CONFIDENTIAL UNTIL PUBLISHED

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

Nintedanib for treating idiopathic pulmonary fibrosis

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 17 July 2015

Source of funding: This report was commissioned by the NIHR HTA Programme as project number 14/148/03.

Acknowledgements

We are very grateful to Professor Ann Millar, Emeritus Professor, Academic Respiratory Unit, University of Bristol who offered clinical advice and comments on the draft report:

We also thank: Karen Welch, Information Scientist, SHTAC, for appraising the literature search strategies in the company's submission, running an update of the company's clinical effectiveness search and searching for ongoing studies; and Geoff Frampton, Senior Research Fellow, SHTAC, for acting as internal editor for the ERG report.

Declared competing interests of the authors

None from the authors. Professor Ann Millar has have undertaken clinical trials with nintedanib, with no personal financial gain but funds going to her department, in addition to personal funding to attend clinical conferences

Rider on responsibility for report

The views expressed in this report are those of the authors and not necessarily those of the NIHR HTA Programme. Any errors are the responsibility of the authors.

This report should be referenced as follows:

Cooper K, Kalita N, Rose M, Streit E, Pickett K, Picot J, Shepherd J. Nintedanib for treating idiopathic pulmonary fibrosis: A Single Technology Appraisal. Southampton Health Technology Assessments Centre (SHTAC), 2015.

Contributions of authors: Keith Cooper (Senior Research Fellow) critically appraised the health economic systematic review, critically appraised the economic evaluation and drafted the report. Neelam Kalita (Research Fellow) critically appraised the health economic systematic review, critically appraised the economic evaluation and drafted the report. Micah Rose (Research Fellow) critically appraised the health economic systematic review, critically appraised the economic evaluation and drafted the report. Elke Streit (Research Fellow) critically appraised the clinical effectiveness systematic review and drafted the report. Karen Pickett (Research Fellow) critically appraised the clinical effectiveness systematic review and

drafted the report. Joanna Picot (Senior Research Fellow) critically appraised the clinical effectiveness systematic review, critically appraised the NMA, drafted the report and project managed the review. Jonathan Shepherd (Principal Research Fellow) critically appraised the clinical effectiveness systematic review, critically appraised the NMA, drafted the report and is the project guarantor.

Word count: 41,061

Commercial or academic in confidence information has been redacted: [redacted]

TABLE OF CONTENTS

1 Introduction to ERG Report 15
2 BACKGROUND 15
2.1 Critique of the company's description of the underlying health problem 15
2.2 Critique of the company's overview of current service provision 15
2.3 Critique of the company's definition of decision problem 16
3 CLINICAL EFFECTIVENESS 21
3.1 Critique of the company's approach to the systematic review 21
3.2 Summary statement of company's approach 50
3.3 Summary of submitted evidence 50
3.4 Summary 74
4 ECONOMIC EVALUATION 78
4.1 Overview of the company's economic evaluation 78
4.2 Critical appraisal of the company's submitted economic evaluation 81
4.3 Additional work undertaken by the ERG 113
4.4 Summary of uncertainties and issues 118
5 End of life 119
6 Innovation 119
7 DISCUSSION 120
7.1 Summary of clinical effectiveness issues 120
7.2 Summary of cost effectiveness issues 120
8 REFERENCES 121

LIST OF TABLES

Table 1 Summary of the key features of the three included RCTs 25
Table 2 Between-trial differences in patients' smoking history 27
Table 3 Ongoing trials 29
Table 4 Company and ERG assessments of trial quality 30
Table 5 Outcomes synthesised by meta-analysis and/or NMA 41
Table 6 Intervention and comparator trials identified for inclusion in the NMA 43
Table 7 Summary of NMA evidence scenarios 45
Table 8 ERG appraisal of NMA approach 48
Table 9 Quality assessment (CRD criteria) of CS review 50
Table 10 Lung function: Change in FVC 53
Table 11 NMA Loss of lung function: Contributing evidence and NMA outcomes 55
Table 12 Overall survival (defined as all-cause mortality) 56
Table 13 NMA Overall survival (defined as all-cause mortality): Contributing evidence and NMA outcome 58
Table 14 Acute exacerbations within 52 weeks 59
Table 15 NMA Acute exacerbations: Contributing evidence and NMA outcomes 61
Table 16 PFS evidence 62
Table 17 6MWT distance 62
Table 18 NMA 6MWT distance: contributing evidence and NMA outcomes 63
Table 19 Lung function: Change in SpO_2 63
Table 20 Lung function: Change in DLco 64

Table 21 HRQoL 65
Table 22 Lung function: Subgroup analysis FVC% predicted ≤90% versus >90%..... 66
Table 23 Adverse events..... 67
Table 24 Serious cardiac events 69
Table 25 NMA serious cardiac events: Contributing evidence and NMA outcomes..... 70
Table 26 Serious gastro-intestinal events 71
Table 27 NMA serious gastro-intestinal adverse events: Contributing evidence and NMA outcomes..... 71
Table 28 Overall discontinuations 72
Table 29 NMA overall discontinuation: Contributing evidence and NMA outcomes 73
Table 30 Discontinuation due to adverse events..... 73
Table 31 NMA discontinuation due to adverse events: Contributing evidence and NMA outcomes..... 74
Table 32 Base case results of cost effectiveness analyses at the nintedanib list price (CS Table 165 p.266) 80
Table 33 Base case results of cost effectiveness analyses at the nintedanib PAS price (CS Table 166 p.267) 81
Table 34 Critical appraisal checklist for the economic evaluation 83
Table 35 NICE reference case requirements 84
Table 36 OR values obtained from the NMA as used in the company's economic model 96
Table 37 Summary of quality of life values used in the company's cost effectiveness analysis 98
Table 38 Impact of pirfenidone discount rate on the ICER, nintedanib at list price (CS Table 189, p.294) 109
Table 39 ASCEND-like population analysis results (CS Table 188, p. 293)..... 110
Table 40 Base case deterministic and probabilistic results of the CS model (derived directly from the model)..... 111
Table 41 Base case analysis..... 113
Table 42 One way sensitivity analyses using 95% CI of nintedanib efficacy OR 114
Table 43 Scenario analyses conducted by the ERG 115
Table 44 Scenario analyses conducted by the ERG 117
Table 45 Combined scenario analysis conducted by the ERG of analyses 1,2,4 and 5 118

LIST OF FIGURES

Figure 1 Model Structure (Figure 37, p. 160, CS)[35] 85
Figure 2 Fit of parametric models of the pooled overall survival data from the Kondoh study (Company's clarification response, Fig 23) 90
Figure 3 Comparison of overall survival of the INPULSIS-BSC arm against Nathan and colleagues and Kondoh and colleagues..... 92

LIST OF ABBREVIATIONS

6MWD	6-Minute Walk Distance
6MWT	6-Minute Walk Test
AE	Adverse event
AIC	Akaike Information Criterion
ANCOVA	Analysis of covariance
BD	Twice daily
BSC	Best supportive care
CASA-Q	The Cough and Sputum Assessment Questionnaire
CG	Clinical Guideline
CI	Confidence interval
CrI	Credible interval
CS	Company's submission
CSR	Clinical study report
DIC	Deviance Information Criterion
DLco	Carbon monoxide Diffusing Capacity
DSA	Deterministic sensitivity analysis
EMA	European Medicines Agency
EQ-5D	EuroQoL five dimension questionnaire
ER	Emergency room
ERG	Evidence Review Group
FDA	Food and Drug Administration
FVC	Forced Vital Capacity
GI	Gastro-intestinal
HCRU	Health Care Research Unit
HRQoL	Health-related quality of life
ICER	Incremental cost-effectiveness ratio
ICU	Intensive Care Unit
IPF	Idiopathic pulmonary fibrosis
ITC	Indirect treatment comparison
ITT	Intention-to-treat
LOCF	Last observation carried forward
LYG	Life year gained
MCID	Minimal clinically important difference
NAC	N-acetylcysteine
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NMA	Network meta-analysis
OD	Once daily
OR	Odds ratio
PaO_2	Partial pressure of oxygen in arterial blood
PAS	Patient Access Scheme
PFS	Progression-free survival
PGI-C	Patient's Global Impression of Change
PSA	Probabilistic sensitivity analysis
PSS	Personal Social Services

QALY	Quality-adjusted life year
RCT	Randomised controlled trial
SD	Standard deviation
SE	Standard error
SGRQ	St. George's Hospital Respiratory Questionnaire
SGRQ-I	IPF-specific St. George's Hospital Respiratory Questionnaire
SmPC	Summary of Product Characteristics
SpO2	Oxygen saturation by pulse oximetry
STA	Single Technology Appraisal
TA	Technology Appraisal
UCSD-SOBQ	University of California in San Diego Shortness of Breath Questionnaire
VBA	Visual basic
VC	Vital Capacity
WMD	Weighted mean difference

SUMMARY

Scope of the company submission

The company's submission (CS) reflects the scope of the appraisal issued by the National Institute for Health and Care Excellence (NICE). The submission assesses the clinical effectiveness and cost effectiveness of nintedanib compared to pirfenidone and to best supportive care for the treatment of adults with IPF.

Summary of submitted clinical effectiveness evidence

The company's systematic review of clinical effectiveness identified three relevant RCTs of nintedanib:

- The TOMORROW trial[1] (phase II) compared four doses of nintedanib [50mg once daily (OD) to 150mg twice daily (BD)] to placebo, over 52 weeks.
- The INPULSIS-1 and INPULSIS-2 (phase III) replicate RCTs[2] compared nintedanib (150mg BD) to placebo, over 52 weeks.

In these trials placebo was considered to be similar to current best supportive care (BSC). The trials were similar in terms of patient inclusion criteria and design (apart from the TOMORROW trial which was a multi-arm dose escalation study). The primary outcome in all three studies was the annual rate of forced vital capacity (FVC) decline (in L or mL). The trials were judged by the Evidence Review Group (ERG) to be of generally good methodological quality. The ERG believes that it is likely that the company has identified all relevant RCTs.

As there were no head-to-head RCTs of nintedanib and pirfenidone the company conducted a Bayesian network meta-analysis (NMA) to perform an indirect treatment comparison. The NMA compares the treatments via a common comparator (placebo), and therefore does not contain both direct and indirect evidence for the treatment comparisons. Results are presented for all studies included in the NMA, and also for a range of scenarios to assess the impact of excluding a study or studies from the analysis (e.g. due to heterogeneity). The NMA presents results for nine outcomes relevant to the scope and decision problem, six of these are used to inform the economic model (sometimes based on the all studies included in the NMA, and sometimes based on scenarios with certain studies excluded). The ERG considers that the NMA appears to be of good methodological quality, though the description of the methods used was brief. The ERG's key concerns are that a small number of trials contributed data for some outcomes and

there is the potential for bias in favour of nintedanib via the exclusion of certain studies from the NMA, and also due to the different length of follow-up between studies, discussed below.

The CS reports the effects of nintedanib treatment across a range of outcomes relevant to the NICE scope and decision problem, summarised below (For the TOMORROW trial, results for nintedanib are only given in the ERG report for the licensed 150mg BD dosage trial arm).

Various measures of lung function were reported in the CS. For the primary outcome of annual rate of FVC decline the INPULSIS trials reported a significant reduction over 52 weeks for nintedanib compared to placebo. In the TOMORROW trial, the difference between nintedanib treated patients and those treated with placebo was of lower magnitude and statistical significance varied according to which analysis method was used. The NMA for loss of lung function was not based on the primary outcome of the nintedanib trials but on a 10-point decrease in FVC% predicted, by the end of study follow-up. This outcome was from a post-hoc analysis of observed data that was not supplied to the ERG. The NMA conducted using all the available evidence indicated little difference between nintedanib and pirfenidone, however there was heterogeneity. Heterogeneity was investigated by excluding one of the pirfenidone studies (ASCEND trial by King and colleagues[3]) and it was the output from this NMA, which indicated a greater benefit from nintedanib than pirfenidone (although it could not be concluded that the difference was statistically significant), that contributed to the economic model.

There was a reduction in all-cause mortality with nintedanib when compared against placebo in all trials, although no p-values were reported for the individual trials. For the pooled INPULSIS-1 and -2 analysis the difference was not statistically significant. Up to 8.1% of patients in the nintedanib groups died, as compared to 10.3% of those treated with placebo. 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). There was no statistically significant difference between nintedanib and pirfenidone in overall survival based on the results of the NMA [Median odds ratio (OR) 1.00; 95% credible interval (CrI) 0.55 to 1.85] and it was the results from the 'all evidence' scenario that contributed to the economic model.

In terms of acute exacerbations of IPF, there was a significant decrease in the number of patients with at least one investigator-reported exacerbation in the nintedanib arm of the INPULSIS-2 trial, as compared to patients treated with placebo. However, no significant

difference in investigator-reported exacerbation rates was found in INPULSIS-1. There was a decrease in number of patients with at least one exacerbation in the TOMORROW trial but no p-value is reported. There was heterogeneity in the all evidence scenario of the NMA. Consequently the economic model drew on NMA data from a scenario that excluded three pirfenidone studies conducted in Japanese patients. This scenario, which contributed data to the economic model, increased the difference in favour of nintedanib in comparison to pirfenidone (but it cannot be concluded that the differences between nintedanib and pirfenidone are statistically significant).

There was no difference, based on a post-hoc analysis of individual patient data from the INPULSIS trials matched to one of the pirfenidone trials, in progression-free survival between nintedanib and pirfenidone OR 1.00 (95% CrI 0.71 to 1.39). This outcome did not contribute to the economic model.

The remaining clinical effectiveness outcomes (6 minute walk test distance, absolute change in oxygen saturation and change in carbon monoxide diffusion capacity) did not contribute to the economic model. Apart from the TOMORROW trial where change in oxygen saturation was statistically significant in favour of nintedanib, no statistically significant differences between nintedanib and placebo were reported for the other outcomes.

Health-related Quality of Life (HRQoL) was reported in terms of changes in scores of the St. George's Respiratory Questionnaire (SGRQ). In the TOMORROW trial there was a significant difference between the nintedanib and the placebo group in favour of nintedanib in terms of adjusted mean absolute change score from baseline. In INPULSIS-2 the mean change in SGRQ from baseline was significantly smaller for nintedanib compared with placebo, favouring nintedanib. No significant difference between groups was measured in INPULSIS-1. In a pooled analysis of the two INPULSIS trials the difference in SGRQ between nintedanib and placebo was not statistically significant. There was no NMA for HRQoL.

Three subgroup analyses (two for different thresholds of FVC% predicted, one for presence/absence of emphysema at baseline) were reported in the CS. No statistically significant differences were found by subgroup.

The proportion of patients with adverse events (AEs) was generally similar between nintedanib and placebo (around 90%). Common events included diarrhoea, cough, and nausea. The proportion of patients with serious AEs was similar between trial arms. The proportion of patients experiencing serious cardiac events was low and generally similar between trial arms, with the exception of the TOMORROW trial where a higher proportion of placebo patients experienced an event. The proportion of fatal cardiac events was low, but was more than double in the placebo arm than the nintedanib arm (reported for INPULSIS only, pooled analysis nintedanib 0.5%; placebo 1.4%). NMA was conducted for two adverse events, serious cardiac events and serious gastro-intestinal (GI) events, and both contributed data to the economic model. The NMA for serious cardiac events used in the model excluded the TOMORROW trial (due to observed heterogeneity) and indicated a greater benefit from nintedanib than pirfenidone although it cannot be concluded that the difference between treatments is statistically significant. The NMA for serious GI events indicated a greater benefit from pirfenidone than nintedanib (OR 3.96 95% CrI 1.18 to 14.51).

Data on overall discontinuations and discontinuations due to adverse events were both outcomes analysed by NMA but only the overall discontinuation data contributed to the economic model. A smaller proportion of the placebo arms discontinued than in the nintedanib arms of the nintedanib trials but the differences in proportions were less than 10%. The proportion of patients discontinuing due to AEs was generally similar between nintedanib and placebo, though in INPULSIS-1 discontinuations for nintedanib patients were almost double that of placebo.

Summary of submitted cost effectiveness evidence

The CS includes:

i) A review of published economic evaluations of nintedanib for IPF

ii) An economic evaluation undertaken for the NICE STA process. The cost effectiveness of nintedanib is compared with that of pirfenidone and best supportive care.

A systematic search of the literature was conducted by the company to identify economic evaluations of nintedanib for the treatment of IPF. The initial review did not identify any relevant studies. However, an additional non-systematic search identified one relevant study of the cost effectiveness of IPF treatments conducted in the UK.

The economic evaluation used a Markov model (developed in Microsoft Excel) to assess the cost effectiveness of nintedanib compared with pirfenidone and best supportive care (BSC) in adult patients with IPF. The model adopted a lifetime horizon to capture lifetime costs and health outcomes, with a cycle length of 3 months. Disease progression was measured by change in FVC% predicted and treatment efficacy was accounted through change in mortality risk, acute exacerbations and decline in lung function. The model used pooled data from the nintedanib phase II (TOMORROW) and phase III (INPULSIS) trials. Results from NMA were used to estimate the relative effectiveness of nintedanib and pirfenidone compared to BSC.

Results of the economic model were presented as incremental cost per quality-adjusted life years (QALY); incremental cost per life years gained and incremental cost per exacerbation avoided for nintedanib vs pirfenidone and nintedanib vs BSC at the nintedanib list price and with the nintedanib patient access scheme (PAS) respectively. The results of the cost effectiveness analysis of nintedanib vs pirfenidone at the nintedanib list price showed that nintedanib dominated pirfenidone, i.e. nintedanib was more effective and less costly than pirfenidone. For nintedanib vs BSC, the estimated incremental cost effectiveness ratio (ICER) was £149,361 at nintedanib list price and [redacted] with a PAS incorporated in the nintedanib price.

The company performed a range of deterministic and probabilistic sensitivity analyses to assess model uncertainty. For the comparison of nintedanib vs pirfenidone, the company's deterministic analyses showed that nintedanib dominated pirfenidone, except for one scenario in which a stopping rule was applied in the pirfenidone arm where patients would discontinue treatment if they declined by more than 10%FVC predicted in one year. For nintedanib compared to BSC, the key drivers of the base case results were mortality. The results from the PSA indicated that the probability of nintedanib being cost-effective compared to pirfenidone was 60% at any willingness-to-pay threshold.

Commentary on the robustness of submitted evidence

Strengths

- The company's systematic review of clinical effectiveness was generally of good methodological quality. The ERG does not consider that any key RCTs are missing. Three well conducted RCTs of reasonably good quality provide evidence for the effectiveness of nintedanib versus placebo (considered to be similar to current BSC) in adults with IPF.

- 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 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.

SUPERSEDED - SEE ERRATUM

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 -0.014
- 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 –0.014, 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.

SUPERSEDED – SEE ERRATUM

1 Introduction to ERG Report

This report is a critique of the company's submission (CS) to NICE from Boehringer Ingelheim Ltd on the clinical effectiveness and cost effectiveness of nintedanib for the treatment of adults with idiopathic pulmonary fibrosis (IPF). It identifies the strengths and weaknesses of the CS. A clinical expert was consulted to advise the ERG and to help inform this review.

Clarification on some aspects of the CS was requested from the company by NICE and the ERG on 27th May 2015. A response from the company via NICE was received by the ERG on 11th June 2015 and this can be seen in the NICE committee papers for this appraisal.

2 BACKGROUND

2.1 Critique of the company's description of the underlying health problem

The CS provides a clear and accurate overview of IPF.

2.2 Critique of the company's overview of current service provision

The CS generally provides a clear and accurate overview of how IPF is managed in current clinical practice. The company's description of current practice reflects the recommendations made in NICE's clinical pathway for IPF,[5] clinical guidance (CG) 163[6] and technology appraisal (TA) 282.[7] The company accurately states that the current recommended treatment options for IPF are limited to best supportive care and pirfenidone. In line with the recommendations in TA 282,[7] the company states pirfenidone can only be used with patients who have a percentage predicted forced vital capacity (FVC% predicted) of between 50% and 80%, and that this treatment needs to be withdrawn if a patient shows a decline in FVC% predicted of equal to or greater than 10% in a 12 month period, which indicates disease progression.

