Evidence Review Group Report commissioned by the NIHR HTA Programme on behalf of NICE

Tofacitinib for moderately to severely active ulcerative colitis

ERRATUM

Replacement pages for factual inaccuracies in Evidence Review Group report

14 August 2018

Produced by Southampton Health Technology Assessments Centre (SHTAC)

tofacitinib groups with statistically significant differences between the 5mg and the 10mg tofacitinib arms versus placebo.

Remission, mucosal healing and sustained corticosteroid-free remission did not contribute data to the economic model.

Clinical remission is an outcome with an almost identical definition to the primary outcome of remission. The difference being that the rectal bleeding sub-score of the Mayo score does not have to be zero to achieve clinical remission. The outcomes of clinical remission and clinical response contribute data to the economic model.

Using locally read data (which were used in the base case economic evaluation) in OCTAVE 1, the mean difference between the tofacitinib group and the placebo group was 13.3 percentage points (95% CI 6.5 to 20.2, p=0.0017). The corresponding data for OCTAVE 2 were a mean difference from placebo of 15.6 percentage points (95% CI 9.9 to 21.3, p=0.0002). At week 52 in the OCTAVE Sustain maintenance trial the results for clinical remission also favoured tofacitinib (difference versus placebo 35.1%, 95% CI 26.7 to 43.5, p<0.0001 (10mg BID); 26.8%, 95% CI 18.5 to 35.1, p<0.0001 (5mg BID), both using locally read data).

Clinical response at both week 8 (OCTAVE Induction trials) and week 52 (OCTAVE Sustain trial) was also statistically significantly higher among participants who received tofacitinib.

Subgroup analyses according to prior TNFi-exposure status were conducted for the main clinical effectiveness outcomes. The results were consistent regardless of prior TNFi-exposure status.

HRQoL was reported using generic (EQ-5D and SF-36) and disease specific (IBDQ and WPAI-UC) instruments. HRQoL was typically improved by tofacitinib treatment however for some HRQoL measures the ERG was uncertain about the impact of missing data. Data from the EQ-5D-3L did not inform the base-case economic model but were included in a scenario analysis.

Safety data for tofacitinib in patients with moderate to severely active ulcerative colitis comes from the Phase II trial, the three Phase III OCTAVE trials and the ongoing OCTAVE Open extension study. Rates of adverse events of any type were broadly similar for the tofacitinib and placebo arms within OCTAVE Induction 1 and 2 and OCTAVE Sustain with serious adverse events affecting fewer than 10% of patients. Ulcerative colitis was the most frequent serious adverse event and most other serious adverse events were related to ulcerative colitis. Serious infections were uncommon (data on serious infections were included in the economic model). Overall, and in comparison with evidence from the use of tofacitinib in patients with rheumatoid arthritis, no new safety signals were identified.

There are no head-to-head RCTs of tofacitinib versus the comparators defined in the company's decision problem. Therefore the company used NMA to estimate the relative effectiveness and safety of tofacitinib in both the induction and maintenance phases of treatment in comparison to TNF-alpha inhibitors (infliximab, adalimumab and golimumab), vedolizumab and conventional therapies. The company's systematic review identified 21 RCTs that were considered for inclusion in the NMA. Four of these were the tofacitinib RCTs listed above, a further 14 were included in one or more NMA networks and three studies could not be included in any of the NMA networks.

Table 1 NMAs conducted by the company

	TNFi-naïve population subgroup	**TNFi-exposed population subgroup**
Induction phase	Clinical response and clinical remission Mucosal healing	Clinical response and clinical remission Mucosal healing
	Safety outcomes (discontinuation due to AEs, SAEs, serious infections)	
Maintenance phase	Clinical response and clinical remission Mucosal healing	Clinical response and clinical remission Mucosal healing

The ERG judged the NMAs to be generally well conducted but identified nine issues:

- Use of the probit scale to model clinical response/clinical remission is an improvement on a previous approach in NICE guidance TA342 but a multinomial logit model could have been considered.
- Potential inconsistency in a closed loop of the maintenance TNFi-naïve network was not examined

Table 3 Cost effectiveness: Company base case, with prior TNFi (with tofacitinib PAS)

Strategy	Total		Incremental analysis			Pairwise ICERs tofacitinib vs. comparator (£/QALY)
	QALYs	Costs (£)	QALYs	Costs (£)	ICER (£/QALY)	
Conventional	[redacted]	[redacted]	[redacted]	[redacted]	-	£10,302
Tofacitinib	[redacted]	[redacted]	[redacted]	[redacted]	£10,302	-
Vedolizumab	[redacted]	[redacted]	[redacted]	[redacted]	£7,838,238	£7,838,238

A range of uncertainty analyses were conducted by the company, but they have been selective in the scenarios they present

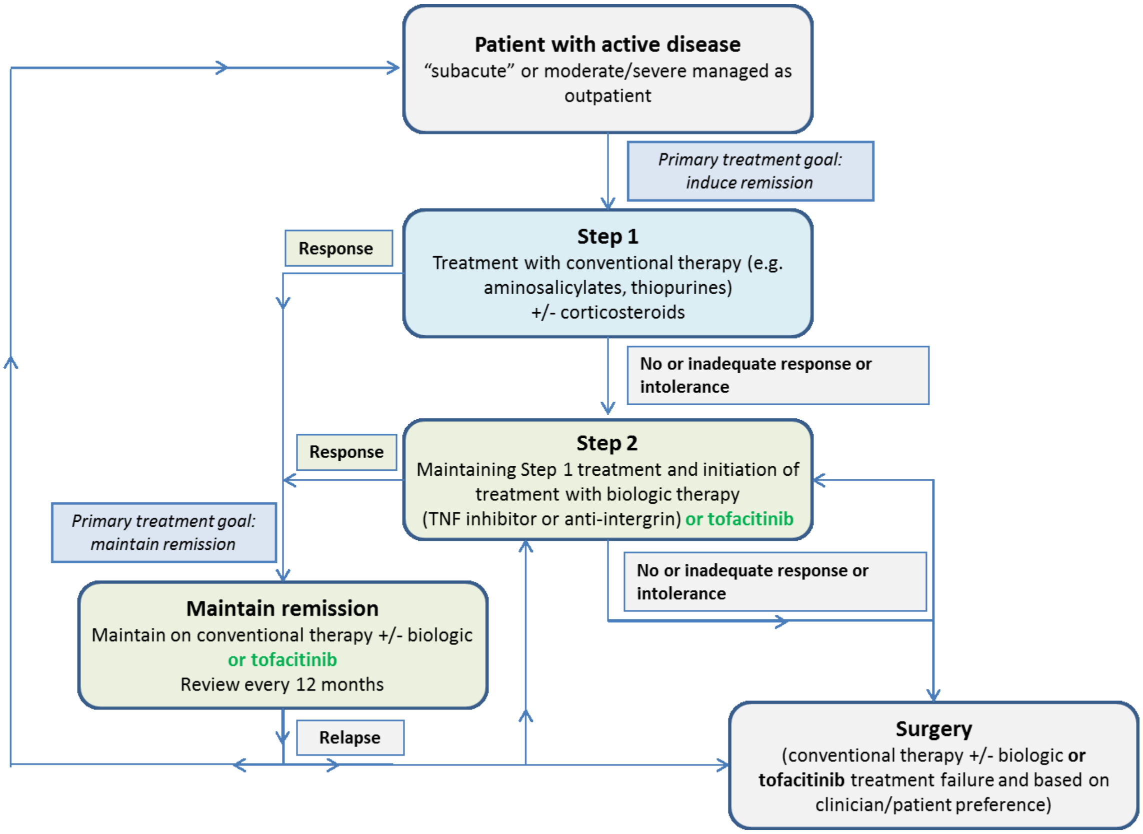
The company performed a range of deterministic-, probabilitistic- and scenario analyses to assess the methodological as well as parameter uncertainty of their base case analyses. The ERG agrees with their assumptions for DSA and PSA and their results, in general. However, we identified errors in the scenarios relating the use of central read NMA results and tofacitinib maintenance using [redacted] split. The company corrected the error in the latter scenario in their response to clarification question. For the scenario analyses, we view that the company has been selective in the scenarios they present.

