)

The relationship between EULAR response and HAQ is based on analyses conducted by Stevenson et al. (2016) using the BSRBR database. Their analysis is based on predictions from a mixture model with covariates set to sample means. Moderate and good EULAR responses are associated with -0.317 (SE = 0.048) and -0.672 (SE = 0.112) changes in HAQ scores respectively (Table 7).

Table 8 compares the impact of treatment on HAQ using model structures **H1-H3**. There is considerable variation, which suggests that the choice of model structure has important consequences for cost-effectiveness. The effect of treatment on the change in HAQ is largest when modeled directly

Table 7: Relationship between EULAR response and HAQ

EULAR response	Mean	SE
None	0.000	0.000
Moderate	-0.317	0.048
Good	-0.672	0.112

(**H2**) and smallest when modeled through both ACR and EULAR response (**H3**). 95% credible intervals are also considerably larger when the change in HAQ is modeled indirectly (**H1** and **H3**), then when it is modeled directly (**H2**).

Table 8: Simulated change in HAQ at 6 months under different model structures

	H1	**H2**	**H3**
cDMARDs	-0.17 (-1.01, 0.00)	-0.23 (-0.25, -0.21)	-0.19 (-0.80, 0.00)
ABT IV + MTX	-0.39 (-1.14, -0.00)	-0.44 (-0.58, -0.29)	-0.30 (-0.84, 0.00)
ADA + MTX	-0.40 (-1.14, -0.00)	-	-0.31 (-0.84, 0.00)
ETN + MTX	-0.48 (-1.16, -0.00)	-0.52 (-0.65, -0.37)	-0.34 (-0.85, 0.00)
GOL + MTX	-0.43 (-1.15, -0.00)	-0.50 (-0.63, -0.34)	-0.32 (-0.85, 0.00)
IFX + MTX	-0.49 (-1.16, -0.00)	-0.43 (-0.64, -0.23)	-0.35 (-0.85, 0.00)
TCZ + MTX	-0.39 (-1.15, -0.00)	-0.49 (-0.59, -0.38)	-0.30 (-0.84, 0.00)
CZP + MTX	-0.58 (-1.17, -0.00)	-0.55 (-0.66, -0.43)	-0.38 (-0.86, 0.00)
ABT SC + MTX	-0.40 (-1.15, -0.00)	-0.43 (-0.64, -0.18)	-0.31 (-0.84, 0.00)
RTX + MTX	-0.40 (-1.15, -0.00)	-0.51 (-0.63, -0.37)	-0.31 (-0.84, 0.00)
TOF + MTX	-0.44 (-1.15, -0.00)	-	-0.33 (-0.85, 0.00)

Notes: **H1**, **H2**, and **H3** are the Treatment → ACR → HAQ, Treatment → HAQ, and Treatment → ACR → EULAR → HAQ pathways respectively. 95% credible intervals are in parentheses. Estimates are based on 6-month simulations of 1,000 patients and 1,000 parameters sets for each therapy. ΔHAQ denotes a change in the HAQ score at 6 months from baseline; a negative value indicates a reduction in the HAQ score. cDMARDs = conventional disease-modifying antirheumatic drugs; MTX = methotrexate; ABT IV = abatacept intravenous; ADA = adalimumab; ETN = etanercept; GOL = golimumab; IFX = infliximab; TCZ = tocilizumab; CZP = certolizumab pegol; ABT SC = abatacept subcutaneous; RTX = rituximab; TOF = tofacitinib. ACR = American College of Rheumatology.

5.4 HAQ progression in the absence of bDMARD treatment

The natural course of HAQ progression in the absence of bDMARDs develops over time according to an estimated natural course for patients remaining on cDMARDs or following discontinuation of the last bDMARD of the sequence (i.e., on NBT). The natural course of HAQ can either be assumed to change at a constant linear rate or be modeled using a LCGM that accounts for non-linear progression and heterogeneity across patients.

5.4.1 Constant linear rate of progression

The rate of progression in the linear case is based on the observational study by Wolfe and Michaud (2010). They assessed the development of HAQ over time at six month intervals for up to 11 years among 3,829 RA patients who switched from non-biologic treatment to biologic treatment and participated in the National Data Bank for Rheumatic Diseases (NDB) longitudinal study of RA outcomes. The annual HAQ progression rate prior to biologic therapy was 0.031 (95% confidence

interval (95%CI): 0.026 to 0.036) and is assumed to reflect the course of progression of HAQ in the absence of bDMARDs.

Based on the same data, Michaud et al. (2011) reported overall and age-specific specific HAQ progression rates. The differences between the overall and age specific rates are as follows: <40: -0.020 (95%CI: -0.0223 to -0.0177); 40-64: -0.008 (95%CI: -0.0101 to -0.0059); ≥ 65 0.017 (95%CI: 0.0136 to 0.0204). These estimates are applied to the overall progression rate of 0.031 to obtain age specific HAQ progression rates (see Section D.1).

Table 9: Annual linear progression of HAQ in the absence of bDMARDs beyond 6 months

		95% CI		
	Estimate	Lower	Upper	Reference
Overall progression rate				
MTX or non-biologic treatment	0.031	0.026	0.036	Wolfe and Michaud (2010)
Change in overall progression rate by age				
<40	-0.020	-0.028	-0.012	Michaud et al. (2011)
40-64	-0.008	-0.010	-0.006	Michaud et al. (2011)
65+	0.017	0.013	0.021	Michaud et al. (2011)

Notes: 95% confidence intervals are calculated using a normal distribution. Confidence intervals for changes in HAQ progression rates by age assume no covariance between the overall progression rate and the age-specific rates reported by Michaud et al. (2011).

5.4.2 Latent class growth model

We also model the rate of HAQ progression in the absence of bDMARDs using a mixture model approach that has increasingly been used to model HAQ progression over time (Stevenson et al. 2016; Norton et al. 2013, 2014). These models suggest that different subgroups have distinct HAQ trajectories and that the rate of worsening of HAQ progression decreases over time. We use the LCGM estimated by Norton et al. (2014) and since we aim to model trajectories for cDMARDs and NBTs we chose the specification based on data from the Early Rheumatoid Arthritis Cohort Study (ERAS) cohort, which has a high percentage of patients receiving methotrexate and a very small percentage receiving biologics. Complete details of the LCGM are provided in Section D.2.

The Norton et al. (2014) LCGM determined that there are four classes of patients and thus four distinct HAQ trajectories. The probability of class membership depends on 7 variables: age, gender, DAS28, disease duration, rheumatoid factor, the ACR 1987 criteria for RA, and a measure of socioeconomic status. Age, gender, and the DAS28 are relevant to the way the population is defined within our model (see Section 4) and are therefore important determinants of the HAQ trajectory. Other variables (disease duration, rheumatoid factor, ACR criteria, and socioeconomic status) are not defined within our population. We consequently set disease duration (8.2 months), rheumatoid factor (0.73), and the socioeconomic status variable (0.49) equal to their mean values with the ERAS cohort. The ACR criteria was set to 1.

HAQ trajectories (in levels) by class are shown Figure 4. The dotted lines plot observed mean values. There are clear distinguishable classes as both the level of the HAQ score and its slope vary between groups. Norton et al. (2014) refer to the groups as "low", "moderate", "high", and "severe" groups, in order from the lowest to highest HAQ scores. The observed trends for the low, medium, and high groups follow a J-shaped pattern with a sharp drop following treatment initiation and an upward slope thereafter, while the severe group experiences persistently high HAQ scores. Since

our model separates the initial treatment phase from the maintenance phase, we are only concerned with HAQ progression following the initial drop. As in Stevenson et al. (2016), we consequently only predict values from year 2 onward. The fitted values are the solid upward sloping lines in the plot.

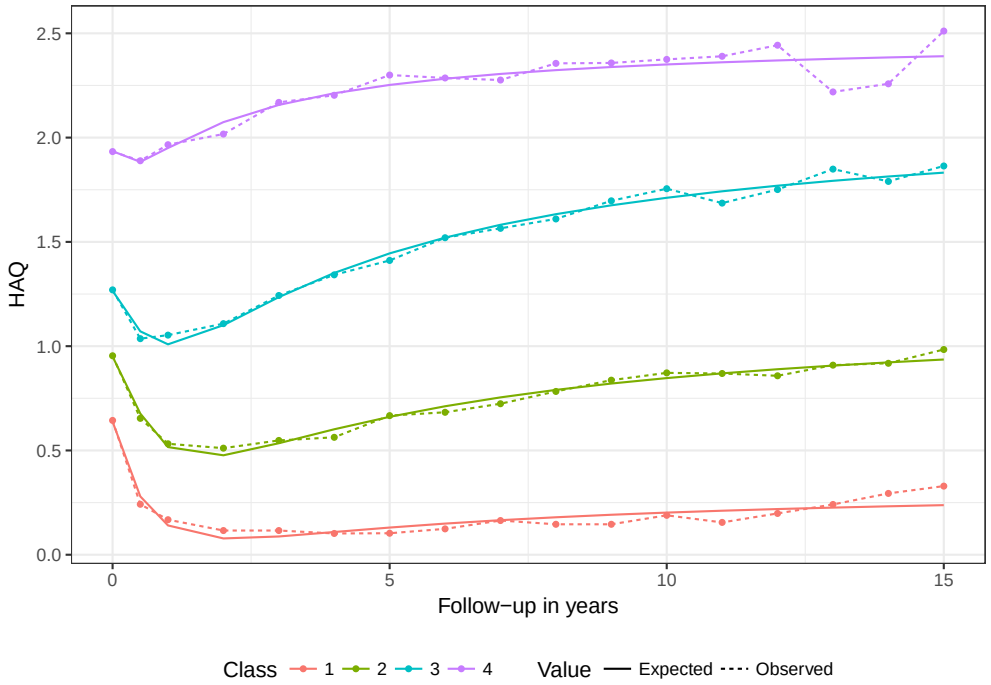


Figure 4: Observed and predicted HAQ trajectories in the ERAS dataset from the latent class growth model

Notes: The first three data points corresponds to years 0, 0.5, and 1, respectively; all other data points are spaced 1 year apart.

An important question for cost-effectiveness modeling in RA is how the rate of progression within each class in the LCGM compares to a constant linear trajectory. We examine this question in Figure 5, which compares yearly rates of changes in HAQ using the LCGM and with constant annual rates of change (0.031 per year) based on the Wolfe and Michaud (2010) analysis. The LCGM was simulated over 30 years and differences between year t and year $t-1$ were used to assess changes in HAQ score from one year to the next.

In the moderate, high, and severe groups the rate of HAQ progression is higher initially in the LCGM than in the Wolfe and Michaud (2010) analysis; however, the LCGM modeled rate of HAQ progression declines over time and eventually begins to approach zero. In the low group, HAQ increases at a rate less than 0.031 per year and the rate of increase declines over time.

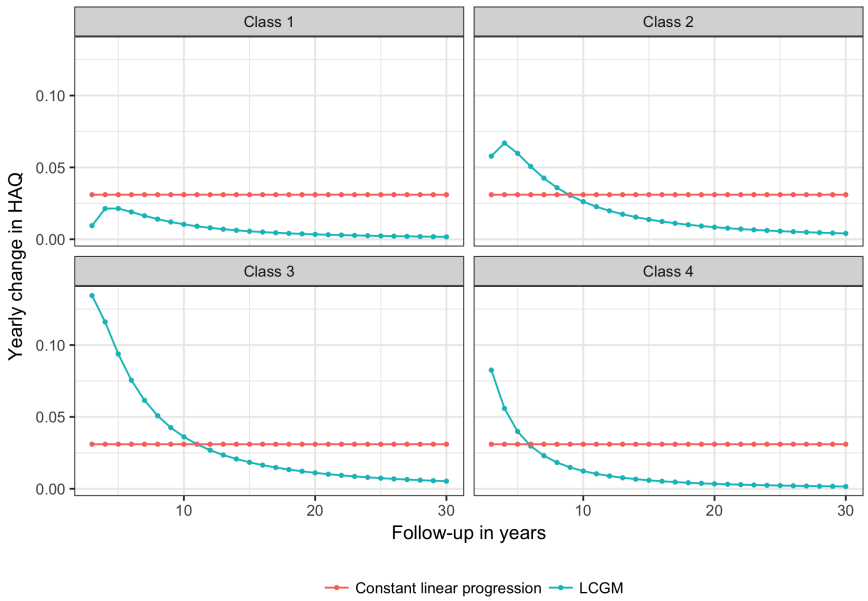


Figure 5: A comparison of predicted yearly changes in HAQ between a latent class growth model and constant linear progression from year 2 onwards

5.5 HAQ trajectory with bDMARD maintenance treatment

Based on the NDB longitudinal study, Wolfe and Michaud (2010) estimated the overall annual HAQ progression rate among RA patients who had switched to biologic treatment at -0.001 (95CI: -0.004 to 0.002). In a separate analysis, also based on NDB data, Michaud et al. (2011) reported annual HAQ progression rates by treatment adjusted for baseline HAQ score, age, sex, education, smoking, BMI, comorbidity, and RA onset. The average HAQ rate among patients on a biologic was -0.001 as well, which instills confidence that the reported HAQ progression rates for different bDMARDs as reported by Michaud et al. (2011) can be directly compared with the overall annual HAQ progression rate of 0.031 reported by Wolfe and Michaud (2010). Accordingly, bDMARD specific HAQ progression rates by Michaud et al. (2011) are used in the model. For bDMARD treatments evaluated in the model for which no HAQ progression rate was reported by Michaud et al. (2011), the overall biologic rate of -0.001 is used.

5.6 Duration of maintenance treatment

Time to treatment discontinuation in the maintenance phase depends on the pathway (**S1-S6**) used to model treatment switching. If **S1** is selected, a single treatment discontinuation curve based on an analysis from the CORRONA database is used for all patients. In **S2-S5**, time to treatment discontinuation is stratified by the level of disease activity, and in **S6** treatment duration depends on EULAR response.

5.6.1 Treatment duration in the US

We based our estimates of treatment duration during the maintenance phase for patients in the US with analyses of the CORRONA database (Strand et al. 2013). The analysis sample consisted of 6,209 patients age 18 or older treated between 2002 and 2011 receiving either TNF inhibitors or

other bDMARDs. The mean age was 57.6 years, 43% of patients were biologic naive, the mean CDAI was 16, and just over 26% of patients had high disease activity (CDAI $\geq$ 22).

7 parametric survival models (exponential, Weibull, Gompertz, gamma, log-logistic, lognormal, and generalized gamma) were estimated on individual patient data reconstructed from a Kaplan-Meier curve from the CORRONA analysis using the algorithm developed in Guyot et al. (2012). We compared fit using the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The generalized gamma had the lowest AIC and BIC, so we consider it to be the preferred model. A plot of of the generalized gamma distribution against the Kaplan-Meier curve is shown in Figure 6. As can be seen in the plot, the shape of the survival curve estimated using a generalized gamma distribution tracks the Kaplan-Meier curve closely.

Table 10: AIC and BIC for parametric models of treatment duration from the CORRONA database

Distribution	AIC	BIC
Exponential	33,240	33,246
Weibull	33,182	33,196
Gompertz	32,963	32,977
Gamma	33,222	33,236
Log-logistic	32,848	32,861
Lognormal	32,650	32,663
Generalized gamma	32,507	32,527

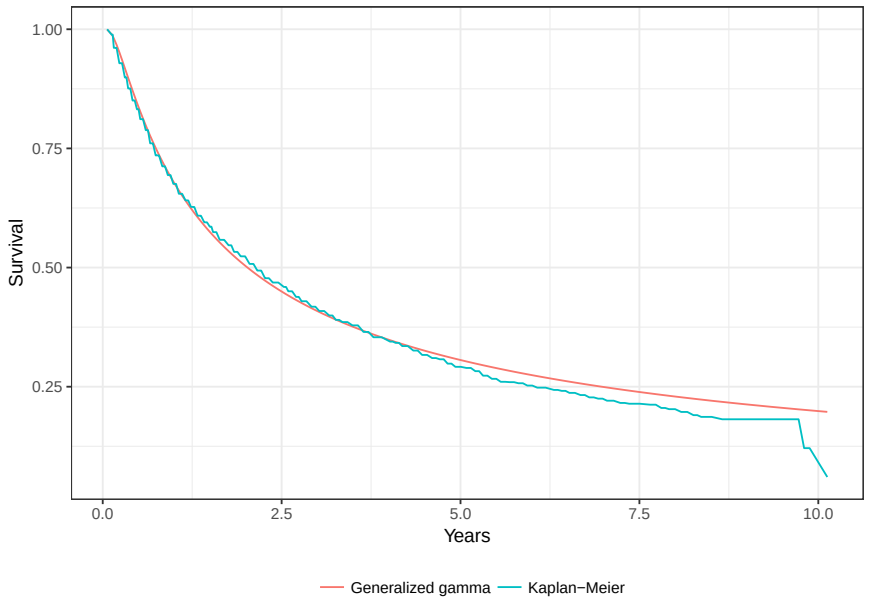


Figure 6: Generalized gamma and Kaplan-Meier time to treatment discontinuation curves using reconstructed individual patient data from the CORRONA database

We considered estimating separate time to discontinuation curves for each therapy, but did not for a number of the reasons cited in Stevenson et al. (2016). The majority of the literature focuses on anti-TNFs (e.g., infliximab, etanercept, and adalimumab) (e.g. Gomez-Reino and Carmona 2006;

Yazici et al. 2009; Pan et al. 2009), which makes it difficult to estimate discontinuation curves for the other therapies. Furthermore, studies comparing rates of discontinuation across therapies tend to be observational because clinical trials are of short duration and do not reflect real-world patient populations. However, although observational studies provide accurate predictions on time to discontinuation, it is difficult to avoid bias from confounding when estimating differences across treatments because patients are not randomized into treatment and control groups (Souto et al. 2015) .

We also lack data on treatment duration for patients on cDMARDs. Following Stevenson et al. (2016), we assume that, conditional on continuing treatment at 6 months, treatment duration for bDMARDs is applicable to treatment duration for cDMARDs. This is, in turn, based on the assumption that cDMARDs are not likely to be more toxic than biologics used in combination with cDMARDs.

5.6.2 Treatment duration by disease activity level

When **S2-S5** are selected, treatment duration is stratified by the level of disease activity. Since patients in the CORRONA database tended to have moderate disease activity (mean CDAI = 16), we use the CORRONA survival curve to model treatment duration for patients with moderate disease activity. We adjust this curve for patients in remission or low disease activity using the odds ratios reported in Zhang et al. (2011), which imply that patients in remission or with low disease activity have .52 times the odds of stopping treatment as patients with moderate disease activity. In particular, we adjust the probability of treatment failure at each point in time using the methodology described in Section A.1. As with the analysis described in Section 5.6.1, we then fit 7 parametric survival models to individual patient data reconstructed from the adjusted survival curve using the Guyot et al. (2012) algorithm. Generalized gamma time to treatment discontinuation curves stratified by disease activity level are shown in Figure 7.

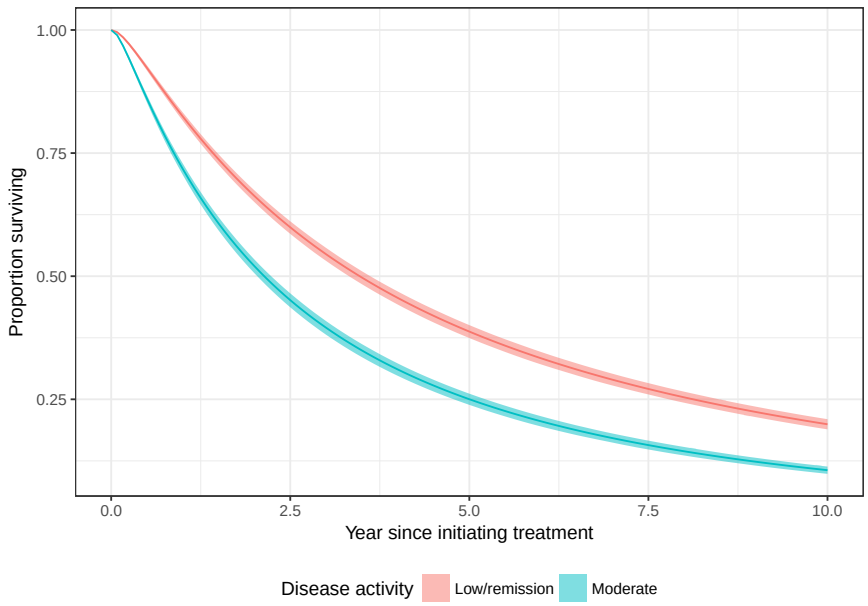


Figure 7: Generalized gamma time to treatment discontinuation curves by disease activity level

5.6.3 Treatment duration by EULAR response

In **S6**, we stratify time to treatment discontinuation by EULAR response based on analyses of the British Society for Rheumatology Biologics Registers (BSRBR) database (Stevenson et al. 2016). We again fit 7 parametric survival models using reconstructed individual patient data. The survival curves reported in Stevenson et al. (2016) were used to create the patient data. The AIC and BIC of each model by EULAR response category are shown in Table 11.

Table 11: AIC and BIC for parametric models of treatment duration by EULAR response

	Moderate EULAR response		Good EULAR response	
Distribution	AIC	BIC	AIC	BIC
Exponential	38,840	38,847	15,126	15,132
Weibull	38,478	38,492	15,090	15,101
Gompertz	38,099	38,112	15,066	15,077
Gamma	38,587	38,600	15,098	15,110
Log-logistic	38,142	38,155	15,062	15,073
Lognormal	37,988	38,001	15,047	15,059
Generalized gamma	37,869	37,889	15,048	15,065

One concern is that the BSRBR is representative of the UK but not the US. As a result, we also estimate "adjusted" survival models appropriate for US based analyses. The adjustment is made in six steps using the analyses from the CORRONA database described in Section 5.6.1.

1. Calculate a hazard function based on a survival curve from an analysis of the CORRONA database. In particular, reconstruct individual patient data from the survival curve (Guyot et al. 2012) and fit a spline-based survival model. Then use the spline-based model to estimate the hazard function $h(t)_{corrona}$.

2. Calculate a hazard function based on the BSRBR. To do so, first calculate hazard functions for both moderate and good EULAR responders using the same method described in step 1. Then calculate an overall hazard function with the proportion of moderate and good responders in the BSRBR analysis. Given that the number of moderate responders is $5,492$ and the number of good responders is $2,417$ the overall hazard function is $h(t)_{bsrbr} = \frac{5,492}{7,909} h(t)_{bsrbr,moderate} + \frac{2,417}{7,909} h(t)_{bbsrbr,good}$.

3. At each point in time, calculate the ratio of the CORRONA and BSRBR hazard functions: $HR(t) = h(t)_{corrona}/h(t)_{bbsrbr}$.

4. Apply the hazard ratio in step 3 to the BSRBR hazard functions for each EULAR response category. That is $h(t)_{bsrbr,moderate,adj} = h(t)_{bsrbr,moderate} \cdot HR(t)$ and $h(t)_{bsrbr,good,adj} = h(t)_{bsrbr,good} \cdot HR(t)$.

5. Generate survival curves using the hazard functions from step 4. Specifically, given a general hazard function $h(t)$, calculate the cumulative hazard function, $H(t) = \int_{z=0}^{t} h(z)dz$, convert this to a survival function using $S(t) = exp(-H(t))$, and reconstruct individual patient data using the survival curve.

6. Fit parametric survival models to the individual patient data generated in step 5.

Both adjusted and unadjusted survival curves by EULAR response fit using a generalized gamma distribution are shown in Figure 8. AIC and BIC for the parametric models fit in step 6 to the adjusted individual patient data are shown in Table 12.