The ERG notes that CG 163[6] also suggests that clinicians can discuss the option with patients of taking off-label N-acetylcysteine. The company has not described this option in their overview of current service provision, but have included it as a comparator in the inclusion criteria for the systematic review and network meta-analysis (NMA) presented in the CS (see below for more details). Clinical expert advice to the ERG indicates that N-acetylcysteine is not used at all in practice now as a disease modifying agent, but a small number of patients may be still taking it as a mucolytic therapy.

The company suggests that the place of nintedanib in the clinical pathway will be as another treatment option for IPF and as one that can be used regardless of a patient's FVC% predicted.

2.3 Critique of the company's definition of decision problem

Population

The population defined in the company's decision problem is adults with IPF. This is the population specified in the final scope issued by NICE and it is appropriate for the potential use of nintedanib in the NHS.

Intervention

In line with the final scope, the intervention described in the company's decision problem is nintedanib (brand name: Ofev). Nintedanib received its marking authorisation in January 2015. As outlined in the company's submission, the summary of product characteristics (SmPC)[8] states that nintedanib is approved for the treatment of adults with IPF and it is administered orally at a dose of 150mg BD. The company states in CS Tables 2 (CS p. 17) and 5 (CS p. 25) that a reduced dose of 100mg BD can be used to manage adverse events and that the patient can return to the 150mg BD dose when the adverse event is resolved. The ERG notes, however, that the SmPC[8] more specifically states that the reduced dose can be used in patients who cannot tolerate the higher dose and that adverse events can be managed by dose reduction or temporary discontinuation of nintedanib, in addition to measures to control symptoms. The SmPC[8] states that upon the resolution of the adverse event, the patient may return to either dose, as appropriate. If a patient cannot tolerate the 100mg BD dose, then nintedanib should be discontinued. The SmPC[8] does not state the length of treatment, but the company suggests in CS Table 5 (CS p. 25) that nintedanib should be used until disease progression or the need to discontinue due to unacceptable adverse events. Overall, the intervention described in the decision problem is appropriate for the NHS.

Comparators

In line with the final scope, the company has listed pirfenidone and best supportive care as the comparators of interest. These are the only two treatment options currently recommended for IPF by NICE (CG 163[6] and TA 282[7]), and therefore are appropriate for the NHS. The company, however, has in practice also included N-acetylcysteine (NAC) monotherapy as a comparator in the inclusion criteria for the systematic review and NMA presented in the CS. The company

states on CS p. 32 and p. 68 that this was because it was a comparator that was discussed at the draft decision problem meeting with NICE and because the draft scope stated that it might be a comparator. While NAC is included in the systematic review and NMA, the company has not included it in the economic model. The ERG considers that this is reasonable, given that NAC does not currently have a marketing authorisation for IPF, its effectiveness is uncertain,[6] it is not widely used in clinical practice and it was not included by NICE in the final scope.

Outcomes

The outcomes stated in the company's decision problem are all those specified to be of interest in the final scope:

- Pulmonary function parameters
- Physical function
- Exacerbation rate
- Progression-free survival (PFS)
- Mortality
- Adverse effects of treatment
- Health-related quality of life (HRQoL).

These outcomes are appropriate and clinically meaningful. The company, however, has not stated in the decision problem or the earlier part of the CS which specific pulmonary or physical function parameters are clinically considered the most important outcomes. The company has also not made it clear how these parameters or acute exacerbations predict patient prognosis. This means that it is unclear which specific outcomes among these are the most clinically meaningful. The company does comment, however, in the NMA section of the CS (CS p. 93) that FVC is a predictor of progression (although they do not provide a reference for this) and is a measure used in clinical practice to assess patients' pulmonary function. They state that FVC% predicted is a standardised measure of FVC that takes into account patient factors (e.g. age, gender and height) that can cause heterogeneity in interpreting FVC outcomes. For the purposes of the NMA, the company has defined loss of lung function as a 10 percentage point reduction in FVC% predicted by the end of the trial and states that, based on the literature and clinical expert opinion, this decrease represents a clinically important difference (see CS p. 93 and CS Table 39, p. 103). Based on a study by du Bois and colleagues (2011)[9] (which was cited in the CS and was sponsored by InterMune) that examined the utility of FVC as a clinical marker, the ERG notes that a 5% to 10% reduction in FVC% predicted over 6 months is associated with an increased mortality risk in IPF. du Bois and colleagues (2011)[9] suggest that,

based a number of different methods of estimation, the minimal clinically important difference (MCID) on this outcome lies between a change of 2% and 6% in FVC% predicted. A clinical expert consulted by the ERG concurs with the company's position that a 10 percentage point reduction represents a clinically meaningful change. The ERG therefore agrees with the company's approach to defining this outcome in the NMA.

In the NMA section of the CS (CS Section 4.10, p. 66), the company has defined acute exacerbations of disease using criteria employed in the INPULSIS trials[2] which were based on those provided by Collard and colleagues (2007).[10] The ERG notes that acute exacerbations are associated with an increased risk of mortality.[11] A clinical expert advised the ERG that the Collard and colleagues[10] definition is rarely fully applied in practice and that, in practice, acute exacerbation is a less well defined phenomenon and more vague.

The company has included the distance walked during the 6 minute walk test (6MWT) as an outcome in an NMA. The ERG notes that the literature shows that baseline results for this outcome and changes in it can predict mortality risk.[11] The company states in the NMA section of the CS (Section 4.10, p. 96) that the 6MWT has limited value in clinical practice, because it is challenging to standardise the requirements for the test across settings. The company states it is not clear if the measures from this test taken in clinical trials are reproducible in the clinical setting and that therefore interpretation of the meaning of this outcome in clinical trials in terms of the relative efficacy of treatments is challenging. The ERG suggests that the company's conclusion about the utility of this test is reasonable and concurs that there can be variation in its implementation in practice and notes that patient learning and motivation effects might impact on the results of the test.[12] A clinical expert consulted by the ERG agreed that the test has limitations, but indicated that it is a clinically valuable measure.

The ERG agrees with the company's statement on CS p. 95 that there is no current consistent definition of progression-free survival in IPF.[13] For the purposes of the NMA, the company has defined progression-free survival as "Time to confirmed ≥10% decline in FVC% predicted, confirmed ≥15% decline in carbon monoxide diffusing capacity (DLco) % Predicted, or death" (CS Table 39, p. 103). A clinical expert consulted by the ERG agreed that the company's definition of PFS is reasonable.

In terms of which outcomes are the most clinically meaningful in IPF, expert clinical advice to the ERG was that opinion on this varies, but the opinion of our expert was that that PFS is the most clinically important outcome. Of the physical function measures, the clinical expert indicated that again there is no consensus about which are the key clinical ones, but it was suggested that in clinical practice, most clinicians would use the 6-minute walk distance (6MWD) and ability of patients to perform activities of daily living (such as washing and dressing). Of the pulmonary function measures, the clinical expert suggested that DLco and desaturation on exercise (during the 6MWT) are the key clinical ones. Of the outcomes considered the most important by the clinical expert we consulted, only activities of daily living is not included in the CS. The trials do not appear to have measured this outcome. The company therefore appears to have considered and provided results for the most clinically important outcomes in the CS, with the exception of activities of daily living.

In summary, the outcomes selected by the company are appropriate, match those specified in the final scope and are adequately defined. The company has included the most clinically meaningful outcomes in the CS, with the exception of activities of daily living, which was not measured in the trials nor specified as an outcome to be considered in the final scope.

Economic analysis

The economic analysis proposed in the decision problem matches the final scope and is appropriate for the NHS. The company has used a Markov model with an NHS and Personal Social Services perspective and a lifetime horizon (defined as 50 years from the start of the model). The ERG suggests that a shorter time horizon would have been more appropriate in this instance (see section 4.2.1 of this report for more details).

The final scope specifies that the economic model should take into account any cost discounts that are available through patient access schemes (PAS) for both the comparators and the intervention. The ERG notes that the company submitted a PAS for nintedanib at the same time as preparing this STA submission and that pirfenidone is also available through a PAS.[7] In the economic model in the CS, the company has taken into account the PAS for nintedanib, presenting ICER results for the base case both with and without the PAS applied. The company has additionally carried out scenario analyses where PAS cost discounts for both nintedanib and pirfenidone were applied.

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.

SUPERSEDED – SEE ERRATUM

The final scope did not identify any equity or equality issues and the company also did not identify any in its decision problem in the CS. The ERG also did not identify any potential equity or equality issues related to the implementation of nintedanib in the NHS.

3 CLINICAL EFFECTIVENESS

3.1 Critique of the company's approach to the systematic review

3.1.1 Description of the company's search strategy

All search strategies were grouped in one appendix, enabling easy access to the searches and the company used an acceptable set of bibliographic databases. All years were recorded as searched, however, exact dates could have been specified, as access to years can depend on database subscription type. The search strategies contained a mix of descriptor and free text terms. Some of the lines contained complex bracketing which could have gained benefit from being split into separate lines for greater transparency. An RCT trial filter was not applied to limit the clinical searches to RCTs, which is deemed appropriate to capture a variety of clinical trial types. The economic, HRQoL and resource searches have used relevant filters. The documentation of the searches contains several mistakes:

- Ofev, the tradename for the IPF product has not been used in any of the searches. The tradename Vargatef for an alternative indication of nintedanib (in non-small cell lung carcinoma) has been used instead. The ERG checked the search results returned for Ofev on Medline and Embase and no additional relevant items were found.
- The use of ADJ3 in all the Cochrane searches implies that it was not searched directly as stated, since NEAR/3 is the appropriate syntax.
- There is inaccuracy in the linking of the economic search sets in Embase and it is possible that an incorrect table has been included within the CS [CS Appendix A, Search strategy (4): Embase (Ovid®)]. Search terms for lines 1- 34 are recorded. Sets 10-34 should have been combined and then sets 9 and 35 should have been linked whereas the search strategy displayed combining sets 17-41 which is beyond the lines recorded and then linking sets 16 and 42. The recording of the economic search sets for Medline was accurate with the correct sets linked [CS Appendix A, Search strategy (5): Medline, Medline In-Process (Ovid®)].
- In the Resource Use searches it is noted that for Medline [CS Appendix A, Search strategy (13): Medline and Medline In-Process (Ovid®)], the company possibly mapped

the search terms, rather than using the correct MESH descriptors directly. The use of the syntax ".tw,ab.", is a tautology as "tw" by itself instructs searching in the title or abstract.

- Although the HRQoL search filter appeared acceptable, the ERG noted in Embase that Set 36 is missing from the list (or/8-35 is the assumption that has to be made as combining the sets 7 and 36 would then be correct) [CS Appendix A, Search strategy (12): Embase (Ovid®)]. In Medline sets 8-33 are combined however that renders sets 34 and 35 appertaining to the respiratory questionnaires redundant from the search [CS Appendix A, Search strategy (13): Medline and Medline In-Process (Ovid®)]. Other specific terms that could have been used but were not were UCSD-SOBQ (University of California in San Diego Shortness of Breath Questionnaire) and CASA-Q (Cough and Sputum Assessment Questionnaire) and PGI-C (Patient's Global Impression of Change).

Despite the mistakes in the recording of the search strategies the ERG does not believe that any more relevant records would have been produced by more accurate representation or more detailed searching due to the known limited number of trials in this topic area.

All searches were out of date by 7 to 9 months. The ERG elected to re-run the clinical searches (which were 8 months out of date) along with searches for ongoing trials from UKCRN, ISRCTN, and WHO ICTRP. Only clinicaltrials.gov, recent conferences and the regulatory agencies were documented in the submission as having been searched for ongoing studies. The updated clinical search results were checked by one ERG researcher and two additional references[14;15] matching the inclusion criteria were identified. However one was a pharmacokinetic study[15] and amongst the study population just 11 patients received the licenced dose of nintedanib for 28 days (adverse events reported but no efficacy outcomes) and the other was a pre-specified subgroup analysis of Asian participants in the INPULSIS trials.[14] The ERG does not believe that either of these studies contribute anything substantial to the evidence base for this STA. One ongoing study was identified (see section 3.1.3 of this ERG report). Although the economic searches were 9 months out of date the ERG elected not to run them, the resource use or the HRQoL searches due to the known lack of availability of data relating to nintedanib.

3.1.2 Statement of the inclusion/exclusion criteria used in the study selection.

Inclusion and exclusion criteria are stated separately for the systematic review of RCTs containing nintedanib, and for the systematic review underpinning the NMA.

Study design in both CS systematic reviews was limited to phase II, III, and IV RCTs. Phase I RCTs and non-RCT studies of any design were excluded from the CS systematic reviews, as were reviews (systematic or otherwise), case reports, critical appraisals, updates or commentaries on data published elsewhere, notes, letters, and editorials. Only English-language studies were included. Further exclusion criteria were clearly stated for population, comparators and outcomes.

To be included in the NMA trials had to meet the eligibility criteria provided in CS Table 21 (p. 67). Inclusion and exclusion criteria were identical to those in CS Table 6 for the systematic review (p. 35), with the exception of studies not containing nintedanib could now be included.

No limits were placed on inclusion criteria relating to the quality of RCTs. Blinded and non-blinded RCTs were eligible for inclusion, as well as other designs (including parallel design extensions, post-hoc and pooled analyses of RCTs, and studies published as abstracts or conference presentations if adequate data were provided) as described in CS Tables 6 (p. 35) and 21 (p. 67). Setting was not used as an eligibility criterion.

To be included in the systematic review of RCTs containing nintedanib, trials had to meet the eligibility criteria provided in CS Table 6 (p. 35).

The ERG notes that the inclusion criteria are generally appropriate. The CS included all the outcomes specified in the scope and the decision problem in the eligibility criteria for the systematic review and NMA, and the company does not appear to have omitted any important outcomes.

Two flow diagrams are provided with the numbers of titles and abstracts included or excluded from the search at each stage, and with reasons for exclusion.

- The first diagram (CS Figure 5, p. 38) demonstrates the identification of relevant studies of the intervention to be assessed, based on the systematic review inclusion and exclusion criteria stated in CS Table 6 (p. 35). Thirteen records of 3 relevant studies were identified. The sums of included and excluded items are correct.

- The second flow diagram (CS Figure 11, p. 69) demonstrates the identification of relevant studies for both the intervention and the comparators for the NMA. Forty-one records of 12 relevant studies were identified.
  The diagram states that the search produced 3341 hits. However, the sum of titles and abstracts listed is 3338. It appears that 3 clinical trial reports (data on file) may not be listed here (those were shown in the first diagram / CS Figure 5, p. 38).
  All other sums of included and excluded items are correct.

The company has not explicitly considered bias at the study inclusion step, but, as discussed above, the company limited the study design to either blinded or open-label Phase II, III and IV RCTs in their inclusion criteria. Additionally, the company did not provide a rationale for their choice of the following exclusion criteria:

- Non-English language publications were excluded from the systematic review and the NMA (see CS Tables 6 and 21, p. 35 and p. 69). As described on CS p. 33, the search strategy was not limited by language, but the company states that studies published in languages other than English were not reviewed in full. The exclusion of non-English language publications was not explained by the company, and the resulting potential language bias was not discussed. However, the ERG notes that it is unlikely that there are relevant studies in non-English languages, and the potential language bias is therefore considered low.
- Phase I RCTs and studies with non-randomised designs were also excluded. The company did not limit the searches to RCTs and it is unclear from the CS why the company then excluded non-RCTs at the study screening stage. However, the ERG agrees that it is reasonable to have limited the inclusion criteria to Phase II, III and IV RCTs.

Excluded references that contained nintedanib are presented in CS Table 9 (p. 40) and described on CS pages 41-42, with all exclusions discussed and justified. Recently completed studies were excluded because data are not yet available. A pharmacokinetic study of nintedanib alone or in combination with pirfenidone[16] was excluded because it did not report on any of the outcomes relevant to the decision problem. The ERG agrees that these exclusions are reasonable. The CS does not discuss or list the excluded studies containing the comparator treatments for the NMA although the flow chart (CS Figure 11 p.69) does categorise reasons for exclusion.

The CS includes studies of NAC in the NMA because the initial draft scope suggested NAC might be included as a comparator. As discussed in section 2.3 of the ERG report, the company states that the NMA process was already underway when the final scope was received which did not include NAC and the company confirmed that NAC was not included in the cost-effectiveness model.

3.1.3 Identified studies

The systematic review identified and included three relevant RCTs of nintedanib: the TOMORROW (phase II), the replicate INPULSIS-1, and the INPULSIS-2 (phase III) trials. These are reported in two journal articles,[1;2] three clinical study reports and in five conference abstracts. All three trials compared nintedanib to placebo. The CS states (CS p. 148) that patients in the INPULSIS trials were allowed to use background medication for acute exacerbations or disease decline after an initial 6 months on therapy and that this is similar to current best supportive care (BSC). The key features of the three RCTs[1;2] are shown in Table 1. In the remainder of the ERG report only the nintedanib 150mg BD arm of the trial is reported on because this the licensed dose.

Table 1 Summary of the key features of the three included RCTs

	Trial arms	Number enrolled	Primary outcome measure	Length of follow-up
TOMORROW[1]	Nintedanib 50mg OD	86	Annual rate of FVC decline	52 weeks
	Nintedanib 50mg BD	86		
	Nintedanib 100mg BD	86		
	Nintedanib 150mg BD	85		
	Placebo	85		
INPULSIS-1[2]	Nintedanib 150mg BD	309	Annual rate of FVC decline	52 weeks
	Placebo	204		
INPULSIS-2[2]	Nintedanib 150mg BD	329	Annual rate of FVC decline	52 weeks
	Placebo	219		

Summary details of the RCTs were provided in the CS.

- Trial design, intervention, population, and duration are reported in CS Table 10 (p. 41).

- Patient numbers are shown in CS Figure 8 for the TOMORROW trial, Figure 9 for INPULSIS-1, and Figure 10 for INPULSIS-2 (CS p. 54-56), including numbers screened, randomised, and treated. The numbers of patients who prematurely discontinued the trial medication are also reported in CS figures 8-10 (CS p. 54 to 56). For the INPULSIS trials, reasons for drop-out are provided in the patient flow diagrams. CS Figure 8 does not include reasons for discontinuation. NICE and the ERG sought clarification from the company and an updated flow diagram was provided (clarification A2).
- Eligibility criteria are reported in CS Table 13 (p. 48) for all three nintedanib trials
- Primary and secondary outcomes are presented in CS Table 12 (p. 44-47).
- The statistical analyses of the nintedanib trials is described in CS Table 14 (p. 53) and includes the hypothesis, objective, techniques of statistical analyses, sample size and power calculation, and data management, including analysis of patient withdrawals.
- The company states in CS Table 15 (p. 58) that the TOMORROW[1] and the INPULSIS trials[2] were analysed by the intention to treat (ITT) principle. Methods to account for missing data are described in this table and in CS Table 14 (p. 53).
- The company identified three subgroups for which subgroup analyses were undertaken, using pooled data from the INPULSIS trials (CS p. 65-66): These subgroups are patients with baseline FVC ≤70% predicted value, as compared to >70%; patients with baseline FVC ≤90% predicted value, as compared to >90%; and patients with or without emphysema at baseline. NICE and the ERG sought clarification from the company about the rationale for these subgroups which was provided (clarification A3) and is discussed earlier in this ERG report (ERG report section 2.3 'Other relevant factors').

The TOMORROW trial[1] was a dose escalation study that investigated different dosing regimens, including the licensed dosage of 150mg BD. Otherwise key characteristics are comparable across the three trials.

The company provided the published RCT reports electronically, but did not provide the clinical study reports (CSRs) for the TOMORROW and the INPULSIS trials.