Commentary on the robustness of submitted evidence

Strengths

- The model structure is consistent and follows the conventional design for ulcerative colitis appraisals.
- The model generally adheres to the NICE scope for this appraisal.
- The perspective of the analysis aligns with the NICE guide to the methods of Technology Appraisal.
- The model uses a lifetime time horizon to allow estimation of all relevant costs and quantity of life impairment.
- The model uses appropriate sources for costs and resource use and in line with other technology appraisals
- The model allows the flexibility to incorporate treatment sequencing which provides a closer reflection of clinical practice.
- The ERG agrees with the company's approach to modelling surgery and its related risks, source of costs and utilities for the base case and mortality.

history and clinical decision making on the appropriateness of therapies, and therefore may not adequately capture the nuances of clinical practice when comparing to the NICE scope" (clarification response A3). In their clarification response the company provided a simplified version of CS Figure 1 in order to better represent the position of tofacitinib in the treatment pathway in relation to the NICE scope (reproduced in Figure 1).


Patient with active disease
"subacute" or moderate/severe managed as outpatient
Primary treatment goal: induce remission
Response
Step 1
Treatment with conventional therapy (e.g. aminosalicylates, thiopurines) +/- corticosteroids
No or inadequate response or intolerance
Response
Step 2
Maintaining Step 1 treatment and initiation of treatment with biologic therapy (TNF inhibitor or anti-intergrin) or tofacitinib
Primary treatment goal: maintain remission
Maintain remission
Maintain on conventional therapy +/- biologic or tofacitinib
Review every 12 months
No or inadequate response or intolerance
Relapse
Surgery
(conventional therapy +/- biologic or tofacitinib treatment failure and based on clinician/patient preference)

Source: company's clarification response A3

Figure 1 Proposed position of tofacitinib within the treatment pathway

Outcomes

The outcomes included in the CS are clinically meaningful and are consistent with the NICE scope and EMA guidance on methods for clinical trials in ulcerative colitis.[12] The primary outcome in the phase 3 OCTAVE trials was remission whilst the primary outcome in the phase 2 trial was clinical response. HRQoL was a secondary outcome in all the tofacitinib trials, and mucosal healing was a secondary outcome in the phase 3 trials. Details of the outcome selection are discussed further below in section 3.1.4. In summary, the key issues noted by the ERG are:

In all four RCTs the comparator was placebo. All these RCTs were used in support of the company's application for a marketing authorisation and were sponsored by Pfizer, the manufacturer of tofacitinib.

The Phase II trial is not described in detail in the CS but it is included in the company's NMA (CS section B.2.9) and data from this trial are also included in the adverse events section (CS Appendix F Table 166). As the Phase II trial was a small dose-finding study with 194 patients, of whom only 33 received the licensed 10 mg BID dose (company clarification response A16), the CS focuses on the Phase III trials. The ERG agrees that this is reasonable and accordingly the current ERG report also focuses primarily on the Phase III trials.

It was unclear to the ERG from the description of the Phase II trial population reported both in the CS and in the trial publication whether this matched the NICE scope. The company confirmed that it does match the scope, as "*patients were only included if they continued to have moderate to severe disease despite previous treatment*" (clarification response A2). In addition, the company provided a table detailing the failed drug treatments at baseline (clarification response Table 1) and full details of the inclusion and exclusion criteria (clarification response Appendix A).

The number of centres in the studies ranged from 51 (Phase II trial) to 297 (OCTAVE Sustain), but it should be noted that a number of centres in the Phase III trials randomised just one patient (16 centres in OCTAVE 1; 25 centres in OCTAVE 2; and 66 centres in OCTAVE Sustain[19]). While each study included some patients from the UK, this number was low [redacted].

OCTAVE 1 and 2 were double-blind, randomised placebo-controlled tofacitinib induction trials with an 8 week treatment phase, and used identical methods (see Table 6).

In addition to the criteria listed above, patients had to have moderately to severely active disease (6 to 12 on the Mayo score, with a rectal bleeding sub-score of 1 to 3 and an endoscopic sub-score of 2 or 3). Prohibited therapies included TNFi therapies within 8 weeks of baseline; azathioprine, methotrexate, and 6-mercaptopurine within 2 weeks; anti-adhesion-molecule therapy taken within 1 year; and ciclosporin and intravenous corticosteroids (CS Tables 9 and 10). Permitted concomitant medications for ulcerative colitis included oral