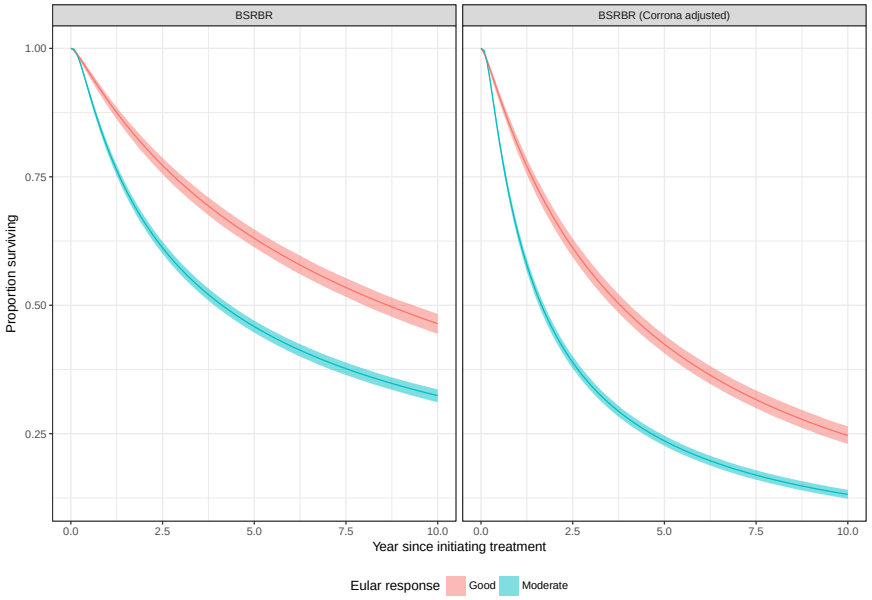


Figure 8: Generalized gamma survival curve of treatment duration using reconstructed individual patient data based on analyses from Stevenson et al. (2016) by EULAR response category

Table 12: AIC and BIC for CORRONA adjusted parametric models of treatment duration by EULAR response

	Moderate EULAR response		Good EULAR response	
Distribution	AIC	BIC	AIC	BIC
Exponential	42,304	42,310	18,098	18,103
Weibull	41,946	41,959	18,051	18,062
Gompertz	41,569	41,582	18,039	18,050
Gamma	42,098	42,111	18,063	18,074
Log-logistic	41,406	41,419	18,037	18,049
Lognormal	41,235	41,248	18,004	18,016
Generalized gamma	41,110	41,129	18,000	18,017

5.7 Rebound post treatment

Since no data exists on the size of the HAQ rebound post treatment, we vary its size as a proportion of the initial 6-month HAQ decline. 1 is used as an upper bound, which implies that the HAQ rebound is equal to the improvement experienced at the end of the initial 6-month period with that treatment. 0.7 is currently used as a lower bound.

5.8 Serious infections

Based on the NMA by Singh et al. (2011) and in accordance with Stevenson et al. (2016), we assume a rate of 0.035 (95% CI: 0.027 to 0.046) infections per person-year with all bDMARDs and a rate of 0.026 (no CI reported) infections per person-year with cDMARDs. The rate of infection is assumed to be equal across bDMARDs because the published results for specific bDMARDs are estimated with very little precision. The standard error on the infection rate for bDMARDs is assumed to be the same as the standard error for cDMARDs since no standard error was reported for bDMARDs in Singh et al. (2011).

A patient in the IPS has a serious infection if the simulated time to serious infection occurs before the simulated time of treatment discontinuation. Table 13 shows the probability of this occurring when treatment duration is modeled using a generalized gamma distribution. The probability of a serious infection is relatively rare as only 0.06 of patient on cDMARDs and 0.11 using bDMARDs have serious infections. However, differences between cDMARDs and bDMARDs are not insignificant as the probability of a serious infection is almost 5 percentage points higher with bDMARDs than with cDMARDs.

Table 13: Probability of serious infection

	Probability		
		95% CI	
	Mean	Lower	Upper
cDMARDs or NBT	0.0574	0.0390	0.0790
bDMARDs	0.1116	0.0790	0.1480

Notes: Probabilities are estimated by simulating 1,000 patients and 1,000 parameter sets. Treatment duration is simulated using a generalized gamma distribution.

An important question related to the sensitivity of cost-effectiveness to the model specification is whether the probability of serious infections depends on the distribution used to model time to treatment discontinuation. Table 14 consequently compares serious infection probabilities by the time to treatment discontinuation distribution. There are very small differences across distributions, suggesting that the treatment duration distribution has almost no impact on the probability of serious infections.

Table 14: Probability of serious infection with cDMARDs by distribution used to model treatment duration

Distribution	Mean probability
Exponential	0.0572
Weibull	0.0570
Gompertz	0.0575
Gamma	0.0573
Log-logistic	0.0578
Lognormal	0.0574
Generalized gamma	0.0574

Notes: Probabilities are estimated by simulating 1,000 patients and 1,000 parameter sets.

5.9 Utility

Two algorithms can be used to map HAQ to an EQ-5D utility score. Each is used to simulate utility for every patient in the model to obtain a distribution of utility over time. Our preferred algorithm is the mixture model developed by Alava et al. (2013), which is described in detail in Appendix F. The second algorithm uses the logistic regression equation reported in Wailoo et al. (2006). The regression coefficients from Wailoo et al. (2006) are shown in Table 15 and are used to predict utility with the inverse logit function.

Table 15: Logistic regression coefficient from Wailoo utility algorithm

	Estimate	Standard error
Intercept	2.0734	0.0263
Age	0.0058	0.0004
Disease duration	0.0023	0.0004
Baseline HAQ	-0.2004	0.0101
Male	-0.2914	0.0118
Number of previous DMARDs	0.0249	0.0028
Current HAQ	-0.8647	0.0103

Notes: Coefficients are from the logistic regression reported in Wailoo et al. (2006).

Figure 9 compares results from the two algorithms. Mean utility scores from the Alava et al. (2013) mixture model lie above those from the Wailoo et al. (2006) equation for all values of HAQ. Moreover, the slope of utility curve produced from the mixture model is steeper (although less so for the commonly observed HAQ scores between 1 and 1.5), implying that changes in HAQ from the mixture model predict larger changes in utility. Given that the mixture models have been shown to predict utility more accurately (Alava et al. 2012, 2013; Hernández Alava et al. 2014), this suggests that standard models underestimate the quality-adjusted life-year benefits, and hence, the cost-effectiveness of treatments.

The utility score depends on serious infections in addition to HAQ. In particular, disutility due to serious infections is assumed to be 0.156 for the duration of the month of infection based on prior

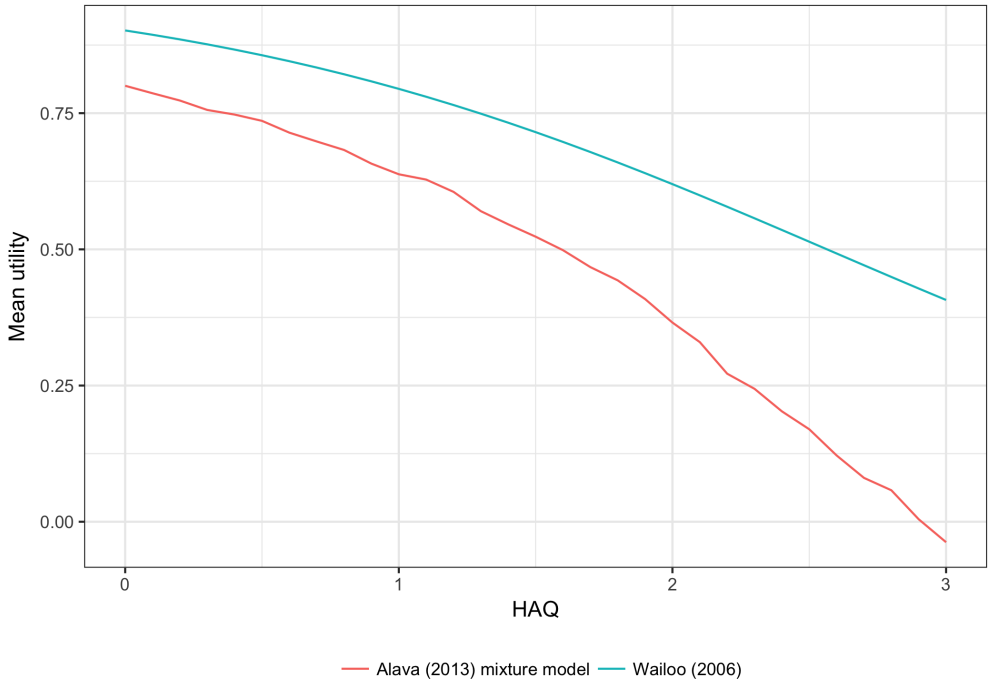


Figure 9: Simulated mean utility by current HAQ

studies (Stevenson et al. 2016; Oppong et al. 2013). However, given the weak evidence for this estimate, the disutility of an infection is allowed to vary by 20% in either direction.

5.10 Mortality

The probability of death is simulated as a function of age/sex specific mortality from U.S. lifetables (Arias 2015), baseline HAQ, and changes in HAQ from baseline. Wolfe et al. (2003) estimate an odds ratio for the effect of HAQ on mortality of 2.22, which is applied to the absolute mortality rates of the general population (HAQ score of 0). To capture the effect of treatment on mortality, we assume that, for every 0.25-unit increase in HAQ score, subsequent 6-month mortality increases according to the hazard ratios reported in Michaud et al. (2012). Parameter estimates are shown in Table 16.

Table 16: Mortality parameters

		95% CI		
	Estimate	Lower	Upper	Reference
Impact of baseline HAQ on mortality				
Log odds of mortality	0.798	0.582	1.012	Wolfe et al. (2003)
Impact of 0.25-unit change in HAQ from baseline on mortality				
Log hazard ratio 0-6 months	0.113	0.077	0.157	Michaud et al. (2012)
Log hazard ratio >6-12 months	0.148	0.104	0.191	Michaud et al. (2012)
Log hazard ratio >12-24 months	0.148	0.095	0.191	Michaud et al. (2012)
Log hazard ratio >24-36 months	0.191	0.131	0.247	Michaud et al. (2012)
Log hazard ratio >36 months	0.174	0.104	0.239	Michaud et al. (2012)

Notes: 95% confidence intervals are calculated using normal distributions on the log odds and log hazard ratio scales.

Figure 10 plots survival curves by gender for 1,000 patients with a baseline age of 55. Survival was simulated by setting the log odds ratios and log hazard ratios from Table 16 equal to their point estimates. Three scenarios are considerd. In scenario one, patients do not have RA (i.e., HAQ score of 0). In the second scenario, patients have baseline HAQ score of 1 but it does not increase over time. In the third scenario, patients still have a baseline HAQ score of 1, but it increases by 0.03 per year. The third scenario, therefore, utilizes the relationship between changes and HAQ and mortality from Michaud et al. (2012) while the second scenario does not.

Mean survival for females without RA was 82.5 years and declined to 77.0 for females with a constant baseline HAQ of 1 and to 72.4 when HAQ increased by 0.03 per year. Mean survival for males in the first, second, and third scenario were 79.4, 73.2, and 70.1 years respectively. Overall, the figure suggests that RA increases mortality and that larger increases in HAQ over time increase mortality by even more.

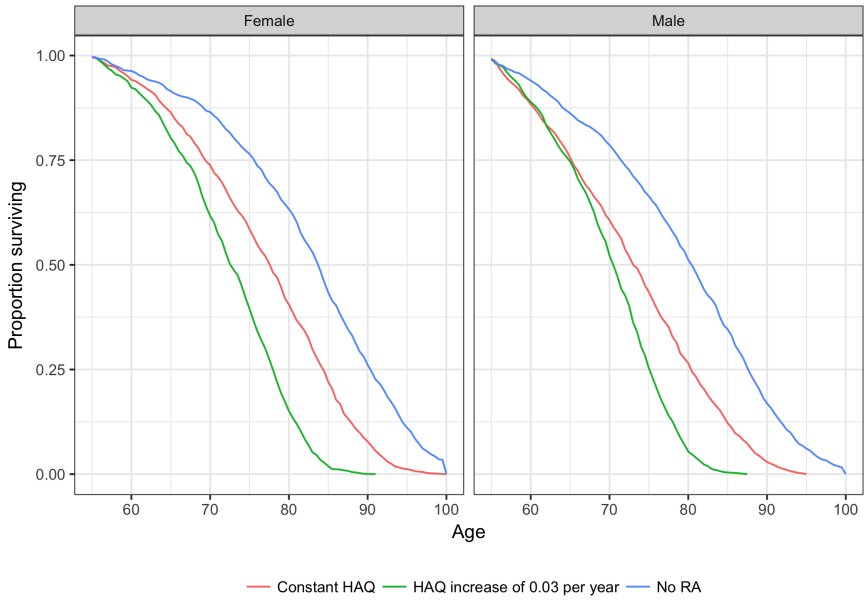


Figure 10: Simulated survival curve for a patient age 55

Notes: Baseline HAQ is 1 for the "Constant HAQ" and "HAQ increase of 0.03 per year" scenarios; baseline HAQ is 0 for the "No RA" scenario.

5.11 Cost

An overview of drug acquisition and administration costs is presented in Table 17. Costs are a function of dose and frequency of administration, strength and dosage form, the wholesale acquisition cost (WAC), and infusion costs. Since infliximab dosing depend on patient weight, the costs for infliximab reported in the table average over a patient population that is 21% male. The WACs in the table do not include discounts or rebates so they may be higher than actual drug costs. The methodology used to calculate drug acquisition and administration costs is described in more detail in Appendix G.

Parameters associated with resource use are show in in Table 18. Costs related to physician visits, chest X-rays, tuberculosis tests, and outpatient follow-up are based on Claxton et al. (2016). The cost per hospital day and the relationship between the HAQ score and the annual number of hospital days are from Carlson et al. (2015). Cost of any serious infection are assumed to be equal to the

Table 17: Drug acquisition and administration cost

Drug	Dose and frequency of administration	Strength and dosage form	Number of doses first 6 months	Number of dosees per year beyond the first 6 months	WAC per unit	Infusion cost	Cost for the first 6 months	Cost per year beyond the first 6 months
Etanercept	50 mg QW	50 mg/0.98 mL syringe or pen injector	26	52	1,110.50	0	28,873	57,746
Adalimumab	40 mg EOW	40 mg/0.8 mL syringe or pen injector	13	26	2,220.62	0	28,868	57,736
Infliximab	3 mg/kg at 0, 2, and 6 weeks, 3mg/kg Q8W, 6 mg/kg Q6W after 6 months	100 mg vial	5	8.67	1,113.27	164	17,519	51,669
Golimumab	50 mg QM	50 mg/0.5 mL syringe or pen injector	6	12	3,811.18	0	22,867	45,734
Certolizumab pegol	400 mg at weeks 0, 2, 4 then 200 mg Q2W	400 mg kit or syringe kit (200 mg 2)	8	26	3,679.87	0	29,438	47,838
Abatacept IV	750 mg IV at weeks 0, 2, 4 then Q4W	250mg vial	8	13	931.16	164	23,659	38,447
Abatacept SC	125 mg SC QW with IV loading dose	125mg/ml syringe	26	52	957.14	0	24,885	49,771
Tocilizumab	162 mg SC EOW	162 mg/0.9 mL syringe	13	26	898.31	0	11,678	23,356
Rituximab	1000 mg at weeks 0, 2; then Q24 W	500 mg/50ml vial	4	4.33	4,176.10	164	34,064	36,903
Tofaticinib citrate	5 mg BID	5mg tablet	364	728	63.26	0	23,026	46,053
Methotrexate monotherapy	15mg QW	15 mg injection	26	52	32.42	0	842	1,685
Hydroxychlorquine sulfate	400mg daily	200 mg tablet	182	364	3.18	0	1,157	2,315
Sulfazalazine	1-2 g daily	500 mg tablet	182	364	0.47	0	342	684

Notes: Costs do not include rebates or discounts. Cost for infliximab are calculated by assuming that 21% of patients are male and that the weight of men and women are 89 kg and 75 kg respectively. Tocilizumab is dosed weekly if weight is greater than 100 kg; costs for tocilizumab reported in the table are for patients weighing less than 100 kg. IV = intravenous; SC = subcutaneous; WAC = wholesale acquisition cost.

cost of pneumonia hospitalization at $5,873, based on Medicare reimbursement rates. Wolfe et al. (2005) provide an estimate of annual income loss in relation to HAQ scores: $4,372 (95% CI: 2,078 to 6,607; 2002 dollars) change per unit HAQ change, which are inflated to 2016 dollars.

Table 18: Resource use parameters

		95% CI		
	Estimate	Lower	Upper	Reference
Days in hospital per year				
HAQ: 0-<0.5	0.260	0.000	1.725	Carlson et al. (2015)
HAQ: 0.5-<1	0.130	0.000	1.409	Carlson et al. (2015)
HAQ: 1-<1.5	0.510	0.015	1.850	Carlson et al. (2015)
HAQ: 1.5-<2	0.720	0.092	1.979	Carlson et al. (2015)
HAQ: 2-<2.5	1.860	1.013	2.960	Carlson et al. (2015)
HAQ: >2.5	4.160	3.238	5.196	Carlson et al. (2015)
Cost per day in hospital	1,251	904	1,652	Carlson et al. (2015)
General management cost				
Chest x-ray	109	97	121	Claxton et al. (2016)
X-ray visit	53	45	61	Claxton et al. (2016)
Outpatient follow-up	187	159	215	Claxton et al. (2016)
Mantoux tuberculin skin test	30	30	30	Claxton et al. (2016)
Productivity loss				
Linear regression coefficient - HAQ	5,853	2,861	8,845	Wolfe et al. (2005)

Notes: 95% confidence intervals for hospital days per year by HAQ score and hospital cost per day are calculated by using the methods of moments to generate the parameters of the gamma distribution given a mean and standard error. The 95% confidence intervals for general management costs are based on normal distributions as assumed in Claxton et al. (2016). 95% confidence interval for productivity loss are calculated using a normal distribution and inflated to 2016 dollars.

6 Simulation and uncertainty analysis

6.1 Individual patient simulation

The IPS is a discrete-time simulation that simulates individual patients one at a time. Model cycles, denoted by t, were chosen to be 6-months long to be consistent with most RCT and real-world data evidence. Algorithm 1 describes the main components of the IPS for a single patient and a single treatment. The full simulation cycles through each treatment in a treatment sequence and through each simulated patient.

6.2 Parameter uncertainty

Parameter uncertainty is quantified using PSA, which propagates uncertainty in the model input parameters throughout the model by randomly sampling the input parameters from their joint probability distribution (Baio and Dawid 2015; Claxton et al. 2005). Probability distributions are determined according to the distributional properties of the statistical estimates, which, in turn, depend on the statistical techniques used and the distributions of the underlying data. We, for the most part, use normal distributions for sample means, gamma distributions for right-skewed data (e.g., hospital costs), and Dirichlet distributions for multinomial data. The multivariate normal distribution is used for regression parameters estimated using frequentist techniques, provided that the variance-covariance from the statistical analysis is available. For parameters estimated using a Bayesian NMA, we fit multivariate normal distributions to the posterior distribution of the parameters generated from the Markov-Chain Monte-Carlo (MCMC) algorithm using sample means

Algorithm 1 Main components of the individual patient simulation

1. **First 6 months** ($t = 0$)
   (a) Simulate treatment switching using **S1-S6**, time to serious infection T_{si}, and death (Appendix E).
      i. **If S1-S6** leads to a treatment switch or if the sampled time to serious infection occurs during cycle 0 (i.e., $T_{si} = 0$), **then** stop treatment. It is assumed that HAQ does not change.
      Else, continue treatment. Simulate change in HAQ using **H1-H3** and time to treatment discontinuation T.
      ii. **If** patient died, **then** move to next patient.
2. **Maintenance phase** (for $t > 0$ and $t \leq T$)
   (a) Simulate death and change in HAQ.
   (b) **If** patient died, **then** move to next patient.
   (c) **If** $t = T$, **then** switch treatment. Treatment switch caused by a serious infection if time to serious infection occurred during or before cycle T (i.e., $T_{si} \leq T$).

and the sample covariance matrix. When we lack evidence on a parameter, we typically assume a uniform distribution with lower and upper limits that reflect the degree of uncertainty in the parameter. The PSA parameter distributions are summarized in Table 19.

6.3 Structural uncertainty

We consider structural uncertainty due to two factors:

- The relationship between health states within the model.
- The statistical model used to estimate parameters.

Table 20 summarizes the competing model structures, which are conditional on the perspective of the decision maker. In total, there are 12 x 2 x 7 x 2 = 336 possible model structures. The choice of model structure for the initial treatment phase (**H1-H3** and **S1-S6**) depends on the preferred measures of disease activity included in the model as well as whether statistical relationships should be modeled directly or indirectly. Likewise, model structures related to HAQ progression, treatment duration, and converting HAQ to utility all reflect uncertainty in the appropriate statistical model.

6.4 Implementation

We begin by describing the simulation procedure conditional on model structure, which uses PSA to capture uncertainty within but not between models. The procedure proceeds in two steps: first, model parameters are sampled from their joint probability distribution (Section 6.2), and second, for each parameter set, model outcomes are simulated one at a time for individual patients in the specified population (Section 4).

Table 19: Probabilistic Sensitivity Analysis Parameter Distributions

Parameter(s)	Distribution
Rebound factor	Uniform
NMA parameters - ACR response	Multivariate normal
NMA parameters - DAS28	Multivariate normal
NMA parameters - HAQ	Multivariate normal
Drug acquisition and administration cost	Fixed
Survival model parameters for treatment duration during maintenance phase	Multivariate normal
US lifetable mortality rates	Fixed
Mortality probability odds ratio - baseline HAQ	Normal
Mortality probability hazard ratio - change in HAQ from baseline	Normal
ACR response to EULAR response mapping	Dirichlet
ACR response to SDAI mapping	Uniform
ACR response to CDAI mapping	Uniform
ACR response to HAQ mapping	Normal
EULAR response to HAQ mapping	Normal
Linear HAQ progression - by therapy	Normal
Linear HAQ progression - by age	Normal
Latent class growth model for HAQ progression	Normal
Utility model - Alava et al. (2013) mixture model	Multivariate normal
Utility model - Wailoo et al. (2006)	Normal
Hospital costs - hospital days by HAQ	Gamma
Hospital costs - hospital costs per day	Gamma
General management cost	Gamma
Serious infection - survival parameters	Normal
Serious infection - cost per infection	Uniform
Serious infection - utility loss	Uniform

Analysts who wish to expand the analysis to capture uncertainty between models can follow the approach described in Bojke et al. (2009). In particular, for each randomly sampled parameter set, each model structure (or a subset of plausible model structures) can be simulated. The distribution

Table 20: Competing model structures

Component of model structure	Possible combinations
Initial effect of treatment on HAQ (**H1-H3**) and switching (**S1-S6**)	12
HAQ trajectory	2
Probability distribution for treatment duration	7
Utility algorithm	2

of simulated outcomes across parameters and models will then reflect uncertainty both within and between models.

It's important to note that simulation output for an individual patient captures differences in outcomes across patients due to random variation (often referred to as first order uncertainty). This information might be useful to patients since it is needed to predict the distribution of their future outcomes conditional on their characteristics, but less useful to a decision maker concerned with making treatment decisions for a population or subset of a population. Analysts wishing to use the model for CEA should therefore estimate mean outcomes by averaging over the simulated patients for each parameter set and model structure. The number of simulated patients should be sufficiently large so that mean outcomes are stable across model runs (i.e., so that first order uncertainty is eliminated).

Although CEA is concerned with mean outcomes, that does not imply that it does not account for heterogeneity. Instead, since outcomes depend on the characteristics of each patient, model averages are a function of the population analyzed. Subgroup analyses can be used to examine differences in cost-effectiveness across subgroups by simulating patients with certain shared characteristics.

Parameter and structural uncertainty imply decision uncertainty, or the degree to which decisions are made based on imperfect knowledge. Indeed, with the aim to maximize health outcomes for a given budget, the optimal decision with current information is to choose the policy that maximizes the expected NMB; however, due to uncertainty, the incorrect policy may be considered the most cost-effective. To characterize this uncertainty, standard summary measures including 95% credible intervals for NMBs and other model outcomes, cost-effectiveness planes, and cost-effectiveness acceptability curves, and the expected value of perfect information can be calculated from the simulated output. Since the expected value of partial perfect information is computationally costly, it can be approximated using meta-modeling techniques (Jalal et al. 2013, 2015; Heath et al. 2016).