The TOMORROW trial and the two INPULSIS trials were sponsored by the company Boehringer Ingelheim Ltd. The INPULSIS trials were additionally supported by funding from the National Institute for Health Research (NIHR), Southampton Respiratory Biomedical Research Unit at the University Hospital Southampton NHS Foundation Trust and from the NIHR

Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust and Imperial College London.

The CS does not include any non-randomised studies. The company states that no non-randomised or non-controlled studies were identified in the systematic literature review.

Baseline patient characteristics of the included nintedanib trials are reported in CS Tables 16 and 17 (p. 60-61) and were reported separately for each arm of the TOMORROW and the INPULSIS trials. Baseline characteristics for the total trial population were also reported for the TOMORROW trial.

The company states that baseline characteristics, including age, gender, time since diagnosis of IPF, and key outcome measurements were similar across treatment groups in all nintedanib trials, but the ERG identified some differences between the intervention and the placebo arms of the INPULSIS trials with regards to the proportion of current smokers enrolled (Table 2). However, the ERG feels that these are unlikely to impact on outcomes, given the overall small proportion of current smokers that participated in these trials. The ERG also observed differences in smoking history between trials, in that INPULSIS-1 enrolled a higher proportion of former and current smokers than the INPULSIS-2 and the TOMORROW trials. These are summarised in Table 2 below. There were also more men in the INPULSIS trials (between 77.8 and 81.2% depending on trial arm) as compared to the TOMORROW trial (74.8%).

Table 2 Between-trial differences in patients' smoking history

	TOMORROW[1]			INPULSIS-1[2]		INPULSIS-2[2]	
Smokers	150mg BD arm (n=85)	Placebo arm (n=85)[c]	Total (N=428)[a]	Treatment N=309	Placebo N=204	Treatment N=329	Placebo N=219
Former	60.0%	64.0%	62.9%	70.2%	70.6%	66.3%	63.5%
Current	7.1%	4.7%	4.2%	6.8%	4.4%	2.4%	4.1%
Total[b]	67.1%	69.4%	67.1	77.0%	75.0%	68.7%	67.6%

[a] All participants in the TOMORROW trial i.e. including three trial arms not included in the ERG report because they do not reflect the licensed dose of nintedanib. Calculated by the ERG from data in CS Table 16.

[b] Calculated by the ERG from data in CS Tables 16 (TOMORROW) and 17 (INPULSIS trials) on p. 60-61.

[c] The ERG calculated that the total n for this arm is 86 from the data reported in CS Table 16 (p. 60) and not 85 as reported in the top row of the CS table. This minor error in the CS, however, does not affect the ERG's statement on the between-trial differences in smoking history.

The ERG was concerned whether the trial participants were representative of the UK patients in clinical practice. NICE and the ERG therefore asked the company in their clarifications request to confirm the number of UK participants in each trial and provide their baseline characteristics (clarification response A1).

Analysis of UK patients from TOMORROW could not be provided by the company in the time available. Overall, 45 UK patients were enrolled in INPULSIS-1, 33 to nintedanib and 12 to placebo. No UK patients were enrolled in INPULSIS-2. There are some differences in baseline characteristics between UK patients and the total INPULSIS-1 trial population, but it is not clear to the ERG whether these are significant:

- there was a smaller proportion of men in the UK group (UK: 75.6% vs. INPULSIS-1 total 80.7%)
- more UK patients had a smoking history (UK: 80% vs. INPULSIS-1 total: 76.2%).
- UK patients had higher FVC values than INPULSIS-1 participants overall (FVC% predicted, UK nintedanib: 83.7, placebo: 87.6 vs. INPULSIS-1 nintedanib: 79.5, placebo: 80.5).

The age of the UK trial participants appear consistent with that of the INPULSIS-1 trial overall and clinical advice received by the ERG indicated that the UK trial participants were younger than those seen in typical practice in the UK.

All the included RCTs included in the systematic review appear to meet the inclusion criteria, and the ERG believes that it is likely that the company has identified all relevant RCTs.

Ongoing trials

The CS lists the ongoing TOMORROW and INPULSIS extension trials in CS Table 11 (CS p.41) but does not comment on whether any other nintedanib studies are ongoing. Some studies are listed among the excluded studies that contained nintedanib as 'Study in progress, no data' (CS Table 9, p. 40) but no further details of these were provided. The ERG searched for ongoing

studies and identified just one trial (Table 3) that did not appear to be related to the existing TOMORROW and INPULSIS studies.

Table 3 Ongoing trials

Trial identifier, sponsor	Design, Country	Intervention, comparator, patient group	Expected end date
NCT01979952 Boehringer Ingelheim	Multicentre RCT US, Canada, Turkey.	Nintedanib 150 mg BD vs placebo. Patient aged = 40 years at visit 1, with IPF, DLCO 30% to 79% predicted of normal and FVC = 50% predicted of normal at visit 1 and visit 2.	July 2017 (July 2016 for final data collection for primary outcome measure)

3.1.4 Description and critique of the approach to validity assessment

The company critically appraised the included trials using the NICE criteria and presents a summary of findings on CS p. 57 and in CS Table 15 (p. 58). The replicate INPULSIS-1 and -2 were assessed together as one, presumably because the same methods were applied in both trials, some endpoints (e.g. exacerbation, number of deaths) were analysed from pooled data, and both trials were reported in one single publication,[2] although this is not explicitly stated in the CS.

The ERG agrees with the company assessment for most criteria (see Table 4). For the TOMORROW trial the ERG assessment differs from the industry assessment for question 5 (imbalances in drop-outs), as the ERG feels that there was an imbalance in drop-outs between the placebo and nintedanib 150mg BD arms which was not discussed in the CS or the publication.[1]

For question 6 the ERG identified that the TOMORROW trial had measured time to progression (CS Table 12, p. 47), but that this outcome was not reported in the trial paper or CS.

The ERG assessment also differs for question 7 (ITT analysis and methods used to account for missing data). The ERG feels that the last observation carried forward (LOCF) approach to estimate missing data for the ITT analysis in the TOMORROW trial could potentially bias the outcomes in favour of nintedanib. In the INPULSIS trials, the company did not explain the

assumptions used in their approach to missing data for the primary outcome. Therefore the ERG was uncertain whether the methods used in the CS were appropriate.

Table 4 Company and ERG assessments of trial quality

		TOMORROW[1]	INPULSIS-1[2]	INPULSIS-2[2]
1. Was randomisation carried out appropriately?	CS:	Yes	Yes	
	ERG:	Yes	Yes	Yes
Comment:				
2. Was concealment of treatment allocation adequate?	CS:	Yes	Yes	
	ERG:	Yes	Yes	Yes
Comment:				
3. Were groups similar at outset in terms of prognostic factors?	CS:	Yes	Yes	
	ERG:	Yes	Yes	Yes
Comment: In all of the three trials, smoking history differed between the trial arms, but these differences were small and overall groups appear similar.				
4. Were care providers, participants and outcome assessors blind to treatment allocation?	CS:	Yes	Yes	
	ERG:	Yes[a]	Yes[a]	Yes[a]
[a] Patients, investigators, adjudication committee members and the study sponsor were blinded to treatment allocation. The ERG presumes that investigators were care providers and outcome assessors.				
5. Were there any unexpected imbalances in drop-outs between groups?	CS:	No	No	
	ERG:	Uncertain[b]	No	No
[b] In the TOMORROW trial the drop-out rate was highest in the group receiving the highest dose of nintedanib (150 mg BD / drop-out 37.6%). In the remaining arms, drop-out rates were highest in the group receiving the lowest dose (50mg OD / 27.9%) or placebo (28.2%) respectively, and lowest in the 50mg BD arm (16.3%). These variations were not discussed in the CS or publication and no reasons for dropout were provided in the CS or the study paper.[1] However, an updated patient flow diagram provided by the company in their response to clarification questions (clarification response A2) showed that the majority of patients who did not complete the trial withdrew due to adverse events. Drop-outs from treatments were similar in both arms in INPULSIS-2, whereas the proportion of drop-outs in INPULSIS-1 were higher in the intervention group as compared to the placebo group, due to adverse events. There were no imbalances in drop-out rates in relation to completion of planned observation time in the INPULSIS trials.				
6. Is there any evidence that authors measured more outcomes than reported?	CS:	No	No	
	ERG:	Yes[c]	No	No
[c] Summaries of predefined primary and secondary end points are provided in the published articles, with detailed results for most outcomes provided in separate appendices. However, for the TOMORROW trial differences in DLco and for distance achieved in the 6MWT were only reported narratively as non-significant, but no outcome data were provided for these end points. The ERG notes that the TOMORROW trial measured time to progression (CS Table 12, p. 47), but that this outcome was not reported in the trial paper or CS. The ERG additionally notes that in the CS, DLco has been reported differently to how it was pre-specified in the trial protocol (as described in section 3.1.5 below).				

7. Did the analysis include an ITT analysis? If so, was this appropriate and were appropriate methods used to account for missing data?	CS:	Yes/Yes	Yes/Yes	
	ERG:	Yes/No[d]	Uncertain[e]/ Uncertain[e]	Uncertain[e]/ Uncertain[e]

[d] Efficacy and safety analyses were conducted on all patients who were randomised to treatment. The ERG notes that the TOMORROW trial used the last observation carried forward (LOCF) approach to estimate missing data for the ITT analysis of secondary outcomes (for the analysis of the primary outcome no replacement of missing data was planned). The ERG considers LOCF an inappropriate method to use in a progressive disease such as IPF, because a patient's score on an outcome measure may be more favourable earlier in a trial than later when they drop out. Given the higher rate of dropouts in the 150mg BD compared to the placebo arm, the use of LOCF could potentially bias the outcomes in favour of nintedanib.

[e] Efficacy and safety analyses were conducted on patients who were randomised to treatment and received ≥1 dose of study medication and a small number of patients did not receive ≥1 dose (INPULSIS-I: 2; INPULSIS-II: 3). However, for the "primary analysis" (CS Table 14, p. 53), the INPULSIS trials assumed data were missing at random and so missing data were not imputed. The company has not provided any information about how this assumption was tested and therefore it is uncertain if this was an appropriate approach to take. The company also conducted sensitivity analyses, using multiple imputation, which is an appropriate approach. It is unclear which of these analyses are presented in the CS.

3.1.5 Description and critique of the company's outcome selection

Overview of outcomes reported in the trials

In both the TOMORROW and INPULSIS trials, the primary endpoint was the annual rate of FVC decline, measured in L or mL per year.[1;2] As shown in CS Table 12, CS p. 44 to 47, a range of secondary outcomes were also assessed in the trials. These included a number of pulmonary function, physical function, survival and acute exacerbation measures, as well as time to progression, adverse events and HRQoL. None of the trials appears to have measured progression-free survival (PFS) in line with the definition used in the CS for the PFS NMA, although the TOMORROW trial measured (but did not report) time to progression (and this measure included death; see below for more information about this).[1;2] Outcomes were assessed at a variety of time points throughout the trials, with the longest follow-ups at 52 weeks in all the trials.

Outcomes included in the company's systematic review

In their systematic review in the CS, the company has presented the results for a selection of the outcomes measured in the trials, including the annual rate of decline in FVC, change in FVC% predicted, absolute change in DLco (but this is presented differently to how it is defined

in the TOMORROW trial protocol, please see below for details), 6MWT results (but this is presented differently to how it is presented in the trial paper; again please see below for details), number and % of patients with at least one exacerbation, HRQoL, mortality and adverse events (see CS Section 4.7, CS p. 62, for all the outcomes reported). Therefore the company has included in the CS all the outcomes specified in the NICE final scope and the company's decision problem, except for PFS, which was not measured in the trials. (Please note, though, that, as described below, the company has included PFS as an outcome in an NMA.) However, as discussed in Section 2.3 of this report, the company has not made it clear which of the outcomes presented are the most important or clinically meaningful. As also discussed in section 2.3 of this report, the company has included in the CS all the outcomes considered to be the most clinically important by our clinical expert advisor (PFS, 6MWD, DLco and desaturation on exercise on 6MWT), except for activities of daily living. The latter outcome, however, was not specified as an outcome of interest in the final scope issued by NICE and did not appear to be measured in the trials.

In the CS, the company has presented the results for the DLco outcome (which was considered by a clinical expert we consulted to be a key clinical measure of pulmonary function) as 'absolute change in DLco' (CS Table 18, p. 63). The ERG checked how this outcome was pre-specified in the trial protocol. The protocol states that change in DLco from baseline at 6 and 12 months would be assessed according to the following categories:

- Decrease of > 15% or > 1 mmol min^{-1} kPa^{-1}
- Increase of > 15% or > 1 mmol min^{-1} kPa^{-1}
- Change within ≤ 15% and ≤ 1 mmol min^{-1} kPa^{-1}

Given that the DLco results in the CS are reported differently to how this outcome was pre-specified in the trial protocol, the results presented in the CS may be at risk of bias. The DLco results presented in the CS for the INPULSIS trials appear to be reported in line with how this outcome was pre-specified in the protocol for the trials.

The results of the 6MWT are reported as "Absolute change in worst SpO_2 during 6MWT from baseline % (SE)" (Section 4.7, CS p. 62) in the systematic review, and this differs to how outcomes from this test were defined in CS Table 12 (CS p. 44 to 47) and the TOMORROW trial paper.[1] In CS Table 12, outcomes from the 6MWT were defined as 1) "Change from baseline in distance walked (m)" (CS p. 47) on 6MWT and 2) "Dyspnoea rating on Borg scale: change from baseline" (CS p. 47). In the paper, outcomes from this test were presented solely as "the

distance achieved on the 6-minute walk test" (Richeldi and colleagues, 2011: p. 1081[1]). It is not clear from the CS why the company has chosen to present the 'Absolute change in worst SpO_2 during 6MWT from baseline %' rather than these outcomes. A clinical expert consulted by the ERG indicated that lowest O_2 saturation would be the most clinically informative measure from this test. Change in distance walked on the 6MWT test, however, is an outcome examined in an NMA in the CS, and the relevant results from the TOMORROW trial are included in this. Another endpoint that was measured in the trial papers, but not reported in the company's systematic review in the CS was: 'an SpO_2 decrease of more than 4 percentage points' (TOMORROW[1]). A clinical expert consulted by the ERG indicated that this outcome is of minor importance, and so the ERG suggests that it is reasonable that the company has not included it.

The ERG additionally notes that the CS and the trial protocol state that time to progression was measured as an outcome in the TOMORROW trial (as shown in CS Table 12, p. 44 to 47) and the definition includes death. This differs from typical definitions of time to progression which, at least in the field of oncology, would not include death. Results for time to progression as defined in the CS are not reported in the CS nor in the published TOMORROW paper.[1] The ERG notes that the definition of time to progression differs to the definition of PFS used in the NMA and therefore that these appear to be different outcomes.

In the trials, HRQoL was measured by the:

- St. George's Respiratory Questionnaire (SGRQ) (INPULSIS-1, INPULSIS-2 and TOMORROW)
- IPF-specific SGRQ (SGRQ-I) (INPULSIS-1 and INPULSIS-2)
- EuroQol 5-Dimensional Quality of Life Questionnaire (EQ-5D) (INPULSIS-1 and INPULSIS-2)

The SGRQ is a validated measure.[17] The SGRQ has been validated in people with diseases associated with chronic airflow limitation.[17] It provides a total score and measures of symptoms, activity and impacts. The SGRQ-I is a modified version of the SGRQ, specifically for use with patients with IPF.[18] It also measures symptoms, activities, impacts and a total score. Although the SGRQ-I was developed in an IPF population,[18] the ERG could not find evidence that it has been externally validated with IPF patients. The EQ-5D[19] is a validated, generic measure of HRQoL and is NICE's favoured HRQoL measure.[20] Overall, the HRQoL measures used in the trials and reported in the company's systematic review in the CS are appropriate, although as

the SGRQ is not a disease-specific measure for IPF, it may not fully reflect changes in HRQoL in IPF. Only results from the SGRQ are reported in the CS systematic review.

Other patient reported outcomes measured in the trials and mentioned in the CS are the University of California in San Diego Shortness of Breath Questionnaire (UCSD-SOBQ), Patient's Global Impression of Change (PGI-C) and the Cough and Sputum Assessment Questionnaire (CASA-Q). The CS does not provide detailed results for these and these outcomes were not reported in the trial papers.

Outcomes included in the NMA

The company conducted an NMA for each of the nine outcomes listed below (CS p. 91 onwards, outcomes defined in CS Table 39 p. 103) – see Section 2.3 of this ERG report for further information on how some of these outcomes were defined and measured and the ERG's commentary on this:

- "Overall survival" – which the company has defined as all-cause mortality and has measured as events rather than time to death in the NMA
- Acute exacerbations (events) (using investigator reported rather than adjudicated events for the INPULSIS trials, which the ERG agrees is reasonable, given that the investigator reported results are less favourable to nintedanib than the adjudicated results; see CS Table 19, CS p. 64)
- Pulmonary function – a 10 percentage point decrease in FVC% predicted (as the company states that this represents a clinically meaningful change, based on the literature and clinical expert opinion)
- PFS
- 6MWD
- AE – serious cardiac events
- AE – serious gastro-intestinal (GI) events
- Treatment tolerability – discontinuation due to AEs
- Treatment tolerability – overall discontinuation

The company considered an NMA of the HRQoL outcome, but decided that this was not possible due to differences in the HRQoL measures used in studies. The ERG agrees that this is reasonable. The company has used utility values in the economic model derived from the EQ-5D data from the INPULSIS trials (see Section 4.2.5 of this report).

To include the INPULSIS trials[2] in the NMA for the PFS outcome, the company carried out a post-hoc analysis of PFS using individual patient data. The company did not include data from the TOMORROW trial[1] in this analysis and it is unclear from the CS whether or not a similar post-hoc analysis of PFS could have been conducted for this trial from individual patient data for use in the NMA, as the company does not discuss this possibility. This outcome is therefore at a risk of bias.

Regarding the NMA of the FVC% predicted outcome, the company states on CS p. 93 that the results used in the NMA from the TOMORROW and INPULSIS trials were those based on post-hoc analyses of observed data only, with no imputation of missing data, including for those participants who dropped out of the study. The ERG therefore considers that this NMA outcome may be subject to some bias, given that proportionally more patients in the nintedanib 150mg than the placebo arm did not complete the TOMORROW trial [n = 32 (38%) versus n = 24 (28%), respectively].

The ERG noted that the incidence of acute exacerbation data from the TOMORROW trial used in the NMA (shown in CS Table 30, p. 93) were not available in the trial publication.[1] NICE and the ERG therefore asked the company in their clarifications request to provide a citation and reference for these data (clarification A10). In their response, the company stated that AEs reported as “progression of IPF” in Table 2 of the trial paper[1] were used as a proxy measure for acute exacerbations. The company state that selection of this outcome as a proxy was based on the definition of acute exacerbations used in the trial and that the use of this proxy does not affect the results. The ERG has not been able to check this nor whether these outcomes are similarly defined, as “progression of IPF” is not defined in the trial paper.[1] Additionally, data on the incidence of acute exacerbations in the trial paper[1] were presented as number of events per 100 patient-years, which makes it difficult to directly compare the results of this outcome with those of the “progression of IPF” outcome, which were presented as the number and proportion of patients who experienced progression. The ERG, however, considers that overall, based on the company’s statement in their clarifications response, that the use of “progression of IPF” as a proxy is unlikely to be an issue.

Overall, the company’s outcome selection is appropriate and the company has included the outcomes that the ERG’s clinical expert considered to be the most clinically important, with the

exception of activities of daily living (which was not specified as an outcome of interest in the scope and which was not measured in the trials). The ERG is concerned, however, that the PFS outcome analysed in the NMA (but not a contributor to the economic model) may be subject to some bias because data from the TOMORROW trial were not included and no rationale is provided by the company for this omission (for INPULSIS data came from a post-hoc analysis of individual patient data but a similar analysis was not reported for the TOMORROW study). The ERG is additionally concerned that the results for the 'absolute change in DLco' outcome presented in the CS for the TOMORROW trial may be at risk of bias, as this definition of the DLco outcome differs to how it was pre-specified in the trial protocol (this outcome does not contribute to the economic model).