Table 6 Summary characteristics of tofacitinib RCTs

Phase II trial[11] **(efficacy/dose RCT)**		**OCTAVE 1[19]** **(induction RCT)**		**OCTAVE 2[19]** **(induction RCT)**		**OCTAVE Sustain[19]** **(maintenance RCT)**		**OCTAVE Open[20]** **(extension study)**
Tofacitinib 0.5 mg (n=31) 3 mg BID (n=33) 10 mg BID (n=33) 15 mg BID (n=49)	Placebo (n=48)	Tofacitinib 10 mg BID (n=476)[a]	Placebo (n=122)	Tofacitinib 10 mg BID (n=429)[a]	Placebo (n=112)	Tofacitinib 10 mg BID (n=197) 5 mg BID (n=198)	Placebo (n=198)	Tofacitinib[b] 10 mg BID ([redacted]) 5 mg BID ([redacted])
Design: randomised, double-blind, placebo-controlled trial (2:2:2:3:3 ratio tofacitinib 0.5 mg: 3mg: 10 mg: 15 mg: placebo)		*Design:* identical randomised, double-blind, placebo-controlled trials (4:1 ratio tofacitinib: placebo, stratified according to previous treatment with TNFi therapies, glucocorticoid use at baseline, and geographic region)				*Design*: randomised, double-blind, placebo-controlled trial (1:1:1 ratio tofacitinib 5 mg: tofacitinib 10 mg; placebo)		*Design:* open-label extension
Location: 51 sites worldwide (UK = 2, [redacted][d])		*Location:* 144 sites worldwide (UK = 2, [redacted])		*Location:* 169 sites worldwide (UK = 3, [redacted])		*Location:* 297 sites worldwide (UK = 5, [redacted])		*Location:* 215 sites worldwide (UK = 5)
Inclusion: • age ≥18 years • confirmed diagnosis of UC for ≥3 months		*Inclusion:* • age ≥18 years • confirmed diagnosis of UC for ≥4 months				*Inclusion:* • entry criteria for the Induction trials • completed 8 weeks induction therapy		*Inclusion:* • completed or demonstrated treatment failure in

		2.5 mg per week until the dose was 0 mg).	
Length of follow-up: 8 weeks of treatment and 4 weeks follow-up	*Length of follow-up:* 9 weeks (primary efficacy endpoint at 8 weeks)	*Length of follow-up:* 53 weeks (primary efficacy endpoint at 52 weeks)	*Length of follow-up:* up to 6 years (12-month interim results reported)

Sources Sandborn et al.[11], CS Table 7, 8, 9, 10, 13 and 14, and B.2.6.3.1

BID, twice daily; NR, not reported; UC, Ulcerative colitis.

[a] 15 mg BID tofacitinib treatment was discontinued based on feedback from regulatory authorities (OCTAVE 1: n=38, OCTAVE 2: n=18)

[b] Three subpopulations received tofacitinib 10 mg ([redacted]) in the open label extension study: Induction non-responders tofacitinib 10 mg [redacted]; maintenance completers tofacitinib 10 mg [redacted], maintenance treatment failures 10 mg ([redacted] comprising participants from OCTAVE Induction 1 and 2 who withdrew from OCTAVE Sustain due to treatment failure on tofacitinib (5 mg, [redacted]; 10 mg,[redacted];) or placebo ([redacted]). One subpopulation received tofacitinib 5 mg in the open label extension study: Maintenance: remission tofacitinib 5 mg [redacted]. Note that there appears to be a typographical error in CS Table 8 where the number of patients receiving tofacitinib 10 mg is given as n=[redacted].

[d] [redacted]

3.1.5 Description and critique of company's outcome selection

The outcomes included in the CS match those in the NICE final scope and appear appropriate. However, time to surgical intervention, although specified in the NICE final scope, was not included, as this was not assessed in the OCTAVE trials.

In clinical trials of therapies for ulcerative colitis the Mayo Score is widely used and was used within the OCTAVE trials (CS Section B1.3.1 and CS Table 3). There are four components to the Mayo score, one of which is 'Endoscopic findings'. In the OCTAVE trials the Mayo endoscopic sub-score was assessed both locally (by the study site investigator) and centrally (from a video recording). Consequently the outcomes in the CS that utilise the endoscopic sub-score were reported separately using the local or the central read of the endoscopic data. The ERG notes that the FDA[22] state that central reading is the preferred approach and the OCTAVE clinical trial programme is the first in ulcerative colitis to use central reads (CS Section B.2.3.1.2.4).

The primary outcome in OCTAVE 1 and 2 was remission at week 8 based on centrally read endoscopic Mayo sub-scores, and at week 52 in OCTAVE Sustain (for definition of remission see Table 10). Higher Mayo scores indicate more severe disease. The company also defined key secondary outcomes: mucosal healing (OCTAVE 1 and 2: week 8; OCTAVE Sustain: week 52), and for OCTAVE Sustain only, sustained corticosteroid-free remission among patients in remission at baseline (measured at weeks 24 and 52). Mucosal healing is associated with lower rates of hospitalisation and surgery,[23] while the use of corticosteroids long-term is not suitable due to side effects so a corticosteroid-free remission is important.[24]

Clinical response and clinical remission based on Mayo sores (for definitions see Table 10) were reported for all three trials (OCTAVE 1 and 2: week 8; OCTAVE Sustain: week 52). As can be observed from Table 10 the difference between the primary outcome of remission and the secondary outcome of clinical remission is that for the former the rectal bleeding sub-score must be zero whereas this is not necessary for the outcome of clinical remission. Clinical response and clinical remission were the only clinical effectiveness outcomes included in the economic model (the primary outcome did not contribute to the economic model), as they were thought to ensure comparability with trials of biological therapies for ulcerative colitis.

which has a recall period of 1 week (in OCTAVE 1 and 2 assessed at baseline and week 8; in OCTAVE Sustain assessed at baseline and weeks 24 and 52). Higher scores indicate better HRQoL. A systematic review[30] of the SF-36 in patients with ulcerative colitis suggests that a group-level clinically important difference threshold of 3 points for both summary scores and responder-level thresholds of 3.1 for PCS and 3.8 for MCS based on the SF-36v2 manual.[31]

- The WPAI-UC score, based on a 6-item questionnaire (version 2) assessing work productivity, is also reported by all three OCTAVE RCTs (OCTAVE 1 and 2 at baseline and week 8; OCTAVE Sustain at baseline and week 52). The questionnaire yields four scores expressed as impairment percentages: absenteeism; presenteeism; work productivity loss; non-work activity impairment. A higher score indicates greater impairment.[32] As part of the response to NICE and the ERG's clarification question A12, the company states that it is not aware of any validated MCID for this outcome in patients with ulcerative colitis. However the company also state that extrapolating from Crohn's Disease suggests a 7% decrease is the MCID for the WPAI.[33,34]

Table 10 Clinical effectiveness outcomes and outcome definitions of the OCTAVE RCTs