7 Validation

We aim to validate the model using the five types of validation described by Eddy et al. (2012). Currently, we are able to use the first three validation types. First, we have checked the model for face validity by ensuring that simulated outcomes are consistent with current science and evidence. Second, we performed unit tests to verify that the individual units of code that are used to simulate the model return expected results. Third, we compared simulated results for key outcomes such as mortality, HAQ over time, and time to treatment discontinuation with real-world data and our

underlying parameter values. In particular, we ran the model online under various scenarios using our R Shiny web application and checked the simulated outcomes.

In the future, we plan to use both external validation and predictive validation to help fine tune our model. External validation will be performed by comparing outcomes simulated using our model to real-world outcomes and predictive validity will involve using our model to forecast future events and comparing our forecasted outcomes to the observed outcomes.

8 Limitations and areas for improvement

The IVI-RA model is an open-source model that is meant to be updated and improved over time. We believe that there are number of potential areas for improvement.

- **Adverse events other than serious infections:** The current model does not account for side effects other than serious infections even though these are important to patients and can result in treatment switching.
- **Adverse events that vary across biologics:** The model allows the serious infection rate to differ between cDMARDs and bDMARDs but assumes that the infection rate is equal among bDMARDs. Future model versions might want to reconsider the evidence underlying this assumption.
- **Time to treatment discontinuation:** Our time to treatment discontinuation curves are based on scanned data and combine information from multiple sources. Direct analyses of databases like the CORRONA database or the National Data Bank for Rheumatic Diseases (NDB) could generate more accurate estimates of treatment duration as well the effect of treatment response or disease activity level on discontinuation rates.
- **Patient preferences:** In the current model, patient utility is a function of the HAQ score and varies according to age, gender, and unobserved patient-specific factors. In other words, utility depends on treatment (through the effect of treatment on HAQ) and the characteristics of the patient. Future iterations of the model should consider other ways that treatment influences utility and that utility varies across patients. For example, disease activity level or the number of previous therapies might help predict utility conditional on HAQ. Furthermore, surveys could be used to estimate the effect of treatment attributes such as route of administration or frequency of administration on utility. Finally, since unobserved patient-specific factors are very important predictors of utility, the model could be run for specific classes of patients within the mixture model (e.g., subgroups where HAQ has the largest effect on utility), although it might be difficult to identify these patient subgroups in a real-world setting.
- **Treatment effect modifiers:** There is currently little evidence (that we are aware of) suggesting that treatment effects vary across patients. When there is sufficient evidence in the literature related to treatment response heterogeneity, we will allow treatment response at 6 months to depend on the characteristics of the patient.
- **Treatment effects after treatment failure:** There are two main limitations in the model related to reductions in treatment response after failing a biologic; first, there are not enough RCTs to reliably estimate bDMARD-specific treatment effects for bDMARD experienced patients using a NMA, and second, treatment response likely does not only depend on whether

a patient is bDMARD naive or experienced, but on the number of previous failures as well. Our current approach is to assume that treatment response is reduced for bDMARD experience patients based on evidence from Carlson et al. (2015). One possible extension is to use a Bayesian NMA approach in which the Carlson et al. (2015) results are used to generate priors for the bDMARD experienced group. As new RCTs become available, the posterior distributions from the Bayesian analysis would move further from the prior and closer to estimates from the trials. The estimates could be further improved by combing NMA results with real-world data and modeling reductions in treatment response as a flexible function of the number of failed biologics.

- **A LCGM for the progression of bDMARDs over time:** The LCGM can be used to model HAQ progression for patients using cDMARDs or on NBT; however, we only have estimates of constant linear progression of HAQ for patients on bDMARDs. Future studies that use non-linear mixture models to model the long-term progression of disease for patients using bDMARDs are needed.
- **The patient population:** Our population characteristics are based on summary data reported in the published literature. As a result, the sampled patient populations within the model do not account for correlations across all of the variables. Distributions estimated from patient databases like the CORRONA database or the NDB would yield more realistic patient populations.
- **Estimating the rebound effect:** One of the most important predictors of cost-effectiveness is the degree to which the HAQ score increases following treatment failure. Most models currently assume that the HAQ score increases by the same amount as the initial 6 month decline in the HAQ score, but there is little evidence to support this. Studies that attempt to quantify the rebound effect are critical.

Appendices

A Rates, probabilities, and standard errors

A.1 Using odds ratios to adjust probabilities

Let p_1 be a baseline probability, β be a vector of log odds ratios, and x be a vector of regressors. We apply the log odds ratios to p_1 to generate a new probability p_2 with the logistic equation,

$$p_2 = \frac{1}{1 + \exp\left[-\left(\text{logit}(p_1) + x^T\beta\right)\right]}, \tag{A1}$$

where,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{A2}$$

A.2 Converting rates and probabilities

Given a *constant* rate r during a given time period, we estimate the probability of an event occurring before time t using the exponential distribution,

$$p(\tau < t|r) = 1 - e^{-rt}. \tag{A3}$$

Given a probability p, the rate parameter is recovered by applying the log transformation,

$$r = \frac{-\ln(1-p)}{t}. \tag{A4}$$

A.3 Calculating standard errors from confidence intervals

Journal articles often report confidence intervals rather than standard errors. However, given that regression coefficients are asymptotically normally distributed, standard errors can be calculated from a confidence interval using the normal distribution. In particular, given a coefficient estimate β (e.g., a log hazard ratio, log odds ratio, or linear regression coefficient) and an upper bound u and lower bound l of a two-sided 95% confidence interval, we calculate the standard error as,

$$SE(\beta) = \frac{u - l}{2 \cdot \Phi^{-1}(0.975)}, \tag{A5}$$

where $\Phi^{-1}(p)$ is the quantile function of the normal distribution.

B Heterogeneous populations

When generating heterogeneous patient populations, we sample binary variables from binomial distributions, continuous uncorrelated variables from normal distributions, and continuous correlated variables from multivariate normal distributions. Truncated distributions are used when variables are restricted to lie within certain intervals.

In particular, the proportion of the female population is drawn from a binomial distribution while age, disease duration and the number of previous DMARDs are drawn from truncated normal distributions. Each sampled value of the number of previous DMARDs is rounded to the nearest integer. Baseline HAQ and three disease activity measures (DAS28, SDAI, and CDAI) are drawn from truncated multivariate normal distributions. The covariance matrix is calculated using the correlations reported in Aletaha et al. (2005) (Figure A1).

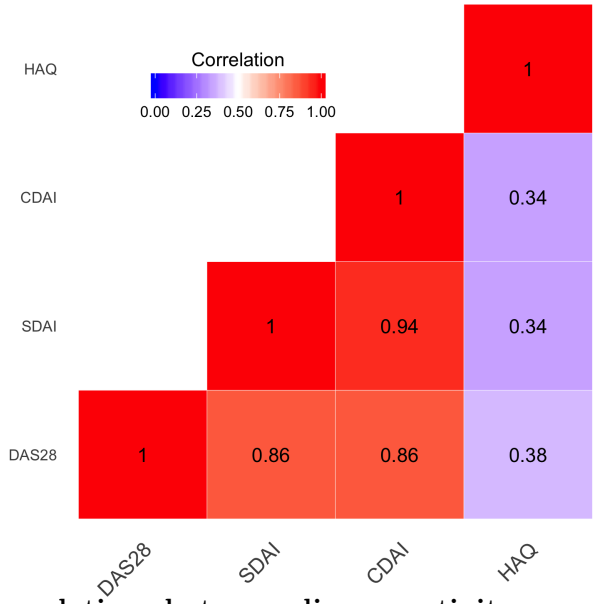


Figure A1: Correlations between disease activity measures and HAQ

We used the correlations from the routine cohort (during visit 1) rather than correlations in the inception cohort (at baseline) since the correlation between HAQ and the disease activity measures were more similar to those from the Leflunomide database (Smolen et al. 2003). That said, correlations between the three disease activity measures were nearly identical in each cohort. The one exception was that the correlation between SDAI and CDAI of 1 in the routine cohort seemed unreasonably high so we used the value of 0.94 from the inception cohort.

We used this sampling procedure to simulate 1,000 patients. Summary statistics from a simulated patient cohort of size 1,000 are shown in Table A1.

Table A1: Summary of characteristics for 1,000 simulated patients

		95 CI%	
	Mean	Lower	Upper
Age	54.95	29.83	77.97
Male	0.24	0.00	1.00
Weight (kg)	78.30	75.00	89.00
Previous DMARDs	3.42	0.00	7.00
DAS28	6.00	3.64	8.13
SDAI	42.95	18.67	66.95
CDAI	41.02	16.86	64.19
HAQ	1.50	0.23	2.67

C Mapping ACR response to changes in disease activity

Let DA denote disease activity, n_1 the number of patients with ACR 20 to <50 response, n_2 the number of patients with ACR 50 to <70 response, n_3 the number of patients with ACR ≥70

response, and N the number of patients with an ACR response greater than or equal to 20%. Mean changes in SDAI, CDAI, and DAS28 by overlapping ACR response categories are converted to mean changes by mutually exclusive ACR response categories as follows:

- **ACR 70**: Mean changes by ACR $\geq$70 were reported directly in Aletaha and Smolen (2005).
- **ACR 50 to <70**: Mean change in disease activity given ACR 50 to <70 response is calculated by solving for $\mathbb{E}[DA|50 \leq ACR < 70]$:

$$\mathbb{E}[DA|ACR \geq 50] = \frac{n_2}{N} \cdot \mathbb{E}[DA|50 \leq ACR < 70] + \frac{n_3}{N} \cdot \mathbb{E}[DA|ACR \geq 70]. \tag{A6}$$

- Mean change in disease activity given ACR 20 to <50 response is calculated by solving for $\mathbb{E}[DA|20 \leq ACR < 50]$

$$\mathbb{E}[DA|ACR \geq 20] = \frac{n_1}{N} \cdot \mathbb{E}[DA|20 \leq ACR < 50] + \frac{n_2 + n_3}{N} \cdot \mathbb{E}[DA|ACR \geq 50]. \tag{A7}$$

D HAQ progression

D.1 Effect of age on linear HAQ progression

Michaud et al. (2011) report an overall rate of linear HAQ progression and rates for three age groups (<40, 40-64, $\geq$ 65). Let β be the overall rate of progression and β_a be the rate of progression for age group a. To estimate the effect of age on the progression rate, we calculated the difference between the overall progression rate and the age specific rate, $\delta_a = \beta - \beta_a$. We estimated the standard error of this quantity assuming no covariance between β and β_a,

$$SE(\delta_a) = \sqrt{SE(\beta)^2 + SE(\beta_a)^2}. \tag{A8}$$

D.2 HAQ trajectory with a latent class growth model

Norton et al. (2014) model HAQ progression using a LCGM. The probability that individual i is a member of class c at time t is modeled using a multinomial logistic regression,

$$P(C_{it} = c) = \frac{\exp(w_{it}^T \delta_c)}{\sum_{s=1}^{4} \exp(w_{it}^T \delta_s)}, \tag{A9}$$

where δ_s is the vector of regression coefficients associated with class s and w_{it} is the corresponding vector of regressors. The variables included in w_{it} are age, gender, baseline DAS28, symptom duration, rheumatoid factor, ACR criteria, and socioeconomic status. Regression coefficients for classes 2-4 relative to class 1 are shown in Table A2. Older age and female gender are especially important predictors of membership in higher risk classes; a worse DAS28 score, rheumatoid factor

Table A2: Determinants of class membership in the ERAS cohort

		95% CI	
	Coefficient	Lower	Upper
Class 2: moderate			
Intercept	-3.496	-4.715	-2.277
Age at onset	0.025	0.011	0.039
Female gender	0.841	0.457	1.225
Disease duration (months)	0.304	0.147	0.461
DAS28 score	0.032	0.001	0.063
Rheumatoid factor positive	0.214	-0.251	0.679
ACR criteria for RA	0.278	-0.163	0.719
Socioeconomic status	0.993	0.276	1.710
Class 3: high			
Intercept	-6.686	-7.980	-5.392
Age at onset	0.037	0.023	0.051
Female gender	1.694	1.275	2.113
Disease duration (months)	0.573	0.424	0.722
DAS28 score	0.046	0.013	0.079
Rheumatoid factor positive	0.315	-0.175	0.805
ACR criteria for RA	0.413	-0.050	0.876
Socioeconomic status	1.119	0.449	1.789
Class 4: severe			
Intercept	-12.055	-14.215	-9.895
Age at onset	0.082	0.060	0.104
Female gender	1.976	1.449	2.503
Disease duration (months)	0.800	0.631	0.969
DAS28 score	0.042	0.001	0.083
Rheumatoid factor positive	0.298	-0.270	0.866
ACR criteria for RA	0.939	0.320	1.558
Socioeconomic status	1.429	0.682	2.176

Notes: Class 1, or the "low" group, is the reference category.

positivity, fulfillment of the 1987 ACR criteria, lower socioeconomic status, and longer disease duration are also predictors of membership in classes with worse HAQ progression.

The HAQ trajectory for a given class can be written as,

$$y^*_{itc} = \beta_{0c} + \beta_{1c}x_t + \beta_{2c}x_t^2 + \beta_{3c}x_t^3 + \epsilon_{it} \tag{A10}$$

$$y_{itc} = \begin{cases} 0 & \text{if } y^*_{itc} < 0 \\ y^*_{itc} & \text{if } 0 \leq y^*_{itc} \leq 3 \\ 3 & \text{if } y^*_{itc} > 3, \end{cases} \tag{A11}$$

where y_{itc} is the HAQ score, x_t is a variable that is a function of time, the β_{jc} are polynomial regression coefficients for members of class c, and ϵ_{it} is an error term.

Sam Norton generously provided us with statistical estimates of the 4 class LCGM used in Norton et al. (2014) from MPlus. Like Stevenson et al. (2016), we noted that the coefficient estimates the MPlus resulted in large fluctuations in the predicted HAQ scores, likely because three decimal places was not precise enough for the cubic term in Equation A10. We consequently used the coefficient estimates to predict the probability of class membership—which are less likely to be influenced by the number of reported decimal places—but estimated Equation A10 using the observed HAQ values reported in Figure 2 in Norton et al. (2014). However, since standard errors were artificially high using grouped data, we standard errors in Equation A10 were based on those reported in the original paper. Moreover, since we are only interested in the HAQ trajectory following the HAQ decline during the initial treatment phase, we limited our analysis to HAQ values from year 2 and onwards. Using the post year 2 data, we estimated Equation A10 using separate linear regressions with cubic polynomials for each class (Table A3). Like Norton et al. (2014), we set x_t equal to a reciprocal transformation of time,

$$x_t = 1 - \frac{1}{t+1} \tag{A12}$$

In the cost-effectiveness model, we simulate the HAQ score at 6 months as a function of the baseline HAQ score and the change in HAQ during the initial treatment phase. Since the Norton et al. (2014) model is not conditional on the HAQ score in the previous period, we use it to predict changes in HAQ rather than the level of the HAQ score. More precisely, for a patient in a given class, we model the change in HAQ as,

$$\begin{aligned} \Delta y^*_{itc} &= y^*_{i,t,c} - y^*_{i,t-1,c} \\ &= \beta_{1c}(x_t - x_{t-1}) + \beta_{2c}(x_t^2 - x_{t-1}^2) + \beta_{3c}(x_t^3 - x_{t-1}^3) + (\epsilon_{i,t} - \epsilon_{i,t-1}). \end{aligned} \tag{A13}$$

Since Equation A10 was estimated on aggregated data, we did not have reliable estimates of ϵ_{it}. We consequently set $\epsilon_{i,t} - \epsilon_{i,t-1}$ equal to 0, which implies that we are generating a mean response rather than a predicted response. In other words, we are not simulating the random variation associated with each individual, but are still accurately simulating mean outcomes across populations or subpopulations.

Table A3: LCGM HAQ trajectory coefficients

	Coefficient	Standard error
Class 1: low		
Intercept	0.638	0.058
Linear	-1.009	0.074
Quadratic	-0.649	0.027
Cubic	1.355	0.003
Class 2: moderate		
Intercept	0.950	0.058
Linear	-0.109	0.020
Quadratic	-3.368	0.002
Cubic	3.699	0.064
Class 3: high		
Intercept	1.265	0.064
Linear	-0.132	0.056
Quadratic	-2.531	0.021
Cubic	3.538	0.002
Class 4: severe		
Intercept	1.935	0.063
Linear	-0.540	0.073
Quadratic	1.196	0.027
Cubic	-0.109	0.003

Notes: Class 1, or the "low" group, is the reference category.

E Simulating death

Death is simulated for each patient during each model cycle based on age, gender, baseline HAQ, and change in HAQ from baseline. A 0/1 death indicator is randomly drawn using the following procedure:

1. Find q_{xg}, the probability that a patient of gender g and age x will die before age $x+1$, from lifetables.

2. As described in Section A.1, adjust q_{gx} using the effect of a change in baseline HAQ on the odds of mortality, OR,

$$p_m = \frac{1}{1+\exp\left[-(\text{logit}(q_x)+\log(OR)\cdot HAQ)\right]}. \tag{A14}$$

3. Following Section A.2, convert the mortality probability, p_m, into a mortality rate, r_m.

$$r_m = -log(1-p_m). \tag{A15}$$

4. Adjust the mortality rate, r_m, using the estimated log hazard ratio of mortality, HR, of a

change in HAQ from baseline, Δ HAQ.

$$r_m = r_m \cdot \exp[\log(HR) \cdot \Delta HAQ] \tag{A16}$$

5. Following Section A.2, convert the mortality rate into a probability given a 6-month cycle length,

$$p_m = 1 - \exp[-r_m * (6/12)]. \tag{A17}$$

6. Randomly draw a 0/1 death indicator, d, given the probability of death, p_m,

$$d \sim \text{Bin}(1, p_m). \tag{A18}$$

F Simulating utility with a mixture model

The mixture model estimated by Alava et al. (2013) simulates utility in two stages. In the first stage, we sampled pain for a given individual in a particular model cycle based on the HAQ score. In the second stage, we simulated utility as a function of HAQ, pain and age/sex.

F.1 Simulating pain

To simulate pain from HAQ, we used the summary statistics for pain and HAQ reported in Sarzi-Puttini et al. (2002). Pain was measured with the visual analog scale (VAS) with mean $\mu_{pain} = 61.65$ and standard deviation $\sigma_{pain} = 19.10$, while HAQ was reported to have mean $\mu_{haq} = 1.39$ and standard deviation $\sigma_{haq} = 0.59$.

We then estimated the correlation between pain and HAQ by digitally scanning the curve depicting the (linear) relationship between pain and HAQ (Figure 114) shown in Stevenson et al. (2016). Using the scanned data, we regressed pain on HAQ using simple ordinary least squares (OLS). The correlation between pain and HAQ, estimated as $\rho = 0.52$, was calculated by rearranging the OLS estimate for the slope, β, of the regression model,

$$\rho = \beta \cdot \frac{\sigma_{haq}}{\sigma_{pain}}. \tag{A19}$$

Pain was simulated using these parameters by assuming that pain was normally distributed conditional on HAQ,

$$pain|haq = h \sim N\left(\mu_{pain} + \rho\frac{\sigma_{pain}}{\sigma_{haq}}(h - \mu_{haq}), \sigma^2_{pain}(1 - \rho^2)\right). \tag{A20}$$

However, since the VAS is constrained to lie between 0 and 100, pain was drawn from a truncated normal distribution with a lower limit of 0 and an upper limit of 100.

F.2 Simulating utility

After simulating pain, we simulated utility with a mixture model. Within each class c, the HAQ score for patient i in period t was modeled as,

$$y_{it|C_{it}} = \begin{cases} 1 & \text{if } y^*_{it|C_{it}} > 0.883 \\ y^*_{it|C_{it}} & \text{otherwise} \end{cases} \tag{A21}$$

$$y^*_{it|C_{it}} = \alpha_{ic} + x_{it}^T \beta_c + \epsilon_{it} \tag{A22}$$

$$\alpha_{ic} = \gamma_c + z_i^T \kappa + \mu_i, \tag{A23}$$

where ϵ_{it} is a random error term and β_c is a vector of regression coefficients corresponding to the vector of variables x_{it}. α_{ic} is a random intercept for individual i and class c that is predicted by a class-specific intercept, γ_c, a vector of individual-specific variables z_i, a coefficient vector κ, and an error term, μ_i. Variables included in x_{it} are HAQ, HAQ^2, $Pain/100$, $Age/10$, and $Age/100$; z_i contains a single indicator variable, $Male$, equal to 1 if the patient is male and 0 if female.

The probability of class membership was modeled using a multinomial logit model,

$$P(C_{it} = c) = \frac{\exp(w_{it}^T \delta_c)}{\sum_{s=1}^{4} \exp(w_{it}^T \delta_s)}, \tag{A24}$$

where there are four possible classes and δ_c is a vector of coefficients corresponding to the vector of variables, w_{it} (which includes an intercept). Variables included in w_{it} other than the intercept are HAQ, $Pain/100$, and $Pain/100^2$.

We sampled from the mixture model as follows.

1. For each individual i, sample the error term, $\mu_i \sim N(0, \sigma_\mu^2)$.
2. For each individual i and time-period t:
   (a) Sample class membership conditional on w_{it}; that is, sample $C_{it} \sim \text{Cat}(\text{p}_1, \text{p}_2, \text{p}_3, \text{p}_4)$ where p_c is the probability of being in class c.
   (b) Predict the intercept α_{ic}.
   (c) Sample the error term, $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$.
   (d) Predict the HAQ score, y_{it}.

G Drug acquisition and administration costs

Drug acquisition and administration costs are calculated separately during the initial treatment phase and the maintenance phase since dosing typically differs. Costs are separated into acquisition costs and infusion costs. Infusion costs are calculated by multiplying the number of doses in a 6 month period by the cost of an infusion and acquisition costs are calculated as,

$$cost = \left\lceil \frac{dose_{amt}}{strength_{amt}} \right\rceil \cdot doses_{num} \cdot WAC, \tag{A25}$$

where $\lceil \cdot \rceil$ is the ceiling function and implies that products cannot be reused after opening, $dose_{amt}$ is the recommended dose of the drug, $strength_{amt}$ is the strength of the drug, $doses_{num}$ is the number of doses in a 6 month period, and WAC is the WAC. For example, as shown in Table 17, both the strength and the dose of adalimumab are 50 mg, so costs for the initial 6 month period are calculated by multiplying the number of doses (13) by the WAC ($2,220.62).

When dosing depends on weight, costs are calculated separately for each patient in the simulation. In particular, costs are calculated as,

$$cost = \lceil weight \cdot dose_{amt} / strength_{amt} \rceil \cdot doses_{num} \cdot WAC, \tag{A26}$$

where weight is patient weight, $dose_{amt}$ is the dose per weight, and $strength_{amt}$, WAC, and $doses_{num}$ are defined in the same way as in the non-weight based scenario. To illustrate, the acquisition cost for infliximab after the first 6 months is calculated by multiplying each patient's weight by the dose (6 mg/kg) and dividing by the size of a vial (100 mg), and then multiplying by the number of doses (8.67) and the WAC ($1,113.27).

H Network Meta-Analysis

H.1 Systematic literature review

Population

- Adult (>18 years) patients with moderate to severe RA who have had inadequate response to cDMARDs

Interventions and comparators

- Biologics as monotherapy or in combination with cDMARDs (adalimumab, certolizumab pegol, etanercept, golimumab, infliximab, abatacept, rituximab, tocilizumab, sarilumab, tofacitinib)
- cDMARDs alone or in combination (MTX, HCQ, SSZ or LEF)

Outcomes

- ACR20/ACR50/ACR70
- DAS28
- HAQ-DI score

Study design

- Randomized controlled trials

Other

- Studies published in English
- Primary study available as full text published manuscript only; no study available as a conference abstract only was included.

