A Description of the IVI-RA Model v1.0

Devin Incerti* Jeroen P. Jansen*

October 6, 2017

Contents

*Innovation and Value Initiative

List of Figures

List of Tables

Executive summary

The Open-Source Value Project

This document describes version 1.0 of the Innovation and Value Initiative's (IVI's) individual patient simulation model for rheumatoid arthritis (RA) (the IVI-RA model). The IVI-RA model is part of IVI's Open Source Value Project (OSVP), which is building an open, collaborative, and consensus-based process for the development of tools for value assessment. Models developed by the OSVP process are iterative, evolving as the science of value assessment advances and as new evidence becomes available.

Version 1.0 is IVI's first release of the IVI-RA model, and is designed to provide a starting point for open debate. As such, the model is very flexible and allows users to choose from a large number of the plausible model structures supported by clinical practice and prior decision-analytic modeling research in RA. General users may run the model online with our web-based user interfaces.

To facilitate transparency, understanding, and debate, OSVP models are released along with the following:

- Source code (on GitHub)
- Code documentation
- Model documentation
- Web-based user interfaces

The IVI-RA model is released specifically as an R package with documentation available online. Technical users can view the source code for the model, run the model using the R package, or even modify the original source code. This technical documentation provides detail on the model structure, statistical methods for parameter estimation, and source data.

Following version 1.0, each major release of an OSVP model (i.e., version 2.0) will occur following a three-step process:

1. **Public comment period**: IVI will collect evidence-based suggestions for improvement.
2. **Technical expert panel**: IVI will collate the suggestions and a technical expert panel comprised of leading scientists in the field will determine which improvements should be incorporated into the next release of the model.
3. **Model updates**: Based on the feedback from the technical expert panel, IVI will update the GitHub source code, create a new version of the model documentation, and release a new version of the R package.

Over time, the number of model structures may shrink as the OSVP process moves toward scientific consensus. To be sure, the OSVP process will not eliminate all the variation in results of value assessment since perspectives on value will vary and disagreements about relevant clinical evidence may persist. But the consensus-based approach will allows users to better understand legitimate and intrinsic reasons why value estimates vary.

Guiding principles

IVI aims to design models under the OSVP process that are:

- **Decision-focused:** Since there are always be gaps in the available evidence and the appropriate scientific assumptions, it is important to quantify uncertainty. OSVP models quantify parameter uncertainty using probabilistic sensitivity analysis (PSA) and quantify structural uncertainty by estimating value across a range of scientifically defensible model structures.
- **Patient-centered:** Models should capture differences in individual characteristics, preferences, circumstances, and response to treatment.
- **Adaptable:** Models should be flexible enough to meet the specific needs and perspectives of diverse different decision-makers including insurers, patients, providers, and policymakers. Analyses can be tailored to specific populations and estimated from multiple perspectives (Sanders et al. 2016; Ioannidis and Garber 2011).
- **Consensus-driven:** Models should reflect a range of scientifically defensible approaches. Potentially controversial elements are highlighted so that they can be debated and introduced as part of the OSVP process.

Toward these ends, the IVI-RA model incorporates:

- **Structural uncertainty:** The model includes 336 possible model structures that reflect different assumptions about the effect of treatment on morbidity, reasons for switching treatment, time on treatment, the progression of disease over time, and patient preferences.
- **Patient heterogeneity:** The progression of disease, mortality, and preferences for treatment vary according to individual characteristics.
- **Preferences for treatment attributes other than efficacy/safety**: Most decision-analytic models capture preferences for living in particular disease states and from adverse events; the IVI-RA model incorporates preferences for other treatment attributes such as ease of use and the time a medication has been on the market.
- **Subpopulation treatment effects:** Although current evidence is scarce, users can adapt the model so that treatment effects vary across patients (e.g., as a function of patient characteristics or prognostic factors).
- **The value of treatment to the healthy:** Conventional value assessments focus on value to the sick, but recent research provides a framework for valuing technology for the healthy (i.e., "insurance value") as well Lakdawalla et al. (2017). The IVI-RA model allows users to optionally incorporate insurance value, but we note that it is less well established than conventional approaches.
- **Real-world evidence:** To ensure that simulated clinical and economic outcomes reflect outcomes in routine practice, we model "baseline event rates" (i.e., disease progression, mortality, time on treatment), patient preferences, and costs using real-world data. To minimize bias, relative treatment effects (i.e., differences in safety and efficacy across treatments) are, when possible, based on randomized clinical trials (RCTs), and then applied to the baseline event rates.

Value assessment

Cost-effectiveness analysis (CEA) is a well-established approach for value assessment grounded in economic theory and widely used in the scientific literature (Briggs et al. 2006; Meltzer et al. 2011; Drummond et al. 2015). CEAs calculate cost per quality-adjusted life-years (QALYs) for a medical intervention relative to the previous standard of care and deem an intervention "cost-effectiveness" when total (incremental) benefits outweigh total (incremental) costs. Some researchers have suggested multiple-criteria decision analysis (MCDA) as an alternative to CEA (Thokala and Duenas 2012), although, like CEA, static versions of this framework are often limited by the fact that there is often significant disagreement on the relative value of different MCDA components depending on the perspective of the decision-maker.

The IVI-RA model is not a value assessment framework but a model that simulates clinical and economic outcomes for each individual in the population of interest. It can therefore be used with any value framework preferred by the user. Currently, the web-based user interface supports both CEA and MCDA (where users can enter their own decision weights for each criterion). IVI has also developed an R package, `hesim`, for health-economic simulation modeling and decision analysis that can be used to perform individualized CEA (Basu and Meltzer 2007; Ioannidis and Garber 2011; Espinoza et al. 2014) on simulation output from the IVI-RA model.

Model overview

The IVI-RA model is a discrete-time individual patient simulation that simulates clinical and economic outcomes for individual patients. Model cycles are 6-months long, which is consistent with clinical trial evidence. Parameter uncertainty is assessed using probabilistic sensitivity analysis (PSA) and structural uncertainty can be assessed by estimating value across the 336 possible model structures. The model simulates the progression of the health assessment questionnaire disability index (HAQ), a measure of functional status in RA.

Serious infection rates and changes in HAQ score during the first 6 months from baseline are based on clinical trial evidence. The change in HAQ can be modeled indirectly as a function of the American College of Rheumatology (ACR) response to treatment, the European League Against Rheumatism (EULAR) response to treatment, or directly as a function of the treatment. Patients switch treatment during the initial 6 months if they have a serious infection. Additionally, the user can chose whether treatment switching should be based on disease activity level or treatment response.

After the first 6 months on a new treatment, the HAQ score progresses over time at a rate based on observational data. Progression can either be assumed to be linear (Wolfe and Michaud 2010; Michaud et al. 2011) or modeled using a non-linear mixture model (Norton et al. 2014).

Patients remain on treatment until treatment discontinuation or death. Time to treatment discontinuation is based on parametric survival analyses of real-world data. Seven possible distributions (exponential, Weibull, Gompertz, log-logistic, lognormal, and generalized gamma) can be chosen by the user. Male and female mortality is based on US lifetables and increases with the HAQ score at baseline and the change in the HAQ score from baseline.

Health care sector costs consist of drug acquisition and administration costs, hospital costs (which increase with the HAQ score), general management costs, and costs caused by serious infections. Non-health care sector costs are those due to lost wages.

Users wishing to calculate utility for CEA can map HAQ and individual characteristics to utility using the logistic regression algorithm of Wailoo et al. (2006) or the Alava et al. (2013) mixture model. With both the Wailoo et al. (2006) and Wailoo et al. (2006) mappings, utility is calculated as a function of the HAQ and individual patient characteristic mapping, serious infections, and preferences for treatment attributes unrelated to safety and efficacy. QALYs combine life expectancy with per cycle utility.

We have found that model outcomes are especially sensitive to certain parameters and model structures, which highlights the importance of a flexible and consensus-based model. Primary sources of uncertainty include:

- The effect of treatment on the change in HAQ from baseline during the first 6 months of treatment
- The long-term progression of HAQ
- The reduction in treatment response after previous treatment failures
- The extent to which the HAQ score "rebounds" to its initial level after failing treatment
- Time on biologic treatment
- The relationship between HAQ and quality of life

1 Open-source consensus-based models for value assessment

The continuing increase in US healthcare costs has stimulated the introduction of initiatives to promote the use of high-value care. Decision-analytic models can be used to inform efficient use of health care resources, but are only relevant when deemed credible by different stakeholders, are representative of the local context and patient population, and can be easily updated without duplication of effort.

The nature of simulation modeling often leads to scientific disagreements and mistrust among decision-makers. Models are typically complex and difficult to understand. Even modeling experts may not be able to fully understand a model without public source code and detailed model documentation. Furthermore, efforts to make models accessible to non-experts are lacking. Models also become quickly outdated as new evidence arises or new scientific approaches are developed, which means that previous finding quickly become irrelevant to decision-makers.

The OSVP aims to increase understanding and relevance to diverse stakeholders by developing open-source consensus-based models. The hope is that these efforts can increase confidence in efforts to base reimbursement and policy decisions on value.

OSVP models are released and updated using a four step process:

1. Develop a flexible, open-source decision-analytic model reflecting a range of plausible modeling approaches based on clinical practice and prior scientific literature.
2. Invite feedback and suggested improvements to the model in a public comment period.
3. A panel of experts determines which of the evidence-based suggestions for improvement suggested in Step 2 should be implemented by means of peer-review and a formal voting process.
4. Revise the model based on the feedback from the technical expert panel in Step 3.

To provide a starting point for debate, the initial release of each OSVP model (i.e., version 1.0) must be flexible and allow users to choose from a large number of plausible model structures and approaches based on clinical practice and previous modeling efforts. The four-step process is designed to be repeated many times so that the scientific approach and evidence considered can be refined over time.