Outcome	**Definition**	**When assessed, week**		**Used in Model**
		OCTAVE 1 & 2	**OCTAVE Sustain**	
Primary: Remission based on centrally-read endoscopic sub-scores	Mayo score ≤2, no individual sub-score >1, rectal bleeding sub-score = 0	8	52	No
Key secondary: Mucosal healing	Mayo endoscopic sub-score ≤1	8	52	No
Key secondary: Sustained corticosteroid-free remission among patients in remission at baseline	Remission (as defined above for the primary outcome) plus no treatment with steroids for ≥4 weeks before the 24-week and 52-week visits	Not assessed	24, 52	No
Clinical response	Mayo score decrease from baseline ≥ 3, and ≥ 30%, with a decrease in rectal	Week 8	52	Yes

Table 13 Company choice of base-case and ERG preference

	Company base-case model	ERG favoured model
Clinical response/clinical remission, Induction TNFi-naive	Random effects	Random effects
Clinical response/clinical remission, Induction TNFi-exposed	Fixed effects	Random effects
Clinical response/clinical remission, Maintenance TNFi-naive	Fixed effects	Random effects
Clinical response/clinical remission, Maintenance TNFi-exposed	Fixed effects	Fixed effects
Serious infections, Induction	Random effects	Fixed effects

In the induction phase TNFi-exposed subgroup, the fixed effects model was preferred despite similar DIC and similar total residual deviance. The ERG would have selected the random effects model as the more conservative analysis. Whilst the base case models are presented in the main NMA results (CS Table 25) the alternative model is not reported. We would prefer to have seen this explored as a sensitivity analysis.

Similarly, the company preferred the fixed effects model in the maintenance phase TNFi-naïve population for clinical response/remission as it deemed the random effect results implausibly imprecise because no treatment was predicted to be significantly better than placebo. The ERG would have chosen the random effects model for both the lower DIC and total residual deviance. The ERG would prefer to have seen this explored as a sensitivity analysis.

Finally, the company chose the random effects model for serious infections. In response to a clarification request the company provided the random effect standard deviation (1.82, 95%CrI 0.15, 4.59) (clarification question A22). This wide CrI indicates weak support for the random effects model which has a similar DIC, thus we might have favoured the fixed effects model. The ERG would prefer to have seen the fixed effects model included in a sensitivity analysis.

Table 14 and Table 15 show the results of the ERG validation and exploratory analysis for the response and remission analyses. The ERG ran the same number of chains, burn-in and simulations reported by the company (section D.1.3.3). Models converged and our results concur to two decimal places.

The alternative choice random effects models show wider credible intervals and some variation in the median estimates for adalimumab and golimumab in the maintenance analysis for the TNFi-naïve population as smaller studies are given more weight under the random effects than the fixed effects model.

Table 14 ERG replication and additional analysis on model choice - clinical response and clinical remission for TNFi-naïve subgroup

Comparator	Treatment effect vs placebo, median (95% CrI), probit scale[a]		
	Company base-case (fixed effects)	ERG replication of base-case (fixed effects)	ERG alternative model selection (random effects)
Maintenance phase			
Tofacitinib 5 mg	[redacted]	[redacted]	[redacted]
Tofacitinib 10 mg	[redacted]	[redacted]	[redacted]
Infliximab 5 mg/kg	[redacted]	[redacted]	[redacted]
Adalumimab 40 mg Q2W	[redacted]	[redacted]	[redacted]
Golimumab 50 mg	[redacted]	[redacted]	[redacted]

Source of company base-case (fixed effects) is CS Table 26

[a] On the probit scale, negative coefficients indicate improvement over placebo. Where the upper and lower CrI are both negative, treatments show strong evidence of benefit versus placebo.

Table 15 ERG replication and additional analysis on model choice - clinical response and clinical remission for TNFi-exposed subgroup

Comparator	Treatment effect vs placebo, median (95% CrI), probit scale[a]		
	Company base-case (fixed effects)	ERG replication of base-case (fixed effects)	ERG alternative model selection (random effects)
Induction phase			
Tofacitinib 10 mg	[redacted]	[redacted]	[redacted]
Adalumimab 160/80/40 mg	[redacted]	[redacted]	[redacted]

Vedolizumab 300 mg	[redacted]	[redacted]	[redacted]

Source of company base-case (fixed effects) is CS Table 25

[a] On the probit scale, negative coefficients indicate improvement over placebo. Where the upper and lower CrI are both negative, treatments show strong evidence of benefit versus placebo.

However, when we attempted to replicate the serious infections results there was a higher level of uncertainty around the coefficients particularly for tofacitinib (Table 16). The wider credible intervals persisted under the fixed effects model conducted by the ERG.

Table 16 ERG replication and additional analysis on model choice - serious infections

Comparator	Treatment effect vs placebo, median (95% CrI), logit scale		
	Company base-case (random effects)	ERG replication of base-case (random effects)	ERG alternative model selection (fixed effects)
Tofacitinib 10 mg	[redacted]	[redacted]	[redacted]
Infliximab 10 mg/kg	[redacted]	[redacted]	[redacted]
Adalumimab 160/80/40 mg	[redacted]	[redacted]	[redacted]
Golimumab 200/100 mg	[redacted]	[redacted]	[redacted]
Vedlizumab 300 mg	[redacted]	[redacted]	[redacted]
Azathioprine	[redacted]	[redacted]	[redacted]

Source of company base-case (fixed effects) is CS Table 34

The very wide credible intervals for tofacitinib are caused by the lack of any serious infections across placebo arms in the three tofacitinib studies, hence the difficulty to estimate a relative treatment effect compared to placebo (Table 17). There was also considerable autocorrelation in the tofacitinib coefficient despite thinning and running an extended number of simulations.

The reasons for the difference in our results are unclear, particularly how the company arrived at their estimate for tofacitinib.

case which combined TNFi-exposed data for tofacitinib and adalimumab with TNFi-failure data for vedolizumab. Our scenario analysis at least included comparable data for tofacitinib and vedolizumab.

In the event, as Table 18 shows, use of TNFi-failure data makes little difference to the response/remission results for tofacitinib.

Table 18 ERG scenario analysis using TNFi-failure data from both OCTAVE Sustain and GEMINI 1

Comparator	Treatment effect vs placebo, median (95% CrI), probit scale[a]		
	Company base-case (fixed effects)	ERG replication of base-case (fixed effects)	ERG exploratory scenario analysis (fixed effects)
Maintenance phase			
Tofacitinib 5 mg	[redacted]	[redacted]	[redacted]
Tofacitinib 10 mg	[redacted]	[redacted]	[redacted]
Adalumimab 40 mg Q2W	[redacted]	[redacted]	[redacted]
Vedolizumab 300 mg Q8W	[redacted]	[redacted]	[redacted]
Vedolizumab 300 mg Q4W	[redacted]	[redacted]	[redacted]

Source of company base-case (fixed effects) is CS Table 28

[a] on the probit scale, negative coefficients indicate improvement over placebo. Where the upper and lower CrI are both negative, treatments show strong evidence of benefit versus placebo.