H.2 Criteria for studies to be selected from the systematic literature review and included in the NMA

The following criteria were used to select relevant studies to be included in the NMA:

Population

- Adult (>18 years) patients with moderate to severe RA who have had inadequate response to cDMARDs and are bDMARD-naive

Interventions

- Biologics as monotherapy or in combination with cDMARDs (adalimumab, certolizumab pegol, etanercept, golimumab, infliximab, abatacept, rituximab, tocilizumab, sarilumab, tofacitinib)

Comparators

- cDMARDs
- Any active comparator that allows for an indirect comparison between the bDMARDs of interest

Outcomes

- ACR20/ACR50/ACR70 at 6 months follow-up
- Change in DAS28 from baseline at 6 months follow-up
- Change in HAQ-DI score from baseline at 6 months follow-up

H.3 Identified evidence base

The evidence network and inclusion criteria to go here.

H.4 Statistical models for network-meta analysis

ACR response, 6 month change in HAQ from baseline, and 6 month change in DAS28 from baseline were estimated using a Bayesian (random effects) network meta-analysis approach. The four mutually exclusive ACR response categories were estimated using a ordered probit model appropriate for ordered categorical data (Dias et al. 2013). The model assumes that there is an underlying continuous variable (ACR20/50/70) categorized by specifying different cutoffs corresponding to the point at which an individual moves from one category to the next in each trial. The advantage of this approach over an analysis that considers ACR categories separately is that all possible outcomes are analyzed simultaneously based on the same randomized controlled trials, allowing for consistent estimates by category. The relative treatment effects for each bDMARD versus cDMARDs estimated on the probit scale were transformed into absolute probabilities of the non overlapping ACR response categories by combining them with the average results for cDMARDs.

Changes in HAQ and DAS28 from baseline at 6 months were estimated using a Bayesian (random effects) network meta-analyses model for continuous data (Dias et al. 2013). The models use a normal likelihood (since the sample mean is approximately normally distributed by the central limit theorem if the sample size is reasonably large) and an identity link.

To avoid influencing the observed results by prior belief, uninformative prior distributions were used for the estimated model parameters. The posterior distributions of parameters of interest were summarized by the mean as a reflection of the point estimate and 95% credible intervals, constructed from the 2.5 and 97.5 percentiles. Analyses were performed with the Markov chain Monte Carlo method using the JAGS software package (http://mcmc-jags.sourceforge.net/).

H.5 Comparing the IVI NMA to the NICE NMA

To help ensure that differences in cost-effectiveness estimates from our model relative to others are not driven by the NMA results, we compared our NMA estimates to estimates reported by NICE in Stevenson et al. (2016). We focus on ACR response, since the NICE report and other models use treatment pathways similar to **H1** and **H2** and rarely use DAS28 to inform treatment duration. As shown in Table A4, our results are similar and the NICE point estimates are generally within the 95% credible intervals surrounding our point estimates.

Table A4: A comparison of NICE and IVI estimates of ACR response probabilities

	IVI			NICE		
	ACR20	ACR50	ACR70	ACR20	ACR50	ACR70
cDMARDs	0.265 (0.248, 0.283)	0.102 (0.093, 0.113)	0.032 (0.028, 0.036)	0.298	0.123	0.042
ABT IV + MTX	0.559 (0.479, 0.636)	0.312 (0.241, 0.385)	0.142 (0.100, 0.192)	0.573	0.328	0.156
ADA + MTX	0.565 (0.485, 0.637)	0.318 (0.249, 0.385)	0.146 (0.104, 0.191)	0.615	0.368	0.183
ETN + MTX	0.648 (0.527, 0.751)	0.400 (0.282, 0.516)	0.204 (0.123, 0.293)	0.713	0.472	0.263
GOL + MTX	0.599 (0.451, 0.733)	0.352 (0.224, 0.492)	0.171 (0.089, 0.273)	0.642	0.395	0.202
IFX + MTX	0.662 (0.417, 0.862)	0.425 (0.195, 0.672)	0.230 (0.074, 0.448)	0.595	0.348	0.169
TCZ + MTX	0.562 (0.373, 0.741)	0.320 (0.168, 0.503)	0.151 (0.062, 0.283)	0.706	0.464	0.256
CZP + MTX	0.740 (0.538, 0.887)	0.510 (0.292, 0.712)	0.296 (0.130, 0.491)	0.564	0.319	0.150
ABT SC + MTX	0.567 (0.434, 0.703)	0.322 (0.210, 0.455)	0.150 (0.082, 0.244)	0.638	0.391	0.199
RTX + MTX	0.570 (0.415, 0.716)	0.325 (0.196, 0.473)	0.152 (0.076, 0.257)	0.573	0.328	0.156
TOF + MTX	0.608 (0.447, 0.767)	0.362 (0.219, 0.535)	0.179 (0.087, 0.313)	-	-	-

Notes: ACR20/50/70 categories are the probability of at least a 20/50/70% improvement. 95% credible intervals are in parentheses. IVI estimates are based on 6-month simulations of 1,000 patients and 1,000 parameters sets for each therapy. NICE estimates are from Table 37 in Stevenson et al. (2017). cDMARDs = conventional disease-modifying antirheumatic drugs; MTX = methotrexate; ABT IV = abatacept intravenous; ADA = adalimumab; ETN = etanercept; GOL = golimumab; IFX = infliximab; TCZ = tocilizumab; CZP = certolizumab pegol; ABT SC = abatacept subcutaneous; RTX = rituximab; TOF = tofacitinib. ACR = American College of Rheumatology.

References

Alava, M. H., Wailoo, A., Wolfe, F., and Michaud, K. (2013). The relationship between eq-5d, haq and pain in patients with rheumatoid arthritis. *Rheumatology*, 52(5):944–950.

Alava, M. H., Wailoo, A. J., and Ara, R. (2012). Tails from the peak district: adjusted limited dependent variable mixture models of eq-5d questionnaire health state utility values. *Value in Health*, 15(3):550–561.

Aletaha, D., Nell, V. P., Stamm, T., Uffmann, M., Pflugbeil, S., Machold, K., and Smolen, J. S. (2005). Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. *Arthritis research & therapy*, 7(4):R796.

Aletaha, D. and Smolen, J. (2005). The simplified disease activity index (sdai) and the clinical disease activity index (cdai): a review of their usefulness and validity in rheumatoid arthritis. *Clinical and experimental rheumatology*, 23(5):S100.

Anderson, J., Caplan, L., Yazdany, J., Robbins, M. L., Neogi, T., Michaud, K., Saag, K. G., O'dell, J. R., and Kazi, S. (2012). Rheumatoid arthritis disease activity measures: American college of rheumatology recommendations for use in clinical practice. *Arthritis care & research*, 64(5):640–647.

Arias, E. (2015). United states life tables, 2011. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 64(11):1–62.

Baio, G. and Dawid, A. P. (2015). Probabilistic sensitivity analysis in health economics. *Statistical methods in medical research*, 24(6):615–634.

Bojke, L., Claxton, K., Sculpher, M., and Palmer, S. (2009). Characterizing structural uncertainty in decision analytic models: a review and application of methods. *Value in Health*, 12(5):739–749.

Carlson, J. J., Ogale, S., Dejonckheere, F., and Sullivan, S. D. (2015). Economic evaluation of tocilizumab monotherapy compared to adalimumab monotherapy in the treatment of severe active rheumatoid arthritis. *Value in Health*, 18(2):173–179.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., Brazier, J., and O'Hagan, T. (2005). Probabilistic sensitivity analysis for nice technology assessment: not an optional extra. *Health economics*, 14(4):339–347.

Claxton, L., Jenks, M., Taylor, M., Wallenstein, G., Mendelsohn, A. M., Bourret, J. A., Singh, A., Moynagh, D., and Gerber, R. A. (2016). An economic evaluation of tofacitinib treatment in rheumatoid arthritis: Modeling the cost of treatment strategies in the united states. *Journal of managed care & specialty pharmacy*, 22(9):1088–1102.

Curtis, J. R., Jain, A., Askling, J., Bridges, S. L., Carmona, L., Dixon, W., Finckh, A., Hyrich, K., Greenberg, J. D., Kremer, J., et al. (2010). A comparison of patient characteristics and outcomes in selected european and us rheumatoid arthritis registries. In *Seminars in arthritis and rheumatism*, volume 40, pages 2–14. Elsevier.

Deighton, C., Hyrich, K., Ding, T., Ledingham, J., Lunt, M., Luqmani, R., Kiely, P., Bukhari, M., Abernethy, R., Ostor, A., et al. (2010). Bsr and bhpr rheumatoid arthritis guidelines on eligibility criteria for the first biological therapy. *Rheumatology*, 49(6):1197–1199.

Dias, S., Sutton, A. J., Ades, A., and Welton, N. J. (2013). Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617.

Eddy, D. M., Hollingworth, W., Caro, J. J., Tsevat, J., McDonald, K. M., and Wong, J. B. (2012). Model transparency and validation: a report of the ispor-smdm modeling good research practices task force–7. *Medical Decision Making*, 32(5):733–743.

Gibson, L., Alava, M. H., and Wailoo, A. (2015). Progression of disease in people with rheumatoid arthritis treated with non biologic therapies. Technical report.

Gomez-Reino, J. J. and Carmona, L. (2006). Switching tnf antagonists in patients with chronic arthritis: an observational study of 488 patients over a four-year period. *Arthritis research & therapy*, 8(1):R29.

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12(1):9.

Heath, A., Manolopoulou, I., and Baio, G. (2016). Estimating the expected value of partial perfect information in health economic evaluations using integrated nested laplace approximation. *Statistics in medicine*, 35(23):4264–4280.

Hernández Alava, M., Wailoo, A., Wolfe, F., and Michaud, K. (2014). A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. *Medical Decision Making*, 34(7):919–930.

Institute for Clinical and Economic Review (2017). Targeted immune modulators for rheumatoid arthritis: Effectiveness & value. Technical report.

Jalal, H., Dowd, B., Sainfort, F., and Kuntz, K. M. (2013). Linear regression metamodeling as a tool to summarize and present simulation model results. *Medical Decision Making*, 33(7):880–890.

Jalal, H., Goldhaber-Fiebert, J. D., and Kuntz, K. M. (2015). Computing expected value of partial sample information from probabilistic sensitivity analysis using linear regression metamodeling. *Medical Decision Making*, 35(5):584–595.

Madan, J., Ades, T., Barton, P., Bojke, L., Choy, E., Helliwell, P., Jobanputra, P., Stein, K., Stevens, A., Tosh, J., et al. (2015). Consensus decision models for biologics in rheumatoid and psoriatic arthritis: recommendations of a multidisciplinary working party. *Rheumatology and Therapy*, 2(2):113–125.

Michaud, K., Vera-Llonch, M., and Oster, G. (2012). Mortality risk by functional status and health-related quality of life in patients with rheumatoid arthritis. *The Journal of rheumatology*, 39(1):54–59.

Michaud, K., Wallenstein, G., and Wolfe, F. (2011). Treatment and nontreatment predictors of health assessment questionnaire disability progression in rheumatoid arthritis: a longitudinal study of 18,485 patients. *Arthritis care & research*, 63(3):366–372.

Norton, S., Fu, B., Scott, D. L., Deighton, C., Symmons, D. P., Wailoo, A. J., Tosh, J., Lunt, M., Davies, R., Young, A., et al. (2014). Health assessment questionnaire disability progression in early rheumatoid arthritis: systematic review and analysis of two inception cohorts. In *Seminars in arthritis and rheumatism*, volume 44, pages 131–144. Elsevier.

Norton, S., Sacker, A., Dixey, J., Done, J., Williams, P., and Young, A. (2013). Trajectories of functional limitation in early rheumatoid arthritis and their association with mortality. *Rheumatology*, page ket253.

Oppong, R., Kaambwa, B., Nuttall, J., Hood, K., Smith, R. D., and Coast, J. (2013). The impact of using different tariffs to value eq-5d health state descriptions: an example from a study of acute cough/lower respiratory tract infections in seven countries. *The European journal of health economics*, 14(2):197–209.

Pan, S. M. D., Dehler, S., Ciurea, A., Ziswiler, H.-R., Gabay, C., and Finckh, A. (2009). Comparison of drug retention rates and causes of drug discontinuation between anti–tumor necrosis factor agents in rheumatoid arthritis. *Arthritis Care & Research*, 61(5):560–568.

Prevoo, M., Van't Hof, M., Kuper, H., Van Leeuwen, M., Van De Putte, L., and Van Riel, P. (1995). Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis & Rheumatology*, 38(1):44–48.

Ramiro, S., Sepriano, A., Chatzidionysiou, K., Nam, J. L., Smolen, J. S., van der Heijde, D., Dougados, M., van Vollenhoven, R., Bijlsma, J. W., Burmester, G. R., et al. (2017). Safety of synthetic and biological dmards: a systematic literature review informing the 2016 update of the eular recommendations for management of rheumatoid arthritis. *Annals of the rheumatic diseases*, pages annrheumdis–2016.

Sarzi-Puttini, P., Fiorini, T., Panni, B., Turiel, M., Cazzola, M., and Atzeni, F. (2002). Correlation of the score for subjective pain with physical disability, clinical and radiographic scores in recent onset rheumatoid arthritis. *BMC musculoskeletal disorders*, 3(1):18.

Singh, J. A., Saag, K. G., Bridges, S. L., Akl, E. A., Bannuru, R. R., Sullivan, M. C., Vaysbrot, E., McNaughton, C., Osani, M., Shmerling, R. H., et al. (2016). 2015 american college of rheumatology guideline for the treatment of rheumatoid arthritis. *Arthritis & rheumatology*, 68(1):1–26.

Singh, J. A., Wells, G. A., Christensen, R., Tanjong Ghogomu, E., Maxwell, L. J., MacDonald, J. K., Filippini, G., Skoetz, N., Francis, D. K., Lopes, L. C., et al. (2011). Adverse effects of biologics: a network meta-analysis and cochrane overview. *The Cochrane Library*.

Smolen, J., Breedveld, F., Schiff, M., Kalden, J., Emery, P., Eberl, G., Van Riel, P., and Tugwell, P. (2003). A simplified disease activity index for rheumatoid arthritis for use in clinical practice. *Rheumatology*, 42(2):244–257.

Souto, A., Maneiro, J. R., and Gómez-Reino, J. J. (2015). Rate of discontinuation and drug survival of biologic therapies in rheumatoid arthritis: a systematic review and meta-analysis of drug registries and health care databases. *Rheumatology*, 55(3):523–534.

Stephens, S., Botteman, M. F., Cifaldi, M. A., and van Hout, B. A. (2015). Modelling the cost-effectiveness of combination therapy for early, rapidly progressing rheumatoid arthritis by simulating the reversible and irreversible effects of the disease. *BMJ open*, 5(6):e006560.

Stevenson, M., Archer, R., Tosh, J., Simpson, E., Everson-Hock, E., Stevens, J., Hernandez-Alava, M., Paisley, S., Dickinson, K., Scott, D., et al. (2016). Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for the treatment of rheumatoid arthritis not previously treated with disease-modifying antirheumatic drugs and after the failure of conventional disease-modifying antirheumatic drugs only: systematic review and economic evaluation. *Health Technology Assessment*, 20(35):1–610.

Stevenson, M. D., Wailoo, A. J., Tosh, J. C., Hernandez-Alava, M., Gibson, L. A., Stevens, J. W., Archer, R. J., Simpson, E. L., Hock, E. S., Young, A., et al. (2017). The cost-effectiveness of sequences of biological disease-modifying antirheumatic drug treatment in england for patients with rheumatoid arthritis who can tolerate methotrexate. *The Journal of Rheumatology*, pages jrheum–160941.

Strand, V., Williams, S., Miller, P., Saunders, K., Grant, S., and Kremer, J. (2013). Op0064 discontinuation of biologic therapy in rheumatoid arthritis (ra): Analysis from the consortium of rheumatology researchers of north america (corrona) database. *Annals of the Rheumatic Diseases*, 72(Suppl 3):A71–A72.

Tosh, J., Brennan, A., Wailoo, A., and Bansback, N. (2011). The sheffield rheumatoid arthritis health economic model. *Rheumatology*, 50(suppl 4):iv26–iv31.

Wailoo, A., Brennan, A., Bansback, N., Nixon, R., Wolfe, F., and Michaud, K. (2006). Modeling the cost effectiveness of etanercept, adalimumab and anakinra compared to infliximab in the treatment of patients with rheumatoid arthritis in the medicare program. *Rockville, MD: Agency for Healthcare Research and Quality*.

Wailoo, A. J., Bansback, N., Brennan, A., Michaud, K., Nixon, R. M., and Wolfe, F. (2008). Biologic drugs for rheumatoid arthritis in the medicare program: a cost-effectiveness analysis. *Arthritis & Rheumatology*, 58(4):939–946.

Wolfe, F. and Michaud, K. (2010). The loss of health status in rheumatoid arthritis and the effect of biologic therapy: a longitudinal observational study. *Arthritis research & therapy*, 12(2):R35.

Wolfe, F., Michaud, K., Choi, H. K., and Williams, R. (2005). Household income and earnings losses among 6,396 persons with rheumatoid arthritis. *The Journal of rheumatology*, 32(10):1875–1883.

Wolfe, F., Michaud, K., Gefeller, O., and Choi, H. K. (2003). Predicting mortality in patients with rheumatoid arthritis. *Arthritis & Rheumatism*, 48(6):1530–1542.

Yazici, Y., Krasnokutsky, S., Barnes, J. P., Hines, P. L., Wang, J., and Rosenblatt, L. (2009). Changing patterns of tumor necrosis factor inhibitor use in 9074 patients with rheumatoid arthritis. *The Journal of rheumatology*, 36(5):907–913.

Zhang, J., Shan, Y., Reed, G., Kremer, J., Greenberg, J. D., Baumgartner, S., and Curtis, J. R. (2011). Thresholds in disease activity for switching biologics in rheumatoid arthritis patients: experience from a large us cohort. *Arthritis care & research*, 63(12):1672–1679.

---------- Forwarded message ----------
From: **Carole Wiedmeyer** <carolewied@me.com>
Date: Tue, Jul 11, 2017 at 1:15 PM
Subject: RA Model input
To: "Shafrin, Jason"
Cc: Suepattra May-Slater

Hi Jason,
Here is my feedback on the protocol. Overall, it's a very impressive piece of work - extremely thorough, and with a clear commitment to transparency and continuous improvement.

Answers to your questions are listed below. Please see the attached version of the document for specific feedback in tracked changes and comments, focusing on critiques and suggestions for improvement.

1a. Additional outcomes that would matter to patients (may not be possible to incorporate, but since you asked):

- Side effects minimized/avoided
- Permanent joint damage avoided

1b. Additional treatment attributes that would matter to patients:

- Cost to patient (out of pocket); this is a really huge issue once you start on biologics. Even with insurance the cost may be out of reach, especially for those on Medicare (since manufacturer co-pay programs are not allowed). If this is not accounted for, I think patients will be skeptical of the value of the model, potentially viewing it as a way to justify higher medication prices. The model accounts for health sector costs, but that seems more general and not geared to costs borne by patients.
- Time the medication has been on the market. More time means more is known about safety as well as effectiveness.
- As part of administration, patients may have preferences one way or the other for: oral vs needle; injection vs infusion; self-administered vs. clinical setting.

2. Methodology clarity:

- Not surprisingly, this is a very dense read for patients. I was able to follow most of it, but some of the sections on statistical methods are over my head. As you will see in my comments, I was confused in a number of spots. Therefore, somereaders might look for a review or stamp of approval of some sort from a

trusted source. I'm not sure what that would be, but perhaps once this is released you can get reviewers to publish their opinions about it.
- It was unclear to me how/if side effects can lead to switching in the model. In practice, side effects, even "less serious" ones, lead to switching.
- I would like to see model results for various scenarios, and ideally, compared to other models, with explanations about what accounts for differences. Perhaps a section showing output could be inserted, including some representation of uncertainty around the results. A sensitivity analysis on key input variables would be helpful as well. Without results, it is all very theoretical and a bit hard to wrap your head around.

3. Methodology critiques:

- My main concerns/critiques are: A) that AEs do not lead to switching (at least as far as I could tell; see comment on pg. 19), and B) the assumed low probability of switching (see comment on pg. 23).
- What is the internal process for model review to minimize errors? (Aside from public release.) Explaining this would enhance confidence in the methods.

4. Model release suggestions for ease of understanding:

- That depends on the audience. I think this document works well for econometricians, statistical researchers, and the like. For other less technical stakeholders, a video and/or webinar summarizing key features, what differentiates this from other model approaches, etc., would be helpful.

5. Model release promotion suggestions:

- Again, this depends on the audience. As far as patients go, I think buy-in from rheumatologists or rheumatology researchers would be helpful. Patients trust them to have their best interest in mind, especially when it comes to advocating for affordable access to therapies.
- I'd be happy to discuss options with you once I have a better understanding of your target audience(s) and planned methods for rollout.

Please let me know if you have any questions or need anything else.

Thanks for the opportunity to review.
Carole

IVI INNOVATION AND VALUE INITIATIVE

11100 Santa Monica Boulevard, Suite 500
Los Angeles, CA 90025

2 Bethesda Metro Center, Suite 850
Bethesda, MD 20814

Phone: 310 984 7793
Fax: 310 982 6311

Protocol

IVI026 An Open Source Consensus-Based Family of Cost-Effectiveness Models for Rheumatoid Arthritis

Version 2

Administrative structure

Innovation and Value Initiative	Darius Lakdawalla PhD; Executive Director Jason Shafrin PhD; Director of Research 11100 Santa Monica Boulevard, Suite 500 Los Angeles, CA 90025 Phone: 310 984 7793 info@thevalueinitiative.org
Project Lead	Devin Incerti PhD; Research Economist devin.incerti@precisionhealtheconomics.com
Study Team	Devin Incerti, PhD Jeroen P Jansen, PhD Jeffrey Curtis, MD MS MPH

Version History

Version	Changes	Date issued
1	--	May 12, 2017
2	Updates to model structure	July 6, 2017

Table of Contents

List of Tables

List of Figures

Abbreviations

95%CI	95% confidence interval
ACR	American College of Rheumatology
AE	Adverse event
bDMARDs	Biologic DMARDs
BSRBR	British Society for Rheumatology Biologics Registers
cDMARD	Conventional DMARD
EQ-5D	EuroQol-5 Dimension Health Questionnaire
EULAR	European League Against Rheumatism
EVPI	Expected value of perfect information
EVPPI	Expected value of partial perfect information
HAQ	Health Assessment Questionnaire
ICER	Incremental cost-effectiveness ratio
IPS	Individual patient simulation
JAK	Janus kinase
NDB	National Data Bank for Rheumatic Diseases
NICE	National Institute for Health and Care Excellence
NMA	Network meta-analysis
NMB	Net-monetary benefit
PSA	Probabilistic sensitivity analysis
QALYs	Quality adjusted life years
RA	Rheumatoid arthritis
RCT	Randomized controlled trial
TNF	Tumor necrosis factor
WAC	Wholesale acquisition cost
WTP	Willingness to pay

1. Introduction

Value assessments of medical technology vary in modeling structure, perspective, evidence analyzed, and populations considered. Different organizations provide different estimates of value, without clearly explaining to marketplace stakeholders the scientific or contextual reasons for the variation. The result is often confusion and controversy that impedes efforts to reimburse based on value, particularly in a decentralized decision-making environment. Furthermore, when new evidence arises, it is very cumbersome, if not impossible, for someone other than the original model developer to update a model and produce new value estimates. What is needed is a strategy for gaining consensus when uncertainty resolves, identifying the reasons why value estimates vary when consensus does not resolve, and easily updating value assessments when new information becomes available. To fill this need, IVI will create flexible disease-specific open source decision analytic models capable of estimating value from multiple perspectives and quantifying all sources of uncertainty (i.e., structural uncertainty, parameter uncertainty) that will be updated over time in a consensus-driven manner as new evidence arises. Importantly, IVI will not develop a single model structure but will build multiple model structures that reflect the range of scientifically defensible approaches. These models will be developed under IVI's Open Source Value Project (OSVP).