To help increase understanding and accessibility, each version of an OSVP model is released with the following components:

- Source code (on GitHub)
- Code documentation
- Model documentation
- Web-based user interfaces

This document comprises the model documentation, which includes details on the model structure, statistical methods for parameter estimation, and source data. The GitHub source code and code documentation are designed for researchers with programming experience. The code documentation

provides tutorials and references to functions so that users can run the model. For example, the IVI-RA model is released as an R package and can be run using the R programming language (see Section 2.2 for more details). Users can also modify the source code directly after "cloning" a GitHub repository, although a model version will not be considered an official IVI model unless the modifications were recommended by the technical expert panel.

IVI develops web-based graphical user interfaces to make its models accessible to non-technical users. Stakeholders can modify the models (e.g., by changing the model structure, parameter values, time-horizon, discount factor, etc.) using a point and click interface. Users can choose from a simpler interface in which some default values are preselected or a more flexible interface where nearly all model inputs can be modified.

Researchers may collaborate with IVI in at least two ways. First, they can provide feedback during the public comment period. Second, programmers can make direct changes to the source code by making a "pull request" on GitHub. IVI will review the proposed changes. Code modifications that affect the scientific approach or evidence considered will only be incorporated after a review by the technical panel but other changes such as bug fixes or performance improvements may be immediately accepted.

2 Overview of the IVI-RA model

2.1 Why IVI is modeling rheumatoid arthritis

Treatment for rheumatoid arthritis (RA) is well suited for the OSVP approach for three reasons. First, modeling methods and assumptions vary considerably across existing simulation models (Brennan et al. 2003; Wailoo et al. 2008; Tosh et al. 2011; Carlson et al. 2015; Stephens et al. 2015; Athanasakis et al. 2015; Stevenson et al. 2016; Institute for Clinical and Economic Review 2017; Stevenson et al. 2017). Predicting disease progression is complex and there are a number of different measures of treatment response and morbidity (Madan et al. 2015). Analyses have, not surprisingly, been performed using different modeling approaches and have reached different conclusions about the cost-effectiveness of treatments for RA.

Second, RA is an area of significant innovation. There have been important advancements in the treatment of RA over the past decade, which suggests that there is an increasing need for tools to assess the cost-effectiveness of these treatments.

Third, not only have new treatments come to market recently, but evidence on existing RA treatments is growing rapidly. Thus, there is a strong need for models that can be updated in a straightforward manner as the evidence base evolves.

2.2 Software

The model is available as the `iviRA` R package with documentation available online. The source code can be viewed or downloaded at our GitHub repository. The simulation 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 web-based user interface.

2.3 About the model

The IVI-RA model is a discrete-time individual patient simulation (IPS) with 6 month cycles that simulates patients one at a time. The model accounts for both parameter and structural uncertainty. Since the range of defensible 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).

To ensure that simulated clinical and economic outcomes reflect outcomes in routine practice, we model "baseline event rates" (i.e., disease progression, mortality, time on treatment), patient preferences, and costs using real-world data. To minimize bias, relative treatment effects (i.e., differences in safety and efficacy across treatments) are, when possible, based on randomized clinical trials (RCTs), and then applied to the baseline event rates.

The IPS approach allows us to take an "individualized" modeling approach that captures both observable and unobservable patient heterogeneity. Disease progression, mortality, and preferences all vary across patients. In addition, although the evidence base is limited, users of the R package can model treatment effects as a function of any combination of patient characteristics (e.g., demographics, prognostic factors). Finally, the model incorporates preferences for treatment attributes unrelated to safety and efficacy.

As recommended by the Second Panel on Cost-Effectiveness in Health and Medicine (Sanders et al. 2016), costs are simulated from both a health care sector perspective and a societal perspective. Productivity losses from lost earnings are included in the societal perspective but not the health care sector perspective. As discussed below (Section 2.4), our individualized approach implies that the model could also be tailored to fit the perspective of a patient or provider.

2.4 Intended use of the model

The model simulates clinical and economic outcomes for each individual in the population (see Section 5), conditional on the intervention specified. As described in Section 6 users can model any sequence of biologic treatments and conventional disease-modifying antirheumatic drugs (cDMARDs).

The model can therefore be used for a number of purposes, conditional on the population of interest and the perspective of the decision maker. Here we describe a few possibilities.

The first and most obvious use of the model is for value assessment. Two approaches, cost-effectiveness analysis (CEA) and multiple-criteria decision analysis (MCDA), are discussed in more detail in Section 3. Within the CEA approach, cost-effectiveness can be evaluated from the conventional perspective of a sick individual or from the perspective of a healthy individual using the "insurance value" framework developed by Lakdawalla et al. (2017).

Second, the model can be used to evaluate the consequences of clinical guidelines such as the current treat-to-target guidelines in the US (Singh et al. 2016) or guidelines based on treatment response like in the UK (Deighton et al. 2010). Unlike most previous models, our flexible framework allows treatment switching decisions to depend on disease activity level or treatment response, so outcomes under different decision rules can be simulated.

Third, although the model is currently desnged for population level decision-making, it could, in principle, be used to predict longterm health and economic consequences for patients. The predicted outcomes could, for example, be used to inform patient and providers decision making. For instance, Ioannidis and Garber (2011) argue that cost-effectiveness has relevance to patients spending their own money on health care services, particularly as out-of-pocket costs grow. Likewise, providers have a growing interest in cost-effectiveness models to demonstrate the value of their care whether through participation in Accountable Care Organizations (ACOs), to ensure coverage of medical interventions for their patients, or to reduce unwanted variability in management.

3 Value assessment

The IVI-RA model simulates clinical and economic outcomes for each individual in a given population of interest. Outcomes can be simulated over a particular time horizon or over a lifetime.

Although simulation output can be used with any value assessment framework, IVI tools currently support two methodologies for decision analysis: CEA and MCDA. Cost-effectiveness results and MCDA value scores are automatically generated when users run IVI's web-based user interfaces. In addition, IVI has developed an R package, `hesim`, for health-economic simulation modeling and decision analysis that can be used to perform individualized CEA (Basu and Meltzer 2007; Ioannidis and Garber 2011; Espinoza et al. 2014).

3.1 Cost-effectiveness analysis

CEA is a well-established technique for health technology assessment grounded in economic theory (Meltzer et al. 2011). In general, CEA can be thought of as a methodology for maximizing health or well being subject to a resource constraint (Garber and Phelps 1997). The total value of a new health technology relative to a comparator is typically assessed using the incremental net monetary benefit (INMB),

$$INMB = k \cdot \Delta e - \Delta p, \tag{1}$$

where $e = e_1 - e_0$ is a measure of the incremental health benefits from the new technology relative to the comparator, $p = p_1 - p_0$ is a measure of the incremental cost of the new technology, and k is the willingness to pay for a one-unit health gain. The new technology can be deemed cost-effective if the $INMB > 0$, or equivalently, in terms of the incremental cost-effectiveness ratio (ICER), if,

$$\frac{\Delta p}{\Delta e} < k. \tag{2}$$

Incremental health benefits are typially measured in terms of health gains or patient well-being. Since treatments can affect both morbidity and mortality, CEAs typically use the quality-adjusted life-year (QALY). Since costs and benefits vary across patients, some researchers have argued for individualized CEA (Basu and Meltzer 2007; Ioannidis and Garber 2011; Espinoza et al. 2014) so that INMBs and ICERs are calculated separately for different subpopulations. It can be shown that if treatment response varies across the population, then making separate decisions in different populations will increase social welfare (Basu and Meltzer 2007).

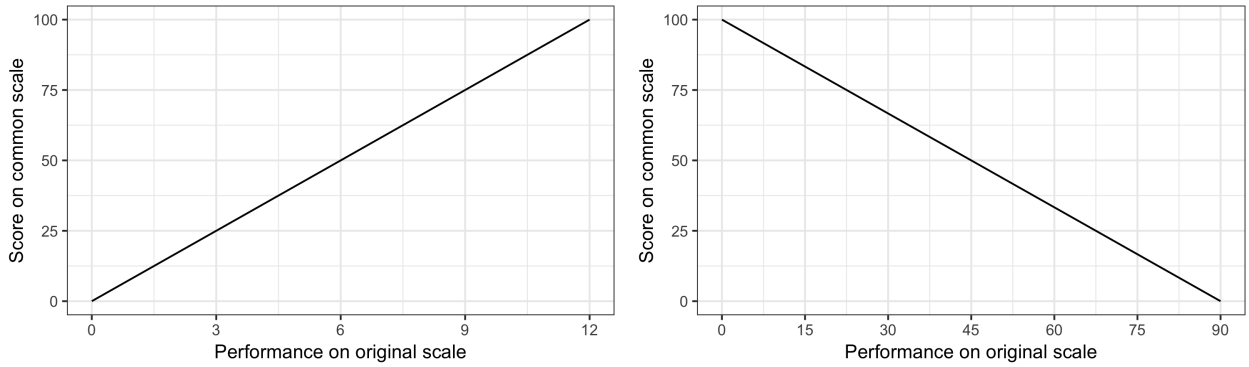
In practice, costs and health benefits are subject to statistical uncertainty. We quantify this uncertainty using probabilistic sensitivity analysis (PSA) and structural uncertainty analysis, which is described in more detail in Section 9. This approach allows us to generate standard measures of uncertainty in CEA including cost-effectiveness planes (Black 1990), cost-effectiveness acceptability curves (CEACs) (Van Hout et al. 1994; Briggs et al. 1999; Fenwick et al. 2001), and estimates of the expected value of perfect information (EVPI) (Fenwick et al. 2001).

3.2 Multiple-decision criteria-analysis

An alternative approach to CEA is MCDA. Keeney and Raiffa (1993) define MCDA as "an extension of decision theory that covers any decision with multiple objectives. A methodology for appraising alternatives on individual, often conflicting criteria, and combining them into one overall appraisal..." We use a similar approach, which implies that separate criteria are aggregated into a single measure of value.

There are many approaches to MCDA; here, we discuss the approach used by IVI in the web-based user interface, which is based on the discussion in Thokala et al. (2016). First, decision-makers must select the relevent criteria for the analysis. These criteria are based on the costs, health outcomes, and risks simulated from the underlying health-economic model. We discuss the criteria relevant to the IVI-RA model in Section 7.8.