3.1.7.4 Baseline response models – uncertainty around absolute probabilities

To estimate absolute probabilities of each event, treatment effects from the NMA were combined with an estimate of the placebo (baseline) response from the placebo arms of included studies. In response to clarification request A17 the company provided the data, priors and output (meanA, precA) in WinBUGs code format for the probit baseline models. We were able to replicate selected median estimates for the baseline calculations. However, despite running the CS code [validated against NICE DSU Technical Support Document (TSD) 2[46]] and data we were unable to replicate the baseline credible intervals used in the probit or logit models. The company models tended to lead to wider credible intervals compared to our calculations, thus would lead to conservative results. A summary of the differences in our findings is provided in Table 19 below.

Table 19 ERG replication of baseline (placebo) response results

Comparator	**Treatment effect vs placebo, median (95% CrI)**	
	Company baseline	**ERG replication of company baseline**
Induction TNFi-exposed, probit scale		
Response/remission	[redacted]	[redacted]
Maintenance TNFi-naïve, probit scale		
Response/remission	[redacted]	[redacted]
Induction, logit scale		
Serious Infections	[redacted]	[redacted]
Serious adverse events	[redacted]	[redacted]

3.1.7.5 Inclusion of the tofacitinib phase II trial

The Sandborn 2012 Phase II (induction) tofacitinib trial[11] is less well described in the CS despite being included in the NMAs. Furthermore, the company state:
All studies, except for one [Sandborn 2012], were conducted in patients with moderately to severely active ulcerative colitis who had an inadequate response to or had failed to tolerate one or more of the following conventional therapies: oral or intravenous corticosteroids, azathioprine, and/or 6-mercaptopurine (CS section B.2.9.1.1).

The ERG thus questioned the eligibility of this trial. The company confirmed that the Phase II trial met the inclusion criteria for the NMA and they also provided selected NMA results obtained with the Phase II trial excluded from the NMA (Table 7 in clarification response A16). These results for response and remission for the TNFi-naïve and TNFi-exposed populations in the induction period were similar to the base case (CS Table 25).

Base case results without the Phase II trial were not provided for the safety outcomes. However, given the relatively high serious infection rate in the tofacitinib arms of the Phase II trial compared to the OCTAVE trials (6% [2/33] patients had an event compared to 1% [6/476] in OCTAVE Induction 1 and none in OCTAVE Induction 2), the Phase II trial may healing at week 52 and sustained mucosal healing at weeks 24 and 52 were reported for the 5 mg and 10 mg tofacitinib maintenance doses in comparison to the placebo arm of the trial.

Sustained corticosteroid-free remission among those in remission at baseline (a further key secondary outcome) in the OCTAVE Sustain trial, was statistically significantly greater in the tofacitinib 5 mg and 10 mg arms than in the placebo arm.

Clinical remission, which has a very similar definition to the primary outcome of remission, contributed data to the economic model via the NMA. Due to the similarity of outcome definition the results from the OCTAVE trials were almost identical to those reported above for remission, favouring tofacitinib.

The outcome of clinical response also contributes data to the economic analysis via NMA. The percentage difference between the tofacitinib group and the placebo group in favour of tofacitinib was statistically significant in both OCTAVE induction trials and the OCTAVE Sustain maintenance trial and for both the central and locally read data.

HRQoL was reported using both generic (EQ-5D and SF-36) and disease specific (IBDQ and WPAI-UC) instruments. Results showed HRQoL was typically improved by tofacitinib treatment; however, for some HRQoL measures we are uncertain about the impact of the missing data. Data from the EQ-5D-3L do not inform the base-case economic model but were included in a scenario analysis.

Subgroup analyses focused on results according to prior TNFi-exposure. Note that this is a more restricted subgroup than that of prior biologic therapy (which would also include other biological therapies such as vedolizumab) which is listed in the NICE scope. The OCTAVE trials were not powered to test the statistical significance of subgroup analyses so the results should be interpreted cautiously. Overall, the results were consistent regardless of prior TNFi-exposure status.

Safety data for tofacitinib in patients with moderate to severely active ulcerative colitis comes from the Phase II tofacitinib trial, the two OCTAVE Induction trials, the OCTAVE Sustain trial and the ongoing OCTAVE Open extension study. In total tofacitinib has been evaluated in 1157 patients with ulcerative colitis with a maximum exposure to tofacitinib of 4.4 years.

Rates of adverse events of any type were broadly similar for the tofacitinib and placebo arms within OCTAVE Induction 1 and 2 and OCTAVE Sustain. Serious adverse events

For comparison, the median age at diagnosis of ulcerative colitis in the 2016 RCP audit was 32 years (interquartile range (IQR) 24 to 45) and the median age at initiation of biologic treatment was 39 years (IQR 28 to 52).[64] The gender distribution in the audit was 59% males (529/903), similar to that in the OCTAVE trials.

We consider that the gender mix, initial age and weight of the model cohort should be assumed similar for people with and without prior exposure to TNFi drugs. In ERG analysis, we assume 59% males, initial age 41 years and weight 73.5 kg, based on means for both arms in the OCTAVE Induction trials. We conduct scenario analysis to assess the impact of age (28 to 58) and body weight (range 70 kg to 80 kg) on the results.

4.3.2.2 Comparators

The model assumes that patients start treatment with tofacitinib or the biologic comparators with an induction phase of treatment. Patients who respond during induction continue to receive maintenance treatment with the same drug (with concomitant use of conventional drugs) until loss of response or an acute exacerbation requiring surgery. Patients who do not respond to induction treatment and those who relapse during maintenance continue to receive conventional treatment alone, until planned or emergency surgery, or death.

Inclusion of comparators in economic analysis

Tables 40 and 41 in the CS (page 130) outline the comparators used in the company's economic analysis:

- **TNFi-naïve subgroup**, all comparators specified in the scope (infliximab, adalimumab, golimumab, vedolizumab, tofacitinib and conventional therapy (CT));
- **TNFi-exposed subgroup**, only vedolizumab, tofacitinib and CT are included. Cost-effectiveness is not reported for infliximab, adalimumab or golimumab.

