CONFIDENTIAL UNTIL PUBLISHED

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

Nintedanib for treating idiopathic pulmonary fibrosis

ERRATUM

Replacement pages following the factual accuracy check by Boehringer Ingelheim

Produced by Southampton Health Technology Assessments Centre (SHTAC)

Authors
Keith Cooper, Senior Research Fellow, SHTAC
Neelam Kalita, Research Fellow, SHTAC
Micah Rose, Research Fellow, SHTAC
Elke Streit, Research Fellow, SHTAC
Karen Pickett, Research Fellow, SHTAC
Joanna Picot, Senior Research Fellow, SHTAC
Jonathan Shepherd, Principal Research Fellow, SHTAC

Correspondence to
Joanna Picot
Southampton Health Technology Assessments Centre
University of Southampton
First Floor, Epsilon House
Enterprise Road, Southampton Science park
Southampton SO16 7NS

Date completed 23 July 2015

- The economic model presented in the CS uses an appropriate approach for the disease area.

Weaknesses and Areas of uncertainty

- The three nintedanib RCTs enrolled participants with an FVC that was 50% or more of the predicted value thus these trials do not provide evidence for patients starting therapy with an FVC of less than 50% predicted.
- Due to a lack of head-to-head evidence comparing nintedanib to pirfenidone the CS provides a NMA. Although the NMA is considered to be of reasonable methodological quality there are limitations in using indirect evidence, particularly in the absence of any direct evidence for comparison. The company has explored the effects of study heterogeneity through excluding certain studies in NMA scenario analyses. The economic model is informed by a number of the NMA outcomes, and in some cases scenario analyses were used instead of all of the evidence. Given that there were some differences in results according to which scenario was used, this may potentially bias the results of the cost-effectiveness analysis.
- The NMA includes trials which measured outcomes over different periods of time. Data for nintedanib came from a 52 week time point whereas the trials contributing data on pirfenidone had follow-up periods ranging from 36 weeks to 72 weeks. For a highly progressive disease such as IPF if trials enrol participants at the same point in their disease course then those with a shorter follow-up might be expected to observe fewer negative outcomes (e.g. exacerbations, decline in lung function, deaths) whilst trials with a longer follow-up would be expected to observe worse outcomes. In some of the NMA outcomes data for 52 weeks of nintedanib were compared against 72 week data for pirfenidone. There is potential for these results to disadvantage pirfenidone.
- The population used in the economic model may not fully represent the clinical population treated in the UK because they have included patients with FVC% predicted more than 80% which represents IPF that is milder than would typically be seen in current UK practice.
- The NMA results presented in the clinical effectiveness review include both fixed effect and random effects models but the economic model used results only from fixed effect models. The company did not provide sufficient justification for model choices.

Summary of additional work undertaken by the ERG

The ERG has conducted the following analyses:

- A series of one way analyses exploring the upper and lower bounds of ORs for nintedanib vs. placebo efficacy parameters while leaving pirfenidone OR fixed
- Limiting the population to FVC% predicted 50-79.9 patients
- Using ORs from the NMA all evidence scenario analysis (fixed effect model)
- Using ORs from the NMA all evidence scenario analysis (random effects model)
- Using a utility decrement for new exacerbations of -0.14
- Using adverse event data from the RECAP study for rash,[4] with rash assumed to last for one month
- An alternative base case analysis that combined limiting the population, using the all evidence scenario fixed effects OR, a utility decrement of –0.14, and using rash data from RECAP[4] with a one month duration of AE

The model results were robust to any modification with both drugs at list price. Nintedanib dominated pirfenidone in all analyses, except when nintedanib's OR vs placebo for overall survival was set to 1.095. However, the degree by which nintedanib was the dominant option between pirfenidone and nintedanib was significantly narrowed by using the alternative OR derived from scenario 1 in the NMA. Using rash rates from the RECAP study with shorter duration for rash and photosensitivity SAEs lowered pirfenidone's ICER compared to BSC by £8,248 per QALY. The alternative base case analysis further narrowed the difference between the ICERs of nintedanib and pirfenidone vs. BSC to a difference of only £3000 between the ICERs. Additionally, with all the ERG model changes in place, pirfenidone produces 0.008 more total QALYs than nintedanib.