Since different criteria may be measured using different units, performance on each criterion is converted into a common scale, for instance, ranging from 0 to 100. There are a number of techniques for creating a common scale; we use a simple linear partial value function to translate scores, which assumes a linear relationship between performance on the original scale of a given criterion and the common scale. To illustrate, Figure 1 demonstrates two mappings between the original scale and the common scale.



(a) Criterion where high performance is better **(b) Criteria where low performance is better**

Figure 1: Linear partial value functions

Performance on the first criterion, shown in Figure 1a, ranges from 0 to 12 on the original scale, with higher scores denoting better performance. In contrast, performance on the second criterion, shown in Figure 1a, ranges from 0 to 90, with lower scores denoting better performance. The relationship between performance on the original scale and the score on the common scale is therefore positive for the first criterion and negative for the second criterion. In both cases, the relationship follows a straight line because we assume a linear relationship.

Each criterion is assigned points, say ranging from 0 to 10, by the decision maker, and weighted by dividing each criterion's points by the sum of points across all criteria. For example, if there were 3 criteria and each criterion was given a score of 5, then each criterion would receive a weight of 1/3. If, on the other hand, the three criteria were given scores of 2.5, 5, and 7.5, then they would be given weights of .167, .33, and .5, respectively.

To aggregate results, we assume an additive model. In other words, the total score for a given treatment sequence is calculated by multiplying each criterion by the simulated standardized score

and summing across criteria.

As with CEA, MCDA results are subject to statistical uncertainty. In the user interface, users choose a single model structure at a time, so uncertainty in MCDA outcomes is quantified using PSA. This produces a probability distribution around the simulated total score for each treatment sequence, which can be used to derive quantities of interest such as Bayesian credible intervals around the total score or the probability that each treatment sequence obtains a particular ranking among relevant treatment sequences.

4 Broader concepts of value

Garrison et al. (2017) suggest five concepts of value that researchers should consider adding to the standard cost per QALY based CEA: (1) a reduction in uncertainty from a diagnostic test; (2) insurance value for healthy patients due to reduction against physical risk; (3) the value of hope for individuals who become risk-loving and would rather pay for a therapy with a long right survival tail than a therapy with a shorter right survival tail but an equivalent (or shorter) expected life-expectancy; (4) real option value when a therapy allows an individual to benefit from future medical innovations; and (5) scientific spillovers when the benefits of an innovation cannot be entirely appropriated by the innovator.

The concept that is arguably most salient to RA is insurance value, which focuses on valuing morbidity-reducing innovations and has the largest effects relative to conventional CEA on treatments for severe diseases where the burden of illness is the greatest. The IVI-RA model allows users to incorporate insurance value into their analyses, while noting that the approach is less well-established than conventional CEA.

Other concepts of value may be incorporated in the future, but likely in future disease areas. For example, real option value is most relevant for innovations that increase longevity and might be particularly well suited to analyses of treatments in oncology. Likewise, survey evidence for the value of hope is based on technologies that increase survival Lakdawalla et al. (2012) rather than those that affect morbidity. Reductions in uncertainty from diagnostic tests are clearly most relevant to diagnostics and scientific spillovers are most relevant to diseases with large externalities such infectious diseases.

Lakdawalla et al. (2017) provide a general mathematical framework for incorporating the effects of medical innovation on physical and financial risk. Conceptually, innovation can lower physical risk to healthy patients who might get sick in the future. New medical technologies act like "insurance policies" that protect a healthy person from all or part of the costs of falling ill. And while innovation certainly increases financial risk, this increase in financial risk can be mitigated by healthcare insurance.

The insurance value framework is an extension of the conventional CEA approach from the perspective of a healthy individual deriving utility from non-health consumption, c and health, h, according to $u(c, h)$. The individual is sick with probability π and well with probability $1 - \pi$. Health when well is h^w and health when sick is $h^s < h^w$. Income is y^w when well and $y^s < y^w$ when sick. The marginal utility of good $j \in c, h$ in state $i \in s, w$ is denoted by u^i_j.

The value of a technology to a healthy consumer (with no health insurance), V^{NHI} is derived implicitly by,

$$\pi u(y^s - p - V^{NHI}, h^s + \delta h) + (1-\pi)u\left(y^w - V^{NHI}, h^w\right) = \pi u(y^s, h^s) + (1-\pi)u(y^w, h^w). \tag{3}$$

The marginal value of the technology, dV^{NHI}, can be shown to be,

$$dV^{NHI} = \pi(k \cdot dh - dp) + \pi(1-\pi)(k \cdot dh - dp)\left(\frac{u_c^s - u_c^w}{\pi u_c^s + (1-\pi)u_c^w}\right) \tag{4}$$

$$= \left[\underbrace{k \cdot dh - dp}_{INMB}\right]\left[\pi + \pi(1-\pi)\left(\frac{u_c^s/u_c^w - 1}{\pi u_c^s/u_c^w + 1 - \pi}\right)\right], \tag{5}$$

where $k = \partial u_h^s / \partial u_c^s$ is the marginal value of a one unit health gain in dollar terms, dh is the marginal health gain from the technology, and dp is the marginal cost of the technology. The term $k \cdot dh - dp$ is equivalent to the INMB in conventional CEA. The insurance value framework can therefore be implemented with knowledge of only two additional parameters beyond those in conventional CEA: the probability of illness, π, and the marginal rate of substitution between the sick and the well states, u_c^s/u_c^w.

The probability of illness can be estimated using incidence of disease in the population of interest (e.g., in the RA population). The second term, u_c^s/u_c^w, is harder to estimate, but we allows users to specify it directly in our model and web-based user interfaces. Intuitively, this term reflects the amount of money the consumer would give up when healthy in exchange for gaining an additional dollar when sick. It rises when the consumer faces greater risks from illness.

It is worth emphasizing that insurance value is only larger than conventional value if the consumer is willing to give up more than \$1 in the well state in exchange for an additional \$1 in the sick state (i.e., $u_c^s/u_c^w > 1$). This is likely to be true, because if the demand for healthcare insurance is positive, then $u_c^s/u_c^w > 1$.

The difference between the insurance value of a technology and its conventional value is even larger when individuals can purchase health insurance. For example, consider an actuarially fair insurance contract that pays the consumer I(p) when she falls sick. In this case, the insurance value of a health technology can be shown to be:

$$dV^{WHI} = dV^{NHI} + \pi(1-\pi)\left(\frac{u_c^s/u_c^w - 1}{\pi u_c^s/u_c^w + 1 - \pi}\right)\frac{dI}{dp}dp. \tag{6}$$

The term dI/dp is the marginal payment made to the insuree per 1 dollar spent on healthcare. In the extreme case where there is no cost-sharing so that $I(p) = p$ and $dI/dp = 1$. Here, health insurance completely eliminates spending risk the value of a technology is equal to its conventional value plus the value of physical risk reduction. More generally, $dI/dp < 1$ and the value of a health technology with health insurance is equal to the sum of its conventional value, the insurance value absent health insurance, and the value of health insurance made possible by the technology.

5 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 1. Details on the algorithm for simulating heterogeneous patients are described in Appendix B.

Table 1: 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

6 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) or Janus kinase/STAT pathway inhibitors (collectively referred to as 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).

7 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 decision-analytic models in RA, version 1 of the model measures changes in disease severity using the Health Assessment Questionnaire (HAQ) Disability Index score (Brennan et al. 2003; Wailoo et al. 2008; Tosh et al. 2011; Carlson et al. 2015; Stephens et al. 2015; Athanasakis et al. 2015; Stevenson et al. 2016; Institute for Clinical and Economic Review 2017; Stevenson et al. 2017). 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 2). In the absence of treatment, HAQ deteriorates at a certain rate as depicted by the dashed line in the figure. For each treatment in a treatment sequence, 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.

7.1 Initial treatment phase

During the initial treatment phase HAQ is modeled as a change from baseline.

- **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 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 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 different pathways **S1-S6**.

- **S1:** Treatment → ACR → Switch
- **S2:** Treatment → ACR → ΔDAS28 → DAS28 → Switch

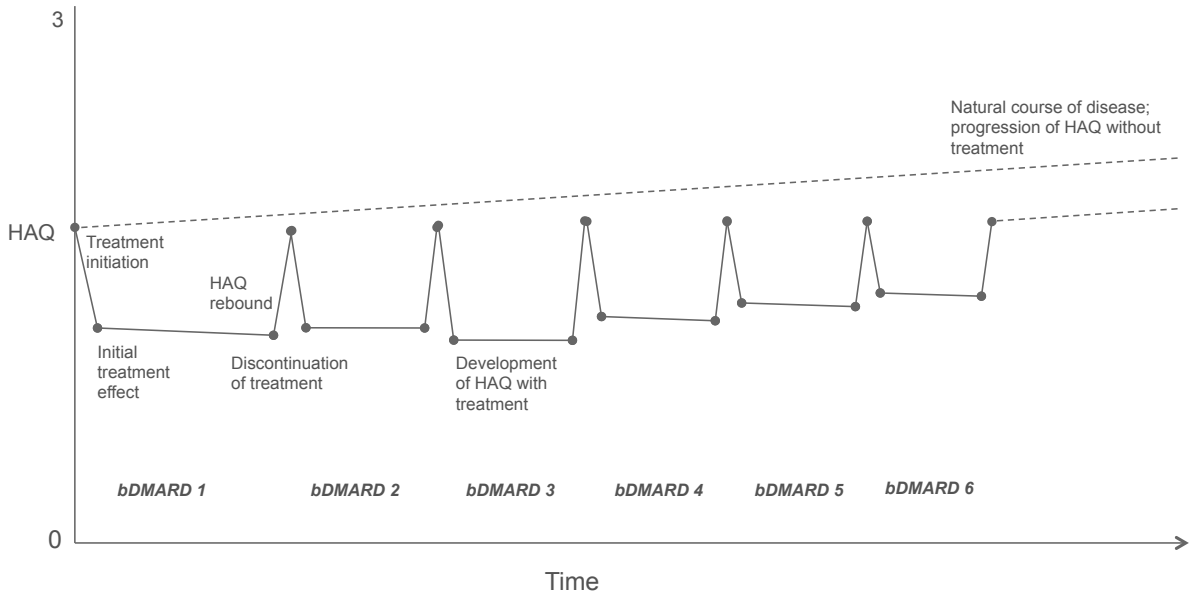


Figure 2: Model structure regarding development of HAQ with sequential biologic treatment

- **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 pathways **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 combinations are outlined in Table 2.