For patients with prior exposure to TNFi drugs, infliximab and golimumab could not be included in the company's NMA due to a lack of trial evidence (CS section B.2.9.2.1). However, the TNFi-exposed NMA does include adalimumab, so the company could have included adalimumab in the cost-effectiveness analysis for this subgroup. The CS does not give a clear rationale for omitting adalimumab for the TNFi-exposed subgroup.

Clinical experts have advised the ERG that treatment with a TNFi would sometimes be considered for a patient with prior exposure to another TNFi. There is a group of patients

Stopping rules for drug treatment

- *Discontinuation due to lack of response to induction therapy*
  CS Table 38 summarises SmPC recommendations about when to stop tofacitinib and biologic drug treatment. These recommendations relate to early assessment of response following induction treatment (from 2 to 16 weeks after initiation). In contrast, the clinical trials provide evidence of response at 6 weeks for golimumab and vedolizumab and at 8 weeks for other comparators, and the model assumes a fixed 8-week induction period followed by immediate cessation of treatment for patients whose disease does not show a response in this time. *If in practice clinicians assess response to induction later than 8 weeks, the average cost of induction therapy will be higher than that estimated by the company model. However, effectiveness may also be higher if some patients have a late response to induction. The direction and magnitude of the bias from assuming a fixed 8-week period of induction for all comparators is unclear.*

- *Discontinuation due to loss of response during maintenance*
  Guidance for the TNF-alpha inhibitors (TA329) and vedolizumab (TA342) recommend annual assessment of response, with treatment continuing only if there is clear evidence of ongoing benefit. Clinical advice to the ERG is that the benefit of biologic treatment is usually considered annually, in line with NICE guidance. However, treatment would usually be withdrawn earlier for patients who lose response, as the patient will seek an appointment when symptoms recur or get worse and this will trigger consideration of changing or stopping treatment.

  The company model applies a constant risk of relapse across each 8-week cycle of maintenance, with treatment stopping immediately when patients lose response. Thus, it assumes that maintenance treatment is stopped within 8 weeks of a loss of response. To achieve this, all patients on maintenance treatment must have fast access to clinical assessment on relapse or be seen routinely every 8 weeks. The company model assumes an average of 2 outpatient visits for patients in remission on maintenance treatment and 4.5 visits per year for patients with a response but no remission.

  The ERG considers that the assumption that treatment will be withdrawn following relapse reflects UK practice. However, we have concerns that the costs of monitoring and follow-up in the company's model do not reflect the full cost of ensuring that treatment can be withdrawn within 8 weeks of a relapse. We consider a scenario with

- **Choice of fixed effects versus random effects**
  The company state that their choice of NMA models was based on model fit statistics. For the induction phase the results and model fit for the fixed and random effects models were comparable for both patients subgroups. In the TNFi-naïve subgroup the model fit diagnostics were slightly better for the random effects model so this was preferred. For the TNFi-exposed subgroup they preferred the simpler fixed effect approach because the DIC statistics were similar (CS B.2.9.2.1.1). In the maintenance phase the fixed effect models were preferred because the company deemed the random effects results implausibly imprecise with no treatment predicted to be significantly better than placebo. Table 56 below summarises the NMA models chosen for the company base case analysis.

Table 56 Selection of response/remission NMA models

	Patient subgroup	**Induction**	**Maintenance**
Company base case	TNFi-naive	Random effects	Fixed effects
	TNFi-exposed	Fixed effects	Fixed effects
ERG preference	TNFi-naive	Random effects	Random effects
	TNFi-exposed	Random effects	Fixed effects *

* Random effects model would not run for the maintenance NMA

The ERG has a general preference for the random effect NMA models, as we believe that the fixed effect models may underestimate uncertainty due to heterogeneity between the studies. We test the impact of different NMA models on cost-effectiveness results in section 4.4.3 below.

- **Combination of TNFi-failed and TNFi-exposed subgroups**
  The base case NMAs combine outcomes for subgroups defined as TNFi-failed for vedolizumab with TNFi-exposed subgroups for tofacitinib and adalimumab (CS Table 22). The company conducted a sensitivity analysis for the TNFi-failure subgroup, which reduced the probit score for tofacitinib by -0.13 in the induction phase, bringing it closer to vedolizumab. (CS Table 28). They reported that results were not available for adalimumab and that the analysis could not be run for the maintenance phase. Therefore, the TNFi-failure NMA sensitivity analysis does not provide the required input parameters and was not used in the economic model.

4.3.8 Model validation

The company describes their approach to model validations in CS section B.3.10. They state that they engaged UK clinical experts, statisticians and health economists to validate model inputs and assumptions in a UK advisory board meeting. Further details on the key aspects of validation are summarised in CS Table 78.

The CS stated that clinical experts validated model methods pertaining to the patient population; subgroup analysis by prior TNFi-exposure; time on treatment and discontinuation rates; costs (including monitoring cost for tofacitinib, health state costs and resource use, including rate of hospitalisation); emergency surgery; quality of life and maintenance dose of tofacitinib. The experts are reported to agree with the company's assumptions in most of these aspects, except for:

- **Patient population:** Although the baseline characteristics of the patient population in OCTAVE reflect UK practice, the duration of disease in OCTAVE trials (which was 6-7 years) is longer than that in clinical practice (which is ~2-4 years).
- **Health state unit costs and resource use, including rate of hospitalisation:** Tsai et al. was confirmed to reflect an accurate representation of unit costs and resource use as per clinical practice. However, the experts suggested that the model base-case assumptions relating to annual medical resource use (CS Table 55) underestimated the resource use per patient per year.
- **Tofacitinib maintenance dose:** Experts observed that the company assumption relating to [redacted] of patients benefitting from maintenance dose of 10mg twice daily may not be limited to patients in the TNFi-exposed group only.

The economic model was quality checked by health economists. For face validity, the company compared the proportion of patients in response and remission predicted by the model against the estimated values from the NMA, shown below in Figure 9.

Further, the model results were compared with previous TA329; however, the CS did not report any comparison of the results in TA329 with those in the current appraisal. We discuss this in detail in section 4.4.1. For internal validity, the CS stated that a second modeller reviewed the model; conducted extreme value tests alongside inspecting model code, formulae and references. An independent health economist was reported to have reviewed the model structure, parameter inputs and core model assumptions.