The ERG analyses are repeated with confidential PAS discounts for both nintedanib and pirfenidone in a separate commercial in confidence appendix.

Other relevant factors

The final scope does not specify any subgroups that should be examined and the company has not specified any in their decision problem in the CS. In the results section of the CS, however, the company presents subgroup analyses by patients' baseline FVC% predicted (≤70% or >70%) (CS p. 65), which was an analysis that was pre-specified in the INPULSIS trials.[2] NICE and the ERG sought clarification from the company about the rationale for the FVC% predicted cut-offs used in this analysis (Clarification question A3). The company responded that there are no accepted thresholds for defining disease severity and these thresholds were selected for consistency with a subgroup analysis performed for the preceding phase II TOMORROW trial.[1] The company additionally presents post-hoc subgroup analyses by patients' baseline FVC% predicted >90% or ≤90% in the CS (p. 66). In their clarifications response, the company indicated that subgroup analyses using a FVC% predicted threshold of 80% have also been conducted and published. The company referred to an analysis published in "Maher et al. ERS 2015" but did not provide a full reference for this source. The ERG was unable to locate this reference and therefore was not able to check the analyses and results provided in it. The ERG notes that results for the 80% threshold subgroup analyses are not presented in the CS. Clinical expert advice to the ERG is that, approximately, a FVC > 80% predicted indicates mild IPF, a FVC of 80 to 50% predicted indicates moderate disease and a FVC of < 50% predicted indicates severe disease. The ERG and a clinical expert consulted by the ERG consider that subgroup analyses according to these thresholds would have been more informative for assessing the efficacy of nintedanib in different patient groups than the 70% and 90% thresholds selected by the company and presented in the CS. Clinical expert advice to the ERG indicates that severity of disease at presentation is a predictor of prognosis in IPF. The TOMORROW[1] and INPULSIS trials[2] recruited patients with a FVC that was 50% or more of the predicted value so consequently there is no evidence about how efficacious nintedanib is in patients with severe disease (<50% FVC% predicted) and who are not eligible for treatment with pirfenidone, the only drug currently approved by NICE for treating IPF. The ERG and a clinical expert consulted by the ERG consider this to be an important limitation to the evidence presented.

The company additionally presented subgroup analyses for the presence of emphysema at baseline (present or not present) (CS p. 65). A clinical expert consulted by the ERG agreed that this is an important subgroup analysis. The ERG has not identified any other key subgroups that should be considered.

Summary of results for overall survival

The CS reports overall survival (defined in CS Table 39 p. 103 as all-cause mortality) for the TOMORROW[1] and the two INPULSIS[2] trials, as presented in Table 1 below. Data from the INPULSIS trials were reported individually and from pooled data. In the narrative the CS also reported results from a pooled analysis of data from the INPULSIS and the TOMORROW trials (only the licensed dose from TOMORROW). In each of the nintedanib trials, death from any cause was measured over the 52-week treatment period, and patients included in the survival analysis were all those randomised to any of the study arms, including the small number of patients who were not treated.

There was a reduction in all-cause mortality with nintedanib vs. placebo across trials, although the difference was not statistically significant. As presented in Table 1 mortality from any cause is reported to be lower in the INPULSIS trials than in the TOMORROW trial. In the INPULSIS trials 5.5% of the participants in the nintedanib groups and 7.8% in the placebo groups died, as compared to 8.1% vs. 10.3% in the TOMORROW trial.

In their narrative the CS also reported results from a pooled analysis of data from the INPULSIS and the TOMORROW trials (CS p. 62). In this analysis the proportion of patients who died was 5.8% in the nintedanib groups vs. 8.3% in the placebo group. No reference is given to the source of the analysis.

Table 1 Overall survival (defined as all-cause mortality)

	Nintedanib	**Placebo**	**HR (95% CI)** **p-value**
TOMORROW[1]	**N=86**[a]	**N=87**[a]	
Mortality, n (%)	7 (8.1)	9 (10.3)	Not reported
INPULSIS-1	**N=309**[b]	**N= 206**[b]	
Mortality, n (%)	13 (4.2)	13 (6.4)	0.63 (0.29 to 1.36)
INPULSIS-2	**N =331**[b]	**N = 220**[b]	
Mortality (%)	22 (6.7)	20 (9.1)	0.74 (0.40 to 1.35)

INPULSIS-1 & 2 pooled data	N=638 [a]	N=423 [a]	
Mortality, n (%)	35 (5.5)	33 (7.8)	0.70 (0.43 to 1.12) p=0.14

[a] The ERG notes that for the TOMORROW trial and for the analyses of pooled data from the INPULSIS trials, participant numbers were reported as the number of randomised patients, i.e. including those who did not receive the trial drug after randomisation.