Table 2: 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 pathway used to relate treatment to HAQ and columns denote the pathway used to determine treatment switching. Each number denotes a unique combination of pathways (i.e., 1 corresponds to H1 and S1, and 8 corresponds to H2 and S3) and the "-" denotes a combination of pathways that is not possible. There are 12 possible model structures for the initial treatment phase.

7.2 Maintenance phase

In the maintenance phase, the long-term progression of HAQ can be modeled in two ways. 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 a function of 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).

7.3 Adverse events

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.

7.4 Mortality

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.

7.5 Utility

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, are 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).

7.6 Costs

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.

7.7 Summary of simulation

The flow diagram in Figure 3 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 4 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 4a focuses on the effect of treatment on disease progression and adverse events while Figure 4b looks at the variables influencing the primary health and cost outcomes.

The model accounts for patient heterogeneity in two ways. First, baseline event rates vary across patients by both observable and unobservable factors. For example, longterm HAQ progression, mortality, and utility depend on patient specific variables including age, gender, and baseline disease level. Moreover, unobserved differences in longterm HAQ progression and utility across patients are modeled using mixture models. Second, relative treatment effects for ACR response, the change in HAQ at 6 months, and the change in DAS28 at 6 months, can be modeled as a function of explanatory variables in the R package.

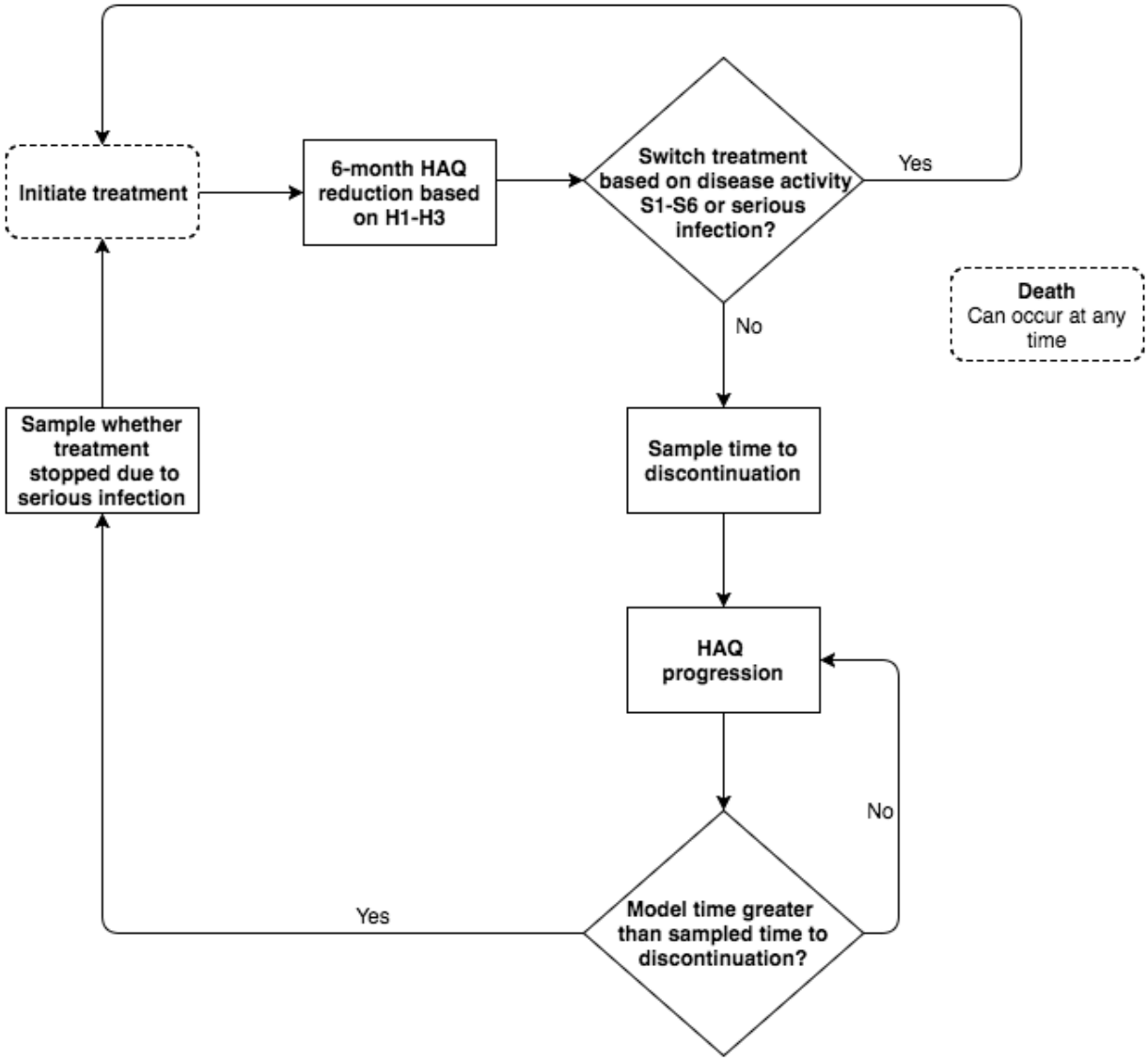
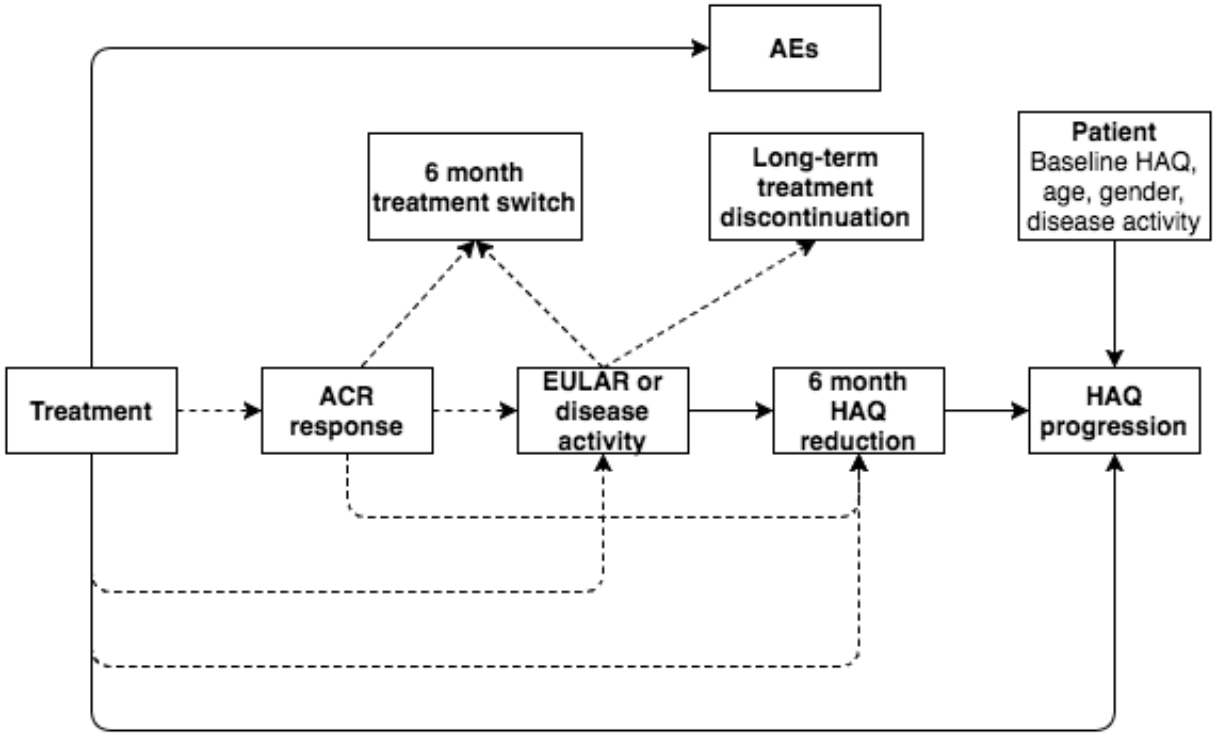
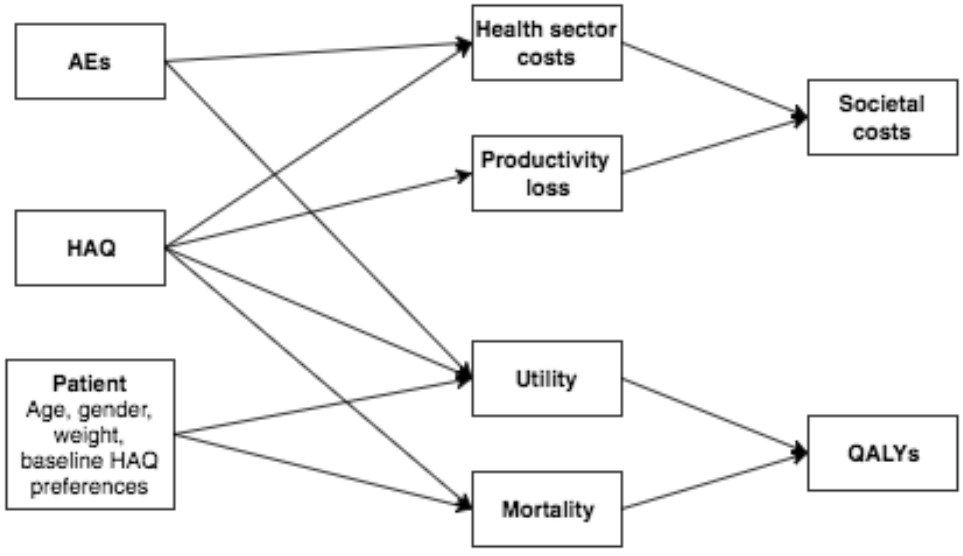


Figure 3: Flow diagram of the simulation for a single patient

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.