IVI's first OSVP model—the IVI-RA model—will be developed for the purpose of assessing the cost-effectiveness of treatments for rheumatoid arthritis (RA). RA is a disease area particularly well suited to IVI's approach. For one, modeling assumptions vary considerably across existing decision-analytic models.[1-7] Cost-effectiveness analyses (CEAs) have been applied to different countries using different perspectives, populations, and model structures and have consequently reached different conclusions about the cost-effectiveness of therapies for RA.[1-7] Second, since the evidence base in RA is rapidly evolving, there is a strong need for RA models that will continue to be updated as the evidence base evolves. Third, there have been significant advancements in the treatment of RA over the past decade, which suggests that there is an increasing need for decision-analytic tools to assess the cost-effectiveness of these therapies.

This document describes how IVI will apply its open source consensus-based approach to RA. Section 2 describes the primary intent of the open-source RA family of models. It explains why it is important to consider different perspectives and methodologies and why an open-source

consensus-driven approach is needed. Section 3 defines the concepts of perspective, heterogeneity, and uncertainty, and explains how IVI will make clear distinctions between them when developing its models. Section 4 describes the treatment strategies, competing model structures, data sources and methods use to estimate model parameters, and the simulation techniques used to implement the model. Finally, Section 5 outlines how IVI will make the model open source and how it will incorporate feedback in a consensus-driven manner.

2. Objectives

The aims of the development of the IVI-RA model are fourfold. First, the model should be able to properly account for differences in perspectives, populations considered, and uncertainty (see Section 0 for further discussion). The model should be capable of estimating value according to the perspective of the decision maker and the preferences and characteristics of a given patient population. Uncertainty must also be considered to quantify the probability of making incorrect decisions based on imperfect knowledge. The second objective is transparency, which is especially important given that models often reach different conclusions without clear explanations from modelers about why this is the case. IVI will therefore make the model publicly available by releasing code, documentation, and interactive user interfaces as described in Section 5.1. Third, the model should be dynamic and based on consensus where possible. The model will therefore be updated over time as new evidence arises based on feedback from the general public and the consensus of experts. The process is described in Section **Error! Reference source not found.**~~5.2~~. Fourth, the model should be designed so that can be used for purposes other than CEA if desired. For example, the model could be used for a patient tool that predicts the distribution of possible future outcomes conditional on characteristics and preferences unique to the patient.

Commented [Office1]: As a reader, I'm wondering what kind of outcomes we are talking about here. Health outcomes? Value outcomes? I feel like I need a basic overview of the model at this point (what goes in, what comes out), so I know what we are talking about here.

3. Importance of perspective, heterogeneity, and uncertainty

There are a number of factors that influence value estimates. First, value depends on the *perspective* of the decision maker, which is why the second panel on cost-effectiveness in health and medicine suggests that cost-effectiveness studies should report results from both a societal and a health sector perspective.[8] However, even given a particular perspective, there is considerable variation in value across patients and significant uncertainty in those estimates.

Variation across patients can be separated into two distinct concepts, known as *heterogeneity* and *variability*.[9] Heterogeneity reflects true differences in value across patients that can, in part, be explained, while variability (often referred to as first order uncertainty), reflects differences in outcomes across patients due to random variation.

Uncertainty should not be confused with heterogeneity or variability. *Parameter uncertainty* (often referred to as second order uncertainty) occurs because the model parameters are uncertain. *Structural uncertainty* concerns the structural assumptions underlying a model.[9] Parameter and structural uncertainty imply decision uncertainty, or the degree to which decisions are made based on imperfect knowledge.

The IVI-RA model will follow the recommendations from the second panel on cost-effectiveness in health and medicine and include options so that value can be estimated from either a health sector or societal perspective. Heterogeneity will be modeled by allowing model outcomes and treatment response to depend on the characteristics and preferences of patients. Value estimates will average over heterogeneous patients within a given population of interest and subgroup analyses will be used to analyze differences in value across patients. Since the primary outcome of interest is mean cost-effectiveness, we will attempt to eliminate variability from the analysis by simulating a sufficient number of patients within a given patient population. Parameter uncertainty will be quantified using probabilistic sensitivity analysis (PSA) as described in Section 4.4.1 and structural uncertainty will be quantified by averaging over the competing model structures within the family of models as discussed in Section 4.4.2.

4. The IVI-RA Model

4.1. Treatment strategies

Since patients typically use multiple treatments over a lifetime, the model is capable of simulating a treatment sequence of any arbitrary length. Treatments that can be included in a sequence include conventional disease-modifying anti-rheumatic drugs (cDMARDs) such as methotrexate as well as the following biologic DMARDS (bDMARDs):

- **Tumor necrosis factor (TNF) inhibitors**: etanercept, adalimumab, infliximab, certolizumab, golimumab
- **Nn~~n~~on-TNF inhibitors**: abatecept, tocilizumab, rituximab
- Inhibitors of Janus kinase/signal transducers and activators of transcription (JAK/STAT) inhibitors: tofacitinib

At the end of a sequence, patient switch to non-biologic therapy (NBT), which encompasses a range of therapies that do not affect the rate of disease progression and are not associated with adverse events.

Commented [Office2]: Should be abatacept

Commented [Office3]: I can't understand this sentence. Doesn't seem to be a complete sentence.

4.2. Competing Model Structures

The IVI-RA model is a discrete-time individual patient simulation (IPS) with 6 month cycles that can be run using a number of different model structures. Like most RA cost-effectiveness models, the model measures changes in disease severity using the Health Assessment Questionnaire (HAQ) Disability Index score. In particular, at the start of the simulation, each patient is assigned a baseline HAQ score. Subsequently, the impact of the disease measured by the HAQ trajectory over time is modeled as a function of a sequence of treatments (). In the absence of treatment, HAQ deteriorates at a certain rate as depicted by the dashed line in the figure. Treatment is separated into two distinct phases: an initial phase of up to 6 months, consistent with data reported from randomized controlled trials (RCTs), and a maintenance phase thereafter until discontinuation.

Commented [Office4]: Not sure why this is here; perhaps a reference is missing.

Commented [Office5]: What figure?

During the initial treatment phase HAQ is modeled as a change from baseline. Three possible model structures labeled **H1-H3** are possible. In **H1**, treatment influences HAQ through its effect on the American College of Rheumatology (ACR) response criteria, which is similar to the

structure used in other US based cost-effectiveness models.[6, 10] ACR response is measured using four mutually exclusive categories: no response (defined as less than 20% improvement), ACR 20-50% improvement, ACR 50-70% improvement, and ACR 70% improvement or greater. The rationale for using ACR response rather than HAQ directly is that the evidence base relating treatment to ACR response is larger than the evidence based relating treatment to HAQ. **H2** follows the National Institute for Health and Care Excellence (NICE) cost-effectiveness model [5, 7] and models the effect of treatment on HAQ indirectly through its effect on ACR response and, in turn, the three categories of the European League Against Rheumatism (EULAR) response (no response, moderate response, or good response). Finally, since modeling the effect of treatment on HAQ through intermediary variables may mediate treatment response, in **H3**, treatment impacts HAQ directly. The three scenarios are summarized below:

- **H1**: Treatment → ACR → HAQ
- **H2**: Treatment → ACR → EULAR → HAQ
- **H3**: Treatment → HAQ

The probability of switching treatment during the initial treatment phase is modeled using 6 possible structures labeled **S1-S6**. **S1** follows a common approach where ACR non-responders discontinue treatment (e.g. Carlson et al. 2015; Institute for Clinical and Economic Review 2017). One drawback of this approach is that it is not consistent with current treat-to-target guidelines in the United States.[11] **S2** and **S3** consequently model treatment switching as a function of disease activity (remission, low, moderate, high).[12] ACR response predicts the change in disease activity from baseline, which, along with baseline disease activity, predicts absolute disease activity. The probability of switching treatment is increasing in the severity of disease (i.e., the probability is lowest in remission and greatest with high disease activity). **S2-S4** measure disease activity using the Disease Activity Score with 28-joint counts (DAS28) [13], Simplified Disease Activity Index (SDAI) [14, 15] and Clinical Disease Activity Index (CDAI) [16] respectively.

S5 is similar to **S2-S4**, but models the effect of treatment on changes in DAS28 directly, rather than indirectly through ACR response. We also aimed to model the direct effect of treatment on SDAI and CDAI, but sufficient clinical trial data is not available. Finally, since in the UK, the British Society for Rheumatology and the British Health Professionals in Rheumatology recommends using the EULAR response [17], treatment switching in **S6** is a depends on

Commented [Office6]: These labels are not mutually exclusive because 50% and 70% are found in more than one category.
Note:
- ACR20 is ≥ 20% improvement
- ACR50 is ≥ 50% improvement
  - ACR50 responders include ACR20 responders
- ACR70 is ≥ 70% improvement
  - ACR70 responders include ACR20 & ACR50 responders

So, suggest rewording your text to say:
ACR 20-49%, ACR 50-69%, and ACR 70+%

Commented [Office7]: This sentence is a bit challenging to understand so shortening it might help. Alternate wording suggestion: "The H1 model structure uses ACR response rather than HAQ because the evidence base is larger for ACR response."

Commented [Office8]: I think this should be "base" (not "based")

Commented [Office9]: Not sure exactly what is meant here by "mediate" treatment response. Does it mean some sort of bias might be introduced?

Commented [Office10]: It might have helped me to see this summary first, before reading the detail above, perhaps with an introductory sentence that explains H1 and H2 have intermediary variables to get to HAQ, and H3 just goes to HAQ. Also suggest putting these bullets in table form.

However, I'm still scratching my head a bit about some things:

...

Commented [Office11]: As a reader, I'm left wondering the reasons why these different structures ...

Commented [Office12]: Hmmm. There are a lot of other scenarios that would lead to discontinuation of treatment. Less than ACR50 response might warrant ...

Commented [Office13]: What "it" are you referring to here? Do you mean discontinuing treatment without switching to something else?

Commented [Office14]: I don't see the difference between S2 and S3…?

Commented [Office15]: Is this EULAR?

Commented [Office16]: None of these measures account for joints in the feet or ankles. See: ...

Commented [Office17]: Not sure if you consider data plural, or not. If so, say "data are".

Commented [Office18]: Change to recommend (plural subject)

Commented [Office19]: Delete "is a"…?

EULAR response. In particular, following the NICE model [5], we assume that EULAR non-responders discontinue treatment while moderate and good responders continue treatment. The reasoning is that rules stipulated by NICE require a DAS28 improvement of more than 1.2 to continue treatment which is associated with moderate or good EULAR response. The 6 treatment switching scenarios are summarized below:

- **S1:** Treatment → ACR → Switch
- **S2:** Treatment → ACR → ΔDAS28 → DAS28 → Switch
- **S3:** Treatment → ACR → ΔSDAI → SDAI → Switch
- **S4:** Treatment → ACR → ΔCDAI → CDAI → Switch
- **S5:** Treatment → ΔDAS28 → DAS28 → Switch
- **S6:** Treatment → ACR → EULAR → Switch

Commented [Office20]: Again, it would have helped me to see this first, or at least have a reference to this at the beginning of the discussion. Suggest changing these bullets to a table format.

Not all model structures **S1-S6** can be used with each of **H1-H3**. If **H1** is used, then **S1-S5** are available, but **S6** is not because EULAR response is not simulated. In **H2**, **S1-S6** are all available while in **H3** only **S5** can be used since ACR response is not simulated. The 12 possible model structures and the number of each structure are outlined in Table 1.

Table 1: Model structures for initial treatment phase

	S1	S2	S3	S4	S5	S6
H1	1	2	3	4	5	-
H2	6	7	8	9	10	11
H3	-	-	-	-	12	-

Notes: Rows denote the model structure used to relate treatment to HAQ and columns denote the model structure used to predict treatment switching. Each number denotes a unique model structure (i.e. 1 corresponds to H1 and S1 and 8 corresponds to H2 and S3) and the "-" denotes a model structure combination that is not possible. There are 12 possible model structures for the initial treatment phase.

In the maintenance phase, two model structures can be used to simulate the long-term progression of HAQ. First, as is common in CEAs of therapies for RA, HAQ is assumed to progress at a constant linear rate over time.[1, 2] However, since emerging evidence suggests that the rate of HAQ progression is non-linear [18], our second scenario simulates HAQ progression using a non-linear mixture model [19] with 4 distinct HAQ trajectories and a rate of HAQ progression that decreases over time within each trajectory. Upon discontinuation of

treatment, the HAQ score rebounds by a proportion of the improvement experienced at the end of the initial 6-month period with that treatment.

The duration of the maintenance phase (i.e., time to discontinuation of maintenance treatment) is simulated using parametric time-to-event distributions. When structure **S6** is used, the time-to-event distributions are stratified by EULAR response category. Patients with good response at the end of the initial treatment phase stay on treatment longer, on average, than patients with a moderate response. In contrast, when **S1** is used, time to treatment discontinuation is simulated using a single time-to-event curve because we have been unable to obtain curves stratified by ACR response categories. Likewise, when **S2-S5** are selected, we use a single time-to-event curve because we have not obtained curves stratified by disease activity level. In each case, time to discontinuation can be simulated using one of 7 possible distributions (exponential, Weibull, Gompertz, normal, gamma, log-logistic, generalized gamma).

In line with the NICE model [5] the adverse events included in the model are limited to serious infections; we assume that only serious infections have a significant cost impact and increased risk over background rates to be meaningful to include.[20] While on a treatment, a patient experiences a serious infection if the individual's sampled time to the adverse event is shorter than the sampled time to treatment discontinuation.

Baseline HAQ scores (and changes in HAQ scores from baseline) are used to determine mortality relative to age/sex specific rates for the US general population (assumed to have a HAQ score of 0). Treatment therefore has an indirect effect on mortality through its effect on HAQ.

Individual HAQ scores at a particular point in time were also used to simulate EQ-5D utility scores (0-1 range), which, in turn, were used to simulate quality-adjusted life-years (QALYs). However, since a number of different methods have been used to convert HAQ into utility, our model contains two different possible mapping algorithms. Our preferred algorithm is Hernandez Alava et al. mixture model [21], which uses a much larger sample size than other statistical models and has been shown to have better predictive accuracy. Other algorithms are typically estimated using clinical trial data [4, 10] and consequently have limited generalizability. The second utility algorithm available within our model is based on a linear regression analysis of real-world data [22] that has been used in a few previous CEAs.[1, 6]

Commented [Office21]: So a higher HAQ score (assume this is HAQ disability) goes up when treatment is discontinued, which in layman's terms, means you get worse when you stop taking the meds. Is that right?

Commented [Office22]: Not familiar with this method – can you briefly describe?

Commented [Office23]: I'm trying to figure out if an adverse event (AE) automatically means you switch; this does not say.

Commented [Office24]: Suggest you define "serious" infections here. Does it mean infections requiring IV antibiotics? Hospitalization?

As stated previously, other side effects (less than serious) can and do result in discontinuation of medication/switching. And because they are more common than serious side effects, switching is more likely to occur in the general patient population for these "less serious" reasons.

Note that less serious infections could also lead to switching if they become chronic. For example, repeated UTI or fungal infections, repeated bouts of shingles.

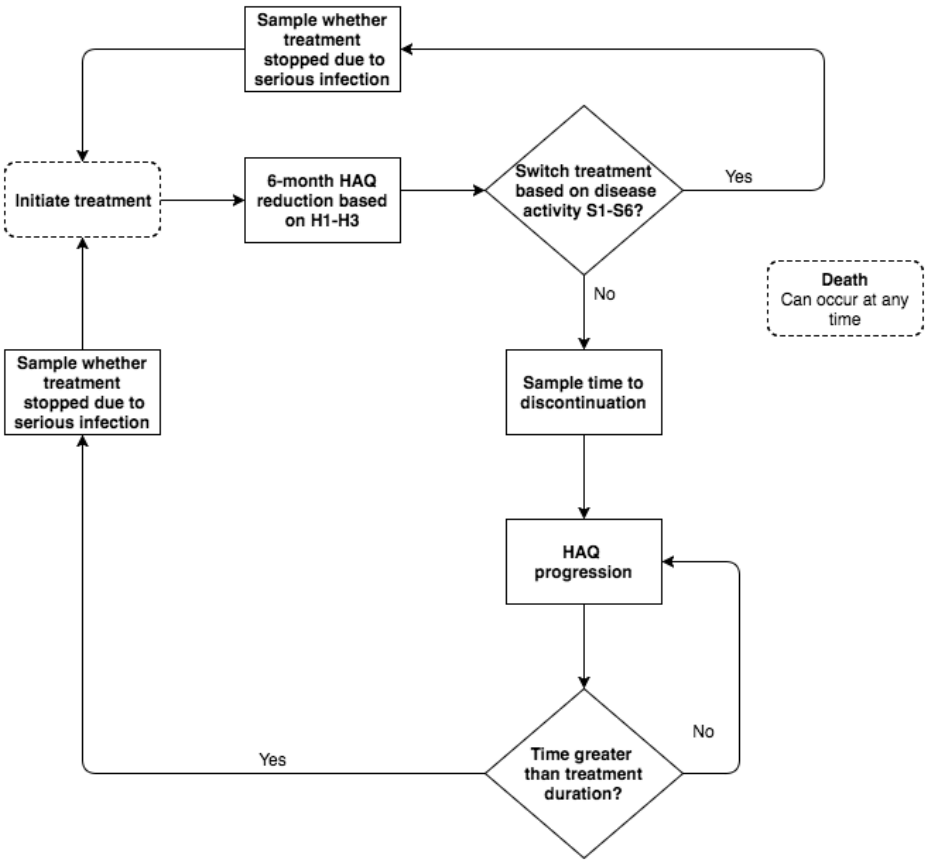
Other serious side effects (other than infections) also can result in switching – for example: eye problems, stomach or intestinal tear, lupus-like syndrome.. These are rare, but there is no doubt that if you experience these, you will switch.

Commented [Office25]: I don't really understand this last sentence. Part of my confusion is because I don't really know how you are using the term "sampled" here. Aren't all times in the model "sampled"?

Annual hospitalization days and productivity losses are simulated as a function of HAQ. Health sector costs considered in the models are related to drug acquisition and administration, adverse events, general management of RA, and hospitalization. Non-health sector costs considered are limited to work related productivity loss.

Patient preferences for treatment attributes have a direct effect on utility as patients with treatments that more closely match their preferences have higher utility while holding treatment efficacy and safety constant. Treatment attributes that are incorporated into the models include route of administration and frequency of administration.

The flow diagram in



Commented [Office26]: No other treatment costs included? What about physical therapy? Changes in opioid prescribing guidelines call for more non-pharmacological interventions (PT, acupuncture, etc.), but these are not accounted for in the model (and they tend to be more costly to patients than medications because they are not well covered by insurance).

Commented [Office27]: Are those the only treatment attributes in the models? Sentence implies there are others. If so, what are they? This will be of interest to patients. The one big one that comes to mind is preference for lower-cost options.

Figure 1~~Figure 1~~ describes the flow of a single patient through the simulation. Each patient simulation by initiating treatment and ends the simulation with death. The rectangles in the figure represent "processes" determining the effect of treatment on disease progression and the diamonds represent "decisions" that determine whether a patient will switch to a new treatment.

The influence diagram in Figure 2 summarizes the assumed structural relationships among different variables in the model. Each arrow represents the direct effect of one parameter on another. Dashed lines represent relationships that depend on the structural assumptions used. Panel (a) focuses on the effect of treatment on disease progression and adverse events while Panel (b) looks at the variables influencing the primary health and cost outcomes.

Model outcomes depend on patient characteristics, which have a direct effect on HAQ progression, mortality, and utility. The primary health outcome is the quality-adjusted life-year (QALY), which depends on mortality and utility. Total costs consist of health care sector costs and productivity losses. The components of health sector costs include drug acquisition and administration costs, general management and monitoring costs, adverse event costs, and hospitalization costs Analyses from a societal perspective would include productivity losses while analyses from a health care sector perspective would not. The value of treatment is estimated using the net-monetary benefit (NMB), which is calculated by multiplying QALYs by a willingness to pay threshold and subtracting costs ($NMB = QALYs \cdot WTP - Costs$).

Commented [Office28]: Ugh. This sentence stopped me cold. There must be some better way to say this. Sounds like all treatments end in death! Is there no other way to end the simulation? Or does this imply that the simulation runs for the (modeled) patient's entire lifespan because the treatment must be continued for the entire patient's life since RA is a chronic condition?

Commented [Office29]: Suggest placing Figure 1 here, before the description of Figure 2. It is confusing to read text about Figure 2 and then see Figure 1.

Commented [Office30]: Why is "structural" needed here? I'm not sure what other kind of relationship there would be between the variables.

Commented [Office31]: This seems a bit buried in the discussion. Suggest putting this at the beginning of the discussion of Figure 1.

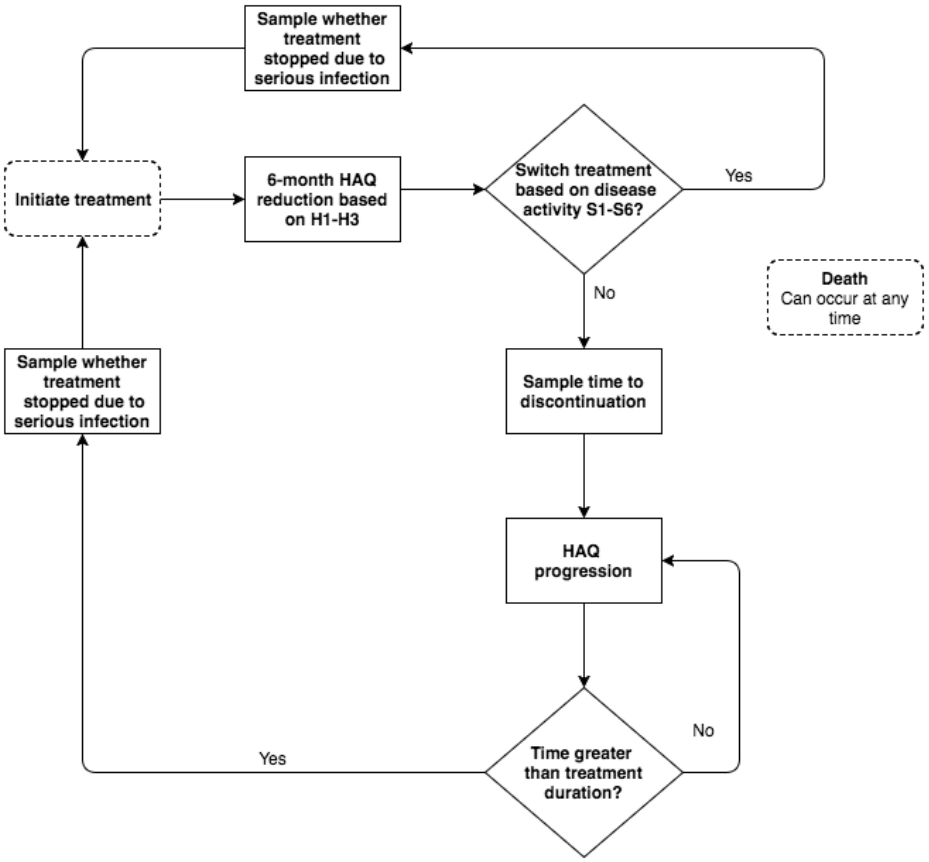
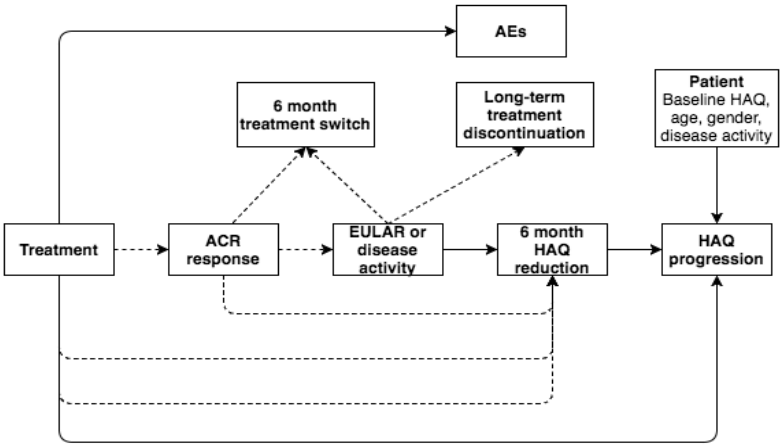


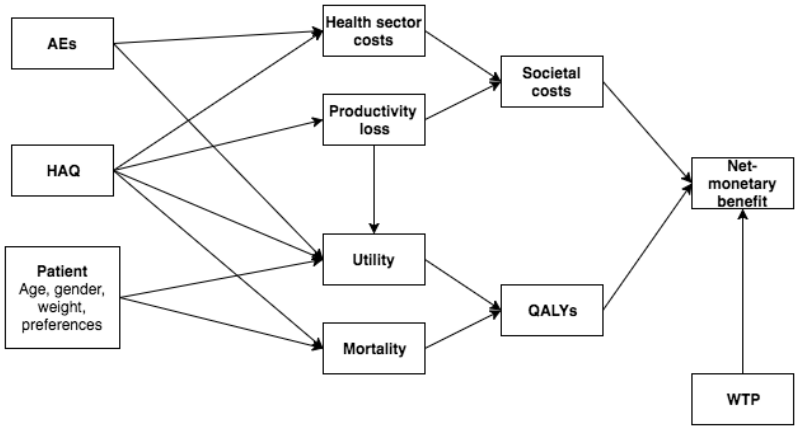
Figure 1: Flow diagram of the simulation for a single patient

Commented [Office32]: Suggest adding a key to explain the meaning of the different shapes (as described in narrative above)

I'm still confused about how "Time greater than treatment duration" means serious infection. I don't really even understand what "Time" is referred to here. Is it "Time to discontinuation"?