Table 74 Company scenario analyses

Company scenarios	Brief rationale/assumption	ICERs for Tofacitinib vs CT (£/QALY)	
		TNFi-naïve	TNFi-exposed
Base case		**£8,554**	**£10,302**
Overall ITT population			£7,805
Tofacitinib maintenance dose mix	[redacted] of patients receiving 5mg; [redacted] of patients receiving 10mg	£12,628	£13,947
Fixed baseline utility instead of age-adjusted	Assumption that patient quality of life stays constant over time.	£8,760	£10,589
OCTAVE trial utilities	EQ-5D data were collected in Tofacitinib Phase III clinical trials	£15,508	£18,276
Swinburn utilities	To compare with previous analyses	£11,932	£14,487
Emergency surgery from any state	Due to the uncertainty on the likely protection from acute events based on the level of response/remission, patients are assumed to undergo emergency surgery regardless of state membership	£8,194	£9,962
Emergency surgery only from active UC	As above but assuming response to treatment offers the same level of protection from acute events, as remission	£8,652	£10,475
No emergency surgery	As above, but assuming no emergency surgery in the model	£8,710	£10,593
Central read NMA results	Central read was the primary endpoint in OCTAVE trials.	£9,469	£10,793
Discounting every cycle	It tested the sensitivity of the model when the discounting of outcomes is applied every 8 weeks.	£8,606	£10,398
Adalimumab maintenance 73% 40 mg Q2W and 27% 40 mg QW	Dose escalation of adalimumab was assumed in Archer *et al*.	£8,554	--
Golimumab 100 mg every 4 weeks in maintenance	A 100 mg Q4W maintenance dose was assessed as part of the clinical trials and is recommended for consideration in some patients, such as those who have experienced a decrease in their response	£8,554	--
Vedolizumab 300 mg every 4 weeks in maintenance	A 300 mg Q4W maintenance dose was assessed as part of the clinical trials and is recommended for consideration in some patients who have a body weight ≥ 80 kg	£8,554	Dominated

Source: CS Table 63 to 66; 71 to 77

4.4 Additional work undertaken by the ERG

4.4.1 ERG model validation

4.4.1.1 Model verification procedures

We checked the economic model for transparency and validity. The visual basic code used within the model was accessible. The NMA code in WinBUGs was provided in Appendix D.1.3.4.

We conducted a range of 'white box' tests to verify model inputs, calculations and outputs:

- Cross-checking of all parameter inputs against values in the CS and cited sources;
- Checking the individual equations within the model;
- A range of extreme value and logic tests to check the plausibility of changes in results when parameters are changed
- Checking the VBA code for treatment sequencing
- Checking all model outputs against results cited in the CS, including the base case, PSA, DSA and manually ran all the scenarios
- Running the NMA code in WinBUGs to replicate selected results (see section 3.1.7).

In addition, we checked the model calculations of patient transitions through the health states, costs and QALYs by re-coding the model independently based on the inputs from the company's submitted model.

Overall, we found the economic model to be of a good quality, with very few errors in input parameters, logic or coding. We identified a few small errors that we correct in section 4.4.2 below, which did not make any substantive difference to the results. We were also successful in replicating outputs from most of the company's NMA models, with the exception of the serious infection NMA (section 3.1.7).

1.1.1.1 External validity

We have tabulated the model predictions against the observed clinical data for the maintenance phase, in Table 75 below.

Table 75 Comparison of the predicted model results of Tofacitinib and Placebo (CT) against the observed clinical data – INDUCTION Phase

Study	Treatment	TNFi-naive		TNFi-exposed	
		Clinical response	Clinical remission	Clinical response	Clinical remission
OCTAVE Induction 1	Placebo	[redacted]	[redacted]	[redacted]	[redacted]
	Tofacitinib	[redacted]	[redacted]	[redacted]	[redacted]
OCTAVE Induction 2	Placebo	[redacted]	[redacted]	[redacted]	[redacted]
	Tofacitinib	[redacted]	[redacted]	[redacted]	[redacted]
Model	Placebo	[redacted]	[redacted]	[redacted]	[redacted]
	Tofacitinib	[redacted]	[redacted]	[redacted]	[redacted]

Source: CS Appendix J.1.2. Table 199

4.4.1.3 Cross validation

In section 4.2 above (page 134), we state that the CS reported previous economic models, including published literature and analyses conducted by ERGs for previous NICE TAs, for patients in ulcerative colitis. Whilst we acknowledge that there are methodological differences between the economic models across these studies, nonetheless we view that they provide sources for cross-validation of results from the company base-case analysis. Of the reported studies, we cross-validate the modelled findings of the current appraisal with 2 previous NICE TAs (TA342 and TA329) and 1 published study as summarised in Table 76. The most relevant analysis for the current appraisal is the final version from the NICE TA of vedolizumab (TA342). This appraisal relates to same patient population as the current appraisal and comparators overlap, except Tofacitinib and surgery.

Table 81 Scenario analyses, company base case (ERG corrected) – TNFi-naive subgroup

Scenarios	Assumption	ICER for tofacitinib vs.	
		CT	Vedolizumab
Base case		**£8,564**	**£615,077**
Tofacitinib maintenance dose mix	[redacted] of patients receiving 5mg; [redacted] of patients receiving 10mg	£12,637	Tofacitinib dominant
Fixed baseline utility instead of age-adjusted	Assumption that patient quality of life stays constant over time.	£8,770	£634,346
OCTAVE trials utilities	EQ-5D data were collected in Tofacitinib Phase III clinical trials	£15,525	£1,079,814
Swinburn utilities	To compare with previous analyses	£11,945	£853,228
Emergency surgery from any state	Due to the uncertainty on the likely protection from acute events based on the level of response/remission, patients are assumed to undergo emergency surgery regardless of state membership	£8,204	£606,872
Emergency surgery from active UC only	As above but assuming response to treatment offers the same level of protection from acute events, as remission	£8,661	£618,151
No emergency surgery	As above, but assuming no emergency surgery in the model	£8,719	£618,068
Central read NMA	Central read was the primary endpoint in OCTAVE trials.	£9,534	£187,809
Discounting every cycle	It tested the sensitivity of the model when the discounting of outcomes is applied every 8 weeks.	£8,616	£617,451
Vedolizumab dose 300 mg Q4W	A 300 mg Q4W maintenance dose was assessed as part of the clinical trials and is recommended for consideration in some patients who have a body weight ≥ 80 kg	£8,564	Tofacitinib dominant