(a) Treatment effects



(b) Model outcomes

Figure 4: 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).

7.8 Model outcomes

The model simulates the health outcomes, costs, and risks associated with treatment. Depending on the model structure, model outcomes include the following:

- **Clinical outcomes during initial treatment phase**: ACR response, EULAR response, DAS28, SDAI, CDAI
- **Longterm clinical outcomes**: HAQ, QALYs
- **Adverse events**: Number of serious infections
- **Health care sector costs**: drug acquisition and administration costs, general management and monitoring costs, adverse event costs, hospitalization costs
- **Non-health care sector costs**: productivity losses

If a CEA approach is used for value assessment, then the value of treatment is estimated using the net-monetary benefit (NMB), as described in Section 3.1. CEA from a societal perspective would include productivity losses while analyses from a health care sector perspective would not.

Any combination of simulated model outcomes can be used for MCDA. In IVI's user interfaces, the MCDA is currently based on the following criteria: (i) QALYs, (ii) total healthcare sector costs, (iii) productivity losses, (iv) number of serious infections, (v) number of treatment switches, (vi) route of administration (oral/injection/infusion) and (vii) time the medication has been on the market.

8 Source data and parameter estimation

8.1 Treatment effects at 6 months

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).

Its important to note that treatment effects for each bDMARD were estimated relative to cDMARDs and then applied to the average response for patients using cDMARDs. A limitation of our current approach is that that the average response for patients using cDMARDs is estimated using data from the clinical trials include in the NMA, and may not reflect outcomes seen in routine practice. Future versions of the model could consider using real-world data instead of clinicial tial evidence to estimate this average response.

Given that there is limited evidence that treatment effects vary across patients in the published literature, treatment response at 6 months for a given treatment does not vary according to patient characteristics. Nonetheless, in our R package, treatment effects for each simulated patient can be modeled as a function of any variables chosen by the user. Our approach to modeling treatment effect heterogeneity is described in Section H.4

Treatment effects for bDMARD experienced patients are reduced by multiplying treatment effects for bDMARD naive patients by a constant k. Based on evidence reported in Carlson et al. (2015),

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, 0.282)	0.102 (0.093, 0.112)	0.032 (0.028, 0.036)	-1.148 (-1.218, -1.083)	-0.232 (-0.253, -0.210)
ABT IV + MTX	0.555 (0.469, 0.643)	0.309 (0.235, 0.390)	0.140 (0.095, 0.195)	-2.265 (-2.690, -1.843)	-0.448 (-0.579, -0.319)
ADA + MTX	0.562 (0.482, 0.641)	0.315 (0.245, 0.391)	0.144 (0.102, 0.195)	-	-0.551 (-0.666, -0.443)
ETN + MTX	0.646 (0.530, 0.751)	0.398 (0.285, 0.518)	0.202 (0.125, 0.293)	-2.666 (-3.214, -2.109)	-0.532 (-0.660, -0.409)
GOL + MTX	0.598 (0.432, 0.742)	0.352 (0.209, 0.507)	0.171 (0.083, 0.286)	-2.584 (-3.207, -1.975)	-0.501 (-0.625, -0.375)
IFX + MTX	0.655 (0.415, 0.864)	0.418 (0.196, 0.674)	0.224 (0.074, 0.450)	-1.965 (-2.485, -1.418)	-0.443 (-0.631, -0.251)
TCZ + MTX	0.555 (0.369, 0.739)	0.314 (0.163, 0.499)	0.146 (0.059, 0.278)	-3.003 (-3.292, -2.706)	-0.500 (-0.593, -0.407)
CZP + MTX	0.744 (0.546, 0.893)	0.516 (0.300, 0.722)	0.301 (0.135, 0.506)	-3.149 (-3.561, -2.740)	-0.559 (-0.660, -0.459)
ABT SC + MTX	0.563 (0.421, 0.691)	0.318 (0.200, 0.444)	0.148 (0.077, 0.236)	-2.286 (-2.951, -1.598)	-0.441 (-0.640, -0.258)
RTX + MTX	0.565 (0.413, 0.707)	0.321 (0.195, 0.460)	0.150 (0.075, 0.248)	-2.512 (-2.964, -2.084)	-0.516 (-0.630, -0.403)
TOF + MTX	0.608 (0.447, 0.756)	0.361 (0.217, 0.521)	0.177 (0.086, 0.297)	-	-0.473 (-0.711, -0.249)

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.

we assume that k is uniformly distributed and ranges between .75 and .92, implying that (rounding up) the average value of k is .84. In other words, reductions in DAS28 and HAQ scores for bDMARD experienced patients are, on average, 84% of the reduction in DAS28 and HAQ scores for bDMARD naive patients, and an ACR response of 60/40/20 for bDMARD naive patients would, on average, be reduced to 50/33/16 for bDMARD experienced patients.

In the simulation, treatment response 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 response for all other treatments in a treatment sequence is reduced using the constant k. For bDMARD experienced patients, treatment response is reduced using k at each line of therapy including the first line. One limitation of this approach is that we are unable to model the relationship between line of therapy and k; that is, treatment response for a patient who has failed at least one biologic is assumed to be reduced by, on average, .84, regardless of line of therapy.

8.2 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.

8.2.1 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. However, note that the referenced publications did not report mean outcomes, so we were unable to generate standard errors for the estimates. We consequently assume allow the estimates to vary by 20% in either direction.

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 5) 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.

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

8.2.2 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.

8.3 Change in HAQ at 6 months

In model structures including **H1**, the impact of treatment on changes in HAQ at 6 months is modeled by first estimating the effect of treatment on ACR response and then mapping ACR response to a change in HAQ. 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).

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 when using **H1-H3**. Results were estimated by simulating 1,000 patients for 6 months and randomly sampling 1,000 parameter sets. For each randomly sampled parameter set, we calculated the average decrease in HAQ at 6 months across the 1,000 patients. Estimates reported in the table are the mean and 95% credible interval of the mean decrease in HAQ at 6 months. To maintain consistency across **H1-H3**, we did not allow HAQ

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)

Table 7: Relationship between EULAR response and change in HAQ at 6 months

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

scores for patients who might have otherwise switched treatments accoring to **S1-S6** to rebound back to their baseline levels (i.e., levels at the start of the simulation) at the end of the 6 month period.

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

	H1	**H2**	**H3**
cDMARDs	-0.24 (-0.34, -0.12)	-0.19 (-0.27, -0.10)	-0.23 (-0.25, -0.21)
ABT IV + MTX	-0.44 (-0.55, -0.33)	-0.31 (-0.41, -0.21)	-0.45 (-0.59, -0.32)
ADA + MTX	-0.44 (-0.56, -0.32)	-0.30 (-0.42, -0.21)	-0.55 (-0.67, -0.44)
ETN + MTX	-0.51 (-0.65, -0.39)	-0.35 (-0.48, -0.23)	-0.52 (-0.62, -0.40)
GOL + MTX	-0.47 (-0.62, -0.33)	-0.33 (-0.45, -0.22)	-0.50 (-0.62, -0.39)
IFX + MTX	-0.50 (-0.73, -0.30)	-0.34 (-0.50, -0.22)	-0.45 (-0.60, -0.25)
TCZ + MTX	-0.45 (-0.64, -0.25)	-0.31 (-0.48, -0.17)	-0.50 (-0.58, -0.42)
CZP + MTX	-0.60 (-0.82, -0.37)	-0.38 (-0.53, -0.25)	-0.55 (-0.66, -0.45)
ABT SC + MTX	-0.44 (-0.57, -0.31)	-0.31 (-0.43, -0.19)	-0.43 (-0.62, -0.24)
RTX + MTX	-0.42 (-0.58, -0.27)	-0.30 (-0.41, -0.19)	-0.51 (-0.60, -0.42)
TOF + MTX	-0.48 (-0.63, -0.34)	-0.33 (-0.45, -0.23)	-0.47 (-0.68, -0.24)

Notes: **H1**, **H2**, and **H3** are the Treatment → ACR → HAQ, Treatment → ACR → EULAR → HAQ, and Treatment → 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. Mean ΔHAQ is calculated for each parameter set by averaging across the 1,000 patients. 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.

Estimates for **H1** and **H3** are generally similar but treatment response is considerably smaller when using **H2**. This suggests that the additional mapping between ACR response and EULAR response attenuates treatment response. Given these varying estimates of the change in HAQ during the

initial treatment phase and the impact of HAQ on other important outcomes within the model including utility and health care costs, the choice of **H1-H3** (and in particular **H2** vs. **H1**/**H3**) appears to have important consequences for value assessment.

8.4 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.

8.4.1 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); $\geq$ 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).

8.4.2 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 5) 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 5. 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.