(a) Treatment effects



(b) Model outcomes

Figure 2: Influence diagram outlining structural relationships

Notes: ACR: American College of Rheumatology; EULAR: European League Against Rheumatism; HAQ: Health Assessment Questionnaire; QALY: AEs: adverse events; QALYs: quality-adjusted life-years; WTP: willingness to pay. Disease activity refers to the Disease Activity Score with 28-joint counts (DAS28) or the Simplified Disease Activity Index (SDAI).

Commented [Office33]: I'm not sure what "Long-term treatment discontinuation" means. Discontinuation of that specific treatment, or of any treatment? At any time, or after 6 months?

It appears that AEs do not trigger switching. Why? They do in real life. Especially given the model only focuses on serious infections, I would think they would definitely result in switching treatments. This is a pretty big concern with the model for me at this point.

Commented [Office34]: What about cost to patient? Co-pays, deductibles, etc.? Is that included under patient preferences? Or is that WTP? If it is WTP, why is it separate, and not connected to the patient? This model makes it look like patients don't really pay, but "society" does.

Commented [Office35]: Delete this extra reference to QALY

Table 2: Competing model structures

Commented [Office36]: Not sure why this is here?

Component of model structure	Possible combinations
Initial effect of treatment on HAQ (**H1**-**H3**) and switching (**S1**-**S6**)	12
HAQ progression linear or non-linear	2
Probability distribution for treatment duration	7
Utility algorithm	2

4.3. Source Data and Parameter Estimation

4.3.1. Comparative treatment efficacy from the network meta analysis

The effects of treatment on ACR response, DAS28, and HAQ at 6 months are estimated using a using a Bayesian network meta-analysis (NMA) of published randomized controlled trials (RCTs).

A systematic literature search was performed to identify published randomized controlled trials (RCTs) comparing the bDMARDs of interest regarding ACR20/50/70 response estimates, the change in HAQ from baseline, and the change in DAS28 from baseline at 6 month follow-up, among a bDMARD-naïve cDMARD-IR population (See **Appendix B**).

Treatment response is estimated using Bayesian network meta-analysis (NMA) techniques. The ACR20/50/70 response probabilities are modeled using the probit model for ordered categorical data described in Dias *et al [23]* while changes in HAQ and DAS28 are modeled using normal linear models. *All* competing interventions equally relevant or indicated for the target population of interest are included in the same network, which implies that both monotherapy and treatment in combination with MTX are included in the network but with separate "nodes". As such, we assume that relative treatment effects differ in the presence or absence of MTX background therapy. Consistency between direct and indirect estimates in the network will be evaluated and adjustments for between-trial differences regarding effect-modifiers or baseline risk will be made if needed. When modeling sequential bDMARD treatments, the relative treatment effects obtained with the NMA for bDMARD naïve patients no longer apply to the modeled population after they failed their first bDMARD. In principle, a NMA of the bDMARD-IR population needs to be performed. However, the bDMARD-IR trials differ regarding the distribution of bDMARD treatment history. The limited number of studies in such a network does not allow us to statistically adjust the reported estimates to match the patients simulated in our model characterized by an increasing number of bDMARDs received as the model progresses. As an alternative, treatment effects among bDMARD inadequate responders are obtained by applying percent reductions to the parameter estimates so that treatment effects decline at later lines of therapy. For example, we reduce the mean of the reference treatment (cDMARDs) and the ACR50/70 cut-points from the NMA among bDMARD naïve patients so that an ACR response of 60/40/20 would approximately be reduced to 50/30/10 among bDMARD experienced patients [24].

Commented [Office37]: Delete repeated phrase "using a"

Commented [Office38]: Need to say this means "inadequate response"

Commented [Office39]: Not sure why they have to be called "competing"? Because you only do them one at a time?

Commented [Office40]: Good; glad to see the model accounts for MTX use/non-use. Though I'm not sure exactly what the use of the term "network" means here.

Commented [Office41]: A bit of a hard sentence to follow.

Commented [Office42]: Why? Based on what? I realize you have to do something in the absence of studies on multiple biologic use, but I'm not sure how to justify reducing the standard of effectiveness for subsequent biologics.

Commented [Office43]: Do you mean 70/50/20?

4.3.2. Treatment switching at 6 months

4.3.2.1. ACR response and change in disease activity

There are currently no established mappings between mutually exclusive ACR response categories and DAS28, SDAI, or CDAI.[25] However, Aletaha and Smolen [15] provides evidence on the relationship between overlapping ACR response categories (ACR 20/50/70) and mean changes in each of the three disease activity measures. Results are reported for three cohorts—the Leflunomide datasets, the inception cohort, and the routine cohort—with 1,839, 91, and 279 patients respectively. We transformed mean changes by overlapping ACR response categories to mean changes by mutually exclusive ACR response categories (ACR < 20, ACR 20-50, ACR 50-70, ACR 70+) by using the number of patients in each mutually exclusive ACR response category. Smolen et al. [14] provided the number of patients in each ACR response category in the Leflunomide dataset and Aletaha et al. [16] provided the number of patients in the inception cohort.

We did not include estimates from the routine cohort for two reasons. First, we were unable to find information on the number of patients in each ACR response category. Second, patients in the routine cohort had considerably lower disease activity levels [16, 26] and our main population of interest consists of patients with high disease activity at baseline. Mean DAS28 in the inception cohort and routine cohort were 5.62, and 4.09 respectively, while the mean DAS 28 ranged from 6.3 to 7 across the clinical trials making up the Leflunomide dataset.

4.3.2.2. ACR response and change in EULAR response

ACR responses were translated into EULAR response probabilities based on evidence of their relationship reported in Stevenson et al. [5] and obtained from the US Veterans Affairs Rheumatoid Arthritis (VARA) registry.

4.3.2.3. Probability of switching treatment

The probability of stopping treatment depends on the model structure used. If **S1** or **S6** is used, then patients stop treatment when there is no ACR response or no EULAR response, and continue treatment otherwise. If switching depends on SDAI, CDAI, or DAS28, then we model the probability of treatment switching and adjust it according to the level of disease activity.

The probability of treatment switching is estimated using a logistic regression equation. Differences in discontinuation by disease activity level are based on the odds ratios from an

Commented [Office44]: Change provides to provide (plural subject)

Commented [Office45]: Use one style – 70/50/20 or 20/50/70 throughout

Commented [Office46]: Again, these are not mutually exclusive categories (50 and 70 appear twice each)

Commented [Office47]: A bit confused by this sentence. Seems like you are talking about how you mapped from DAS28/SDAI/CDAI to ACR. But here is sounds like you are mapping from ACR to ACR.

Commented [Office48]: This is the first reference to this. Why are you limiting this to patient with high disease activity? Current treatment guidelines call for early aggressive intervention, which would imply at least moderate disease activity would be of interest to this model (and to patients if not other stakeholders).

Commented [Office49]: Again, seems like a very low bar. Why?

analysis of the Consortium of Rheumatology Researchers of North America (CORRONA) database.[27] Odds ratios were estimated for three disease activity categories based on CDAI (<10 (low), 10-22 (moderate), and ≥ 22 (high)). We set the intercept in our logistic regression equation so that the probability of treatment switching for a patient with moderate disease activity is 16.3%. This assumption is based on an estimate of the probability of treatment discontinuation at 6 months from a Kaplan-Meier curve estimated using the Consortium of Rheumatology Researchers of North America (CORRONA) database—a population that tended to have moderate disease activity (mean CDAI = 16.0)—and is consistent with ranges from the literature.[28] The standard error of the intercept was set so that the probability of switching with moderate disease activity varied by ± 3%.

In the simulation, we use the absolute level of disease activity at 6 months (not at baseline) to make predictions. We define low, moderate, and high disease activity when using DAS28 and SDAI as in Anderson et al.[12] The predicted probabilities of switching at 6 months for low, medium and high disease activity patients are 9.1%, 16.3%, and 25.4% respectively.

Commented [Office50]: OK, here I see that moderate disease activity does trigger switching.

Commented [Office51]: All of these numbers seem low to me. Why would only 25.4% of those with high disease activity not switch treatments? I would think it would easily be quite a bit more than half. I realize you used CORRONA, but there must be something weird in their data. Maybe they are missing information about switching. I think this is one of my biggest concerns about this model at this point. Why would they continue with something that doesn't work? Were they waiting to see if it worked? If so, then the time horizon may need to be adjusted beyond 6 months. Anyone seeking treatment with options left will certainly try something else in this situation, as long as there are funds/insurance to cover it.

4.3.3. Change in HAQ at 6 months

ACR responses from the NMA were translated into HAQ scores based on evidence from the ADACTA trial.[6, 10] No ACR response, ACR 20-50, ACR 50-70, and ACR 70 are associated mean drops in HAQ of 0.11 (SE = 0.067), 0.44 (SE = 0.056), 0.76 (SE = 0.091), and 1.07 (SE = 0.075) respectively.

Commented [Office52]: Same issue as before with overlapping labels.

The relationship between EULAR response and HAQ is based on analyses conducted by Stevenson et al. using the BSRBR database.[5] Their analysis is based on predictions from a mixture model with covariates set to sample means. Moderate and good EULAR responses are associated with -0.317 (SE = 0.048) and -0.672 (SE = 0.112) changes in HAQ scores respectively.

Commented [Office53]: Would be nice to have a table – easier to compare than going back and forth in the text.

4.3.4. HAQ progression in the absence of bDMARD treatment

4.3.4.1. Constant linear rate of progression

The rate of progression in the linear case is based on the observational study by Wolfe and Michaud (2010).[29] They assessed the development of HAQ over time at six month intervals for up to 11 years among 3,829 RA patients who switched from non-biologic treatment to biologic treatment and participated in the National Data Bank for Rheumatic Diseases (NDB)

longitudinal study of RA outcomes. The annual HAQ progression rate prior to biologic therapy was 0.031 (95% confidence interval (95%CI): 0.026 to 0.036) and is assumed to reflect the course of progression of HAQ in the absence of bDMARD.

Based on the same data, Michaud et al. [30] reported overall and age-specific specific HAQ progression rates. The differences between the overall and age specific rates are as follows: <40: -0.020 (95%CI: -0.0223 to -0.0177); 40-64: -0.008 (95%CI: -0.0101 to -0.0059); ≥ 65 0.017 (95%CI: 0.0136 to 0.0204). These estimates are applied to the overall progression rate of 0.031 to obtain age specific HAQ progression rates.

4.3.4.2. Non-linear mixture model

The rate of progression in the non-linear case is based on a mixture model approach that has increasingly been used to model HAQ progression over time.[5, 19, 31] These models suggest that different subgroups have distinct HAQ trajectories and that the rate of worsening of HAQ progression decreases over time. Parameter estimates are based on the 4-class latent class growth model (LCGM) used in Norton *et al [19],* which validated results from an earlier study [31]. We use the statistical model estimated on the Early Rheumatoid Arthritis Cohort Study (ERAS) cohort, which has a high percentage of patients receiving methotrexate and a very small percentage receiving biologics. Following Stevenson *et al* [5], explanatory variables in the statistical model that are not used in the IPS are set to their mean values in the ERAS cohort.

4.3.5. HAQ trajectory with bDMARD maintenance treatment

Based on an NDB longitudinal study, Wolfe and Michaud [29] estimated the overall annual HAQ progression rate among RA patients who had switched to biologic treatment at -0.001 (95CI: -0.004 to 0.002). In a separate analysis, also based on NDB data, Michaud et al. [30] reported annual HAQ progression rates by treatment adjusted for baseline HAQ score, age, sex, education, smoking, BMI, comorbidity, and RA onset. The average HAQ rate among patients on a biologic was −0.001 as well, which instills confidence that the reported HAQ progression rates for different bDMARDs as reported by Michaud et al [30] can be directly compared with the overall annual HAQ progression rate of 0.031 reported by Wolfe and Michaud.[29] Accordingly, bDMARD specific HAQ progression rates by Michaud et al.[30] are be used in the model alongside the HAQ progression rates in the absence of bDMARD treatment by Wolfe and Michaud.[29] For bDMARD treatments evaluated in the model for which no HAQ progression rate was reported by Michaud et al [30], the overall biologic rate of -0.001 will be used.

Commented [Office54]: I seem to be missing something here. I thought I was heading into a discussion of subgroups (what characteristics do the four different classes have, and how do they affect HAQ progression)? But then we go into a discussion of a statistical model based on the ERAS cohort, and I don't get the connection.

Commented [Office55]: Why include the variables at all if they don't vary in the model?

Commented [Office56]: What about onset? Time since onset to treatment?

Commented [Office57]: Do you mean HAQ progression rate?

Commented [Office58]: Do you mean "can be"?

4.3.6. Duration of maintenance treatment

4.3.6.1. CORRONA database

Time to treatment discontinuation for patients on maintenance treatment for model structures **S1-S5** is based on analyses from the CORRONA database (Strand et al. 2013). The analysis sample consisted of 6,209 patients age 18 or older treated between 2002 and 2011 receiving either TNF inhibitors or other bDMARDs. The mean age was 57.6 years, 43% of patients were biologic naive, the mean CDAI was 16, and just over 26% of patients had high disease activity (CDAI ≥ 22).

7 parametric survival models (exponential, Weibull, Gompertz, gamma, log-logistic, lognormal, and generalized gamma) were estimated on individual patient data reconstructed from a Kaplan-Meier curve from the CORRONA analysis using the algorithm developed in Guyot et al.[32] We compared fit using the Akaike Information Criteria (AIK) and Bayesian Information Criteria (BIC). The generalized gamma had the lowest AIC and BIC, so we consider it to be the preferred model.

Unlike during the initial treatment phase, we do not condition rates of treatment discontinuation on the level of disease activity. The reason for this is that we could not find studies stratifying treatment duration curves by DAS28, SDAI, or CDAI. As a result, model structures **S1-S5** assume that conditional on continuing treatment beyond the first 6 months, duration does not depend on baseline disease activity, the change in disease activity from baseline, or whether the patient was biologic naive or experienced.

Commented [Office59]: If not disease activity, what do you use? Maybe I am misunderstanding this sentence.

We also lack data on treatment duration for patients on cDMARDs. Following Stevenson et al., we assume that, conditional on continuing treatment at 6 months, treatment duration for bDMARDs is applicable to treatment duration for cDMARDs.[5] This is, in turn, based on the assumption that cDMARDs are not likely to be more toxic than biologics used in combination with cDMARDs.

4.3.6.2. BSRBR database

In model structure **S6**, we stratify our time to treatment discontinuation by EULAR response based on analysis of the British Society for Rheumatology Biologics Registers (BSRBR) database.[5] As with the CORRONA based estimates, we fit 7 parametric survival models. One

concern is that the BSRBR is representative of the UK but not the US. We therefore adjust the BSRBR estimates to the US population using a survival curve from an analysis of the CORRONA database. In particular, time-varying hazard ratios are calculated between the CORRONA hazard function and the overall BSRBR hazard function to reflect the difference between the US and UK. (The overall BSRBR hazard function is a weighted average of the hazard functions stratified by EULAR response with weights equal to the average proportion of moderate and good responders of the treatments used in our model.) Next, the time varying hazard ratios for the US versus UK are applied to the stratified UK survival curves to obtain US specific treatment survival curves by EULAR response category.

4.3.7. Rebound post treatment

Since no data exists on the size of the HAQ rebound post treatment, we vary its size as a proportion of the initial 6-month HAQ decline. We use 1 as an upper bound, which implies that the HAQ rebound is equal to the improvement experienced at the end of the initial 6-month period with that treatment. Expert opinion is used to determine the lower bound.

4.3.8. Serious infections

Based on the NMA by Singh *et* al [33] and in accordance with Stevenson *et al* [5], we will assume a rate of 0.035 (95% CI: 0.027 to 0.046) infections per person-year with all bDMARDs and a rate of 0.026 (no CI reported) infections per person-year with cDMARDs. The rate of infection is assumed to be equal across bDMARDs because the published results for specific bDMARDs are estimated with very little precision. The standard error on the infection rate for bDMARDs will be assumed to be the same as the standard error for cDMARDs since no standard error was reported for bDMARDs in Singh *et* al [33].

Commented [Office60]: Not sure why future tense is used here; elsewhere the model is discussed in present tense

Commented [Office61]: This is the same as 3.5%, yes? That would be in line with this reference to this analysis using British Society for Rheumatology Biologics Register (although cDMARDs are shown as safer): http://www.rheumatologynetwork.com/rheumatoid-arthritis/infections-and-biologics-rheumatoid-arthritis

Commented [Office62]: Serious infections?

4.3.9. Utility

Two algorithms can be used to map HAQ to an EQ-5D utility score. Each is used to simulate utility for every patient in the model to obtain a distribution of utility over time. Our preferred algorithm is the mixture model developed by Hernandez Alava et al. [21]. The second algorithm uses the logistic regression equation reported in Wailoo et al.[22] Differences in mean utility by

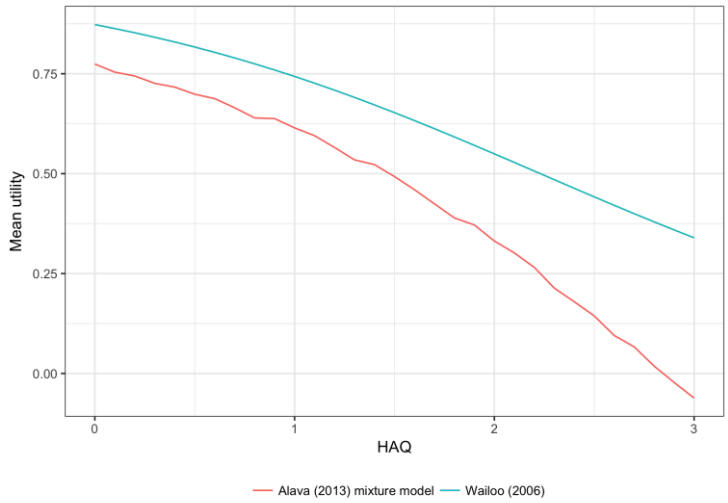


Figure 3: Simulated mean utility by HAQ

HAQ are shown in Figure 3. Mean utility scores from the mixture model lie above those from the Wailoo logistic regression equation for all values of HAQ. Moreover, the slope of utility curve produced from the mixture model is steeper (although less so for the commonly observed HAQ scores between 1 and 1.5), implying that changes in HAQ from the mixture model predict larger changes in utility. Given that the mixture models have been shown to predict utility more accurately [21, 34, 35] this suggests that standard models underestimate the quality-adjusted life-year benefits, and hence, the cost-effectiveness of treatments.

Commented [Office63]: Seems like the reverse? For example, at HAQ of 3, utility for mixture model is near zero, while utility for Wailoo is above .25.

The utility score depends on two other factors in addition to HAQ. First, disutility due to serious infections is assumed to be 0.156 for the duration of the month of infection based on prior studies.[5, 36] However, given the weak evidence for this estimate, the disutility of an infection is allowed to vary by 20% in either direction. Second, patient preferences for treatment unrelated to efficacy or serious infections (e.g., preferences for treatment attributes such as frequency or route of administration) can have a direct effect on utility. Given the limited evidence, we currently assume that this effect is 0, but have programmed the model so that preferences for treatment can easily be incorporated into the model if new evidence emerges. We may try to estimate this effect with expert elicitation or by conducting our own survey.[37]

Commented [Office64]: What if an individual patient wants to indicate preference? Can she do that?

4.3.10. Mortality

Mortality is estimated as a function of age/sex specific mortality from U.S. lifetables [38], baseline HAQ, and changes in HAQ from baseline. Increased odds of mortality in relation to baseline HAQ are based on Wolfe et al.[39] In order to capture the effect of treatment on mortality, every 0.25-unit HAQ increase increases the subsequent 6-month mortality rate according to the hazard ratios reported in Michaud et al.[40] We assume no additional mortality associated with serious infections.

Commented [Office65]: Is that because that is what the evidence tells us, or because there is no evidence either way?

4.3.11. Cost

Commented [Office66]: How will the model account for changes in costs and inflation over time?

Drug costs are based on WACs with discounts assumed to range from 20%-30%. Costs related to physician visits, chest X-rays, tuberculosis tests, and inpatient hospital days will be based on Claxton et al.[41] The annual number of hospital days relates to the HAQ score according to Carlson et al.[10] Cost of any serious infection will be assumed to be equal to the cost of pneumonia hospitalization at $5,873, based on Medicare reimbursement rates. Wolfe et al. provide an estimate of annual income loss in relation to HAQ scores: $4,372 (95% CI: 2,078 to 6,607; 2002 dollars) change per unit HAQ change, which are inflated to 2016 dollars. [42]

4.4. Simulation and uncertainty analysis

4.4.1. Parameter Uncertainty

Commented [Office67]: This section is mostly over my head; I'm not really familiar with the different types of statistical distributions and related assumptions. but I'm glad to see the methods mapped out for statisticians to review.

Parameter uncertainty is quantified using PSA, which propagates uncertainty in the model input parameters throughout the model by randomly sampling the input parameters from their joint probability distribution.[43, 44] Probability distributions are determined according to the distributional properties of the statistical estimates, which, in turn, depend on the statistical techniques used and the distributions of the underlying data. We, for the most part, use normal distributions for sample means, gamma distributions for right-skewed data (e.g., hospital costs), and Dirichlet distributions for multinomial data. The multivariate normal distribution is used for regression parameters estimated using frequentist techniques, provided that the variance-covariance from the statistical analysis is available. For parameters estimated using a Bayesian statistical model, we use the posterior distribution generated from the Markov-Chain Monte-Carlo (MCMC) algorithm. When we lack evidence on a parameter, we typically assume a

uniform distribution with lower and upper limits that reflect the degree of uncertainty in the parameter. The PSA parameter distributions are summarized in Table 3.

Table 3: Probabilistic sensitivity analysis parameter distributions

Parameters(s)	Distribution
Rebound factor	Uniform
NMA parameters – ACR response	Bayesian posterior distribution
NMA parameters – DAS28	Bayesian posterior distribution
NMA parameters - HAQ	Bayesian posterior distribution
Drug acquisition and administration cost	Fixed
Logistic regression coefficients for probability of treatment switch at 6 months	Normal
Survival model parameters for treatment duration during maintenance phase	Multivariate normal
US lifetable mortality rates	Fixed
Mortality probability odds ratio - baseline HAQ	Normal
Mortality probability hazard ratio - change in HAQ from baseline	Normal
ACR response to EULAR response mapping	Dirichlet
ACR response to SDAI mapping	Uniform
ACR response to CDAI mapping	Uniform
ACR response to HAQ mapping	Normal
EULAR response to HAQ mapping	Normal
Linear HAQ progression - by therapy	Normal
Linear HAQ progression - by age	Normal
Non-linear mixture model for HAQ progression	Normal
Utility model - Alava mixture model	Multivariate normal
Utility model – Wailoo	Normal
Hospital costs - hospital days by HAQ	Gamma
Hospital costs - hospital costs per day	Gamma
General management cost	Gamma
Serious infection - survival parameters	Normal
Serious infection - cost per infection	Uniform
Serious infection - utility loss	Uniform

4.4.2. Structural Uncertainty

We consider structural uncertainty due to two factors:

- The relationship between health states within the model.
- The statistical model used to estimate parameters.