Table 85 Scenario analyses, company base case (ERG corrected) – TNFi-exposed

Scenarios	Assumption	ICER for Tofacitinib vs.	
		CT	Vedolizumab
Base case		**£10,311**	**£7,838,381**
Tofacitinib maintenance dose mix	[redacted] of patients receiving 5mg; [redacted] of patients receiving 10mg	£13,956	Tofacitinib dominant
Fixed baseline utility instead of age-adjusted	Assumption that patient quality of life stays constant over time.	£10,599	£6,502,288
OCTAVE trials utilities	EQ-5D data were collected in Tofacitinib Phase III clinical trials	£18,292	Tofacitinib dominant
Swinburn utilities	To compare with previous analyses	£14,501	£7,087,359
Emergency surgery from any state	Due to the uncertainty on the likely protection from acute events based on the level of response/remission, patients are assumed to undergo emergency surgery regardless of state membership	£9,971	£7,612,076
Emergency surgery from active UC only	As above but assuming response to treatment offers the same level of protection from acute events, as remission	£10,485	£6,780,235
No emergency surgery	As above, but assuming no emergency surgery in the model	£10,603	£6,781,118
Central read NMA	Central read was the primary endpoint in OCTAVE trials.	£10,798	Tofacitinib dominant
Discounting every cycle	It tested the sensitivity of the model when the discounting of outcomes is applied every 8 weeks.	£10,408	£8,260,662
Vedolizumab dose 300 mg Q4W	A 300 mg Q4W maintenance dose was assessed as part of the clinical trials and is recommended for consideration in some patients who have a body weight ≥ 80 kg	£10,311	Tofacitinib dominant

olsalazine & sulfasalazine). However, clinical advice to ERG is that most patients receive mesalazine in UK and the doses for active ulcerative colitis are potentially higher than specified in company base case. We view that the net effect on costs from incorporating the changes in base case is likely to be neutral.

Treatment waning of effects and discontinuation

The company assumes treatment effect to be maintained with ongoing treatment and non-responders are given conventional therapy as second-line. The ERG agrees with company's approach to allow discontinuation for failure to respond in induction or loss of response in maintenance, based on the independent economic analysis in NICE TA329. We note this assumes that in practice, patients who experience exacerbations of symptoms can be assessed and, if appropriate, treatment stopped within 8 weeks. However, the model does not reflect NICE recommendations for annual assessment of benefit and need for continued treatment in previous appraisals TA329 and TA342. Clinical advice suggests that withdrawal of treatment for patients in remission is unlikely in practice, and the effects of this are difficult to quantify given the model structure and limited evidence over long-term maintenance of remission.

The company model applies a constant risk of relapse across each 8-week cycle of maintenance, with treatment stopping immediately when patients lose response. Thus, it assumes that maintenance treatment is stopped within 8 weeks of a loss of response. We consider this assumption to reflect UK practice. However, we have concerns that the costs of monitoring and follow-up in the company's model do not reflect the full cost of ensuring that treatment can be withdrawn within 8 weeks of a relapse. We address this by considering additional costs for outpatient visits to enable treatment cessation within 8 weeks of a relapse in our additional analyses.

Source of clinical effectiveness estimates

- *Choice of NMA models for economic analysis*

In general, we agree with company's approach to use locally-read clinical definitions of response and remission in economic model. The company states that their choice of NMA models was based on DIC measures of model fit, but they preferred the simpler fixed effect approach when DIC statistics were similar. In the case of the NMA for the TNFi-naïve population in the maintenance phase the fixed effect model was preferred because the company thought the random effects results were implausibly imprecise with no treatment being predicted to be significantly better than placebo. The ERG has a general preference for the random effect NMA models, as we believe that the fixed effect models may underestimate uncertainty due to heterogeneity between the studies. We test the impact of different NMA models on cost-effectiveness results in our additional analyses.

- *Combination of TNFi-failed and TNFi-exposed subgroups*

The base case NMAs combine outcomes for subgroups defined as TNFi-failed for vedolizumab with TNFi-exposed subgroups for tofacitinib and adalimumab. We consider that combining results for TNFi-failed and TNFi-exposed subgroups is a potential source of bias in favour of tofacitinib. We conduct a scenario analysis using a more like-for-like comparison between tofacitinib and vedolizumab, using data for the TNFi-failed subgroups from the OCTAVE and GEMINI trials.

- *Transformation of NMA results to transition probabilities*

The company transformed the results of the clinical response/remission NMAs from the probit scale to the natural scale and converted to absolute probabilities for use in the model. They take a simpler approach by assuming constant ratio of patients in remission and response throughout maintenance phase and beyond in extrapolation. Clinical advice to the ERG is that these assumptions might not be realistic as clinical -experience indicates the risk is greatest in the first 6-12 months; and falls thereafter. The proportion of patients with response and remission is likely to increase over time as per our clinical advice. This is because responders (without remission) are more likely to stop or switch therapy whereas those in remission would continue with treatment. However, in the absence of evidence it is difficult to adapt the model. Therefore, we conclude that the model assumption of constant risk of loss of response for patients on maintenance treatment does not reflect clinical experience. Extrapolation of relapse and discontinuation rates from the maintenance trials is likely to underestimate the average duration of treatment and hence both the costs and QALYs of active treatments. However, it is not possible to estimate the net direction of bias in ICERs between comparators, because trends in long-term risks may vary between TNFi drugs, vedolizumab and tofacitinib.

- *Exclusion of other serious adverse events*

The company excluded adverse events other than serious infections. We agree that there would have been a risk of double-counting the costs and effects of ulcerative colitis exacerbations had

Table 111 ERG base case: drug sequencing scenarios (tofacitinib PAS, others at list price)

	Treatments	Total costs	Total QALYs	Fully incremental analysis ICER (£ per QALY)
Treatment sequencing	***TNFi- naive***			
	Conventional	[redacted];	[redacted];	[redacted];
	Gol-Ada-CT	[redacted];	[redacted];	[redacted];
	Inf-Ada-CT	[redacted];	[redacted];	[redacted];
	Ada-Ved-CT	[redacted];	[redacted];	[redacted];
	Ada-Tof-CT	[redacted];	[redacted];	[redacted];
	Gol-Ved-CT	[redacted];	[redacted];	[redacted];
	Gol-Tof-CT	[redacted];	[redacted];	[redacted];
	Inf-Ved-CT	[redacted];	[redacted];	[redacted];
	Inf-Tof-CT	[redacted];	[redacted];	[redacted];
	Tof-Ada-CT	[redacted];	[redacted];	[redacted];
	Ved-Ada-CT	[redacted];	[redacted];	[redacted];
	TNFi-exposed			
	Conventional	[redacted];	[redacted];	[redacted];
	Ada-Ved-CT	[redacted];	[redacted];	[redacted];
	Ved-Ada-CT	[redacted];	[redacted];	[redacted];
	Ada-Tof-CT	[redacted];	[redacted];	[redacted];
	Tof-Ada-CT	[redacted];	[redacted];	[redacted];