[b] Participant numbers reported for the individual INPULSIS trials include only those patients who received at least one dose of the study drug. However, the ERG considers the number of untreated patients to be low and therefore unlikely to affect the outcomes.

In addition to all-cause mortality the CS reports death from respiratory causes and on-treatment mortality from pooled data in their narrative (CS p. 62). Across the TOMORROW and INPULSIS trials the proportion of patients who died from respiratory cause was 3.6% in the nintedanib group vs. 5.7% in the placebo group (p=0.0779). The proportion of patients who died while being treated with nintedanib was 3.5% as compared to 6.7% in the placebo group, and this was statistically significant (p=0.0274).

The ERG notes that different time points were applied to the analysis of on-treatment mortality. In the TOMORROW trial on-treatment mortality referred to patients on treatment and up to 14 days after discontinuation of the study drug, whereas in the INPULSIS trials the endpoint was 28 days after the last dose of the study drug. The CS does not comment on this and it is not clear to the ERG whether this may affect the results.

In the NMA for overall survival (defined as all-cause mortality) the 'All evidence' scenario comprised the key nintedanib trials (pooled data from the INPULSIS-1 and -2 RCTs[2] and the TOMORROW RCT[1]) and five trials for the comparator pirfenidone [Noble and colleagues (pooled CAPACITY-1 and -2),[21] King and colleagues,[3] Azuma and colleagues,[22] Taniguchi and colleagues,[23] CS Table 29 CS p.92]. Data for nintedanib came from a 52 week time point whereas the trials contributing data on pirfenidone had follow-up periods ranging from 36 weeks to 72 weeks (Table 1). As already noted, this may have introduced bias in the analysis (with trials of shorter duration potentially observing fewer deaths) although clinical advice to the ERG

(with trials of shorter duration potentially observing fewer acute exacerbations). In the economic model the fixed effect median OR plus 95% CrI for nintedanib versus placebo (OR 0.56 95% CrI 0.35 to 0.89) and pirfenidone versus placebo (OR 1.01 95% CrI 0.22 to 4.50) were used from scenario 3 (Table 2 and CS Appendix B p. 11 of 48). In comparison to the all evidence scenario, scenario 3 which was used in the economic model (where the fixed effect model had the lowest DIC) excluded the Azuma and colleagues[22] and the Taniguchi and colleagues[23] studies. This scenario provided a median OR indicating a benefit with nintedanib whereas there was a wide credible interval for the pirfenidone vs placebo comparison centred around a median OR of 1.01 indicating no difference. Further discussion of the loss of lung function parameters used in the model is available in ERG report section **Error! Reference source not found.Error! Reference source not found.**. The NMA output for the nintedanib vs. pirfenidone comparison in the all evidence scenario (fixed effect) was a median OR of 0.96 (95% CrI 0.36 to 2.58; CS Table 55 p. 120) indicating a small difference in the point estimate in favour of nintedanib whereas the equivalent nintedanib vs. pirfenidone comparison from scenario 3 indicated a greater difference in favour of nintedanib [median OR from the fixed effect model of 0.56 (0.12 to 2.68)]. However, in both cases the credible interval includes one so it cannot be concluded that the differences are statistically significant.

Table 2 NMA Acute exacerbations: Contributing evidence and NMA outcomes

	Contributing evidence – all evidence	
	Nintedanib vs Placebo trials	**Pirfenidone vs Placebo trials**
Median OR (95% CrI)	INPULSIS I & II,[2] 52 wks TOMORROW,[1] 52wks	Noble et al.[21] (CAPACITY I & II) 72wks, Azuma et al.[22] 36 wks, Taniguchi et al.[23] 52 wks
Fixed effect	0.56 (0.35 to 0.89)	0.59 (0.24 to 1.35)
Random effect	0.47 (0.01 to 15.96)	0.37 (0.01 to 4.81)
	Contributing evidence – for model	
	NMA nintedanib vs. placebo	**NMA pirfenidone vs placebo**
Median OR (95% CrI)	INPULSIS I & II,[2] 52 wks TOMORROW,[1] 52wks	Noble et al.[21] (CAPACITY I & II) 72wks
Fixed effect	0.56 (0.35 to 0.89)	1.01 (0.22 to 4.50)
Random effect	0.50 (0.01 to 14.43)	1.00 (0.01 to 140.92)