3.1.6 Description and critique of the company's approach to trial statistics

INPULSIS trials[2]

INPULSIS-1 and -2 were designed to assess the superiority of nintedanib compared to placebo on the annual rate of decline in FVC (ml/year) (primary outcome). A sample size power calculation was performed (90% power to detect a between-group difference of 100ml in the primary outcome (CS section 4.4, Table 14 p. 53 provides more detail).

The primary outcome was analysed using a random coefficient regression model which included gender, age and height as covariates. No rationale is given for these covariates, though they are standard variables used in the calculation of FVC percent predicted.

Efficacy and safety analyses were performed for randomised patients who received at least one dose of study medication (accounting for approximately 99.5% of the study population across the two trials) (NB. see below for description of the ITT analysis). A hierarchical procedure was used to assess superiority of nintedanib, for the primary outcome and two key secondary outcomes (change from baseline in SGRQ total score at 52 weeks; time to first acute IPF exacerbation over 52 weeks). The consecutive steps were considered only if the previous step was statistically significant and results favoured nintedanib. Note, the between group difference in primary outcome was statistically significant in both trials, but there was a difference between two trials in terms of the key secondary outcomes. In INPULSIS-2 the between group difference for both key secondary outcomes was significant allowing formal confirmatory testing for both

key secondary outcomes, however, in INPULSIS 1 neither were statistically significant hence statistical testing was done on a "nominal basis".[2]

The CS makes reference to an intention to treat population (CS p. 57-58) but no other explicit reference is given to ITT either in the CS or the trial journal publication.[2] The CS states that all randomised patients were included in the ITT population (CS page 57) though (as stated above and indicated in ERG report Table 4 the trial journal publication states that efficacy and safety analyses were performed for randomised patients who received at least one dose of study medication (approximately 99.5% of the randomised population).[2] Given the high proportion of patients who received medication this inconsistency in reporting isn't likely to signal bias.

The 'primary analysis' includes all available FVC values from baseline to week 52, including FVC measurements at the four week-follow up for patients who discontinued medication and did not complete study visits through week 52. This analysis assumed that missing data were missing at random and there was no imputation of missing data (other than the inclusion of follow-up data for the aforementioned patients who prematurely discontinued, also see ERG report Table 4). Multiple imputation sensitivity analyses were performed to assess the effects of missing data and to estimate the treatment effect for the primary outcome under a number of different assumptions about missing data (e.g. regarding rates of FVC decline amongst patients who died, and patients who discontinued). The multiple imputation analysis was based on the conservative assumption that missing data were informative rather than random. The results of the sensitivity analyses were consistent with the primary analyses (see supplemental figure S2 in the trial journal publication[2]).

For each trial there was no difference in the proportion of patients with missing data at week 52. The amount of missing data at week 52 (approximately 15%) was considered to be acceptable by the trial authors;[2] however, in the calculation of sample size (CS Table 14 p.53) it was assumed that it would not be possible to evaluate data for 2% of patients and based on this, a sample size of 194 in the placebo arm and 291 in the nintedanib arm was calculated. For each trial the proportion of missing data brought the sample size below these values.

A pre-specified pooled analysis of the two trials was conducted as an additional analysis, in order to improve the precision of the treatment effect estimates for the efficacy endpoints and to increase the size of the safety database. The statistical methods were the same as for the

individual trials, but with the addition of trial as a fixed effect or covariate in the models.[2] The ERG considers that the pooled analyses are appropriate, given the similarity of the trial designs. Note that the pooled data are used in the company's NMA and, in turn, in the economic model (for certain outcomes).

In terms of presentation of results, 95% confidence intervals and p values are provided for differences between nintedanib and placebo. Numbers of patients per trial arm in the analyses are provided (although clinical outcomes appear to be based on numbers randomised and NMA outcomes are based on numbers in receipt of at least one dose of study drug which were very slightly lower). Kaplan-Meier survival curves with accompanying hazard ratios (and 95% CIs, and p values) are given for time to event data (reported in the published trial paper but not in the CS).

Quantification of a clinically important difference is discussed for two outcomes: FVC% predicted, and the SGRQ. A 10 point difference was considered the minimally important difference in FVC% predicted, based on recent studies and consultation with clinical experts (CS p. 93). For the SGRQ it is noted that an MCID in the score has not been established for patients with IPF, but it is noted in the INPULSIS trial journal publication[2] that in patients with chronic obstructive pulmonary disease, this difference is 4 points. In the earlier (2011) journal publication of the TOMORROW trial[1] it is stated that the MCID was recently estimated as between 5 and 8 SGRQ points for IPF.

Although no subgroups were included in the scope and the decision problem, a pre-specified subgroup analyses of patients with baseline FVC ≤70% or >70% of predicted value were conducted using pooled data from the two INPULSIS trials (CS Section 4.8 p. 65-66). Post-hoc sub-group analyses are presented for patients with baseline FVC >90% or ≤90%, and for patients with or without emphysema at baseline. The ERG's view on the appropriateness of the FVC % predicted subgroups is presented in this ERG report section 2.3.

<u>TOMORROW trial</u>[1]

The trial was designed to assess the superiority of at least one of four doses of nintedanib compared to placebo. The primary outcome was the annual decline in FVC (L/year). A sample size power calculation was performed (80% power to detect a between-group difference of 0.1L in the primary outcome). The number of participants required in each group is not stated. A

random coefficient mixed linear regression model was used to calculate the decrease in FVC, based on a linear decrease in FVC over time (with terms for study group, age time, sex, height, and patient). Only on-treatment measurements were used in the primary analysis (no replacement of missing data was planned). A sensitivity analysis was conducted that included all visits (the baseline visit and all follow-up visits, including visits after discontinuation).

Efficacy analyses were based on the randomised set of patients (whether or treated or not, described earlier in ERG report section 3.1.4 and Table 4). Note that only 4 patients (0.9%) were randomised but did not receive treatment (the set omitting these 4 patients is referred to as the 'Treated set'). The ITT principle was used with patients assessed within the dose group to which they were randomly assigned, which is considered particularly appropriate in this trial given the potential for patients to request dose modifications (i.e. switch to another trial arm). All patients were encouraged to remain under their randomly assigned treatment.

The last observation carried forward (LOCF) approach in the case of missing values was not used in the analysis of the change in FVC over time. However, the LOCF approach was used for secondary outcomes when data for the entire 52 week assessment period were not available. No justification is given for use of this approach or its potential limitations. The use of LOCF to account for missing data could lead to favouring the treatment arm with earlier drop outs in a progressive disease such as IPF.

The safety analysis included all patients who received at least one dose of the study drug or placebo (99.1% of the randomised population).

Summary

In summary, the presentation of the results, in terms of use of CIs, numbers of patients and p values is adequate. The statistical procedures used in all three trials appear to be appropriate with the exception of the use of the LOCF approach which may bias in favour of nintedanib treatment.

3.1.7 Description and critique of the company's approach to the evidence synthesis

The company's evidence synthesis comprises a narrative review of the evidence supported by data tables. The evidence sources included clinical trial reports but these were not provided to the ERG (evidence sources for nintedanib are tabulated in CS Table 7 p. 39 and CS Table 8 p. 40). Other data sources were published articles, clinical trials records and conference abstracts which were either provided by the company or could be found via online sources. Where possible, the ERG has checked key data presented in the CS against those in the publications and conference abstracts cited by the company for consistency. However, it should be noted that some outcomes reported in the CS are reported in a different format to the published paper and therefore it was not possible to verify that these data are correct. The CS reports fewer outcomes and analyses than are presented in the published papers but this seems reasonable with the CS appearing to focus on the key outcomes and inputs for the economic model.

As no head-to-head trials comparing nintedanib with pirfenidone (the only pharmacological comparator included in the NICE scope for this appraisal) were identified, the company used NMA in the form of indirect treatment comparisons (ITC) to compare nintedanib with pirfenidone (CS p. 114). Meta-analyses are presented within the NMA results section of the CS (CS pages 115 to 141). The outcomes for which meta-analysis and NMA were undertaken are shown in Table 5 together with an indication of whether the data contributed to the economic model. The order of outcomes is presented with the primary outcome from the nintedanib trials first, followed by other outcomes that contribute data to the economic model and finally the outcomes that do not contribute to the economic model. The remainder of this section of the report will first describe the meta-analyses and then the NMA.

Table 5 Outcomes synthesised by meta-analysis and/or NMA

	Nintedanib trials meta-analysis?	**NMA for nintedanib vs pirfenidone ITC?**	**Input for economic model?**
Annual rate of decline in FVC (1ry outcome)	No	No	No
Other FVC related outcomes	Yes (loss of lung function: 10-points decrease in FVC% predicted)	Yes	Yes
All-cause mortality (described in the CS as overall survival)	Yes	Yes	Yes
Acute exacerbations	Yes	Yes	Yes
Serious cardiac events	Yes	Yes	Yes
Serious GI events	Yes	Yes	Yes
Overall discontinuations	Yes	Yes	Yes
Discontinuation due to AEs	Yes	Yes	No
PFS	No (INPULSIS data only in NMA)	Yes (pairwise comparison, no NAC data)	No
6MWD	No (Only TOMORROW[1] data in NMA)	Yes	No
Lung function - SpO_2	No	No	No
Lung function - Change in DLco	No	No	No
HRQoL	No	No	No

Meta-analyses

As already stated, meta-analyses are embedded within the NMA results section of the CS (CS pages 115 to 141) where they are presented in tabular form, with accompanying forest plots. The loss of lung function, mortality, overall discontinuation, and discontinuation due to AEs outcomes from the TOMORROW[1] and INPULSIS trials[2] were appropriate for meta-analysis because they were defined in the same way. Serious cardiac events and serious GI events

data were obtained from the Summary of Clinical Safety for each trial (relevant tables provided in the company's response to clarification question A19). Events were grouped by system organ class which is the highest level of the reporting hierarchy. So whilst these outcomes were suitable for meta-analysis it must be noted that there may have been heterogeneity in the serious events categorised under this term which would not be captured by the meta-analysis. Finally, there was one obvious difference in the acute exacerbation definitions (CS p.110-111), which was that the definition of acute exacerbation in the TOMORROW study included a decrease in $PaO_2 \geq 10$ mmHg or $PaO_2/FiO_2 < 225$ since last visit but this did not form part of the definition for the INPULSIS trials. However, all other aspects of the definition were similar. Although the methods have not been explicitly stated in the CS, heterogeneity in the meta-analyses of the nintedanib trials has been statistically assessed by means of the I^2 statistic. For five of the outcomes meta-analysed there was no statistical heterogeneity (I^2=0% for overall survival, acute exacerbations, loss of lung function, overall discontinuations and discontinuation due to AEs). There was a small amount of statistical heterogeneity in the meta-analysis of the TOMORROW and INPULSIS trials for the serious GI events outcome but this was not statistically significant (I^2 = 11.8%, chi-squared test p=0.287). Greater statistical heterogeneity was found for serious cardiac events (I^2 67.5%), which was not statistically significant at the conventional 5% cut off but would be considered statistically significant at the alternative 10% cut off (chi-squared test p=0.079).

Results from both fixed effect and random effects models are presented as relative differences (pooled odds ratios with 95% CIs and p-values).

The two INPULSIS RCTs[2] were pooled together as an input for meta-analysis. Therefore there were only two entries in each nintedanib meta-analysis (TOMORROW RCT[1] & pooled INPULSIS RCTs) hence sensitivity analyses for the nintedanib meta-analyses were not undertaken (some sensitivity analyses were undertaken in the context of the NMA as discussed below).

Network meta-analyses

The company used NMA for nine outcomes in the form of ITCs to compare nintedanib with pirfenidone (CS p. 114) in the absence of any head-to-head trials comparing nintedanib with pirfenidone (the only pharmacological comparator included in the NICE scope for this appraisal). Each NMA also included N-acetylcysteine (NAC) because during the initial stages of this STA

NAC was listed as a comparator and the company did not remove it from this part of the submission when the final NICE scope was published (NAC was not included in the economic model however). The trials of comparators contributing data to the NMA were all placebo controlled RCTs and therefore all comparisons were made via placebo (network diagrams are provided in CS figures 12 to 20 on CS p.91-100). The ERG therefore believes that NAC has little influence on the NMA results for nintedanib and pirfenidone. The intervention and comparator trials that were available for inclusion in each NMA are listed in Table 6 however not all trials presented data that could contribute to each NMA outcome. The ERG has not assessed the evidence or NMA results for NAC presented in the CS.

Table 6 Intervention and comparator trials identified for inclusion in the NMA

Nintedanib vs placebo trials	**Pirfenidone vs placebo trials**	***NAC vs placebo trials***
TOMORROW, Richeldi *et al*. 2011[1]	CAPACITY-1, Noble *et al*. 2011[21]	*Martinez et al. 2014*
INPULSIS-1, Richeldi *et al*. 2014[2]	CAPACITY-2, Noble *et al*. 2011[21]	*Homma et al. 2012*
INPULSIS-2, Richeldi *et al*. 2014[2]	ASCEND, King *et al.* 2014[3]	*Tomioka et al. 2005*
	SP2, Azuma *et al*. 2005[22]	
	SP3 Taniguchi *et al*. 2010[23]	

Studies in italic text are not relevant to this assessment because they investigated NAC.

The methodological description of the NMA is limited and not always presented in logical order (e.g. CS Table 28 p. 90 presents a summary of risks of bias in the included trials but the methodological description for this doesn't appear until CS p.114). The NMA appears to broadly follow conventional guidelines for systematic reviews (e.g. systematic search for evidence undertaken, quality of evidence assessed) although none are cited. Justification for some aspects of the analysis is lacking [e.g. the CS describes a feasibility assessment for the NMA (CS p. 103-109) but the purpose of this is not explicitly described]. The NMA was implemented in a Bayesian framework using WinBugs version 1.4.3. The Winbugs code was supplied in response to the NICE and the ERG request for this information (company's response to clarification question A5).

The company assessed the bias risks in the trials contributing to the NMA (CS Table 28, p. 90). The comparison of the ERG and company assessment of the nintedanib trials is presented earlier in this ERG report (Table 4). An assessment of the pirfenidone CAPACITY-1,[21] CAPACITY-2,[21] SP2[22] and SP3[23] trials was undertaken by the ERG for the pirfenidone STA[7]

and although this did not ask questions in the same format as for this current STA it is apparent that no concerns were raised regarding the CAPACITY trials. There was some uncertainty for SP2[22] and SP3[23] regarding the adequacy of allocation concealment and blinding (although both were described as double-blind trials) due to a lack of detail regarding these aspects in the published papers. The use of LOCF to account for missing data in SP2[22] and SP3[23] raised a concern about possible bias in favour of the treatment arm. The ASCEND trial[3] had not been published at the time of the pirfenidone STA and although the ERG have not formally assessed the risks of bias the RCT appears to have been well conducted.

The outcome data for loss of lung function (based on a 10 percentage point decrease in FVC% predicted) and PFS came from post-hoc analyses which are not published and therefore the ERG has been unable to verify these data (these outcomes are discussed earlier in this ERG report section 3.1.5 'Outcomes included in the NMA').

As stated, six of the nine outcomes assessed in the NMA were used in the economic model (mortality, acute exacerbations, loss of lung function, serious cardiac events, serious GI events, overall discontinuations, Table 5). There were differences between nintedanib and comparator (pirfenidone) trials (e.g. in terms of patient characteristics, outcome definitions) and there were also differences in potential effect modifiers (CS Table 40 p. 104) between trials (e.g. disease duration, study duration) which are discussed (CS p.104-109). CS p.109 states that four studies were excluded in sensitivity analyses due to differences in potential effect modifiers and the ERG presumes that the sensitivity analyses mentioned are the scenario analyses presented in CS Appendix B (a summary of the scenario analysis is presented in Table 7. The only discrepancy the ERG has identified is that for the overall mortality NMA, one of the four studies mentioned (Homma and colleagues[24]) is not excluded in any overall mortality scenario analysis however this study investigates NAC which is not included within the final scope for this STA.

The company compared the outcomes from their NMA that were also reported in a published NMA by Loveman and colleagues[25] and comment on observed discrepancies in results (CS p. 88).

Table 7 Summary of NMA evidence scenarios

	NMA Outcome						
Scenario	**loss of lung function**	**overall survival**	**acute exacerbations**	**serious cardiac events**	**serious GI events**	**overall discontinuation**	**discontinuation due to AEs**
1	All evidence	All evidence	All evidence	All evidence	All evidence	All evidence	All evidence
2	Excluding ASCEND (King[3])	Excluding ASCEND (King[3]), SP2 (Azuma[22]) and SP3 (Taniguchi[23])	*Excluding Homma[24]*	Excluding TOMORROW (Richeldi 2011[1])	Excluding TOMORROW (Richeldi 2011[1])	Excluding ASCEND (King[3]) and SP3 (Taniguchi[23])	Excluding ASCEND (King[3]), SP2 (Azuma[22]) and SP3 (Taniguchi[23])
3	Excluding TOMORROW (Richeldi 2011[1]) and ASCEND (King[3])	Excluding TOMORROW (Richeldi 2011[1]), ASCEND (King[3]), SP2 (Azuma[22]) and SP3 (Taniguchi[23])	Excluding SP2 (Azuma[22]) SP3 (Taniguchi[23]) *and Homma[24]*			Excluding ASCEND (King[3]), SP3 (Taniguchi[23]), and TOMORROW (Richeldi 2011[1])	Excluding ASCEND (King[3]), SP2 (Azuma[22]) SP3 (Taniguchi[23]), and TOMORROW (Richeldi 2011[1])
4	Including death (without CAPACITY, Noble[21])	Excluding SP2 (Azuma[22])	Excluding TOMORROW (Richeldi 2011[1]), SP2 (Azuma[22]) SP3 (Taniguchi[23]) *and Homma[24]*			Excluding ASCEND (King[3])	Excluding SP2 (Azuma[22]) and SP3 (Taniguchi[23]) (Japanese studies)

5	Including death (with CAPACITY, Noble[21])	Excluding SP2 (Azuma[22]) and SP3 (Taniguchi[23])	Excluding SP2 (Azuma[22])			Excluding SP3 (Taniguchi[23])	
6			Excluding SP2 (Azuma[22]) *and Homma[24]*				

Studies in italic text are not relevant to this assessment because they investigated NAC.

The analysis of PFS was a pairwise comparison and there was only a single trial in each arm of the 6MWT network (all via placebo) therefore there are no scenarios for these outcomes.