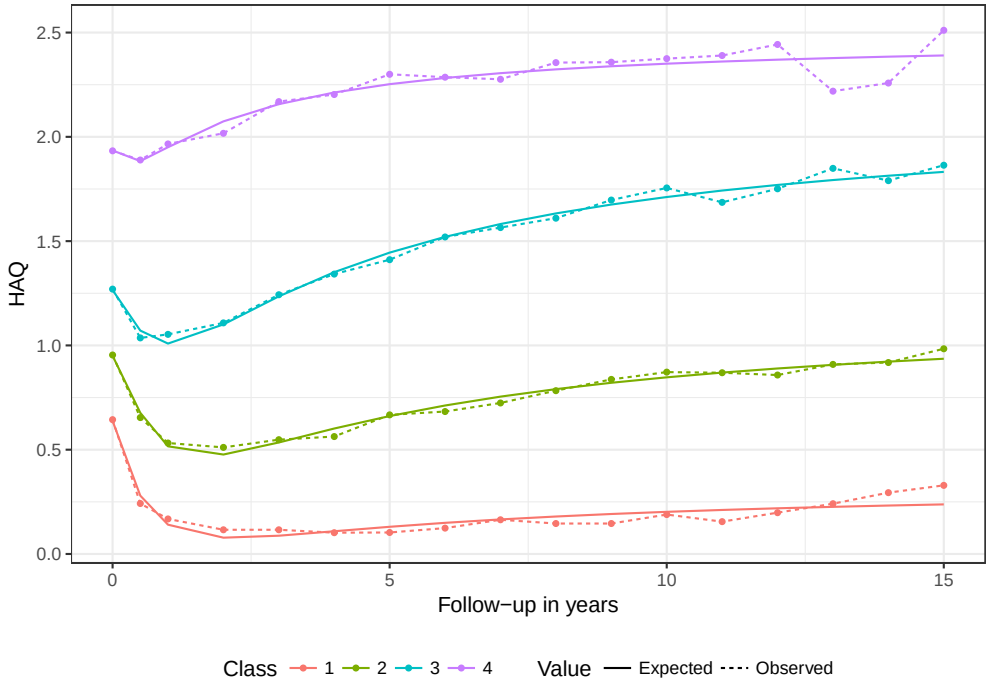


Figure 5: Observed and predicted HAQ trajectories in the ERAS dataset from the latent class growth model

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 modeling disease progression 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 6, 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.

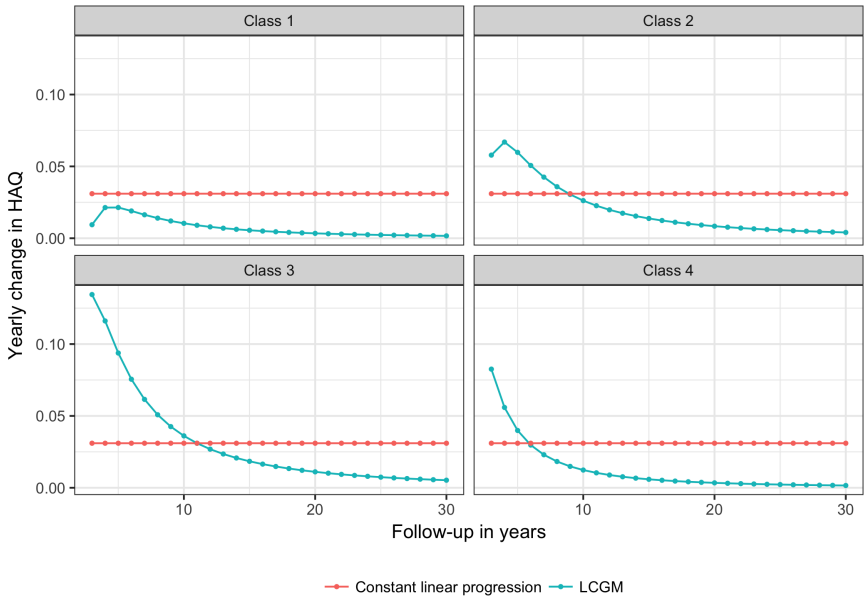


Figure 6: A comparison of predicted yearly changes in HAQ between a latent class growth model and constant linear progression from year 2 onwards

8.5 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.

8.6 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.

8.6.1 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 7. 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

We considered estimating separate time to discontinuation curves for each treatment, 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 treatments. Furthermore, studies comparing rates of discontinuation across treatments 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.

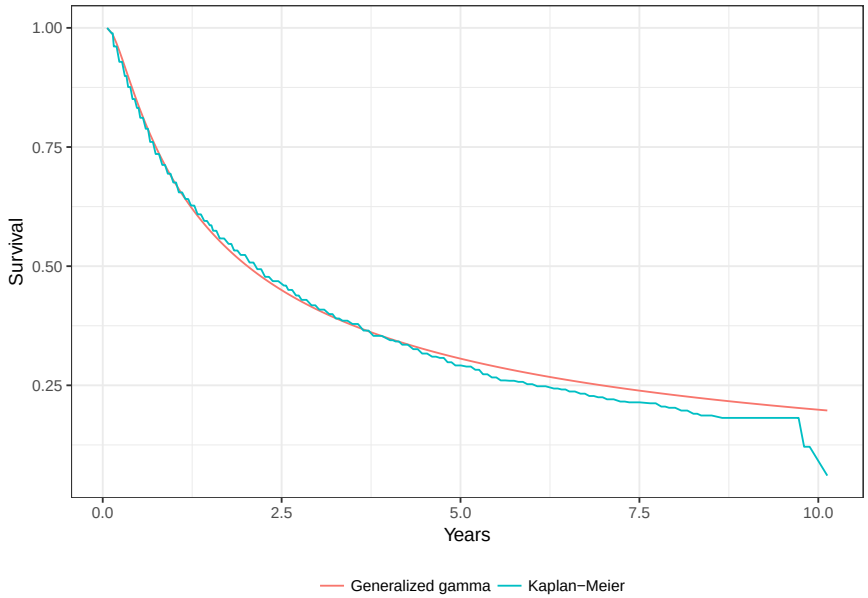


Figure 7: Generalized gamma and Kaplan-Meier time to treatment discontinuation curves using reconstructed individual patient data from the CORRONA database

(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.

8.6.2 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 8.6.1, 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 8. Survival curves for patients with severe disease activity are not displayed because patients with severe disease activity are assumed to switch treatments after the first 6 months of treatment.

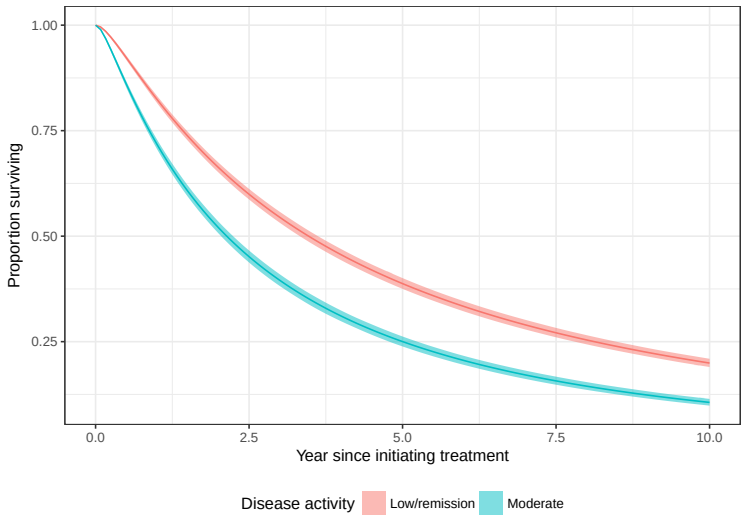


Figure 8: Generalized gamma time to treatment discontinuation curves by disease activity level

Notes: The shaded region denotes the simulation based 95% confidence interval (Mandel 2013).

8.6.3 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 8.6.1.

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 9. AIC and BIC for the parametric models fit in step 6 to the adjusted individual patient data are shown in Table 12.

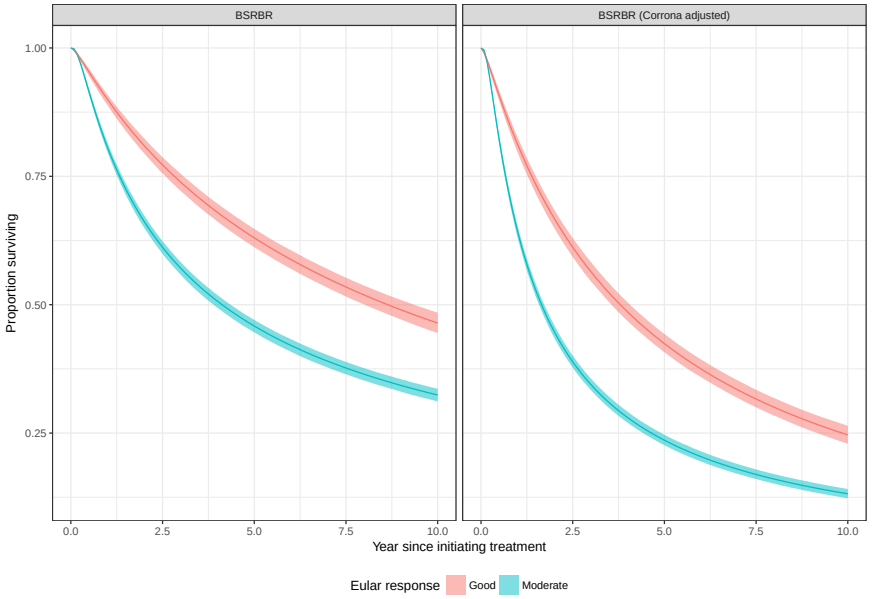


Figure 9: Generalized gamma survival curve of treatment duration using reconstructed individual patient data based on analyses from Stevenson et al. (2016) by EULAR response category

Notes: The shaded region denotes the simulation based 95% confidence interval (Mandel 2013).

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

8.7 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.

8.8 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 3.47% of patients using cDMARDs and 8.10% of patients 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.

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. We consequently simulated time to treatment discontinuation using each of the 7 possible probability distributions. We used the pathway **S1** to model treatment switching, so survival is based on the discussion in Section 8.6.1. Results from the simulation are reported in Table 14. There are very small differences across distributions, suggesting that the treatment duration distribution has almost no impact on the probability of serious infections.

Table 13: Probability of serious infection

	Probability		
		95% CI	
	Mean	Lower	Upper
cDMARDs or NBT	0.0952	0.0670	0.1230
bDMARDs	0.1309	0.0980	0.1680

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

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

Distribution	Mean probability
Exponential	0.0951
Weibull	0.0954
Gompertz	0.0951
Gamma	0.0955
Log-logistic	0.0950
Lognormal	0.0951
Generalized gamma	0.0952

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

8.9 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 Section F.1. The second algorithm uses the logistic regression equation reported in Wailoo et al. (2006). Regression coefficients are reported in Section F.2.