Overall, the ERG considers the company's approach to populate the economic model with clinical effectiveness data to be reasonable, coherent and transparent and in line with the methodologies advocated by NICE. However, the ERG had a few concerns in relation to the NMA outcomes used to inform the economic model. First, there was an inconsistency in the selection of scenarios used to populate the OR values for each of the clinical outcomes (i.e. overall survival, acute exacerbation, loss of lung function, serious adverse events and discontinuation), as shown in **Error! Reference source not found.**. The company performed an "all evidence scenario" for all the outcomes, yet results from this scenario were not used across all the outcomes in the economic model. Secondly, although the company presented results from both the fixed effect and random effects models in the NMA, the company chose estimates from the fixed effect models across all the outcomes to use in the economic model (as favoured in the individual evidence scenarios used) despite the clinical evidence suggesting that random effects models performed better for acute exacerbations and serious cardiac events for the all evidence scenario of the NMA. Due to these uncertainties, the ERG conducted additional analyses whereby the "all evidence scenario" was used for all outcomes in the NMA, along with using both fixed and random effects estimates as shown in section **Error! Reference source not found.**.

1.1.1 HRQoL

The company conducted a literature search for utility values for adult patients with IPF. The search used Medline, Medline In-process and Embase. The inclusion criteria specified generic preference based measures and disease-specific measures, not limited to EQ-5D. Thirty two studies were included in the review (Table 132 CS page 197-221).

Two studies were found that reported EQ-5D scores for patients with IPF, King and colleagues, 2011[41] and Zisman and colleagues, 2010.[42] Both studies were RCTs investigating bosentan and sildenafil treatment respectively. The CS states these studies were considered appropriate but do not contain the same health states as used in the economic model.

The CS states that IPF patients demonstrate impaired HRQoL in many life domains such as physical health. Respiratory symptoms, energy levels and degrees of independence are seriously impacted, and disability increases with the severity of the disease. In addition, IPF also impacts the psychological and emotional well-being of patients.

To further explore uncertainty in the model the ERG conducted a number of scenario analyses. Table 3 provides brief descriptions of these analyses with full descriptions in the paragraphs below. Table 4 provides the results of the scenario analyses.

Table 3 Scenario analyses conducted by the ERG

Analysis	Description
1	Model population 50-79.0 FVC% predicted only
2	NMA Scenario 1 for all efficacy data, fixed effect model
3	NMA Scenario 1 for all efficacy data, random effects model
4	Utility decrements for new exacerbations = 0.14
5	RECAP[4] rash rate with shorter duration of AE

Analysis 1 restricts the model to patients with FVC between 50% and 79.9% of the predicted value. This range corresponds more closely to the range of starting FVC% predicted values used in the pirfenidone model for the pirfenidone STA, TA 282,[7] of 50-80 FVC% predicted. It was the opinion of clinical experts consulted for the pirfenidone STA that patients with FVC% predicted above these values were unlikely to be diagnosed or treated in the UK.[13] The company conducted an analysis of an "ASCEND-like" population with FVC% predicted values between 50 and 89.9. However this analysis may have changed more than is advisable in changing adverse events, and by replacing odds ratios in the model with relative risks and hazard ratios. The ERG believes that conducting an analysis where the population is as close to UK clinical practice as possible is important for assessing validity and external consistency of the CS model results.

Analysis 2 uses OR for overall survival, exacerbations, loss of lung function, serious cardiac events and serious gastrointestinal events exclusively from the fixed effect scenario 1 NMA, whilst Analysis 3 uses OR from the random effects scenario 1 NMA. The company model used various NMA scenarios with various studies removed from the analyses to inform effectiveness in the model, with unclear or no justification for the choices of analysis. In general, the choice of analysis favoured nintedanib. The ERG felt the most appropriate decision was to use NMA scenario 1 for all parameters derived from the NMA as scenario 1 includes all studies. Values from the NMA for overall survival were derived from CS Table 49 (p. 117). Values for acute exacerbations were derived from CS Table 55 (p.120). Values for loss of lung function were

derived from CS Table 61 (p. 123). Values for serious cardiac events were derived from CS Table 72 (p.128). Values for serious gastrointestinal events were derived from CS Table 78 (p. 131).