Table 8 presents the ERG's critical appraisal of the company's NMA. In general the NMA is judged to be of reasonable quality. The key caveats are:

- There are a relatively low number of trials contributing data for some outcomes. In particular, for three outcomes: acute exacerbation, loss of lung function and serious GI events, the comparison is between the three nintedanib trials (two replicate INPULSIS trials pooled data[2] and the TOMORROW trial[1]) and two replicate pirfenidone RCTs (Noble and colleagues[21] CAPACITY-1 & 2 pooled data). For the serious cardiac events outcome the comparison is essentially a pairwise comparison between the replicate INPULSIS and replicate CAPACITY trials.
- Although a rationale is provided for the exclusion of particular studies in the different NMA scenarios no overarching logic for the different scenarios across the outcomes was described. Consequently the ERG has some concerns regarding the potential for selection bias in favour of nintedanib among the outputs from the NMA.
- There are differences in study duration. In particular for nintedanib the replicate INPULSIS trials and the TOMORROW trial measured outcomes at 52 weeks whereas the replicate CAPACITY trials[21] which, as indicated in the preceding bullet point are the sole comparison for four outcomes that contribute data to the economic model, measured outcomes at 72 weeks. The CS itself (CS p.114) indicates that a discrepancy in study follow up length could introduce bias in to the analysis but does not discuss this further and no analyses were undertaken to explore the impact of this. The ERG believes that for a progressive disease such as IPF (where the median survival in the UK is between 2 and 5 years from the time of diagnosis) if trials enrol participants at the same point in their disease course then the trials with a shorter follow-up might be expected to observe fewer negative outcomes (e.g. exacerbations, decline in lung function, deaths) than trials with a longer follow-up. Clinical advice to the ERG indicated that a difference of 20 weeks might be too short to observe a difference in FVC and mortality.
- The choice of fixed or random effects model is based on the Deviance Information Criterion (DIC), whereby the model with the lowest numerical DIC value (indicating parsimony) is favoured. The CS provides NMA results for both fixed and random effects models for the all evidence NMA scenario (and for alternative evidence scenarios in CS Appendix B). In all but two of the NMA outcomes a fixed-effect was favoured, with random effects favoured for acute exacerbations and serious cardiac events (though with wide credible intervals). The ERG notes that the NMA input into the economic

model uses the all evidence NMA scenario for some outcomes, and alternative evidence scenarios (which omit certain trials) for others. In the case of acute exacerbations and serious cardiac events (which used NMA alternative evidence scenarios 3 and 2, respectively) a fixed-effect model was used in the economic model (based on the DIC for those respective evidence scenarios), which is in contrast to the random effects models used in the all evidence scenario (not used in the economic model). Since the point estimates can vary between random and fixed effect models the ERG has conducted a scenario analysis (section 4.3) which investigates the impact on cost-effectiveness by only using the all evidence scenario in the economic model for all outcomes, and for both random and fixed effect models.

Table 8 ERG appraisal of NMA approach

APPRAISAL CRITERIA	
Rationale and searches	
Is the rationale for the NMA and the study objectives clearly stated?	Yes [Executive summary (CS Section 1 p. 15 and CS Section 1.3 p. 19), not in the main clinical effectiveness section of the report.]
Does the reported study follow conventional guidelines for systematic reviews, as well as use explicit search terms, time frames, and avoid ad hoc data?	Yes it appears to although no guidelines are cited.
Are inclusion/exclusion criteria adequately reported?	Yes (CS Table 21 p. 67)
Is quality of the included studies assessed?	Yes (CS Table 28 p. 90)
Methods - Model	
Is the statistical model described?	Yes [The CS briefly indicates that a Bayesian framework was used and provides some description CS p. 114 & 115 (e.g. types of prior distributions given to parameters). The source code was supplied in response to the NICE and the ERG request for this information (company's response to clarification question A5)
Has the choice of outcome measure used in the analysis been justified?	No (Odds ratios were reported for all outcomes except 6MWT distance where weighted mean difference (WMD) is reported. No reasons were given or justification provided for choice of outcome measures however the ERG believes the measures are appropriate.)

Has the choice of fixed or random effects model been justified?	Yes. Fixed -effect and random effects models were used for all outcomes and the most appropriate model in each case was selected based on the DIC (CS p. 115). The DIC provides a numerical measure of goodness of model fit, with lower values favouring a more parsimonious model. The DIC is an appropriate method to select the type of model in Bayesian NMAs.[26] A DIC is reported for each NMA outcome and the accompanying text in the CS suggests which is model is favoured.
Has a structure of the network been provided?	Yes (Network diagrams were provided for each outcome. For PFS there is a pairwise comparison of nintedanib and pirfenidone via placebo and for 6MWT distance omitting NAC as a comparator leaves a pairwise comparison. Some of the tested scenarios which omitted studies become pairwise comparisons including scenario 2 for serious cardiac events which is used in the model.)
Is any of the programming code used in the statistical programme provided (for potential verification)?	Yes [Winbugs code was supplied in response to the NICE and the ERG request for this information (company's response to clarification question A5)].
Methods - Sensitivity analysis	
Does the analysis conduct sensitivity analyses?	Yes (described as scenario analyses)
Results	
Are the results of the NMA presented?	Yes
Does the study describe an assessment of the model fit?	Yes (in CS text for results of each outcome, CS p. 115-136)
Has there been any discussion around the model uncertainty?	Yes (some discussion amongst the results, CS p. 115-136)
Are the point estimates of the relative treatment effects accompanied by some measure of variance such as confidence intervals?	Yes (95% Credible Intervals are reported)
Discussion	
Does the study discuss both conceptual and statistical heterogeneity and incoherence?	Yes [There is some discussion of conceptual heterogeneity and statistical heterogeneity (a discussion of incoherence is not applicable as there was no direct evidence to compare with the indirect evidence)].
Does the discussion flow from the results seen?	Yes
Have the authors commented on how their results compare with other published studies (e.g. NMAs)?	Yes [A brief comparison with Loveman et al. 2015, published shortly before submission of the CS, is provided (CS p. 88-89)]

3.2 Summary statement of company's approach

The ERG's quality assessment of the review in the CS is summarised in Table 9. Processes for inclusion or exclusion of studies and for data extraction are described in the CS for the systematic review and the NMA (CS p. 34 and p. 37 respectively). Included studies were subject to critical appraisal. Overall, the ERG considers the study selection, data extraction and critical appraisal processes are adequate and they appear to follow standard accepted systematic review methodology.

The ERG concludes that the submitted evidence generally reflects the decision problem defined in the CS and considers the overall risk of systematic error in the review to be low.

Table 9 Quality assessment (CRD criteria) of CS review

CRD Quality Item: score Yes/ No/ Uncertain with comments	
1. Are any inclusion/exclusion criteria reported relating to the primary studies which address the review question?	Yes, inclusion and exclusion criteria are clearly stated.
2. Is there evidence of a substantial effort to search for all relevant research? ie all studies identified	Yes. There was substantial effort to search for all relevant studies, but only English-language studies were included in the systematic review and the NMA. The ERG note that there may be potential language bias, but this probably has not resulted in any missing studies.
3. Is the validity of included studies adequately assessed?	Uncertain. The validity of the studies is assessed in the CS using NICE suggested criteria. However, the ERG assessment differed from the CS assessment in two criteria.
4. Is sufficient detail of the individual studies presented?	Yes, overall methodology, patient characteristics and outcomes are described in sufficient detail. The ERG asked the company for details of patient flow (showing reasons for non-completion in the TOMORROW trial) and these data were provided in their clarification letter (clarification A2).
5. Are the primary studies summarised appropriately?	Yes, the primary studies are summarised appropriately, and details are presented in tables and figures.

3.3 Summary of submitted evidence

In this section of the report the ERG concentrates on the main outcomes of the included RCT evidence of nintedanib treatment at the licensed dose (150 mg BD) from the TOMORROW[1] and two INPULSIS RCTs.[2] Data have been reproduced here chiefly from the CS and supplemented

with some data from the trial journal publications.[1;2] The ERG was unable to verify the accuracy of some data presented in the CS because clinical study reports (CSRs) were not provided. There were a few minor discrepancies between the data presented in the CS and the data in the study publications which are noted either in the text or as footnotes to tables. Additional outcomes that were presented in the published papers but which were not included in the CS are not reported here.

The results of the company's NMA are also presented by outcome measure however results for NAC have not been included in this ERG report as NAC was not included as a comparator in the final NICE scope for the STA. Not all of the outcomes for which NMA was performed were used in the company's economic model.

The ERG presents the evidence in the following order:

- Efficacy outcomes that contribute data to the economic model
    - Annual rate of decline in FVC (primary outcome) and other FVC related outcomes
    - All-cause mortality
    - Acute exacerbations
- Efficacy outcomes subject to NMA but which did not contribute data to the model
    - PFS
    - 6MWT distance
- Efficacy outcomes not subject to NMA and which did not contribute data to the model
    - Lung function SpO_2
    - Lung function DLco
    - HRQoL
- Subgroup Analyses results
- Summary of Adverse Events
- Adverse event outcomes subject to NMA and contributing data to the model
    - Serious cardiac adverse events
    - Serious GI adverse events
- Overall discontinuations subject to NMA and contributing data to the model
- Discontinuations due to AEs subject to NMA but not contributing data to the model

Summary of results for lung function: FVC

The TOMORROW[1] and the INPULSIS trials[2] report loss of lung function as the annual rate of decline in FVC from baseline, measured in L or mL, which is the primary outcome used in the systematic review. For the INPULSIS trials, data are reported for the individual trials and from a pre-specified pooled analysis. Data are presented in Table 10 below.

The mean change from baseline was seen to favour nintedanib across all trials. In both of the individual INPULSIS trials and in the pooled INPULSIS analysis patients treated with nintedanib showed a significant reduction in FVC decline over 52 weeks when compared to placebo. In the TOMORROW trial, the difference between nintedanib treated patients and those treated with placebo was less pronounced. There was a non-significant difference in the rate of FVC decline between the nintedanib and the placebo groups when the pre-specified primary analysis method of a closed testing procedure for multiplicity was applied, but a statistically significant reduction was seen using the pre-specified alternate hierarchical testing procedure.

The mean difference in the annual rate of decline in FVC was 109.9 mL (pooled data: 95% CI 75.9 to 144.0, $p<0.001$) in the INPULSIS trials (INPULSIS-1: 125.3 mL; INPULSIS-2: 93.7 mL).

The CS describes narratively the difference in the rate of annual decline in FVC between the nintedanib and the placebo groups as clinically meaningful, in that nintedanib reduced the decline in FVC by 50%, when compared to placebo over 52 weeks (CS p.59).

The CS refers to published data on the natural history and progression of IPF, where the annual FVC decline is reported as 150-200mL in IPF patients as compared to 30-60 mL per year in elderly people without IPF (CS p.59). The ERG noted that the mean annual rate of FVC decline in the nintedanib patients is lower than the expected progression of IPF described above. The CS does not discuss the clinical relevance of the measured decline in FVC. However, the sample sizes of the INPULSIS trials were calculated to provide power for the detection of a between group difference of 100 mL in the annual rate of FVC decline, and a clinical expert consulted by the ERG confirmed that a 100ml decline in FVC is of significant clinical importance.

The TOMORROW[1] and the INPULSIS[2] trials also report various other measures related to FVC, and a number of these were reported in the CS and are presented in Table 10 below. All FVC

related outcomes except one showed significant differences in favour of nintedanib between patients treated with nintedanib and those in the placebo groups.

Table 10 Lung function: Change in FVC

	Nintedanib	**Placebo**	**MD/OR (95% CI) p-value**
TOMORROW	**N=86**	**N=87**	
Annual Rate of Decline in FVC, L/year[a] (SE) [95% CI]	-0.06 (0.04) [-0.14 to 0.02]	-0.19 (0.04) [-0.26 to -0.12]	p<0.05[b]
Absolute change in FVC at 52 weeks, L mean (SE) [95% CI]	-0.06 (0.04) [-0.13 to 0.01]	-0.23 (0.04) [-0.30 to -0.16]	p<0.01[c]
Patients with reduction in mean FVC of >10% or 200mL, n (%)	20 (23.8)	37 (44.0)	p<0.05[a]
Absolute change in FVC% predicted, % mean (SE) [95% CI]	-1.04 (0.99) [-2.98 to 0.91]	-6.00 (1.02) [-8.01 to -4.00]	p<0.001[d]
INPULSIS-1	**N=309**	**N= 206**	
Annual rate of decline in FVC (mL/yr)	-114.7	-239.9	MD: 125.3 (77.7 – 172.8) p<0.001
Adjusted absolute mean change from baseline FVC (mL)	-95.1	-205	MD: 109.9 (71.3 to 148.6) p<0.001
Adjusted absolute mean change from baseline in FVC - % of predicted value	-2.8	-6.0	MD: 3.2 (2.1 to 4.3) p<0.001
Patients (%) with an FVC decline ≤5 percentage points at week 52	163 (52.5%)	78 (38.2%)	OR: 1.85 (1.28 to 2.66) p=0.001
Patients (%) with an FVC decline ≤10 percentage points at week 52	218 (70.6%)	116 (56.9%)	OR: 1.91 (1.32 to 2.79) p<0.001
INPULSIS-2	**N =331**	**N = 220**	
Annual rate of decline in FVC (mL/yr)	-113.6	-207.3	MD: 93.7 (44.8 – 142.7) p<0.001

Adjusted absolute mean change from baseline FVC (mL)	-95.3	-205	MD: 109.8 (70.9 to 148.6) p<0.001
Adjusted absolute mean change from baseline in FVC - % of predicted value	-3.1	-6.2	MD: 3.1 (1.9 to 4.3) p<0.001
Patients (%) with an FVC decline ≤5 percentage points at week 52	175 (53.2%)	86 (39.3%)	OR: 1.79 (1.26 to 2.55) p=0.001
Patients (%) with an FVC decline ≤10 percentage points at week 52	229 (69.6%)	140 (63.9%)	OR: 1.29 (0.89 to 1.86) p=0.18
INPULSIS-1 & 2 pooled data	**N=638**	**N=423**	
Annual rate of decline in FVC (mL/yr)	-113.6	-223.5	MD: 109.9 (75.9 to 144.0) p<0.001

SE = standard error, MD = mean difference, OR = odds ratio.
[a] Although CS table 18 (p. 63) states that decline in FVC is expressed in mL per year, the ERG believes that this is an error and that the TOMORROW trial reports FVC as L per year.
[b] The ERG believes that there may be an error in CS table 18 (p. 63) where the p-value for the difference in the annual rate of decline in FVC between the study arms is reported as p=0.05, whereas in the narrative the CS describes the difference as non-significant and in the trial publication[1] the p-value is reported as p=0.06 from the closed testing procedure for multiplicity.
[c] There is a minor discrepancy between the p-value reported in the CS and reproduced here and the p-values reported in the supplement to the published paper for this outcome from the TOMORROW trial.[1]
[d] This p-value (for comparison with placebo, unadjusted) is not reported in the CS but has been taken from the supplement to the published paper[1]

The NMA for loss of lung function was not based on the primary outcome of the nintedanib trials but instead on a 10-point decrease in FVC% predicted, by the end of study follow-up. These data are not reported in the CS or in the published trial reports but came from a post-hoc analysis of observed data which the ERG has been unable to verify. In the NMA for loss of lung function 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 for the comparator pirfenidone two trials [Noble and colleagues (pooled CAPACITY-1 and -2),[21] and King and colleagues,[3] CS Table 31 CS p.94] (Table 11). However, the all evidence scenario was not used in the economic model. The contributing evidence for the model came from scenario 2 (CS Appendix B, p. 22 of 48) that excluded the King and colleagues[3] study because it introduced heterogeneity into the all evidence results (CS Table 59 CS p. 122). Consequently data for nintedanib came from a 52 week time point whereas the only trials contributing data on pirfenidone had a follow-up period of 72 weeks (Table 11). The CS states (CS p. 114) that "The

discrepancy in the study follow-up duration may have introduced bias in the analysis." However there is no further discussion to indicate which direction this bias might operate and no analysis was undertaken to explore the impact of bias due to study follow-up duration. The ERG believes that for a progressive disease such as IPF (where the median survival in the UK is between 2 and 5 years from the time of diagnosis) if trials enrol participants at the same point in their disease course then those with a shorter follow-up might be expected to observe less loss of lung function than those with longer follow up. However clinical advice to the ERG suggested that a difference of 20 weeks might be too short to observe a difference in FVC. In the economic model the fixed effect median odds ratio (OR) plus 95% credibility interval (CrI) for nintedanib versus placebo (OR 0.54 95% CrI 0.42 to 0.69) and pirfenidone versus placebo (OR 0.69 9% CrI 0.47 to 1.00) were used from scenario 2 (Table 11 and CS Appendix B p. 22 of 48). The fixed effect model was selected because it had the lowest DIC. Further discussion of the loss of lung function parameters used in the model is available in ERG report section 4.2.4iii. The corresponding median OR for the nintedanib vs pirfenidone comparison is 0.78 (95% CrI 0.49 to 1.22) indicating a potentially greater benefit from nintedanib than pirfenidone however as the credible interval includes one it cannot be concluded that the difference between the treatments is statistically significant (CS Appendix B Table 44).

Table 11 NMA Loss of lung function: 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, King et al.[3] (ASCEND) 52 wks
Fixed effect	0.54 (0.42 to 0.69)	0.54 (0.11 to 2.70)
Random effect	0.55 (0.41 to 0.72)	0.54 (0.11 to 2.69)
	Contributing evidence - Scenario 2 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	0.54 (0.42 to 0.69)	0.69 (0.47 to 1.00)
Random effect	0.54 (0.03 to 11.18)	0.69 (0.01 to 47.85)

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 12 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. However, the CS does not explain whether this includes data from only the licensed dose and placebo arms of the TOMORROW trial or from the full study (which included study arms with unlicensed doses). 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 12 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 and it is unclear to the ERG whether these results include data from the licensed dose and placebo arms of the TOMORROW trial only or from the full study.

Table 12 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)

SUPERSEDED – SEE ERRATUM

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. As reported above, it is also unclear to the ERG whether respiratory and on-treatment mortality data included all TOMORROW participants, regardless of nintedanib dose.

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 12). 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

SUPERSEDED - SEE ERRATUM

suggested that a difference of 20 weeks might be too short to observe a difference in mortality. In the economic model the fixed effect median OR plus 95% CrI for nintedanib versus placebo (OR 0.70 95% CrI 0.45 to 1.10) and pirfenidone versus placebo (OR 0.70 95% CrI 0.46 to 1.05) were used from the all evidence scenario (Table 12). Further discussion of the mortality parameters used in the model is available in ERG report section 4.2.4i. In comparison to placebo, the efficacy of nintedanib and pirfenidone were therefore very similar as indicated by the NMA output for the nintedanib vs. pirfenidone comparison where the median OR was 1.00 (95% CrI 0.55 to 1.85; CS Table 49 p. 117).

Table 13 NMA Overall survival (defined as all-cause mortality): Contributing evidence and NMA outcome

	Contributing evidence – all evidence	
	Nintedanib vs Placebo trials	**Pirfenidone vs Placebo trials**
	INPULSIS I & II[2], 52 wks TOMORROW[1], 52wks	Noble et al.[21] (CAPACITY I & II) 72 wks, King et al.[3] (ASCEND) 52 wks Azuma et al.[22] 36 wks, Taniguchi et al.[23] 52 wks
	Contributing evidence – All evidence scenario for model	
	NMA nintedanib vs. placebo Median OR(95% CrI)	**NMA pirfenidone vs placebo Median OR(95% CrI)**
Fixed effect	0.70 (0.45 to 1.10)	0.70 (0.46 to 1.05)
Random effect	0.70 (0.25 to 2.02)	0.70 (0.32 to 1.87)

Summary of results for acute exacerbations

Exacerbation rates were reported for the TOMORROW[1] and the two INPULSIS[2] trials and are presented in Table 14 below. Data from the INPULSIS trials were reported individually and from pooled data.

Acute exacerbation rate was defined as number of patients with at least one exacerbation within the 52-weeks' duration of the three nintedanib trials. The INPULSIS trials measured both investigator-reported and adjudicated acute exacerbations; and both are reported in the CS (tables 18 and 19, CS p. 63-64) and are presented in Table 14 below. The TOMORROW trial did not report how acute exacerbation was confirmed.

There was a significant decrease in the number of patients with at least one investigator-reported acute exacerbation in the nintedanib arm of the INPULSIS-2 trial, as compared to patients treated with placebo. However, no significant difference in investigator-reported acute exacerbation rates was found in INPULSIS-1.

In the TOMORROW trial there was a numerical reduction of acute exacerbation rates in nintedanib treated patients as compared to placebo and this was also observed in the INPULSIS trials, for both investigator-reported and adjudicated acute exacerbations when data from both trials were pooled. The CS does not comment on these data and no information was provided on the statistical significance of the differences observed between acute exacerbation rates in patients treated with nintedanib and those who received placebo treatment.