Both sources of uncertainty are reflected in Table 2. The choice of model structure for the initial treatment phase (**H1-H3** and **S1-S6**) depends on the preferred measures of disease included in the model as well as whether statistical relationships should be modeled directly or indirectly. Likewise, the choice of model for HAQ progression, treatment duration, and converting HAQ to utility all reflect uncertainty in the appropriate statistical model.

Commented [Office68]: Table 2 was inserted earlier in this document; should be here instead

4.4.3. Implementation

We begin by describing the simulation procedure conditional on model structure, which uses PSA to capture uncertainty within but not between models. The procedure proceeds in two steps: first, model parameters are sampled from their joint probability distribution, and second, for each parameter set, model outcomes are simulated one at a time for individual patients in a given population.

Analysts who wish to expand the analysis to capture uncertainty between models can follow the approach described in Bojke et al.[45] In particular, for each randomly sampled parameter set, each model structure (or a subset of plausible model structures) can be simulated. The distribution of simulated outcomes across parameters and models will then reflect uncertainty both within and between models.

Its important to note that simulation output for an individual patient captures differences in outcomes across patients due to random variation (i.e., first order uncertainty). This information might be useful to patients since it is needed to predict the distribution of their future outcomes conditional on their characteristics, but less useful to a decision maker concerned with making treatment decisions for a population or subset of a population. Analysts wishing to use the model for CEA should therefore estimate mean outcomes by averaging over the simulated patients for each parameter set and model structure. The number of simulated patients should

Commented [Office69]: Use either It is or It's (use apostrophe for contractions of It is)

be sufficiently large so that mean outcomes are stable across model runs (i.e., first order uncertainty is eliminated).

Although CEA is concerned with mean outcomes, that does not imply that it does not account for heterogeneity. Instead, since outcomes depend on the characteristics of each patient, model averages are a function of the population analyzed. Subgroup analyses can be used to examine differences in cost-effectiveness across subgroups by simulating patients with certain shared characteristics.

Parameter and structural uncertainty imply decision uncertainty, or the degree to which decisions are made based on imperfect knowledge. Indeed, with the aim to maximize health outcomes for a given budget, the optimal decision with current information is to choose the policy that maximizes the expected NMB; however, due to uncertainty, the incorrect policy may be considered the most cost-effective. To characterize this uncertainty, standard summary measures including 95% credible intervals for NMBs and other model outcomes, cost-effectiveness planes, and cost-effectiveness acceptability curves, and the expected value of perfect information can be calculated from the simulated output. Since the expected value of partial perfect information is computationally costly, it can be approximated using meta-modeling techniques.[46-48].

Commented [Office70]: But are they automatically calculated? How difficult are they to calculate? Seems important to understand the effect of uncertainty on the model.

Our analysis is computationally intensive because it involves an IPS, PSA, and competing model structures. To simulate the model in a reasonable time frame, the model will be developed in C++ and R.[49] The model will be released in the public domain as an R package on GitHub, an online version control repository. An online user interface will also be released for interactive use using R Shiny (see Section 5.1).

5. Open Source Approach

5.1. R Package

The model will be released as an R package on IVI's public GitHub repository, where all R and C++ code will be visible. The package will be released with two types of documentation that are standard for any R package. Both types of documentations will be useful to modelers who wish to use the model for their own analyses. First, a reference manual will document all functions included in the package. Second, a vignette will provide a series of examples showing how the package can be used to conduct cost-effectiveness analyses.

A model description will be provided in addition to the R documentation. The model description will describe the model structure, data and statistical techniques used to estimate model parameters, and the simulation procedure in both conceptual and technical terms so that it can be understood by stakeholders and used by modelers. The technical part of the document will describe the model in mathematical terms and describe the statistical methods used to estimate the parameters in detail. Results related to model validation and the implications of competing model structures will also be included within the model description.

Non-IVI modelers will be able use the R package for their own analyses by installing it from GitHub using standard commands available in the *devtools* package. Modelers can use the package as is or make direct changes to the code by forking (e.g. making a copy) of the repository.

Non-IVI modelers can choose to collaborate with IVI by suggesting that IVI incorporate the changes from their forked repository (e.g. make a "pull request"). IVI will assess the pull request and accept it if it improves the code (e.g. bug fixes, speed improvements) but does not impact model outcomes. Suggestions that impact model outcomes will be assessed by an expert panel as described in Section **Error! Reference source not found.** ~~5.2~~.

5.2. Expert panel

IVI will invite feedback and suggested changes to the family of models from the public. After the initial version of the model is released, feedback will be incorporated into the next revision using a 3-step formal review process.

Commented [Office71]: Will there be a cost?

Commented [Office72]: I would like to see some examples in this paper

Commented [Office73]: Will the process of incorporating feedback go on after the first revision, or is this a one-time thing?

- *Step 1:* Members of the public will be able to anonymously suggest changes to the model on an online forum. Individuals will always be able to post comments to the forum, even while the formal review process is in Step 2 or Step 3; however, comments will only be incorporated into the next model revision if they are posted before Step 2 begins.
- *Step 2:* IVI will review the suggested changes from step 1 and assess their feasibility. If a feasible change does not affect a model outcome, such as a change to the user interface, then IVI will implement that change. If a feasible change does impact a model outcome, IVI will develop precise language stating how it could be incorporated into the model in practice and use this language to create a questionnaire to be used in Step 3.
- *Step 3:* In the third step, a panel of experts will determine whether IVI's identified changes should be incorporated into the model using the Delphi method. The panel will be representative of payers, patient advocacy groups, providers, and life science firms. Panel members will indicate on a four-point Likert scale ("strongly agree" to "strongly disagree") their level of agreement with each item on the questionnaire and provide reasons for their ratings. A "facilitator" will then provide an anonymised summary of the panel members' ratings and reasoning. After reviewing the anonymised summary, panel members will have an opportunity to update their answers. This process will repeat itself one additional time so that an anonymised summary is provided twice and panel members answer the questionnaire three times. After the process is completed, a change will be incorporated into the model if consensus has been reached on it and the panel is in favor of the suggested change. Consensus will be determined by the interquartile deviation.

After the initial implementation, IVI will periodically make small changes to the code that don't influence model outcomes. New major versions of the model will be released every 6 months after completing the formal review process.

Commented [Office74]: Consensus is a high bar. What level of agreement on your scale constitutes agreement?

Commented [Office75]: OK, here is your definition of consensus, but I don't really understand it.

Commented [Office76]: OK, I see you do plan to make updates in the future. Will you follow the same process for seeking input from the public?

References

1. Wailoo AJ, Bansback N, Brennan A, Michaud K, Nixon RM, Wolfe F: **Biologic drugs for rheumatoid arthritis in the Medicare program: a cost-effectiveness analysis**. *Arthritis Rheum* 2008, **58**(4):939-946.
2. Tosh J, Brennan A, Wailoo A, Bansback N: **The Sheffield rheumatoid arthritis health economic model**. *Rheumatology* 2011, **50**(suppl 4):iv26-iv31.
3. Athanasakis K, Tarantilis F, Tsalapati K, Konstantopoulou T, Vritzali E, Kyriopoulos J: **Cost-utility analysis of tocilizumab monotherapy in first line versus standard of care for the treatment of rheumatoid arthritis in Greece**. *Rheumatology international* 2015, **35**(9):1489-1495.
4. Stephens S, Botteman MF, Cifaldi MA, van Hout BA: **Modelling the cost-effectiveness of combination therapy for early, rapidly progressing rheumatoid arthritis by simulating the reversible and irreversible effects of the disease**. *BMJ Open* 2015, **5**(6):e006560.
5. Stevenson M, Archer R, Tosh J, Simpson E, Everson-Hock E, Stevens J, Hernandez-Alava M, Paisley S, Dickinson K, Scott D: **Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for the treatment of rheumatoid arthritis not previously treated with disease-modifying antirheumatic drugs and after the failure of conventional disease-modifying antirheumatic drugs only: systematic review and economic evaluation**. *Health Technology Assessment* 2016, **20**(35):1-610.
6. Institute for Clinical and Economic Review: **Targeted Immune Modulators for Rheumatoid Arthritis: Effectiveness & Value**. In.; 2017.
7. Stevenson MD, Wailoo AJ, Tosh JC, Hernandez-Alava M, Gibson LA, Stevens JW, Archer RJ, Simpson EL, Hock ES, Young A *et al*: **The Cost-effectiveness of Sequences of Biological Disease-modifying Antirheumatic Drug Treatment in England for Patients with Rheumatoid Arthritis Who Can Tolerate Methotrexate**. *J Rheumatol* 2017, **44**(7):973-980.
8. Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, Krahn M, Kuntz KM, Meltzer DO, Owens DK, Prosser LA: **Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine**. *Jama* 2016, **316**(10):1093-1103.
9. Briggs AH, Claxton K, Sculpher MJ: **Decision modelling for health economic evaluation**: Oxford University Press, USA; 2006.
10. Carlson JJ, Ogale S, Dejonckheere F, Sullivan SD: **Economic evaluation of tocilizumab monotherapy compared to adalimumab monotherapy in the treatment of severe active rheumatoid arthritis**. *Value in Health* 2015, **18**(2):173-179.
11. Singh JA, Saag KG, Bridges SL, Akl EA, Bannuru RR, Sullivan MC, Vaysbrot E, McNaughton C, Osani M, Shmerling RH *et al*: **2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis**. *Arthritis Care Res (Hoboken)* 2016, **68**(1):1-25.

12. Anderson J, Caplan L, Yazdany J, Robbins ML, Neogi T, Michaud K, Saag KG, O'Dell JR, Kazi S: **Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice**. *Arthritis Care Res (Hoboken)* 2012, **64**(5):640-647.
13. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL: **Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis**. *Arthritis Rheum* 1995, **38**(1):44-48.
14. Smolen JS, Breedveld FC, Schiff MH, Kalden JR, Emery P, Eberl G, van Riel PL, Tugwell P: **A simplified disease activity index for rheumatoid arthritis for use in clinical practice**. *Rheumatology (Oxford)* 2003, **42**(2):244-257.
15. Aletaha D, Smolen J: **The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis**. *Clin Exp Rheumatol* 2005, **23**(5 Suppl 39):S100-108.
16. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, Smolen JS: **Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score**. *Arthritis Res Ther* 2005, **7**(4):R796-806.
17. Deighton C, Hyrich K, Ding T, Ledingham J, Lunt M, Luqmani R, Kiely P, Bukhari M, Abernethy R, Ostor A *et al*: **BSR and BHPR rheumatoid arthritis guidelines on eligibility criteria for the first biological therapy**. *Rheumatology (Oxford)* 2010, **49**(6):1197-1199.
18. Gibson L, Alava MH, Wailoo A: **Progression of Disease in People with Rheumatoid Arthritis Treated with non-biologic Therapies**. In.: School of Health and Related Research, University of Sheffield; 2015.
19. Norton S, Fu B, Scott DL, Deighton C, Symmons DP, Wailoo AJ, Tosh J, Lunt M, Davies R, Young A: **Health Assessment Questionnaire disability progression in early rheumatoid arthritis: systematic review and analysis of two inception cohorts**. In: *Seminars in arthritis and rheumatism: 2014*: Elsevier; 2014: 131-144.
20. Ramiro S, Sepriano A, Chatzidionysiou K, Nam JL, Smolen JS, van der Heijde D, Dougados M, van Vollenhoven R, Bijlsma JW, Burmester GR *et al*: **Safety of synthetic and biological DMARDs: a systematic literature review informing the 2016 update of the EULAR recommendations for management of rheumatoid arthritis**. *Ann Rheum Dis* 2017.
21. Hernández Alava M, Wailoo A, Wolfe F, Michaud K: **The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis**. *Rheumatology (Oxford)* 2013, **52**(5):944-950.
22. Wailoo A, Brennan A, Bansback N, Nixon R, Wolfe F, Michaud K: **Modeling the cost effectiveness of etanercept, adalimumab and anakinra compared to infliximab in the treatment of patients with rheumatoid arthritis in the Medicare program**. *Rockville, MD: Agency for Healthcare Research and Quality* 2006.
23. Dias S, Welton NJ, Sutton AJ, Ades A: **NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials**. 2011.

24. Rendas-Baum R, Wallenstein GV, Koncz T, Kosinski M, Yang M, Bradley J, Zwillich SH: **Evaluating the efficacy of sequential biologic therapies for rheumatoid arthritis patients with an inadequate response to tumor necrosis factor-α inhibitors**. *Arthritis Research & Therapy* 2011, **13**(1):R25.
25. Madan J, Ades T, Barton P, Bojke L, Choy E, Helliwell P, Jobanputra P, Stein K, Stevens A, Tosh J *et al*: **Consensus Decision Models for Biologics in Rheumatoid and Psoriatic Arthritis: Recommendations of a Multidisciplinary Working Party**. *Rheumatol Ther* 2015, **2**(2):113-125.
26. Aletaha D, Ward MM, Machold KP, Nell VP, Stamm T, Smolen JS: **Remission and active disease in rheumatoid arthritis: defining criteria for disease activity states**. *Arthritis Rheum* 2005, **52**(9):2625-2636.
27. Zhang J, Shan Y, Reed G, Kremer J, Greenberg JD, Baumgartner S, Curtis JR: **Thresholds in disease activity for switching biologics in rheumatoid arthritis patients: experience from a large US cohort**. *Arthritis care & research* 2011, **63**(12):1672-1679.
28. Souto A, Maneiro JR, Gómez-Reino JJ: **Rate of discontinuation and drug survival of biologic therapies in rheumatoid arthritis: a systematic review and meta-analysis of drug registries and health care databases**. *Rheumatology (Oxford)* 2016, **55**(3):523-534.
29. Wolfe F, Michaud K: **The loss of health status in rheumatoid arthritis and the effect of biologic therapy: a longitudinal observational study**. *Arthritis Research & Therapy* 2010, **12**(2):R35.
30. Michaud K, Wallenstein G, Wolfe F: **Treatment and nontreatment predictors of health assessment questionnaire disability progression in rheumatoid arthritis: A longitudinal study of 18,485 patients**. *Arthritis Care & Research* 2011, **63**(3):366-372.
31. Norton S, Sacker A, Dixey J, Done J, Williams P, Young A: **Trajectories of functional limitation in early rheumatoid arthritis and their association with mortality**. *Rheumatology* 2013:ket253.
32. Guyot P, Ades A, Ouwens MJ, Welton NJ: **Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves**. *BMC medical research methodology* 2012, **12**(1):9.
33. Singh JA, Wells GA, Christensen R, Tanjong Ghogomu E, Maxwell LJ, MacDonald JK, Filippini G, Skoetz N, Francis DK, Lopes LC: **Adverse effects of biologics: a network meta-analysis and Cochrane overview**. *The Cochrane Library* 2011.
34. Hernández Alava M, Wailoo AJ, Ara R: **Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values**. *Value Health* 2012, **15**(3):550-561.
35. Hernández Alava M, Wailoo A, Wolfe F, Michaud K: **A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes**. *Med Decis Making* 2014, **34**(7):919-930.
36. Oppong R, Kaambwa B, Nuttall J, Hood K, Smith RD, Coast J: **The impact of using different tariffs to value EQ-5D health state descriptions: an example from a study of acute cough/lower respiratory tract infections in seven countries**. *The European journal of health economics* 2013:1-13.

37. McNamee P, Seymour J: **Incorporation of process preferences within the QALY framework: a study of alternative methods**. *Medical Decision Making* 2008, **28**(3):443-452.
38. **Actuarial Life Table: Period Life Table, 2013** [https://www.ssa.gov/oact/STATS/table4c6.html]
39. Wolfe F, Michaud K, Gefeller O, Choi HK: **Predicting mortality in patients with rheumatoid arthritis**. *Arthritis Rheum* 2003, **48**.
40. MICHAUD K, VERA-LLONCH M, OSTER G: **Mortality Risk by Functional Status and Health-related Quality of Life in Patients with Rheumatoid Arthritis**. *The Journal of Rheumatology* 2012, **39**(1):54-59.
41. Claxton L, Jenks M, Taylor M, Wallenstein G, Mendelsohn AM, Bourret JA, Singh A, Moynagh D, Gerber RA: **An Economic Evaluation of Tofacitinib Treatment in Rheumatoid Arthritis: Modeling the Cost of Treatment Strategies in the United States**. *Journal of Managed Care & Specialty Pharmacy* 2016, **22**(9):1088-1102.
42. Wolfe F, Michaud K, Choi HK, Williams R: **Household income and earnings losses among 6,396 persons with rheumatoid arthritis**. *The Journal of Rheumatology* 2005, **32**(10):1875-1883.
43. Baio G, Dawid AP: **Probabilistic sensitivity analysis in health economics**. *Statistical methods in medical research* 2015, **24**(6):615-634.
44. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, Brazier J, O'Hagan T: **Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra**. *Health economics* 2005, **14**(4):339-347.
45. Bojke L, Claxton K, Sculpher M, Palmer S: **Characterizing structural uncertainty in decision analytic models: a review and application of methods**. *Value in Health* 2009, **12**(5):739-749.
46. Jalal H, Dowd B, Sainfort F, Kuntz KM: **Linear regression metamodeling as a tool to summarize and present simulation model results**. *Medical Decision Making* 2013, **33**(7):880-890.
47. Jalal H, Goldhaber-Fiebert JD, Kuntz KM: **Computing expected value of partial sample information from probabilistic sensitivity analysis using linear regression metamodeling**. *Medical Decision Making* 2015, **35**(5):584-595.
48. Heath A, Manolopoulou I, Baio G: **Estimating the expected value of partial perfect information in health economic evaluations using integrated nested Laplace approximation**. *Statistics in medicine* 2016, **35**(23):4264-4280.
49. R Core Team: **R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013**. In.; 2016.

Appendix A: Systematic literature review

A systematic literature search was performed to identify published studies evaluating relative treatment effects at 6 months. Titles and abstracts of studies identified from electronic databases were reviewed independently by two researchers to determine eligibility according to the inclusion criteria. Potentially relevant studies were independently reviewed by two researchers in full text. Where there is any uncertainty about the inclusion of studies, a third researcher was consulted and provided arbitration.

Commented [Office77]: Might want to mention the databases are listed below

Commented [Office78]: You might mention here that the criteria are listed in the table below.

Network meta-analyses for relative treatment effects

Selection criteria

Relevant studies met the following predefined inclusion criteria:

Population	Aadult RA patients who are cDMARD-IR • bDMARD-naïve • bDMARD experienced (only for web tool, not CEA)
Interventions	bDMARD
Outcomes	The following outcomes at 24/26 weeks of follow-up • ACR 20/50/70 response • Change from baseline in HAQ
Study design	Randomized controlled trials (RCTs)
Language	English
Other	• Full text reports (conference abstracts/presentations excluded)

Electronic databases

EBM Reviews - Cochrane Database of Systematic Reviews 2005 to January 18, 2017

EBM Reviews - ACP Journal Club 1991 to January 2017

EBM Reviews - Database of Abstracts of Reviews of Effects 1st Quarter 2015

CAB Abstracts 1910 to 2017 Week 02

EBM Reviews - Cochrane Central Register of Controlled Trials November 2016

EBM Reviews - Cochrane Methodology Register 3rd Quarter 2012

EBM Reviews - Health Technology Assessment 4th Quarter 2016

EBM Reviews - NHS Economic Evaluation Database 1st Quarter 2015

Global Health 1910 to 2017 Week 02

Ovid Healthstar 1966 to November 2016

Embase 1988 to 2017 Week 04

Epub Ahead of Print, In-Process & Other Non-Indexed Citations,

Ovid MEDLINE(R) Daily

Ovid MEDLINE(R) 1946 to Present (January 24)

As a strategic advisory board partner we are grateful for the advanced opportunity to review and provide comment on the IVI RA model. We see a unique opportunity in your approach to create an open source model for all to have access to. This offers a unique solution to advance the quality of care for all stakeholder and we look forward to this collaboration. In reviewing the documents provided on July 18th and attending the two calls held on August 2nd and 9th we present the following comments. Given the volume and magnitude of our concerns, we would suggest that IVI take the time to address these comments, review the changes made with the advisory board including all stakeholders (most notably the stakeholders invited and in attendance on the calls of August 2nd and 9th), and get the formal ok from them to advance the model forward. We would strongly suggest postponing any public release till these steps are completed. For clarity we have separated the comments into sections 1) general comments 2) model specific comments.

General comments

1) We ask that IVI re-review the principles around value frameworks that have been developed by PhRMA and NPC. These principles should form the underlying basis of both how IVI approaches model development, and what key parameters IVI seeks to advance in its models. Currently, we see significant opportunity for IVI to improve on its alignment against these principles. For example: a key consideration in both PhRMA and NPC's principles revolves around transparency in process. ICER et al largely build models behind closed doors, without gathering feedback from stakeholders as the inputs are selected. When they do release their draft findings, stakeholders are left struggling to understand complex models in often short review periods. Unfortunately, IVI has made the same errors in its process. Throughout the model parameter description there is a significant amount of work that is not shown. Therefore is difficult to understand or comment on the appropriateness of the specific parameters. Of significant note are a lack of information on the efficacy data, data used in the model parameters, and an overall lack of clarity in the technical documents which make using the model online dashboard a confusing process. As well, while the model is recommended to be built on consensus, information on the consensus process has not been provided.

2) We would suggest that a number of external rheumatology focused health economists should be integrated as an additional advisory board to IVI to continue to advance this model. We would be happy to submit the names of number of individuals who could provide independent, objective feedback and support to strengthen the project. This would further separate this work from past efforts.

3) An overriding comment on this work is that it appears to be a reiteration or rebuild of previously completed cost effectiveness analyses (CEA) or disease models; as it stands, the model offers no significant innovation from existing CE models, which use the primary measure of QALYs as the outcome of interest. There is also no discussion of the limitations of such models and alternatives. This is disappointing, as our investment in IVI were made with the explicit hope that the models it put forth would help to begin a dialogue about how value assessment could be done better. The following bullets offer examples of how elements of the model could readily be innovated upon.

    a. Given the maturity of the data on many advanced RA therapies including advanced therapies such as Anti-TNFs, novel Mechanism of Action Biologics, and JAK inhibitors, we would suggest that a truly innovative approach should include a partnership with the

developer of an established observational data set to better achieve the outlined goals of a transparent, market driven, multi-perspective model. Similarly, we recommend that data be requested from stakeholders (industry, payers, and registry owners) which also could inform the model(s) being developed.

b. We note that one of the goals of the IVI project was to develop models that integrated the perspectives of patients, physicians, and primary stakeholders for RA. With respect to this topic, we are concerned that the model only assumes the perspective of the health care provider or society. The work makes minimal effort to integrate the patient's or physician's perspective, and rather seems to focus only the traditional payer perspectives and endpoints of interest.

c. Ignoring the patient/physician viewpoints results in elements such as treatment preference (i.e. method of administration), benefits other than those measured by the HAQ (impact of treatment on activities of daily living, location of administration, monitoring requirements, etc.), not being included in the analysis. Significant research has been conducted on these elements including assessments which have identified utility impacts of the preference of treatments but were not used in this IVI project. Similarly, as outlined in the timeline of the project patient focus groups were held during the development of the model, but it is unclear how their input was used in this model.

d. IVI's objective was to develop models that were market driven; however the model as it is developed does not appear to reflect the US environment and does not seem to integrate such aspects as payer restrictions, step edits, etc. An innovative model should allow for an assessment of these processes and the impact of removing these restrictions on the identified endpoints of interest to patients, rheumatologists, and relevant stakeholders.