Analysis 4 applies a utility decrement of 0.14 to all new exacerbations. The company submission stated that new exacerbations have a utility decrement of 0.14 lasting for one month and a continuing decrement of 0.0780 in subsequent model cycles. The company structured the model to calculate the difference between 0.14 and 0.0780 and apply this to the proportion of patients who had a new exacerbation. However, in the model, the value applied to for new exacerbation disutility is only 0.0987. This is because a multiplier of 1/3 was applied to the additional decrement for the first month of a new exacerbation. We have removed this multiplier.

Analysis 5 applies a risk ratio derived from a comparison of RECAP and CAPACITY rash rates from the RECAP study,[4] and applies a duration of one month to the photosensitivity and rash SAE. Much of the disutility of adverse events for pirfenidone is due to photosensitivity and rash, two interrelated AEs. Since introduction to the market, the company has given preventative instructions to reduce or eliminate these SAEs. In the RECAP study, the rash rate declined from 31% in CAPACITY to 18% in RECAP (RR = 0.58).[4] The study used for the CS model was CAPACITY.[21] Additionally, the ERG consulted a clinical advisor with regards to the duration of adverse events. In the model, the adverse event disutility is calculated based on an annual disutility for skin conditions, whilst the clinical advisor consulted by the ERG indicated that most adverse events in IPF had durations shorter than one month. To incorporate this information, we have applied the ratio of RECAP vs. CAPACITY RR (0.58) to rash rates in the model for pirfenidone, and divided the utility decrement by 12 (equivalent to assuming one month SAE duration with a constant rate). A similar reduction of the disutility for GI adverse events, could also have been applied, but due to the events occurring in both nintedanib and pirfenidone arms and adjustment of the nintedanib OR for GI adverse events having almost no effect on model results, this was not done.

Table 4 Scenario analyses conducted by the ERG

Treatment	Total costs	Total QALYs	ICER vs. BSC	Incremental ICER
Analysis 1: Limiting the population to FVC% predicted 50-79.9				
BSC	£27,960	3.06		
Nintedanib	£87,987	3.45	£153,582	£153,582
Pirfenidone	£90,164	3.39	£184,829	dominated by nintedanib
Analysis 2: NMA using scenario 1 (fixed effect model)				
BSC	£25,359	3.27		
Nintedanib	£85,047	3.67	£149,139	£149,139
Pirfenidone	£87,205	3.66	£157,460	dominated by nintedanib
Analysis 3: NMA using scenario analysis 1 (random effects model)				
BSC	£25,359	3.27		
Nintedanib	£84,972	3.68	£146,860	£146,860
Pirfenidone	£87,045	3.68	£152,191	dominated by nintedanib
Analysis 4: Utility decrement for new exacerbations 0.14				
BSC	£25,359	3.26		
Nintedanib	£85,087	3.66	£148,820	£148,820
Pirfenidone	£87,479	3.61	£176,908	dominated by nintedanib
Analysis 5: Lower disutility and shorter duration for photosensitivity and rash				
BSC	£25,359	3.27		
Nintedanib	£85,087	3.67	£149,361	£149,361
Pirfenidone	£87,381	3.64	£168,022	dominated by NDB

As can be seen by the results of the Table 4, the model results were robust to any modification with both drugs at list price. Nintedanib dominated pirfenidone in all analyses. However, the degree by which nintedanib was the dominant option between pirfenidone and nintedanib was significantly narrowed by using alternative OR derived from scenario 1 in the CS NMA. Using RECAP[4] rash rates and a one month photosensitivity and rash duration lowered perfinidone's ICER vs. BSC by £8,248 (Table 4). It should also be noted that all of these analyses are conducted without PAS submissions from Boehringer-Ingelheim (nintedanib) and Intermune (pirfenidone). In order to further test the effects of these analyses, an alternative base case was created that combined Analyses 1, 2, 4 and 5. The results of the analysis are presented before in **Error! Reference source not found.**.