Table 14 Acute exacerbations within 52 weeks

	Nintedanib	**Placebo**	**HR (95% CI) p-value**
TOMORROW[1]	**N=86**	**N=87**	
Number (%) with ≥1 exacerbations	2 (2.3)	12 (13.8)	Not reported
INPULSIS-1	**N=309**	**N= 206**	
Number (%) with ≥1 investigator reported exacerbations	19 (6.1)	11 (5.4)	1.15 (0.54 to 2.42) p=0.673
Adjudicated acute exacerbations[a], number (%)	7 (2.3)	8 (3.9)	0.55 (0.20 to 1.54)
INPULSIS-2	**N =331**	**N = 220**	
Number (%) with ≥1 investigator reported exacerbations	12 (3.6)	21 (9.6)	0.38 (0.19 to 0.77) p=0.005
Adjudicated acute exacerbations[a], number (%)	5 (1.5)	16 (7.3)	0.20 (0.07 to 0.56)

INPULSIS-1 & 2 pooled data	**N=638**	**N=423**	
Number (%) with ≥1 investigator reported exacerbations	31 (4.9)	32 (7.6)	Not reported
Adjudicated acute exacerbations[a], number (%[b])	12 (1.9)	24 (5.7)	Not reported

[a] Confirmed or suspected adjudicated acute exacerbation events.

[b] Percentage calculated by ERG.

In their narrative (CS p. 59 and p. 62) the company commented on an analysis of pooled data from the INPULSIS trials and stated that there was a non-significant increase in the time to first investigator reported acute exacerbation, whereas a statistically significant increase was found in the time to first adjudicated acute exacerbation. The CS did not report detailed data to support this statement in their summary of clinical outcomes (CS tables 18 and 19). However, the company wrote in the executive summary (CS p. 14) that the significant increase in the time to first acute exacerbation in the nintedanib group was only observed in the INPULSIS-2 trial (HR: 0.38, p=0.005), whereas the increase was non-significant in INPULSIS-1 (HR: 1.15, p=0.67).

In the NMA for acute exacerbation 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 for the comparator pirfenidone three trials [Noble and colleagues (pooled CAPACITY-1 and -2),[21] Azuma and colleagues,[22] Taniguchi and colleagues.[23] CS Table 30 CS p.93] (Table 15). The all evidence scenario however was not used in the economic model. The reason for this is not explicitly stated in the CS but appears to be because of heterogeneity in the meta-analysis of pirfenidone studies, the poor fit of this NMA model and the high level of uncertainty around point estimates in the random effects model which had the lowest DIC (CS p. 119). The contributing evidence for acute exacerbations in the model came from scenario 3 (CS Appendix B p. 11 of 48) that excluded the Azuma and colleagues[22] and Taniguchi and colleagues[23] studies because these trials were conducted in Japanese patients. Consequently data for nintedanib came from a 52 week time point whereas the only trials contributing data on pirfenidone had a follow-up period of 72 weeks (Table 15). As already noted, this may have introduced bias in the analysis

(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 15 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 4.2.4ii. 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 2.85; 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 15 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)

SUPERSEDED SEE ERRATUM

Summary of results for progression-free survival

There were differences in the reporting and definition of PFS across intervention and comparator studies (CS p. 95). Therefore an analysis of individual patient data from the INPULSIS RCTs[2] was conducted by replicating the methods presented in Noble and colleagues[21] (pooled CAPACITY-1 and -2) and by use of their definition of PFS outcome. The CS states (CS p. 112) that the PFS outcomes from the INPULSIS[2] and the CAPACITY[21] trials are therefore comparable however it is not clear to the ERG if or how the difference in length of follow-up between the trials was accounted for. The analysis presented PFS as a hazard ratio with 95% confidence intervals Table 16.

Table 16 PFS evidence

Study	HR vs. placebo	95% CI Lower limit	95% CI Upper limit
Nintedanib (INPULSIS trials, Richeldi 2014[2])	0.74	0.61	0.91
Pirfenidone (CAPACITY trials, Noble 2011[21])	0.74	0.57	0.96

The pairwise comparison of PFS (reported within the NMA section of the CS p. 124) gave an estimated HR of nintedanib vs. pirfenidone of 1.00 (95% CrI 0.71 to 1.39); p-value 0.982. These results indicate no difference in PFS between nintedanib and pirfenidone. This outcome did not contribute to the economic model inputs.

Summary of results for 6-minute walk distance

This outcome was reported as change from baseline in the distance walked during the 6MWT by the TOMORROW trial[1] within the NMA section of the CS (CS Table 33 p. 97) and is reproduced in Table 17.

Table 17 6MWT distance

	Nintedanib	Placebo	Absolute difference (SE) 95% CI; p-value
TOMORROW	**N=86**	**N=87**	
Change in distance, m [baseline mean (SD), m]	-25.15 [437 (13.69)]	-26 [411.1 (15.9)]	6.32 (16.98) 27.08 to 39.72; p=0.7101

SE = standard error, SD = standard deviation

Although three pirfenidone studies [Noble and colleagues (pooled CAPACITY-1 and -2),[21] and King and colleagues[3]] measured this outcome only Noble and colleagues[21] reported data in the format required for the NMA. The CS indicates that the fixed effect model was a poor fit (CS p. 125) and credible intervals were very large (Table 18). This outcome did not contribute to the economic model inputs. Random effects model results are not reported.

Table 18 NMA 6MWT distance: contributing evidence and NMA outcomes

	Contributing evidence – all evidence	
	Nintedanib vs Placebo trials	**Pirfenidone vs Placebo trials**
	TOMORROW,[1] 52wks	Noble et al.[21] (CAPACITY I & II) 72wks
Median WMD (95% CrI)		
Fixed effect	6.2 (-26.5 to 38.8)	23.7 (4.1 to 43.4)
Random effect	Not reported	Not reported

WMD – Weighted mean difference

Summary of results for lung function: SpO_2

Absolute change in oxygen saturation (SpO_2) over 52 weeks was measured in all of the nintedanib trials.[1;2] Changes were generally smaller in the nintedanib treated patients than in those receiving placebo, but the difference between the groups was only significant in the TOMORROW trial (data in Table 19 below).

Table 19 Lung function: Change in SpO_2

	Nintedanib	**Placebo**	**MD (95% CI) p-value**
TOMORROW[a]	**N=86**	**N=87**	
Absolute change from baseline in SpO_2 over 52 weeks % mean (SE) [95% CI]	-0.18%, (0.36%) [-0.89 to 0.53]	-1.29% (0.37%) [-2.03 to -0.56]	not reported p<0.05
INPULSIS-1	**N=309**	**N= 206**	
Absolute change from baseline in SpO_2 over 52 weeks (%)	-0.24%	-0.53%	0.29% (-0.07 to 0.64) p=0.1138

INPULSIS-2	N =331	N = 220	
Absolute change from baseline in SpO_2 over 52 weeks (%)	-0.39%	-0.66%	0.27% (-0.15 to 0.69) p=0.2032

[a] Although the CS (CS Table 18) states this outcome is absolute change in worst SpO_2 during 6MWT the ERG believes this is an error and that these are resting values as indicated in the published paper for the TOMORROW trial.[1] Furthermore the trial protocol (which is available at NEJM.org) does not list SpO_2 during 6MWT as an outcome.

Summary of results for lung function: change in DLco

Carbon monoxide diffusion capacity (DLco) was also reported in the nintedanib trials (data in Table 20 below). Changes in DLco were generally similar between the nintedanib and the placebo groups.

Table 20 Lung function: Change in DLco

	Nintedanib	Placebo	MD (95% CI) p-value
TOMORROW[1]	**N=86**	**N=87**	
Absolute change in DLco[a]	-0.609 (0.1034)	-0.511 (0.1035)	Not reported
INPULSIS-1	**N=309**	**N= 206**	
Absolute change from baseline in DLco over 52 weeks, (mmol/min/kPa)	-0.380	-0.365	-0.015 (-0.191 to 0.161) p = 0.8650]
INPULSIS-2	**N =331**	**N = 220**	
Absolute change from baseline in DLco over 52 weeks, (mmol/min/kPa)	-0.286	-0.400	0.113 (-0.084 to 0.310) p = 0.2600

[a] The CS does not provide any units for this outcome. The ERG assumes that this is DLco mmol/min/kPa reported as mean (SD).

Summary of Health related quality of life

The CS systematic review reported data on health related quality of life from the TOMORROW[1] and the two INPULSIS trials[2] as measured by the SGRQ. These are presented in Table 21 below.

For the TOMORROW trial the CS reported SGRQ adjusted mean absolute change score from baseline and there was a significant difference between the nintedanib and the placebo group in favour of nintedanib. These data are also reported as change in SGRQ score from baseline versus placebo.

Mean change in SGRQ score from baseline was reported for the INPULSIS trials. In INPULSIS-2 the mean change in SGRQ was significantly smaller for nintedanib compared with placebo, favouring nintedanib. No significant difference between groups was measured in INPULSIS-1. However, in the narrative the CS reports a non-significant difference in favour of nintedanib on pooled analysis of the INPULSIS data (CS p. 62).

Table 21 HRQoL

	Nintedanib	**Placebo**	**MD (95% CI) p-value**
TOMORROW[1]	**N=86**	**N=87**	
SGRQ adjusted mean absolute change score from baseline[a] (SE) [95% CI]	-0.66 (1.71) [-4.02 to 2.71]	5.46 (1.73) [2.06 to 8.86]	p=0.007
SGRQ score (change from baseline vs. placebo)	-6.12 (-10.57 to -1.67)	NA	MD not reported p=0.0071
INPULSIS-1	**N=309**	**N= 206**	
SGRQ score (change from baseline)	4.34	4.39	-0.05 (-2.50 to 2.40) p=0.97
INPULSIS-2	**N =331**	**N = 220**	
SGRQ score (change from baseline)	2.80	5.48	-2.69 (-4.95 to -0.43) p=0.02

[a] Adjustment based on an ANCOVA with terms for treatment, baseline, region (all fixed effects)

Sub-group analyses results

Three subgroup analyses of pooled data from the INPULSIS-1 and -2 trials data were presented in the CS (CS section 4.8 p. 66). The CS does not indicate what proportion of the pooled INPULSIS trials population are in each subgroup.

Subgroup analysis: FVC ≤70% vs. >70%

This was a prespecified analysis. In response to clarification question A3 the company indicated that the FVC threshold was chosen to be consistent with subgroup analysis performed in the TOMORROW trial. The analysis was conducted for the primary end point (annual rate of decline in FVC) and what are described as 'key' secondary endpoints which are not listed. Safety was also assessed. No numerical data are provided but the CS states that no statistically significant differences in outcomes were found by subgroup.

Subgroup analysis: FVC ≤90% vs. >90%

This was a post-hoc analysis the purpose of which was to investigate whether patients with marginally impaired FVC receive the same benefit from nintedanib. The analysis appears to have been conducted for the primary end point (annual rate of decline in FVC) and 'key' secondary endpoints which are not listed. Safety was also assessed. Data provided for the primary endpoint are shown in Table 22 and the CS states that there was no significant treatment-by-subgroup interaction for this endpoint (p=0.5300). No further numerical data are presented but the CS states that no statistically significant differences in secondary outcomes were found by subgroup and the frequency of AEs and SAEs was comparable between the treatment arms of each subgroup.

Table 22 Lung function: Subgroup analysis FVC% predicted ≤90% versus >90%

	baseline FVC >90% predicted			baseline FVC ≤90% predicted		
	nintedanib	placebo	difference	nintedanib	placebo	difference
adjusted annual rate of decline in FVC, mL/year	91.5	-224.6	133.1 [95% CI: 68.0, 198.2]	-121.5	-223.6	102.1 [95% CI: 61.9, 142.3]

Subgroup analysis: Emphysema at baseline

This was a post-hoc analysis of patients with or without emphysema at baseline. It is not clear from the CS which outcomes the analysis was conducted for and no numerical data are

presented. The CS states that lung function decline was reduced with nintedanib in both groups and time to first investigator reported acute exacerbation and change from baseline in SGRQ total score were also consistent between those patients with and without emphysema at baseline.

Summary of adverse events

Table 23 reports adverse events, including those classified as severe, serious and fatal. The data are taken from the CS and supplemented with data from the supplements to the trial journal publications. Only key event data are reported here, with results for specific AEs and AEs requiring hospitalisation available in the CS and trial journal publications. For the TOMORROW trial[1] results for nintedanib are only given for the licensed 150mg BD dosage trial arm. AEs leading to study discontinuation are reported in the follow section 'Summary of discontinuations'

The proportion of patients with adverse events was generally similar between nintedanib and placebo. In the TOMORROW trial around 90% of patients reported occurrence of any adverse event. Common events included diarrhoea, cough, and nausea (CS Table 92). There were a higher proportion of fatal adverse events in the placebo arm than the nintedanib 150mg BD arm.

The proportion of patients with any adverse events was also similar between nintedanib and placebo patients in the INPULSIS trials,[2] at around 90%. As with the TOMORROW trial, diarrhoea was the most common AE (CS Table 93). The proportion of patients with serious AEs was around 30% and similar between trial arms. Fatal AEs were slightly higher for placebo than nintedanib patients.

Table 23 Adverse events

	Nintedanib	**Placebo**
	Number of patients (%)	
TOMORROW[1]	**N=85**	**N=85**
Any adverse event	80 (94.1)	77 (90.6)
Severe adverse events[a, b]	19 (22.4)	20 (23.5)
Serious adverse events[c]	23 (27.1)	26 (30.6)
Fatal adverse events	1 (1.2)	12 (14.1)

INPULSIS-1	N=309	N=204
Any adverse event	298 (96.4)	181 (88.7)
Any adverse event, excluding progression of IPF[d]	296 (95.8)	179 (87.7)
Severe adverse events[a]	81 (26.2)	37 (18.1)
Serious adverse events[c]	96 (31.1)	55 (27.0)
Fatal adverse events	12 (3.9)	10 (4.9)
INPULSIS-2	**N=329**	**N=219**
Any adverse event	311 (94.5)	198 (90.4)
Any adverse event, excluding progression of IPF[d]	311 (94.5)	197 (90.0)
Severe adverse events[a]	93 (28.3)	62 (28.3)
Serious adverse events[c]	98 (29.8)	72 (32.9)
Fatal adverse events	25 (7.6)	21 (9.6)

[a] A severe adverse event was defined as an event that was incapacitating or that caused an inability to work or to perform usual activities.

[b] The ERG believes an error has been made in the CS, Table 92 (CS p. 143) which reports 'SAEs' and defines these as serious adverse events. The ERG believes that these data are *severe* adverse events as reported in the trial journal publication.[1]

[c] A serious adverse event was defined as any adverse event that resulted in death, was immediately life-threatening, resulted in persistent or clinically significant disability or incapacity, required or prolonged hospitalization, was related to a congenital anomaly or birth defect, or was deemed serious for any other reason.

[d] Progression of IPF was defined according to the Medical Dictionary for Regulatory Activities, version 16.1, which includes disease worsening and exacerbations of IPF.

For the purposes of the NMA, adverse events of particular significance that occurred in at least one of the studies eligible for the NMA, were identified based on the criteria listed in the CS p.97. Two adverse events were identified, serious cardiac events and serious GI events. As already noted these events are grouped by system organ class and thus (as stated in the company's response to clarification questions A19 and A20) there may be heterogeneity in the serious events categorised under these terms.

The proportion of patients experiencing serious and fatal cardiac events is presented in Table 24. Proportions of serious events were low and generally similar between trial arms, with the exception of the TOMORROW trial where a higher proportion of placebo patients experienced an event. The proportion of fatal cardiac events was low, but was double in the placebo arm than the nintedanib arm (reported for INPULSIS only).

Table 24 Serious cardiac events

	Nintedanib	**Placebo**
	Number of patients (%)	
TOMORROW[1]	**N=85**	**N=85**
Serious cardiac AEs (%)	1 (1.2)	7 (8.2)
INPULSIS-1[a]	**N=309**	**N=204**
Serious cardiac AEs (%)	14 (4.5)	11 (5.4)
Fatal cardiac SAEs (%)	1 (0.3)	2 (1.0)
INPULSIS-2[a]	**N=329**	**N=219**
Serious cardiac AEs (%)	18 (5.5)	12 (5.5)
Fatal cardiac SAEs (%)	2 (0.6)	4 (1.8)
INPULSIS-1 & 2 pooled data	**N=638**	**N=423**
Serious cardiac AEs (%)	32 (5.0[b])	23 (5.4[b])
Fatal cardiac SAEs (%)	3 (0.5)[c]	6 (1.4)[c]

SAEs = serious adverse events

[a] Data for the INPULSIS trials were extracted by the ERG from a supplement to the published INPULSIS paper.[2] The ERG note that fatal cardiac SAEs are not reported in CS Table 34.

[b] Percentage calculated by ERG.

[c] Data pooled and percentage calculated by ERG.

Results of NMA on serious cardiac events

In the NMA for serious cardiac events 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 two comparator pirfenidone trials [Noble and colleagues (pooled CAPACITY-1 and -2),[21] CS Table 34 CS p.98] (Table 25). The all evidence scenario however was not used in the economic model. The contributing evidence for the model came from scenario 2 (CS Appendix B, p. 31 of 48) that excluded the TOMORROW RCT[1] because of heterogeneity in the all evidence results

and to only consider evidence from phase III trials. In common with other outcomes the data for nintedanib came from a 52 week time point whereas the data for pirfenidone came from a 72 week time point (Table 25). This may have introduced bias in the analysis (with trials of shorter duration potentially observing fewer serious cardiac events). In the economic model the fixed effect median OR plus 95% CrI for nintedanib versus placebo (OR 0.92 95% CrI 0.53 to 1.63) and pirfenidone versus placebo (OR 1.27 95% CrI 0.66 to 2.49) were used from scenario 2 (Table 25 and Appendix B, p.31 of 48). The corresponding median OR for the nintedanib vs pirfenidone comparison is 0.73 (95% CrI 0.31 to 1.74) with the point estimate suggesting a greater benefit from nintedanib than pirfenidone however as the credible interval includes one it cannot be concluded that the difference between the treatments is statistically significant (CS Appendix B Table 59).

Table 25 NMA serious cardiac events: 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
Fixed effect	0.76 (0.45 to 1.27)	1.26 (0.65 to 2.49)
Random effect	0.42 (0 to 21.16)	1.26 (0 to 459.98)
	Contributing evidence - Scenario 2 for model	
	NMA nintedanib vs. placebo	**NMA pirfenidone vs placebo**
Median OR (95% CrI)	INPULSIS I & II,[2] 52 wks	Noble et al.[21] (CAPACITY I & II) 72wks
Fixed effect	0.92 (0.53 to 1.63)	1.27 (0.66 to 2.49)
Random effect	0.93 (0 to 527.43)	1.28 (0 to 707.71)

The proportion of patients with serious GI events was low (<5%) but higher amongst nintedanib-treated patients compared to those treated with placebo (Table 26).

Table 26 Serious gastro-intestinal events

	Nintedanib	Placebo
TOMORROW	**N=85**	**N=85**
Number of patients (%)	4 (4.7%)	0 (0.0%)
INPULSIS-1 & 2 pooled data	**N=638**	**N=423**
Number of patients (%)	19 (3.0)[a]	7 (1.7)[a]

[a] Percentage calculated by ERG

Results of NMA on serious GI events

In the NMA for serious GI events 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 two comparator pirfenidone trials [Noble and colleagues (pooled CAPACITY-1 and -2),[21] CS Table 35 CS p.99] (Table 27). The all evidence scenario was used in the economic model and in common with other outcomes the data for nintedanib came from a 52 week time point whereas the data for pirfenidone came from a 72 week time point (Table 27). This may have introduced bias in the analysis (with trials of shorter duration potentially observing fewer serious GI events). In the economic model the fixed effect median OR plus 95% CrI for nintedanib versus placebo (OR 2.35 95% CrI 1.05 to 5.88) and pirfenidone versus placebo (OR 0.60 95% CrI 0.23 to 1.45) were used. The corresponding median OR for the nintedanib vs pirfenidone comparison is 3.96 (95% CrI 1.18 to 14.51) indicating a greater benefit from pirfenidone than nintedanib (CS Table 78).