Figure 10 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 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.

Finally, in the R package, we also allow users to incorporate treatment attributes unrelated to safety

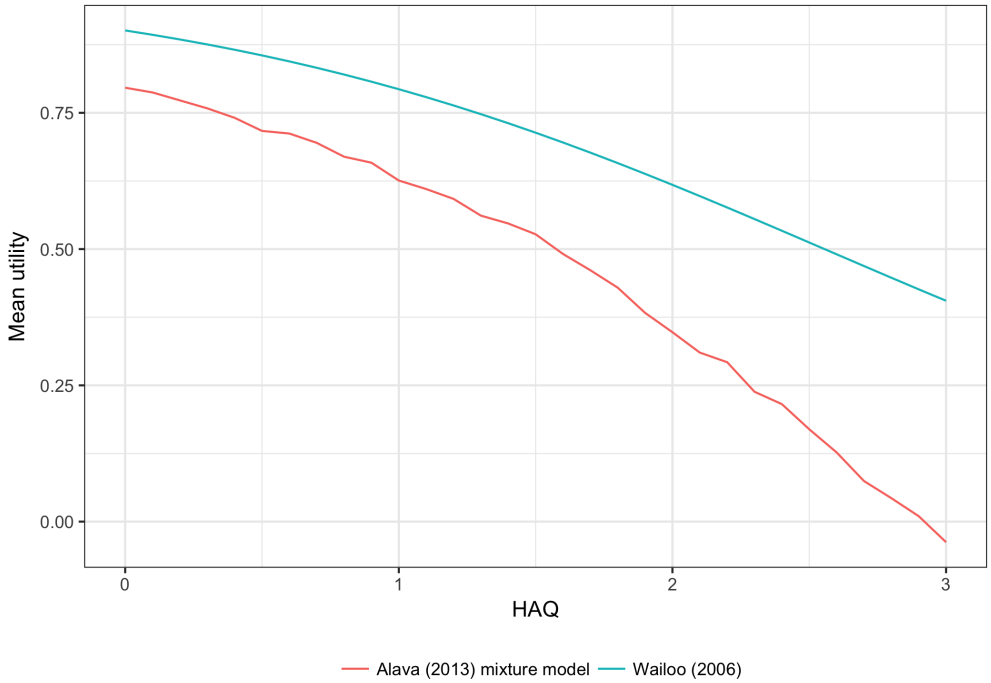


Figure 10: Simulated mean utility by current HAQ

or efficacy that might impact utility. In particular, users can specify a vector of variables and a vector of corresponding coefficients. Each coefficient is the impact of the corresponding variable on utility in a given 6 month period. By default, we include variables related to mode of administration (infusion, injection, oral) and years since FDA approval; however, since we have no evidence on the impact of each variable on utility, the coefficients are set to 0 in our default settings.

8.10 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 15.

Figure 11 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 15 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

Table 15: 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.

suggests that RA increases mortality and that larger increases in HAQ over time increase mortality by even more.

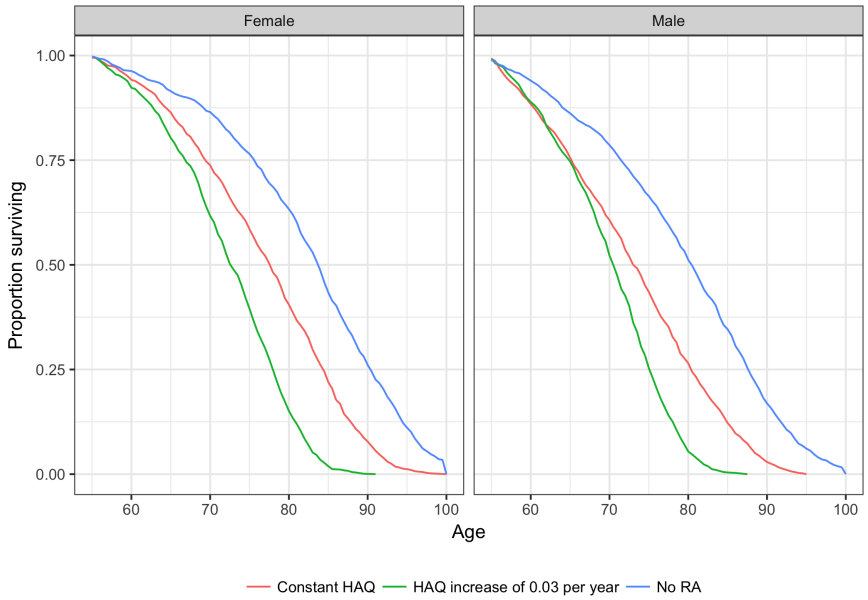


Figure 11: Simulated survival curve for a patient age 55

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.

8.11 Cost

An overview of drug acquisition and administration costs is presented in Table 16. Costs are a function of dose and frequency of administration, strength and dosage form, price, 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 prices in the table are based on the wholesale acquisition cost (WAC) and do not include discounts or rebates so they may be higher than actual drug costs. In the simulation, a unique discount can be used for each drug; currently the discount is assumed to range from 20% to 30%. The methodology used to calculate drug acquisition and administration costs is described in more detail in Appendix G.

Table 16: Drug acquisition and administration cost

Drug	Dose and frequency of administration	Strength and dosage form	Number of doses first 6 months	Number of doses per year beyond the first 6 months	Price 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	0	28,873	57,746
Adalimumab	40 mg EOW	40 mg/0.8 mL syringe or pen injector	13	26	2,220	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	164	17,519	51,669
Golimumab	50 mg QM	50 mg/0.5 mL syringe or pen injector	6	12	3,811	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	0	29,438	47,838
Abatacept IV	750 mg IV at weeks 0, 2, 4 then Q4W	250mg vial	8	13	931	164	23,659	38,447
Abatacept SC	125 mg SC QW with IV loading dose	125mg/ml syringe	26	52	957	164	29,149	58,299
Tocilizumab	162 mg SC EOW	162 mg/0.9 mL syringe	13	26	898	0	11,678	23,356
Rituximab	1000 mg at weeks 0, 2; then Q24 W	500 mg/50ml vial	4	4.33	4,176	164	34,064	36,903
Tofaticinib citrate	5 mg BID	5mg tablet	364	728	63	0	23,026	46,053
Methotrexate monotherapy	15mg QW	15 mg injection	26	52	32	0	842	1,685
Hydroxychlorquine sulfate	400mg daily	200 mg tablet	182	364	3	0	1,157	2,315
Sulfazalazine	1-2 g daily	500 mg tablet	182	364	0	0	342	684

Notes: Costs in the table do not include rebates or discounts, but rebates and discounts are used in the simulation. 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.

Parameters associated with resource use are show in in Table 17. 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 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 17: 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)
Cost per day in hospital	1,251	904	1,652	Carlson et al. (2015)
Cost per day in hospital	1,251	904	1,652	Carlson et al. (2015)
Cost per day in hospital	1,251	904	1,652	Carlson et al. (2015)
Cost per day in hospital	1,251	904	1,652	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	5,965	2,875	9,056	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.

9 Simulation and uncertainty analysis

9.1 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.

9.2 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

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., $T_{si} \leq T$).

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 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 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 18.

9.3 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 19 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.

Table 18: 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

9.4 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,

Table 19: 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

model parameters are sampled from their joint probability distribution (Section 9.2), and second, for each parameter set, model outcomes are simulated one at a time for individual patients in the specified population (Section 5).

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 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 or MCDA 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 and MCDA 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, in CEA, 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 uncertainty within a CEA framework, 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).

10 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.

11 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).

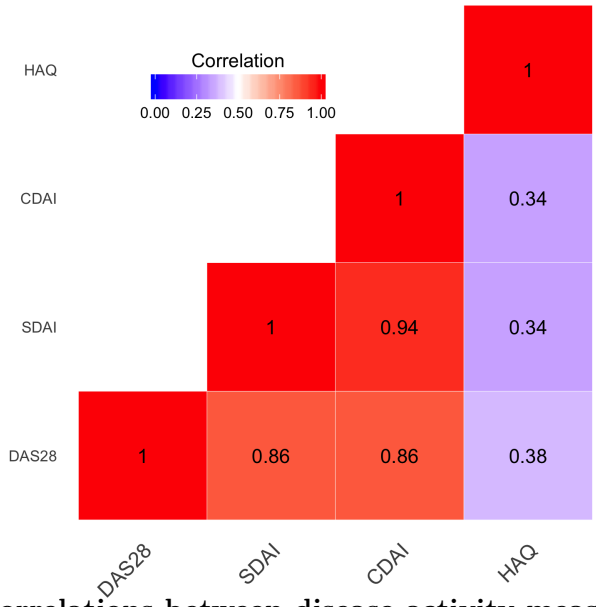


Figure A1: Correlations between disease activity measures and HAQ

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 ≥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 $\mathbb{E}[DA|50 \leq ACR < 70]$:

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

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

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

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, ≥ 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 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,

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.

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

In the simulation 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,

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.

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

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.

E Simulating mortality

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 Simulate utility

F.1 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.1.1 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.1.2 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,

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

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

$$\alpha_{ic} = \gamma_c + z^T_i\kappa + \mu_i, \tag{A23}$$

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,

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

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, $\mu_i \sim N(0, \sigma^2_\mu)$.
2. For each individual i and time-period t:
   (a) Sample class membership conditional on w_{it}; that is, sample $C_{it} \sim \text{Cat}(p_1, p_2, p_3, p_4)$ where p_c is the probability of being in class c.
   (b) Predict the intercept α_{ic}.
   (c) Sample the error term, $\epsilon_{it} \sim N(0, \sigma^2_\epsilon)$.
   (d) Predict the HAQ score, y_{it}.

F.2 Logistic regression model

Wailoo et al. (2006) use a logistic regression equation to predict utility as a function of patient demographics, disease history, and current disease status. The regression coefficients from the model are shown in Table A4 and used to predict utility with the inverse logit function. Specifically, if the vector of coefficients is denoted by β and the corresponding vector of explanatory variables is denoted by the vector x, then predicted utility is given by $1/(1+\exp(-x^T\beta))$.