4) Further, given access to the above data, one should use it and only resort to modeled results when measured outcomes are not present. This would make the result far more valuable to all interested stakeholders and stand out from past efforts to this end.

Model specific comments

1) Model structure
    a. The report refers to the 336 possible model structures, however many of these structural changes are based on the estimation of how HAQ is derived (either by ACR response, or EULAR response) or treatment response is determined (ACR, DAS28, SDAI, or CDAI). This appears to be an attempt to integrate the methods of efficacy of a number of other published CEA models rather than building a consensus on one type of structure/modelling method. Also, as outlined further the transformation from ACR to other variables of effectiveness is flawed.
    b. The issue of taking a categorical variable ACR 20/50/70 response to a continuous variable such as SDAI, CDAI, DAS28, and HAQ is not clearly outlined

or described in the document. Algorithms for mapping the ACR response to SDAI, CDAI, DAS28, EULAR response, and HAQ are taken from a number of different, small mapping trials which may or may not show valid relationships between the measures. This limitation was discussed in detail in the document, but further information and variability in the estimates (and the distributions which could be used in the PsA) should be included in the description and model.

c. The switching methodology and justification for treatment switching is also poorly outlined and does not reflect current clinical practice or how physicians make the decision to change therapy in the US healthcare system. Most notably it ignores the treat to target ideology as recommended by the American College of Rheumatology.
d. It would be preferable to integrate a real world data set/registry which could report on the changes of disease activity and likelihood of switch to better reflect the treatment patterns in the US.
e. Analytical horizon, it is assumed that the model assumes a lifetime time horizon, while this is an accepted time horizon, additional shorter time horizons (i.e. 5 years, 10 years, etc.) should be included.

2) Patient population
    a. The initial data on the patient population were taken from Curtis et al and US301 Clinical trial. Of note is that is unclear which of the US registries reported on in Curtis et al is being used and it appears that difference data points (age, HAQ score, etc.) are taken from different registries without a single registry's data being identified as the seminal piece of patient characteristics. Similarly, reviewing the data from Curtis et al it is unclear of when characteristics were collected- at the time of treatment initiation or registry recruitment. A more appropriate source of data would be a real world registry which would report on the clinical characteristics of a patient population starting an advanced therapy. These data should be easily accessible through collaboration with a US registry such as Corrona.
    b. It also questionable to take the data from a leflunomide (a csDMARD) trial completed published 14 years ago- US301 -to inform a model designed to report on advance therapy utilization. This trial and trial data may not reflect the current patient population or treatment patterns based on disease activity. Furthermore, the work completed by Smolen et al only for reported on the validity of SDAI not CDAI.
    c. The algorithm for distributions used in the heterogeneous patient population also could have been replaced simply by using a patient population taken from US based registries who are initiating an advanced therapy. Similarly, the use of a RWD source of advanced therapy experienced (referred to as a bDMARD experienced) patients could inform on the baseline characteristics for those patients when entering the model.
    d. It is also noted that disease duration is included in baseline characteristics of RCTs and is integrated in CEA models. The justification of omission of this variable in the patient populations should be better articulated. Given that disease duration is identified in the latent class growth model for HAQ

progression as a key variable it is unclear how that has been integrated into the model.

3) Comparators
    a. We question the use of csDMARDs as a comparator for anything other than salvage or last line of therapy. A patient whose disease would necessitate an advanced therapy is likely to receive that therapy and while economic theory does propose a "do nothing" strategy we feel this is both an unethical and unlikely strategy and does not reflect clinical practice. Engagement with the recommended panel of RA specialists would help clarify this point.
    b. As described in the general comments, the structure of the model does not appropriately describe the system in the United States for an RA patient. The current formulary and PBM processes which often include preferential products which impact available treatments. A more appropriate model would include data reflecting that and changing the process to assess the impact of these step edits and restrictions of products.
    c. It is also noteworthy that although many treatments for RA have the opportunity to increase dose or decrease the time between doses, this is not included in the modeling. Again, a number of studies using evidence from registries/observational data has reported on this topic and could be integrated into the model.

4) Efficacy data
    a. Initial treatment response, within the model for the initial 6 months of any treatment the estimated response was taken from a newly completed IVI Network Meta Analysis (NMA) using only using RCT data. However, very little information on the NMA is available, of specific note is the background information on the methods of study selection, the studies and data included in the NMA, and methods employed to run the NMA. It should be recognized that a number of registry, non-RCT, and observational trial data sources exist which include US and non-US populations, however, based on the limited information provided it appears that these data have been excluded from these NMA and therefore are ignored. This appears to go against the ideology of reflecting the current US market/being market driven.
    b. For the initial treatment phase it is further unclear why only RCT data were used when other components of the model integrate observational data (i.e. Adverse Event data, initial populations, relationship between HAQ and costs of RA). Given the long history or data and the integration of other registry data in other aspects of the model (i.e. Adverse events), it is unclear why a NMA was required and one so restrictive of inclusion criteria.
    c. Similarly, it is unclear why a factor was applied for the reduction of initial treatment response when patients in the model are assumed to have previous biologic or advanced therapy exposure (either because they have switched treatments in the model or entered the model with the baseline characteristic of having previous advanced therapy or biologic exposure). A number of RCT studies and observational data sources exist to more accurately estimate the different efficacy for a patient who has been exposed to previous biologic or advanced therapy vs a patient who is naïve to these treatments. These data are

ignored by this model (or else perhaps have been inappropriately integrated into the NMA).

d. In the description of efficacy and progression of disease in the period after the first 6 months the method described a number of sources of data. Many of these described methods have been used in previous cost effectiveness models, but appear to ignore the progression of the disease and the natural path of RA observed in clinical practice.
e. For the data on the HAQ progression (in absence of or in the presence of bDMARD or tsDMARD) these data are taken from a number of different registries and sources of data nationally and internationally. Again, a partnership with a US registry or source of real world data would better articulate the disease progression, as well as provide greater information on the thresholds of change for switching treatments.

5) Adverse event reporting
    a. Using a single adverse event (serious infections) does not represent the many events, many of which may require healthcare resource utilizations, and may or may not require discontinuation of treatment. A number of trials (observational and RCTs) and registries have reported on the safety of the treatments included in the model. Using a single adverse event, having it cause a change in treatment, while this is stated as a model limitation, this does not reflect clinical practice. Furthermore this ignores a number of studies which have been completed in RA patients receiving advanced therapies reporting on the risk and changes in risk reported by treatment for a number of unique adverse events and co-morbidities.
    b. Additionally, it is assumed that the risk of adverse event is identical for all treatments. This also ignores a significant amount of literature on the safety of advanced therapies.

6) Utility measurement
    a. The model reports on two models for transforming HAQ score to utility. It is noted that some other methods were ignored due to the data used coming from clinical trial data. This exclusion of methods seems peculiar since clinical trial data is identified as being the primary tool for efficacy measurement. Similarly, studies measuring the impact of using different utility measures (SF-6D, HUI2/3, etc.) have been completed but were ignored in this model. Similarly a number of RCTs and registries have collected and reported on utility data which could have been integrated into this modeling exercise.

7) HAQ to mortality and healthcare resource utilization
    a. Mortality, RA-related hospitalizations, and productivity costs are all estimated from the patient HAQ score. While this has been used in a number of previous CEA models, again, there is potential to use data from an established registry or data set to continue to improve the reporting of this topic.

8) Prices and costs
    a. Currently the model uses the WAC price which we feel is not appropriate. The WAC, as published by FDB represents the manufacturer's (for purposes of this

Drug Price Policy, the term "manufacturer" includes manufacturers, repackagers, private labelers and other suppliers) published catalog or list price for a drug product to wholesalers as reported to First Databank by the manufacturer. WAC does not represent actual transaction prices and does not include prompt pay or other discounts, rebates or reductions in price. First Databank does not perform any independent investigation or analysis of actual transaction prices for purposes of reporting WAC. First Databank relies on manufacturers to report or otherwise make available the values for the WAC data field.

An alternative source of data would be the NADAC; it is our position that the NADAC is a more appropriate measure of drug price. The purpose of the NADAC is to create a new national price benchmark that is more reflective of the prices that pharmacies pay to acquire prescription and over-the-counter drugs. The statute provides that such prices represent a nationwide average of consumer purchase prices, net of discounts and rebates. The survey data will provide information which CMS expects to use to assure compliance with Federal requirements. A monthly nationwide survey of licensed retail community pharmacies, which will include independent pharmacies and chain pharmacies in the United States, will be performed to collect drug acquisition cost information. To ensure that NADACs are accurate, timely, and robust, the NADACs will be reviewed and updated on a weekly basis.

b. The integrated use of only productivity as the societal measure of cost may not fully articulate the costs of absenteeism, presenteeism, and other non-monetary or intangible costs. Further articulation of the costs used would assist a proper assessment of the model.

PhRMA Recommendations to IVI on Draft RA OSVP

IVI should ensure that the process for revising the proposed OSVP is transparent and provides sufficient opportunity for input from all stakeholders.

- Continually update the Strategic Panel of Advisors on efforts to revise the OSVP and provide opportunities for feedback. We understand that IVI is in the process of revising its model, and look forward to reviewing the next iteration. We strongly recommend that IVI hold weekly teleconferences to provide updates on the OSVP as it is revised, so that stakeholders can provide feedback on the changes.
- Release a detailed outline of the proposed changes IVI intends to make to address concerns voiced on the IVI RA Model Briefing Session call on August 2nd, so that the Panel of Health Advisors can provide more detailed feedback on the revisions.

Clearly and effectively communicate how the OSVP and subsequent IVI research align with IVI's mission and core principles.

- Ensure that the mission and objective of both the OVSP and IVI are communicated clearly and effectively. We appreciate and support IVI's mission to improve the way that value is measured and rewarded in the health care system. IVI's commitment to transparency and the notion that that there is no one single definition of "value" are commendable and consistent with PhRMA's Principles on Value Assessment. However, IVI's mission and what IVI hopes to achieve with the OSVP have not been communicated clearly in the past. Moving forward, IVI should ensure that stakeholders understand the objective of both the OSVP project and IVI, so that IVI maximizes its impact in the conversation around value assessment. IVI should make all materials it develops to promote and explain the OSVP and IVI available to the Panel of Health Advisors for review before release.

Continue to ensure patient engagement is at the core of the OSVP and research, particularly the CEA model, and clearly communicate the approach to patient engagement in research and communication materials.

- Report efforts to integrate patient engagement into the structure of the CEA model. For example, the patient focus groups conducted in preparation are a key resource in developing the models, but stakeholders have no way of knowing how IVI used these patient insights to inform methodology. Wherever patient insights were used to inform the approach and methodology, we recommend noting that in IVI-RA's explanation of the methodology.
- Account for patient preferences for treatment attributes in the model, e.g. mode of administration, side effects.
- Document steps to integrate patient engagement efforts into the development of model in the future and develop innovative ways to capture patient preferences and heterogeneity into the foundation of the model. IVI should make patient engagement a priority as it revises its approach.

Improve methods for determining value and incorporate more value elements that aren't traditionally captured in CEA.

- Define value using more innovative, patient-centered measurements than the QALY. The beta version of IVI-RA proposes the primary health outcome is the QALY, and looks very similar to the models leveraged by organizations focused on traditional HTA methods, such as ICER and NICE. The limitations of QALYs, and their inability to capture important patient differences have been widely recognized by multiple stakeholders. In order for IVI to be truly innovative, it must look beyond QALY-centric models to other methods for measuring value.
- Model outcomes should reflect patient values or real world evidence. The RA model transforms HAQ scores into QALYs and other outcomes. As such, there is no added value to this approach beyond what existing models offer.
- Outline proposed approach to incorporate MCDA. We are encouraged that IVI has proposed to incorporate multi-criteria decision analysis into the revised version of the model. However, it is critical that MCDA also reflect appropriate attributes/ parameters and outcomes.
- Incorporate broader definitions of value not traditionally captured in CEA (e.g. insurance value). There has been significant discussion among thought-leaders and other stakeholders about the need to account for a broader range of elements of value when measuring the value of a treatment. For example, ISPOR's Special Task Force (STF) on Value Assessment Frameworks has identified twelve different components of value in their draft whitepaper – including QALY, net costs, productivity, adherence improving factors, reduction in uncertainty, fear of contagion, insurance value, severity of disease, value of hope, real option value, equity, and scientific spillovers. We understand that in the coming weeks IVI will aim to incorporate 'insurance value' as an additional element to the IVI-RA model, which is a promising first step as the elements identified by ISPOR's STF are critical to assessing value. We believe it is imperative that IVI incorporate 'insurance value' as well as other elements of value into the model as well. The proposed approach to operationalizing new value elements should be outlined and shared with the panel.

Allow flexibility of model to address heterogeneity in patient populations.

- Model should allow for variation in treatment effects by relevant patient characteristics, beyond simple patient demographics.

Improve usability of interface and transparency of open source model

- Improve the usability of the model's interface. We appreciate the difficulty in developing a model and interface that can accommodate a range of users. However, the beta version of the OSVP is difficult to navigate and interpret. For example, the interface uses a range of technical terms that are unlikely to be understood or recognized by lay users. The outcomes presented in the "Results" page of the model are unlikely to mean anything to users who are not savvy in health economics or RA clinicians. The results of the model should be clearly interpreted and displayed in ways that are transparent and accessible to patients. We strongly recommend that IVI work to ensure that the model is usable by a range of stakeholders, including patients.

- The model documentation should more closely align with goals of transparency and be more clear and complete.

Effectively communicate the limitations and achievements of the RA model

- We are encouraged that all limitations of the model will be documented in the “known issues list”. The list should also be accompanied by an agenda to address the limitations in the future. Additionally, IVI should clearly highlight what the model offers beyond other decision tools.

Inspired by patients.
Driven by science.

UCB, Inc. Comments on IVI Rheumatoid Arthritis Open Source Value Model: Beta Version

UCB appreciates the opportunity to review and provide feedback on IVI's Rheumatoid Arthritis (RA) Open Source Value Model beta version. Our comments are based on a review of the model itself, as well as its foundational publication "Cost-effectiveness of sequenced treatment of rheumatoid arthritis with targeted immune modulators" (J. Jansen et. al: J. of Medical Economics 20 (7):703-714 (2017)). UCB applauds IVI's commitment to advancing the science of value assessment through the creation of models that are patient-centered, transparent, and consensus-driven. We understand that IVI's process is an iterative one, and offer the following corrections, comments and suggestions for improving the model.

1. Corrections:

- The cost inputs for certolizumab pegol (Cimzia™) are wrong. The parameters section shows the model cost inputs including total drug acquisition and administration costs (reproduced from Jansen, 2017 Table 4). Focusing only on certolizumab, the units for the first 6 months use appear to be correct at 8 (1 unit is considered 2 X 200 mg lyophilized or syringe kit) with 3 X 400 mg in first month and 5 X 400 mg in subsequent 5 months. However, for the last 6 months of the first year the model assumes a total of 26 units when it should only assume 14. This is a significant error which gets replicated every year the model is run when certolizumab is in the scenario analyzed. This type of error has been made in other published economic analyses of certolizumab where researchers were confused about the cost per certolizumab kit and miscalculated the cost per 200 mg dose of certolizumab. UCB urges IVI to address this error immediately, and before the public comment period.

- The infusion cost may be underestimated. This cost is listed as $164. It is unclear whether this is a cost to the payer or the cost to the provider to deliver the drug to the patient. There is no reference to where this number is derived. In a recent analysis,, the average payer reimbursed amount for an infusion of infliximab in RA was $351. In a recent paper on the cost of providing infusion therapy, the annual cost of infusion services in a hospital outpatient infusion center were $1375 (rituximab), $1723 (infliximab), $1772 (abatacept) and $2092 (tocilizumab) per year even before the application of the 30% allocated overhead costs (Schmier 2017). Either way, this underestimates the cost of infusion in the model and may need to explore the utilization of both 96413 and 96415 CPT codes for a more accurate assessment of cost.

2. Comments:

- These are the first value assessment attempts to take the approach of an "individual patient simulation model over a lifetime" and includes the societal perspective of lost patient productivity costs in the analysis. UCB is very supportive of this innovative patient-centered approach which highlights the impact of heterogeneity in patient-defined aspects of value on the ultimate value assessment. However, we are also mindful of the need for the model to balance this goal with the need for the model to be usable and to resonate with key stakeholders including patients, their employer/benefit provider, and payers. It will be very important to strike the right balance between innovation and usability in order for the model to be successful.

- In addition, the authors make it clear that their study is not intended to compare biologic agents to each other but to cDMARDS. The ambitious aims of the authors are to be commended. However, with this "lifetime" approach comes many technical challenges and drug utilization assumptions which may be lacking in sufficient supporting evidence and data. The authors cite a recent meta-analysis study which shows that discontinuation rates up to 27% have been observed with bDMARDs at 1 year and rises to 52% by year 4 since initiation of TNFi. The authors of the study found that data analyzed of bDMARDs use after 2005 have lower survival times than previous studies. The authors of the study concluded "The advent of treatment strategies aimed at achieving stringent and sustained control of the disease with swift medication changes and higher expectations for patients and doctors are reasons that may explain the lower retention of biologics in recent years." (Souto 2016). From a patient-centric approach, future focus may be placed on identifying individual patient factors that lead to improved durability and consequently lower discontinuation rates as patient segments are better aligned with appropriate therapies.

3. Suggestions

- UCB is supportive of IVI's plan to include additional innovations to the model before its release for the public comment period. We appreciate the explanation provided by IVI on the Q&A call regarding why IVI selected the novel concept of "insurance value" to be integrated into the RA model, and agree that this outcome is particularly salient to RA due to large number of healthy people who pay for RA treatments via premiums or taxes. We suggest that IVI create some transparent criteria and/or a narrative explanation for selecting which novel outcomes to include in various condition-specific models going forward to ensure that it is apparent why one choice is made over competing options.

- UCB supports IVI's plan to expand new perspectives in future iterations of the model, and recommend that IVI specifically include the employer-payer perspectives. Employer/benefit providers are a subset of payers that often have a very different point of view and a longer-term time horizon when it comes to assessing value. Incorporating their perspectives will be important to ensure a more fulsome picture of the variety of opinions on value.

- The model(s) evaluate various treatment strategies accounting in some cases for the impact of treatment strategies (like treat to target). A useful application of the same model could also be to do the opposite and evaluate the value of various guidelines given a set of treatment strategies. Doing so could garner lots of interest in the RA community and raise awareness of the IVI model used in both directions.

- The authors of the R.A. model under evaluation make the model assumption that patients who experience elevated HAQ scores or worsening clinically from baseline will transition immediately to another agent. This is likely not the case as discontinuation of drugs can occur for a variety of other reasons other than clinical worsening of disease. Some of these reasons could include, plan formulary changes, out of pocket costs, adverse reactions, non-adherence and others. In addition, there may be gaps of time in rheumatoid arthritis therapy where patients are not on bDMARDs for various reasons. UCB suggests that IVI recognize this variation. Future iterations may consider the inclusion of factors that prohibit quick changes in biologic therapy as a result formulary restrictions or at least address economic issues that arise due to the inability to change therapy once response is lost. Additionally, how does the model account for dose escalations (either in the form of a higher dose and shortening of intervals) to maintain HAQ progression?

- The nine scenarios proposed are plausible but there is no information presented by the authors to best understand the most likely scenarios over a patient's lifetime much less over a period of 1-5 years.

  Payer policies and country reimbursement schemes can help to inform likely bDMARDs scenarios which can be compared to csDMARDs. Current ACR/EULAR guidelines allow for patients to be initiated on TNFi or non-TNFi bDMARDs. There is also no ACR/EULAR recommendations for specific agents which patients should be transitioned to should a patient not be able to continue their current therapy due to efficacy or safety issues.

  The model should have the capability for a more limited timeframe for analysis (1-3 years) supported by real world evidence of what the progress of transitions between agents might look like given average durations of therapy (at initiation and, where indicated, escalated doses), proportion of patients switching to other bDMARDS, time from initiation to switching and drug discontinuation rates. If the RA model can be adjusted to truncate the time frame of analysis at the discretion of the stakeholder utilizing the results for decision making, the model could have more utility.

- All drug WAC costs were discounted by 30% for the analysis in the cost-effectiveness model. The source of this data is cited. With discounts varying across drugs, applying a single discount rate does not accurately reflect prices paid in the U.S. UCB recommends that IVI allow for customization of any rebates or discounts provided for a decision maker to take these into consideration in the cost-effectiveness analysis. Additionally, for instances where the model user would not know the actual rebates or discounts, UCB recommends that other more accurate sources be used to estimate drug discounts such as the publicly available average sales price (ASP) used in Medicare Part B.

- Alternative scenarios were used to evaluate the cost effectiveness of a regimen. One alternative scenario included dose escalation and its impact on costs. As a threshold matter, UCB notes that dose escalation can also impact adherence, a result that does not appear to be taken into consideration by the model.

  Moreover, the reliance upon the study cited by Moots (2011) is an older study which evaluated charts from 44 sites in 5 European countries to evaluated dose escalation for three biologics (etanercept, adalimumab or infliximab). These data may not reflect U.S. dose escalation patterns for drugs which allow for dose escalation in their package inserts (adalimumab and infliximab). However, a more recent systematic review (Moots 2016) included both U.S. and European studies. In that paper, dose escalation occurred in 42% of infliximab patients, 15% of adalimumab patients and 5% of etanercept patients. Usually these studies evaluated dose escalation from drug initiation and followed patients out to at least one year (Moots 2015).

  Dose escalation tends to be higher in U.S. patients and lower in EU patients. The proportion of patients experiencing dose escalation, the magnitude of dose escalation (especially for infliximab which has a large range up to 10 mg/kg) and the duration of a patient on an escalated dose can have a significant impact on drug administration costs.

  UCB suggests that IVI consider the impact of dose escalation on adherence and incorporate the more recent data on the incidence and impact of dose escalation into the model.

- UCB is concerned that the current timeline for public release may not afford IVI enough opportunity to address all of these corrections, comments and suggestions. We recommend that IVI permit the Strategic Advisory Panel another opportunity to review the revised model before public release, and that sufficient time is allowed for a meaningful review.

Should IVI wish to discuss any of these concerns, please contact Alison Anway, U.S. Director of Public Policy at Alison.Anway@UCB.com; or 404-295-0751.

Thank you,

Patty Fritz
Vice President of US Corporate Affairs
UCB, INC.

References:

1. Souto A, et. al: Rate of discontinuation and drug survival of biologic therapies in rheumatoid arthritis: a systematic review and meta-analysis of drug registries and health care databases. Rheumatology 2016: 55:523-534
2. Harnett J, et. al: Real-world evaluation of TNF-inhibitor utilization in rheumatoid arthritis Journal of Medical Economics 2016;19(2), 91-102
3. Desai RJ, et. al: Temporal trends in the use of biologic DMARDs for rheumatoid arthritis in the United States: a cohort study of publicly and privately insured patients J. Managed Care Specialty Pharmacy 2017: 23(8):809-814
4. Jansen J et. al: Cost-effectiveness of sequenced treatment of rheumatoid arthritis with targeted immune modulators J. of Medical Economics, 2017 20 (7): 703-714
5. Schmier J et. al: Cost of providing infusion therapy for rheumatoid arthritis in a hospital-based infusion center setting Clinical Therapeutics 2017 (online prepublication)
6. Moots RJ et. al: Differences in biologic dose-escalation, non-biologic and steroid intensification among three anti-TNF agents: evidence from clinical practice Clinical Experimental Rheumatology 2011; 29: 26-34
7. Moots RJ et. al: Burden of dose escalation with tumour necrosis factor inhibitors in rheumatoid arthritis: a systematic review of frequency and costs Clinical and Experimental Rheumatology 2015; 33(5):737-745
8. Bedenbaugh A. Data on File at UCB
9. Johnson S et. al: Comparison of healthcare costs between rheumatoid arthritis patients treated with infused biologics after switching from another biologic Drugs-Real World Outcomes 2015; 2:99-199