Table 27 NMA serious gastro-intestinal adverse events: 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
Fixed effect	2.35 (1.05 to 5.88)	0.60 (0.23 to 1.45)
Random effect	3.52 (0.08 to 429.92)	0.59 (0 to 178.99)

Summary of discontinuations

Data on overall discontinuations and discontinuations due to adverse events have been included here because there were outcomes analysed by NMA and the overall discontinuation data contributes to the economic model. For the TOMORROW trial[1] results for nintedanib are only given for the licensed 150mg BD dosage trial arm.

Therapy was discontinued for any reason in a smaller proportion of participants in the placebo arm of the trials than in the nintedanib arms of the trials (Table 28) although the difference in the proportions is small (ranging from 3.6% in INPULSIS-2 to 9.4% in TOMORROW)

Table 28 Overall discontinuations

	Nintedanib	**Placebo**
	Number of patients (%)	
TOMORROW	**N=85**	**N=85**
Overall discontinuation	32 (37.6%)	24 (28.2%)
INPULSIS-1	**N=309**	**N=204**
Overall discontinuation	78 (25.2%)	36 (17.6%)
INPULSIS-2	**N=329**	**N=219**
Overall discontinuation	78 (23.7%)	44 (20.1%)

Results of NMA on overall discontinuation

In the NMA for overall discontinuation the 'All evidence' scenario contributed inputs to the economic model. 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 four comparator pirfenidone trials [Noble and colleagues (pooled CAPACITY-1 and -2),[21] King and colleagues[3] and Taniguchi and colleagues[23] CS Table 36 CS p.100] (Table 29). The data for nintedanib came from a 52 week time point whereas the data for pirfenidone came from either a 52 week time point (2 trials) or a 72 week time point (one trial) (Table 29). The impact of the differences in trial time points on the outcome is unclear. In the economic model the fixed effect median OR plus 95% CrI for nintedanib versus placebo (OR 1.41 95% CrI 1.07 to 1.86) and pirfenidone versus placebo (OR 1.35 95% CrI 1.04 to 1.74) were used. The corresponding median OR for the nintedanib vs pirfenidone comparison is 1.06 (95% CrI 0.73 to 1.54) which shows the credible interval includes one so it cannot be concluded that the difference between the

treatments is statistically significant (Incorrect data were presented in CS Appendix B Table 44, the correct data were supplied as part of the company's response to clarification question A13).

Table 29 NMA overall discontinuation: 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, King et al.[3] (ASCEND) 52 wks Taniguchi et al.[23] (SP3) 52 wks
Fixed effect	1.42 (1.08, 1.87)	1.34 (1.04, 1.73)
Random effect	1.43 (0.79, 2.63)	1.35 (0.83, 2.24)

The proportion of patients discontinuing due to AEs was generally similar between nintedanib and placebo in the TOMORROW[1] and INPULSIS-2 trials[2] (Table 30). However, in INPULSIS-1 discontinuations for nintedanib patients were almost double those of the placebo group. The overall proportion of discontinuations was higher in the TOMORROW trial than it was in the INPULSIS trials (28% compared to 17%, respectively).

Table 30 Discontinuation due to adverse events

	Nintedanib	**Placebo**
	Number of patients (%)	
TOMORROW	**N=85**	**N=85**
Adverse events leading to discontinuation	26 (30.6)	22 (25.9)
INPULSIS-1	**N=309**	**N=204**
Adverse events leading to discontinuation[a]	65 (21.0)	22 (10.8)[b]
INPULSIS-2	**N=329**	**N=219**
Adverse events leading to discontinuation[a]	58 (17.6)[b]	33 (15.1)[b]

[a] Adverse events leading to study-drug discontinuation were reported when they occurred in 2% or more of patients in any study group and are listed according to system organ class. The analysis included adverse events with an onset after administration of the first dose of study medication and up to 28 days

after administration of the last dose. Investigation results refer to the results of clinical laboratory tests, radiologic tests, physical examination, and physiologic tests.

[b] Taken from CS table 93. Figures provided in the trial journal publication are slightly higher.

Results of NMA on discontinuation due to AEs – not in model

The outcomes from this NMA did not contribute to the economic model inputs (Table 31). 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 all five pirfenidone trials [Noble and colleagues (pooled CAPACITY-1 and -2),[21] King and colleagues,[3] Azuma and colleagues,[22] Taniguchi and colleagues.[23] CS Table 36, p. 99]. Nintedanib and pirfenidone were each associated with more discontinuations due to adverse events than placebo.

Table 31 NMA discontinuation due to adverse events: Contributing evidence and NMA outcomes

	Contributing evidence – all evidence	
	Nintedanib vs Placebo trials	**Pirfenidone vs Placebo trials**
Median OR (95% CrI)	INPULSIS I & II, 52 wks TOMORROW,[1] 52wks	Noble et al.[21] (CAPACITY I & II) 72wks King et al.[3] (ASCEND) 52 wks Azuma et al.,[22] 36 wks Tanaguchi et al.[23] (SP3) 52 wks
Fixed effect	1.52 (1.12 to 2.08)	1.73 (1.27 to 2.39)
Random effect	1.50 (0.72 to 2.92)	1.78 (1.09 to 3.35)

3.4 Summary

The ERG considers that the CS presents a generally unbiased estimate of the treatment effect of nintedanib for adults with IPF within the stated scope of the decision problem although there are some exceptions and uncertainties as described below.

The CS is based on a systematic review of clinical effectiveness which includes one phase two RCT (the TOMORROW trial[1]) and two (replicate) phase three RCTs (INPULSIS-1 and -2).[2] All three included trials were placebo controlled RCTs that enrolled adults with IPF who had an

FVC that was 50% or more of the predicted value and they were judged to be of reasonable quality. The final NICE scope specified pirfenidone and best supportive care as comparators but no head to head trials of nintedanib versus pirfenidone were identified by the systematic review therefore an NMA was conducted to provide supporting evidence for this comparison. The NMA includes additional evidence for NAC as a comparator because this was a listed comparator in the draft NICE scope however it was removed for the final NICE scope. The ERG has not assessed this evidence and it does not contribute to the economic model. The ERG believes that the relevant evidence has been identified by the systematic review of clinical effectiveness and by the searches that underpin the NMA.

The NMA consisted of indirect treatment comparisons linking nintedanib and pirfenidone through the common comparator of placebo (hence the ERG believes that the NAC vs placebo trials would have had little influence on the NMA results for nintedanib and pirfenidone). The three nintedanib trials and five pirfenidone trials (CAPACITY-1 and -2;[21] ASCEND,[3] SP2[22] and SP3[23]) were available to contribute data to the NMA however not all trials reported data that could contribute to each NMA outcome. The CS presents NMA results for each outcome from an 'all evidence' scenario' i.e. including all the available evidence. However for most outcomes one or more scenario analyses were conducted in which a trial (or trials) was excluded from the NMA. The scenario analyses conducted varied for the different outcomes and although a rationale was given for excluding certain studies in the different scenarios no overarching logic across the outcomes was described. This creates an impression (potentially falsely) that scenario analyses may have been tried until one was found that provided a favourable result. Consequently the ERG has some concerns regarding the potential for selection bias in favour of nintedanib among the outputs from the NMA.

The results of the RCTs showed that, at the licensed dose of nintedanib (150mg BD) in comparison to BSC (the placebo arm of the trials) there was a statistically significant improvement in the annual rate of decline in FVC from baseline which was the primary outcome for each trial. Statistically significant differences in favour of nintedanib were reported for all but one of the other FVC based outcome measures. The NMA for loss of lung function was not based on the primary outcome of the nintedanib trials but instead on a 10-point decrease in FVC% predicted, by the end of study follow-up. These data are not reported in the CS or in the published trial reports but came from a post-hoc analysis of observed data which the ERG has been unable to verify. The NMA conducted using all the available evidence produced similar

median ORs for the nintedanib vs placebo and for the pirfenidone vs placebo comparison indicating little difference between nintedanib and pirfenidone. However, there was heterogeneity in this scenario due to one of the included pirfenidone studies (ASCEND trial by King and colleagues[3]) so the contributing evidence for the model came from a scenario that excluded this study. The NMA that contributed to the model indicated a greater benefit from nintedanib (median OR 0.54 95% CrI 0.42 to 0.69) than pirfenidone (median OR 0.69 95% CrI 0.47 to 1.00). A statistically significant difference between nintedanib and placebo was not observed consistently across all three trials for any other of the reported outcomes. NMA data for two other effectiveness outcomes (overall survival which was defined as all-cause mortality and acute exacerbations) contributed to the economic model. Neither nintedanib nor pirfenidone have been shown to have a statistically significant impact on overall survival and the NMA demonstrated that the effect of the two drugs was very similar. For acute exacerbations the NMA all evidence model (where the random effects model had the lowest DIC) was a poor fit with a high level of uncertainty. A scenario analysis that excluded three studies conducted in Japanese patients produced a median OR indicating a benefit with nintedanib which was not apparent for pirfenidone.

The proportion of patients with adverse events was generally similar between the nintedanib and placebo groups of the three trials (TOMORROW,[1] INPULSIS-1 and -2[2]). Slightly more fatal adverse events occurred in the placebo arms of the trials than in the nintedanib arms. Proportions of patients experiencing serious cardiac adverse events were low and generally similar between trial arms of the INPULSIS trials[2] but in the TOMORROW trial[1] a higher proportion of events occurred in the placebo arm. Fatal cardiac events were reported for the INPULSIS trials[2] and the proportion occurring was low but double in the placebo arms compared to the nintedanib arms of the pooled analysis. The proportion of patients experiencing a serious GI event was low but higher amongst nintedanib treated patients. A higher proportion of participants in the TOMORROW trial[1] experienced adverse events that led to discontinuation than was observed in the INPULSIS trials.[2] Similar proportions of patients discontinued due to adverse events in the nintedanib and placebo arms of the TOMORROW[1] and INPULSIS-2 trials but in INPULSIS-1 the proportion of nintedanib arm patients discontinuing due to adverse events was double that of the placebo group. [2]

The company's interpretation of the evidence presented in the CS on the effectiveness of nintedanib in comparison to placebo (BSC) is on the whole appropriate. The ERG has identified one area of uncertainty:

- The key clinical trials on the effectiveness of nintedanib enrolled participants with an FVC that was 50% or more of the predicted value thus these trials do not provide evidence for patients with an FVC of less than 50% predicted. However, the ERG acknowledges that there is no restriction in the licence for nintedanib based on severity.

The ERG also has some concerns about the comparison of nintedanib with pirfenidone. The concerns and uncertainties identified by the ERG are as follows:

- There is a lack of any direct evidence comparing nintedanib with pirfenidone therefore the comparison of these two drugs relies on indirect evidence from an NMA for each outcome of interest.
- There is a potential for bias in the selection of evidence contributing to each NMA. This is particularly important to bear in mind for outcomes that contribute to the economic model which didn't use the 'All Evidence' scenario (loss of lung function, acute exacerbation and serious cardiac events). However the ERG acknowledges that there is a potential tension between the inclusion of all available evidence (to reflect diversity and uncertainty) and restricting evidence (e.g. by excluding Japanese studies) to better reflect the characteristics of the UK population within the included evidence.
- There is uncertainty about the impact of the differing lengths of trial follow up among trials contributing to each NMA which could potentially disadvantage pirfenidone (typically 52 weeks for nintedanib but for several outcomes the only pirfenidone evidence is from a 72 week time point).

4 ECONOMIC EVALUATION

4.1 Overview of the company's economic evaluation

The company's submission to NICE includes:

i) A review of published economic evaluations of nintedanib compared with pirfenidone, N-acetylcycteine (NAC) and best supportive care (BSC) (placebo) for patients with IPF.
ii) A report of an economic evaluation undertaken for the NICE STA process. The cost effectiveness of nintedanib is compared with that of pirfenidone and best supportive care.

Company's review of published economic evaluations

A systematic search of the literature was conducted by the company to identify economic evaluations of nintedanib for the treatment of IPF. An additional non-systematic search was performed and found one study by Loveman and colleagues.[27] See section 3.1.3 of this report for the ERG critique of the search strategy.

Cost effectiveness analysis methods

The economic analysis used a Markov model to estimate the cost-effectiveness of nintedanib compared with pirfenidone and BSC in adult patients with IPF. The model adopted a lifetime horizon to capture all the accrued costs and HRQoL over the patients' lifetime, with a cycle length of 3 months. The economic evaluation was conducted from the perspective of NHS and Personal Social Service (PSS). Costs and benefits were discounted at 3.5% per annum and half cycle correction was incorporated. Although NAC was initially scoped as a relevant comparator, the CS does not present any cost effectiveness analysis of nintedanib compared to this treatment strategy.

The economic evaluation used pooled data from the nintedanib phase II and phase III trials: the TOMORROW and INPULSIS trials.[1;2]

Disease progression in the Markov model was measured by FVC% predicted to account for the absolute health state of the patients adjusted for lung capacity, age, gender and height. FVC% predicted was categorised on a 10-point scale which then defined 10 mutually exclusive health states with and without exacerbation. Death was the other health state. Patients entered the

model at different FVC% predicted health states without exacerbation from where they could die, progress to a health state with more severe lung function, suffer an acute exacerbation, progress to a health state with lower FVC% predicted combined with exacerbation, or remain in the same health state. The starting population was based on the characteristics of patients included in the nintedanib phase III INPULSIS trials. The model accounted for treatment efficacy through change in mortality (overall survival), acute IPF exacerbation and decline in lung function. Baseline risks of these parameters were estimated from the placebo arm of INPULSIS[2] and TOMORROW[1] trials and were extrapolated beyond the 52-week trial duration by fitting parametric models. The relative effectiveness of nintedanib and pirfenidone were obtained by applying respective ORs to the baseline risks. The ORs were based on the analyses from the NMA discussed in section 3.1.5 of this report.

The results of the economic evaluation were presented for the following base case assumptions: patients died when their level of FVC% predicted dropped to 30-39.9; disease progression in the baseline was defined as a 10-point drop in FVC% predicted; patients who progressed to a lower FVC% predicted could not move back to health states with higher FVC% predicted; acute exacerbation had no impact on loss of lung function in the base case analysis; liver enzyme elevations were assumed to be asymptomatic for IPF patients; and patients were assumed to receive palliative care at their end of life. A list of other assumptions related to costs and utilities are listed in CS Table 160 (p. 254-260).

Overall baseline treatment discontinuation was based on clinical trial data, and the relative discontinuation risks for nintedanib and pirfenidone were estimated by applying ORs obtained from the NMA. Serious cardiac events, serious GI events along with skin disorders and GI perforations were the adverse events included in the economic analyses.

HRQoL was included in the model through the use of utility values assigned to each health state as defined by FVC% predicted category. These values were obtained from a data analysis based on the INPULSIS trials. In addition, disutilities associated with exacerbation and treatment related adverse events were also incorporated. These values were obtained from a study by Ara and Brazier[28] and an analysis based on the INPULSIS trials.[2]

Costs were included for drug treatments, liver function tests, adverse events, resource use, health state costs, oxygen use, exacerbations and end of life care costs. These were sourced from MIMS,[29] NHS Reference costs 2012/13,[30] PSSRU 2013,[31] and INPULSIS trial analyses.

Deterministic and scenario analyses were performed by the company to check for model uncertainty (CS p.286-301). A probabilistic sensitivity analysis (PSA) was also conducted and the input parameters are described in CS Table 175 (p. 279-81). Validation of the cost effectiveness analysis was conducted through external review by clinical experts and verification by the model developers and the company. Further, validation of overall survival, exacerbation and the FVC% predicted distribution at the end of the first year was also performed.

Cost effectiveness analysis results

Results from the economic model are presented in CS section 5.7.1 and section 5.7.2 (p. 260-267) as incremental cost per quality-adjusted life years (QALY); incremental cost per life years gained and incremental cost per exacerbation avoided for nintedanib vs pirfenidone and nintedanib vs BSC at nintedanib list price and with the nintedanib PAS. Total and incremental costs, life years gained (LYG) and QALYs were also reported, along with a breakdown of total costs. The results of the cost effectiveness analysis of nintedanib vs pirfenidone at nintedanib list price and with the PAS showed that nintedanib dominated pirfenidone. For nintedanib vs BSC, the estimated ICER was £149,361 at nintedanib list price (see Table 32) and [redacted] with PAS incorporated in nintedanib price (see Table 33).

Table 32 Base case results of cost effectiveness analyses at the nintedanib list price (CS Table 165 p.266)

	Total costs (£)	Total LYG	Total QALYs	Incremental costs (£)	Incremental LYG	Incremental QALYs	ICER (£) vs. baseline (vs. BSC) (QALYs)	ICER(£) incremental (QALYs)
BSC	£25,359	4.36	3.27					
PFN	£87,479	4.86	3.62	£62,120	0.49	0.35	£176,081	Dominated by NDB
NDB	£85,088	4.86	3.67	-£2,392	0.00	0.05	£149,361	£149,361

BSC: Best Supportive Care; PFN: Pirfenidone; NDB: Nintedanib; ICER: Incremental cost effectiveness ratio; LYG: Life years gained; QALYs: Quality adjusted life years; All decimals have been rounded to two decimal places

Table 33 Base case results of cost effectiveness analyses at the nintedanib PAS price (CS Table 166 p.267)

	Total costs (£)	Total LYG	Total QALYs	Incremental costs (£)	Incremental LYG	Incremental QALYs	ICER (£) vs. baseline (vs. BSC) (QALYs)	ICER(£) incremental (QALYs)
BSC	£25,359	4.36	3.27					
PFN	£87,479	4.86	3.62	£62,120	0.49	0.35	£176,081	Dominated by NDB
NDB	[redacted]	4.86	3.67	[redacted]	0.00	0.05	[redacted]	[redacted]

BSC: Best Supportive Care; PFN: Pirfenidone; NDB: Nintedanib; ICER: Incremental cost effectiveness ratio; LYG: Life years gained; QALYs: Quality adjusted life years; All decimals have been rounded to two decimal places

In the deterministic sensitivity analyses of nintedanib vs pirfenidone, nintedanib dominated pirfenidone in all the analyses, except for a scenario in which a stopping rule for pirfenidone was applied to discontinue treatment in patients who declined by >10%FVC in one year. Model results were most sensitive to changes in the mortality in the analyses comparing nintedanib vs. BSC. The results from the PSA indicated that the probability of nintedanib being cost-effective was 60% compared to pirfenidone at any willingness-to-pay threshold (CS p. 282).

4.2 Critical appraisal of the company's submitted economic evaluation

The company's review of published economic evaluations

The eligibility criteria for the systematic literature review of economic evaluations conducted by the company are listed in CS section 5.1 (p. 154). The inclusion criteria stated that cost utility analysis, cost benefit analysis, cost effectiveness analysis, cost consequence analysis and cost minimization analysis of nintedanib in comparison with pirfenidone, NAC and BSC in adult patients with IPF were included. Studies that included people aged less than 18 years or healthy individuals were excluded. In addition, studies were excluded that did not contain nintedanib, had no outcomes reported, were reviews or critical appraisals of economic evaluations. The ERG considered the eligibility criteria adopted by the company to be reasonable and appropriate.