Table A4: 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).

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 16, 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 ($).

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 ($).

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

H.4.1 ACR response

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.

More specifically, let r_{jkl} be the number of patients in trial j for treatment k in the mutually exclusive category $l = 1, 2, 3, 4$. The model can be written as,

$$r_{jkl} \sim \text{Multinomial}(p_{jk1}, p_{jk2}, p_{jk3}, p_{jk4}, n_{jk}), \tag{A27}$$

where p_{jkl} is the probability that a patient from trial j and treatment k is in category l and there are n_{jk} patients in trial j and treatment k. Probabilities are modeled using a probit function,

$$\Phi^{-1}(p_{jkl}) = \begin{cases} u_{jb} + z_{jl} & b = A, B, C, \ldots \text{ if } k = b \\ u_{jb} + z_{jl} + \delta_{jkb} & \text{if } k \text{ alphabetically after } b, \end{cases} \tag{A28}$$

where u_j is a trial specific intercept, z_{jl} is a cutpoint for trial j and category l, and δ_{jkb} is the effect of treatment k relative to treatment b. The cutpoint for category c, z_{jc}, is modeled as random,

$$z_{jc} \sim N(v_c, \sigma_z^2). \tag{A29}$$

The study-specific relative treatment effects are also drawn from a common population distribution with mean d_{kb} and variance τ^2,

$$\delta_{jkb} \sim N(d_{kb}, \tau^2). \tag{A30}$$

To generate treatment responses, we estimate the response for treatment A by averaging μ_{jA} across trials containing treatment A. In particular, letting $S_A = \{\mu_{1A}, \ldots \mu_{|S_A|A}\}$ be the set of all trials containing treatment A, we estimate,

$$A = \frac{1}{|S_A|} \sum_{\mu_A \in S_A} \mu_A. \tag{A31}$$

We calculate the probability of ACR $<$ 20% improvement, ACR $<$ 50% improvement, and ACR $<$ 70% improvement with treatment k as,

$$P(ACR_k < 70) = \phi(A + z_3 + d_{kA}) \tag{A32}$$
$$P(ACR_k < 50) = \phi(A + z_2 + d_{kA}) \tag{A33}$$
$$P(ACR_k < 20) = \phi(A + d_{kA}). \tag{A34}$$

The probabilities of overlapping ACR response (i.e., ACR 20/50/70) are then,

$$P(ACR_k > 70) = \gamma \cdot (1 - P(ACR_k < 70)) \tag{A35}$$
$$P(ACR_k > 50) = \gamma \cdot (1 - P(ACR_k < 50)) \tag{A36}$$
$$P(ACR_k > 20) = \gamma \cdot (1 - P(ACR_k < 20)), \tag{A37}$$

where γ is the reduction in treatment response at a given line of therapy. $\gamma = 1$ is a patient is bDMARD naive and on average, equal to .84 after failing a biologic. The mutually exclusive ACR response categories are easily derived from the overlapping categories.

To avoid influencing the observed results by prior belief, uninformative prior distributions were used for the estimated model parameters. Posterior distributions for the model parameters are derived with the Markov chain Monte Carlo method using the JAGS software package (http://mcmc-jags.sourceforge.net/).

H.4.2 Continuous outcomes

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.

More specifically, let y_{jk} be the outcome of interest in trial j and treatment k, and consider the model,

$$y_{jk} \sim N(\theta_{jk}, \sigma^2_{jk}), \tag{A38}$$

where,

$$\theta_{jk} = \begin{cases} \mu_{jb} & b = A, B, C, \ldots \text{ if } k = b \\ \mu_{jb} + \delta_{jkb} & \text{if } k \text{ alphabetically after } b. \end{cases} \tag{A39}$$

δ_{jkb} is modeled using a random effect with $d_{AA} = 0$,

$$\delta_{jkb} \sim N(d_{kb}, \sigma^2). \tag{A40}$$

As with the ACR response model, we allow treatment response to depend on patient characteristics by modeling d_{kb} as a function of covariates for each individual patient i,

$$d_{kb} = x_i^T \beta_{kb}, \tag{A41}$$

In the simulation, we allow for treatment effect modifiers by modeling d_{kb} as a function of covariates for each individual patient i,

$$d_{kb} = x_i^T \beta_{kb}. \tag{A42}$$

Treatment response is estimating by calculating A as in Equation A31. The response for treatment k is then,

$$\gamma(A + d_{kA}), \tag{A43}$$

where γ is defined as in Equation A35. Uninformative priors and the JAGS software package were used to derive posterior distributions.

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 A5, our results are similar and the NICE point estimates are generally within the 95% credible intervals surrounding our point estimates.

Table A5: A comparison of NICE and IVI estimates of ACR response probabilities

	IVI			NICE		
	ACR20	ACR50	ACR70	ACR20	ACR50	ACR70
cDMARDs	0.265 (0.248, 0.282)	0.102 (0.093, 0.112)	0.032 (0.028, 0.036)	0.298	0.123	0.042
ABT IV + MTX	0.555 (0.469, 0.643)	0.309 (0.235, 0.390)	0.140 (0.095, 0.195)	0.573	0.328	0.156
ADA + MTX	0.562 (0.482, 0.641)	0.315 (0.245, 0.391)	0.144 (0.102, 0.195)	0.615	0.368	0.183
ETN + MTX	0.646 (0.530, 0.751)	0.398 (0.285, 0.518)	0.202 (0.125, 0.293)	0.713	0.472	0.263
GOL + MTX	0.598 (0.432, 0.742)	0.352 (0.209, 0.507)	0.171 (0.083, 0.286)	0.642	0.395	0.202
IFX + MTX	0.655 (0.415, 0.864)	0.418 (0.196, 0.674)	0.224 (0.074, 0.450)	0.595	0.348	0.169
TCZ + MTX	0.555 (0.369, 0.739)	0.314 (0.163, 0.499)	0.146 (0.059, 0.278)	0.706	0.464	0.256
CZP + MTX	0.744 (0.546, 0.893)	0.516 (0.300, 0.722)	0.301 (0.135, 0.506)	0.564	0.319	0.150
ABT SC + MTX	0.563 (0.421, 0.691)	0.318 (0.200, 0.444)	0.148 (0.077, 0.236)	0.638	0.391	0.199
RTX + MTX	0.565 (0.413, 0.707)	0.321 (0.195, 0.460)	0.150 (0.075, 0.248)	0.573	0.328	0.156
TOF + MTX	0.608 (0.447, 0.756)	0.361 (0.217, 0.521)	0.177 (0.086, 0.297)	-	-	-

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.

Athanasakis, K., Tarantilis, F., Tsalapati, K., Konstantopoulou, T., Vritzali, E., and Kyriopoulos, J. (2015). Cost-utility analysis of tocilizumab monotherapy in first line versus standard of care for the treatment of rheumatoid arthritis in greece. *Rheumatology international*, 35(9):1489–1495.

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

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, 27(2):112–127.

Black, W. C. (1990). The ce plane: a graphic representation of cost-effectiveness. *Medical decision making*, 10(3):212–214.

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.

Brennan, A., Bansback, N., Reynolds, A., and Conway, P. (2003). Modelling the cost-effectiveness of etanercept in adults with rheumatoid arthritis in the uk. *Rheumatology*, 43(1):62–72.

Briggs, A. H., Claxton, K., and Sculpher, M. J. (2006). *Decision modelling for health economic evaluation*. Handbooks in Health Economic E.

Briggs, A. H. et al. (1999). A bayesian approach to stochastic cost-effectiveness analysis. *Health economics*, 8(3):257–261.

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.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., and Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*. Oxford university press.

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.

Espinoza, M. A., Manca, A., Claxton, K., and Sculpher, M. J. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making*, 34(8):951–964.

Fenwick, E., Claxton, K., and Sculpher, M. (2001). Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health economics*, 10(8):779–787.

Garber, A. M. and Phelps, C. E. (1997). Economic foundations of cost-effectiveness analysis. *Journal of health economics*, 16(1):1–31.

Garrison, L. P., Kamal-Bahl, S., and Towse, A. (2017). Toward a broader concept of value: identifying and defining elements for an expanded cost-effectiveness analysis. *Value in Health*, 20(2):213–216.

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.

Ioannidis, J. P. and Garber, A. M. (2011). Individualized cost-effectiveness analysis. *PLoS medicine*, 8(7):e1001058.

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.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Lakdawalla, D., Malani, A., and Reif, J. (2017). The insurance value of medical innovation. *Journal of public economics*, 145:94–102.

Lakdawalla, D. N., Romley, J. A., Sanchez, Y., Maclean, J. R., Penrod, J. R., and Philipson, T. (2012). How cancer patients value hope and the implications for cost-effectiveness assessments of high-cost cancer therapies. *Health Affairs*, 31(4):676–682.

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.

Mandel, M. (2013). Simulation-based confidence intervals for functions with complicated derivatives. *The American Statistician*, 67(2):76–81.

Meltzer, D. O., Smith, P. C., et al. (2011). Theoretical issues relevant to the economic evaluation of health technologies. *Handbook of health economics*, 2:433–469.

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.

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., Kuntz, K. M., Meltzer, D. O., Owens, D. K., Prosser, L. A., et al. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *Jama*, 316(10):1093–1103.

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.

Thokala, P., Devlin, N., Marsh, K., Baltussen, R., Boysen, M., Kalo, Z., Longrenn, T., Mussen, F., Peacock, S., Watkins, J., et al. (2016). Multiple criteria decision analysis for health care decision making-an introduction: report 1 of the ispor mcda emerging good practices task force. *Value in health*, 19(1):1–13.

Thokala, P. and Duenas, A. (2012). Multiple criteria decision analysis for health technology assessment. *Value in Health*, 15(8):1172–1181.

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

Van Hout, B. A., Al, M. J., Gordon, G. S., and Rutten, F. F. (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health economics*, 3(5):309–319.

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.