The systematic search identified a total of 10 potential relevant studies from screening abstracts. None of these 10 studies met the eligibility criteria; reasons for ineligibility were not stated. No study was included for full review. A non-systematic search identified one study by Loveman and colleagues[27] that met the pre-specified eligibility criteria. This was a UK based study that conducted a systematic review, network meta-analysis and economic evaluation for treatment of patients with initially unprogressed IPF. However, there were distinct differences between the analysis conducted by Loveman and colleagues[27] and the company. First, the economic model developed by Loveman and colleagues[27] included four health states: unprogressed IPF, progressed IPF, lung transplant and death unlike the one developed by the company as described above. Secondly, the NMA performed by Loveman and colleagues[27] did not include the INPULSIS[2] and ASCEND[3] trials and had some methodological differences in how clinical effectiveness was analysed. Finally, Loveman and colleagues[27] did not have the correct list price for nintedanib in their analysis.

The ERG checked the search strategy for the cost effectiveness searches and found them to be reasonably comprehensive, fit for purpose and reproducible. An additional unstructured search was conducted by the ERG which identified a further economic evaluation by Hagaman and colleagues.[32] However, this study did not meet the inclusion criteria of the company submission as it did not include nintedanib. It was therefore justified as excluded from the company submission.

Critical appraisal of the company's submitted economic evaluation

The ERG assessed the methods applied in the economic evaluation in the context of the critical appraisal questions listed in Table 34 below. This list of questions is drawn from common checklists for economic evaluation methods (e.g. Drummond and colleagues[33]).[20;34] Overall, the ERG concludes that the company followed recommended methodological guidance.

Table 34 Critical appraisal checklist for the economic evaluation

Item	Critical Appraisal answer	Reviewer Comment
Is there a well-defined question?	Yes	
Is there a clear description of alternatives?	Yes	
Has the correct patient group / population of interest been clearly stated?	Yes	Discussed in section 4.2.2
Is the correct comparator used?	Yes	
Is the study type reasonable?	Yes	
Is the perspective of the analysis clearly stated?	Yes	
Is the perspective employed appropriate?	Yes	Discussed in sections 4.2.7 for costs and 4.2.5 for outcomes
Is effectiveness of the intervention established?	Yes	Treatment effectiveness shown in TOMORROW and INPULSIS trials
Has a lifetime horizon been used for analysis (has a shorter horizon been justified)?	Yes	Discussed in section 4.2.1
Are the costs and consequences consistent with the perspective employed?	Yes	Discussed in sections 4.2.7 for costs and 4.2.5 for outcomes
Is differential timing considered?	Yes	Described in section 4.1 Discussed in section 4.2.1
Is incremental analysis performed?	Yes	
Is sensitivity analysis undertaken and presented clearly?	Yes	Described in section 4.1. Discussed in section 4.2.9

NICE reference case

The ERG also considered the requirements of the NICE reference case for critical appraisal of the submitted economic evaluation, as shown in Table 35.

Table 35 NICE reference case requirements

NICE reference case requirements:	Included in submission	Comment
Decision problem: As per the scope developed by NICE	Yes	
Comparator: Alternative therapies routinely used in the UK NHS	Yes	Discussed in section 4.2.3
Perspective on costs: NHS and PSS	Yes	
Perspective on outcomes: All health effects on individuals	Yes	Discussed in section 4.2.5
Type of economic evaluation: Cost effectiveness analysis	Yes	
Synthesis of evidence on outcomes: Based on a systematic review	Yes	Discussed in section 4.2.5
Measure of health benefits: QALYs	Yes	Discussed in section 4.2.5
Description of health states for QALY calculations: Use of a standardised and validated generic instrument	Yes	Discussed in section 4.2.5
Method of preference elicitation for health state values: Choice based method (e.g. TTO, SG, not rating scale)	Yes	Discussed in section 4.2.5
Source of preference data: Representative sample of the public	Yes	Discussed in section 4.2.2
Discount rate: 3.5% pa for costs and health effects	Yes	
PSS = personal social services; TTO = time trade off; SG = standard gamble		

Overall, the methods applied in the economic analyses were appropriate and reported transparently. The company's economic evaluation conformed to NICE methodological guidance and met the NICE scope.

4.2.1 Modelling approach / model structure

The company constructed a lifetime state transition Markov cohort model in Microsoft Excel using three month cycles. The three month cycle length was in line with observation periods in the INPULSIS trials and seemed of adequate length to capture relevant clinical events.[2] The model was conducted from the NHS and PSS perspective, with discounting for both costs and benefits at 3.5% annually. Half-cycle correction was employed to account for variable timing of events. The company submission did not explicitly state what 'lifetime' meant within the model, but an inspection of the model reveals that lifetime was assumed to be 50 years from the start of

the model. Given that the age of the patient population is generally 60 and above, with a median survival of approximately 3.5 years, a shorter time horizon of 30 years may have been sufficient. The loglogistic overall survival model predicts that ~0.6% of patients would be alive at 30 years.

Figure 1 shows the CS model schematic (Figure 37, p. 160).[35] The model structure is clearly represented, appears appropriate, and has sufficient justification for the choice of structure.

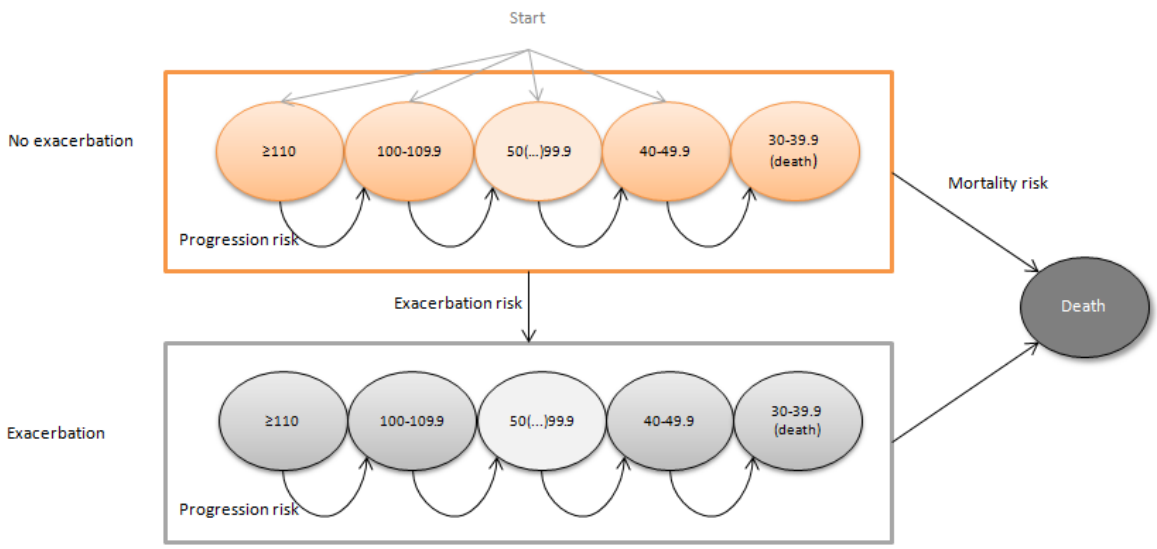

Figure 1 Model Structure (Figure 37, p. 160, CS)[35]

The model represents IPF lung function deterioration using an established clinical measure, FVC% predicted, to define health states. Health states were defined by roughly 10 point percentage intervals in FVC% predicted from ≥110 to 30-39.9 (representing death due to insufficient lung function), a state for death from any cause, and a set of parallel health states for patients who experienced an exacerbation, thus representing a total of 20 distinct health states in the model. Patients could start the model in any live non-exacerbation state, with the distribution of patients defined by the distribution of patients in the INPULSIS 1 and 2 studies.[2] When exacerbations occur, patients move from the no exacerbation health states to the exacerbation health states and cannot return to no exacerbation health states, as shown by Figure 1. Exacerbation health states have different health outcomes and costs than no exacerbation states. FVC% predicted was chosen to represent health states due to consistency with clinical trials in IPF and after consultation with clinicians, the ERG found this to be a reasonable choice for defining health states.

The model allows for clinical events outside of loss of lung function, including: exacerbations (which also affect loss of lung function), cardiac events, and bleeding events (including gastrointestinal perforation). In order to determine events in the model, including progression, exacerbations and serious adverse events, odds ratios were derived from the NMA (CS Sections 4.10 and CS Appendix B) and applied to baseline event rate data from the INPULSIS 1 and 2 trials, assuming a constant risk over time.[2;35] The biological and clinical processes of IPF appear to be sufficiently represented by the model structure.

The model structure was informed by a literature review and checked for face validity through consultation with clinicians. These clinicians are directly identified in the submission. Additionally, the company attended NICE meetings for the technology appraisal for pirfenidone (TA282) to gain modelling insights.[7]

The formulation of the model structure is discussed in detail with thorough referencing to the literature. The explanatory text contains generally good justifications for model structure choices.

The main structural assumptions of the model are as follows:

1. Loss of lung function can be represented as incremental 10-point decreases in FVC% predicted, hence the health states described by these value changes.
2. Lung function can decrease, but not increase.
3. Exacerbation changes the risk of progression.
4. Death occurs if a patient's lung function falls to between 30-39.9% FVC% predicted.
5. Risk of death is independent of exacerbation status.
6. IPF is a progressive disease with no potential for improvement in FVC% predicted. Patient condition deteriorates until death.

There are numerous structural assumptions based on survival curve choices, adverse events included, and the choice of baseline data for placebo (BSC). Justifications were provided for survival curve choices, and evaluated using sensitivity analyses.

The model extrapolates one-year time-to-event data over the lifetime of the model using regression analysis. Multiple potential survival curves were examined, but all are limited by the length of observation in the trial data. Justification for model choices was provided.

Generally, the structural assumptions appear to be justified, and in line with clinical judgment. The company checked validity by comparing CS modelling choices to those made in the NICE pirfenidone model and through consulting specified experts. The documentation of choices and justifications provided are generally of good quality and sound logic.

Compared to models produced by Loveman and colleagues (2015)[27] and the model for the pirfenidone STA,[36] the nintedanib model has more stages of patient progression, assumes independent health states for patients who have had an exacerbation, and has more levels of quality of life due to the increase in health states. The Loveman and colleagues[27] and the nintedanib models are cohort models, whilst the pirfenidone model is a micro-simulation model (individual sampling model);[37] all models have Markov structures with discrete time. The Loveman and colleagues[27] model contains four health states: unprogressed IPF, progressed IPF, lung transplant, and dead. However there was only a 0.6% probability of a lung transplant, so this is unlikely to have a material effect on the model results. The Loveman and colleagues model does not model exacerbations as separate states but acute exacerbations result in transition to the progressed IPF state and a utility decrement and cost. All models assume that patients start with non-progressed disease or disease without exacerbations. All models use some measure of FVC to predict progression using survival analysis. In Loveman and colleagues and the nintedanib submission, this is based on FVC% predicted. In the pirfenidone submission, progression is based on an individual patient regression analysis incorporating FVC and 6MWD as covariates. The structure of the pirfenidone submission model is redacted in both the CS and the pirfenidone ERG report, so full analysis of the structure is not possible.[13;36]

In general, the model approach appears appropriate, comprehensive and well justified. The model has significant sensitivity analysis capabilities and numerous and varied sensitivity analyses were conducted.

4.2.2 Patient group

For the economic model the patient population is based upon phase III trials for nintedanib in IPF. The baseline characteristics are shown in CS Table 159. The patients are described in

terms of the proportion in each FVC% predicted group and have a starting age of 66.75 years. The base case for the economic evaluation comprises the total pooled population recruited into the INPULSIS I and II trials.[2] The patients are described in more detail in section 3.1.3 of this report.

The patient population in the model may not be fully reflective of the target population in current clinical practice or the scope of the appraisal, as these patients may have milder IPF than those typically seen. The analysis includes patients with FVC% predicted higher than 80% (this accounts for about 45% of patients). In the pirfenidone NICE single technology appraisal, these patients were considered to be rarely seen in clinical practice.[7] The ERG conducted a scenario analysis without these milder patients in section 4.3.

The CS also presents a scenario analysis for an 'ASCEND-like' population for nintedanib compared to pirfenidone. The CS states that this was a restricted population representative of the ASCEND trial[3] selection criteria. The ASCEND trial was an RCT for pirfenidone versus placebo and was included in the company's NMA. The restricted criteria for this ASCEND-like subgroup were: IPF diagnosed at least 0.5 years before visit 2, and FVC 50-90% predicted, $FEV_1/FVC \geq 0.8$.

4.2.3 Interventions and comparators

The comparator used for the economic analysis was pirfenidone or best supportive care in agreement with the scope developed by NICE and current clinical practice. The CS included NAC in the NMA, but did not include it within the economic model because it was not within the NICE scope and the CS states that results from the PANTHER-IPF trial[38] demonstrate NAC's lack of effectiveness.

4.2.4 Clinical effectiveness

The clinical effectiveness parameters were used in the model for transition probabilities, serious adverse events and discontinuation.

Transition probabilities

The CS describes the transition probabilities in the model for mortality (overall survival), acute IPF exacerbation and loss of lung function (progression based on FVC% predicted). The base-

case transition probabilities were obtained by fitting parametric models on the patient level data in the placebo arms of the two nintedanib clinical trials.[1;2] These placebo arms of the clinical trials were used to represent best supportive care (BSC) in the company's economic evaluation. Details of the methodologies adopted are discussed below.

i. Mortality (overall survival)

Overall survival was implemented in the model by deriving fitted distributions for the placebo arm and using ORs for the nintedanib and pirfenidone treatment arms. Standard parametric distributions were fitted for the placebo arm of the phase II and phase III clinical trials[1;2] using the exponential, Gompertz, loglogistic, lognormal and Weibull distributions. Based on Akaike Information Criterion (AIC) values, the company stated that Gompertz distribution provided the best fit of these distributions, although the Weibull and log logistic distributions also presented a close fit. The 10 year extrapolated overall survival is presented in CS-Figure 30 (p. 166).

The company validated the fitted models by comparing them against the clinical trial data for 12 months and overall survival from the study by Kondoh and colleagues.[39] This is an observational study that evaluated the frequency, risk factors and impact on survival of acute exacerbation in patients with IPF. The study cohort consisted of 74 patients who were retrospectively followed for more than 3 years. Of these 74 patients, 23 had acute exacerbations and the remaining 51 were without exacerbations. The company fitted survival curves to the data in both the patient groups (i.e. with and without exacerbations) which were then compared against the survival curves fitted to the placebo arm of the INPULSIS trial.

It was stated that the use of Gompertz model underestimated survival in the nintedanib trial compared to the Kondoh study.[39] The company therefore justified the use of the log-logistic curve in the base case by stating that it provided the closest fit to the Kondoh study. However these data were unclear because the data from Kondoh and colleagues[39] were presented for patients with and without acute exacerbations. In response to clarification questions, the company presented a comparison of the pooled data from Kondoh and colleagues with fitted parametric models: exponential, Weibull, Gompertz, lognormal and log logistic. Of these, the log-logistic and log-normal curves provided the closest fit based on AIC (as shown in Figure 2).

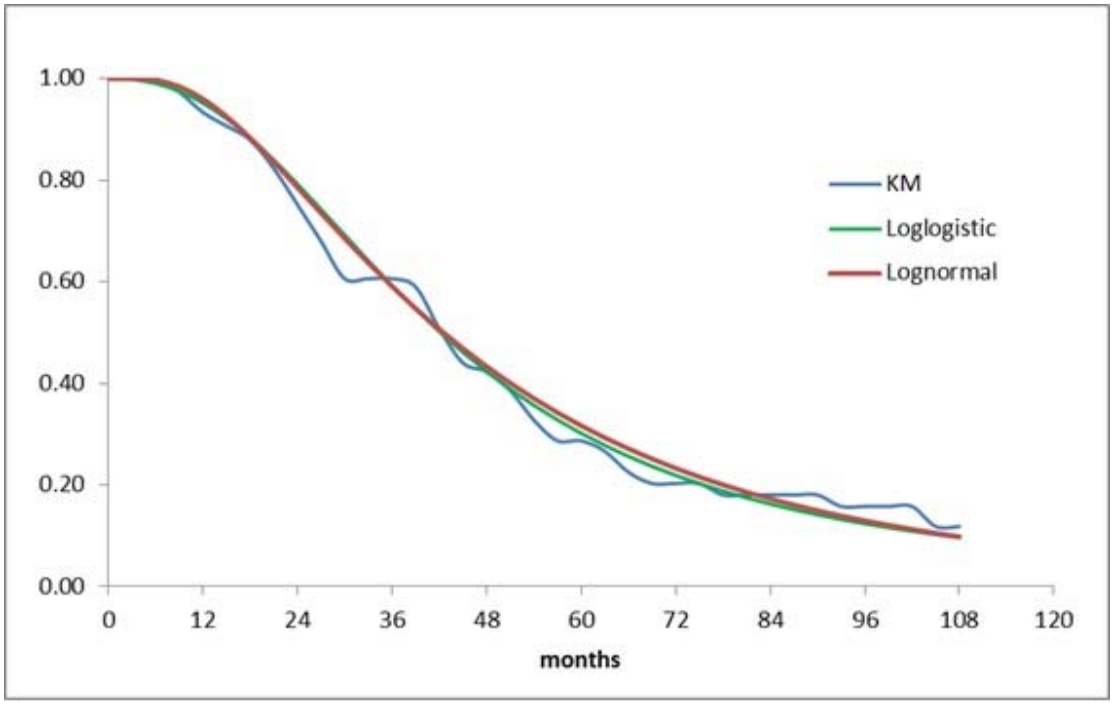

Figure 2 Fit of parametric models of the pooled overall survival data from the Kondoh study (Company's clarification response, Fig 23)

The baseline mortality risk was multiplied by the corresponding OR values for nintedanib and pirfenidone which were obtained from the "all evidence scenario" of the fixed effect NMA. The OR values were 0.70 (95% CrI: 0.45 to 1.10) for nintedanib and 0.70 (95% CrI: 0.46 to 1.05) for pirfenidone respectively.(CS Table 102, p.170) For further details on the ERG critique of the NMA techniques for overall survival see section 3.1.7.

The risk of death was modelled independent of any other outcomes such as exacerbation or progression. In addition, death occurred in patients who reached FVC% predicted level of 30-39.9. The company provided justification for these assumptions, which appeared to be reasonable and consistent. Both one-way sensitivity analyses and PSA were conducted surrounding the estimates of overall survival (CS Tables 175 and 180, p.279, 286).

The ERG considered the approach of applying ORs from the NMA to the base case placebo mortality risks to be consistent with standard modelling methodology. It is to be noted that the OR value obtained by the company for nintedanib vs placebo (0.70, 95% CrI: 0.45 to 1.10) was similar to the value obtained in the NMA conducted by Loveman and colleagues[25] (0.70, 95% CrI: 0.45 to 1.09). However, the OR value obtained by the company for pirfenidone vs placebo

(0.70, 95% CrI: 0.46 to 1.05) differed from that reported by Loveman and colleagues[25] (0.50, 95% CrI: 0.29 to 0.84).This is likely to be because Loveman and colleagues did not include the study by King and colleagues.[3]

The ERG had a few concerns on the base case extrapolation techniques used by the company. First, the ERG observed that beyond the first 12 months, the extrapolated survival models diverged significantly in their predictions for the remaining time-periods where the loglogistic distribution estimated overall survival much greater than the other models and greater than expected survival in the patient population. Changing the parametric model from loglogistic to Weibull had a significant impact on the overall results where the ICER increased by approximately £91,000 as shown in the one-way sensitivity analyses (CS p. 296).

Secondly, the company identified six other studies to validate the extrapolation of base case overall survival but did not provide any detailed information on study characteristics, such as patient characteristics or length of follow up (CS p. 168). Particularly, the ERG observed that a study by Nathan and colleagues[40] had a larger sample size of 357 IPF patients with similar length of follow-up of 10 years compared against Kondoh and colleagues.[39] The company did not provide any justification for choosing the Kondoh study over the Nathan study for validating the extrapolated survival curves. The ERG therefore, conducted a comparative analysis of the INPULSIS trial survival against the survival of patients in Kondoh and colleagues[39] as well as Nathan and colleagues,[40] shown in Figure 3 below.

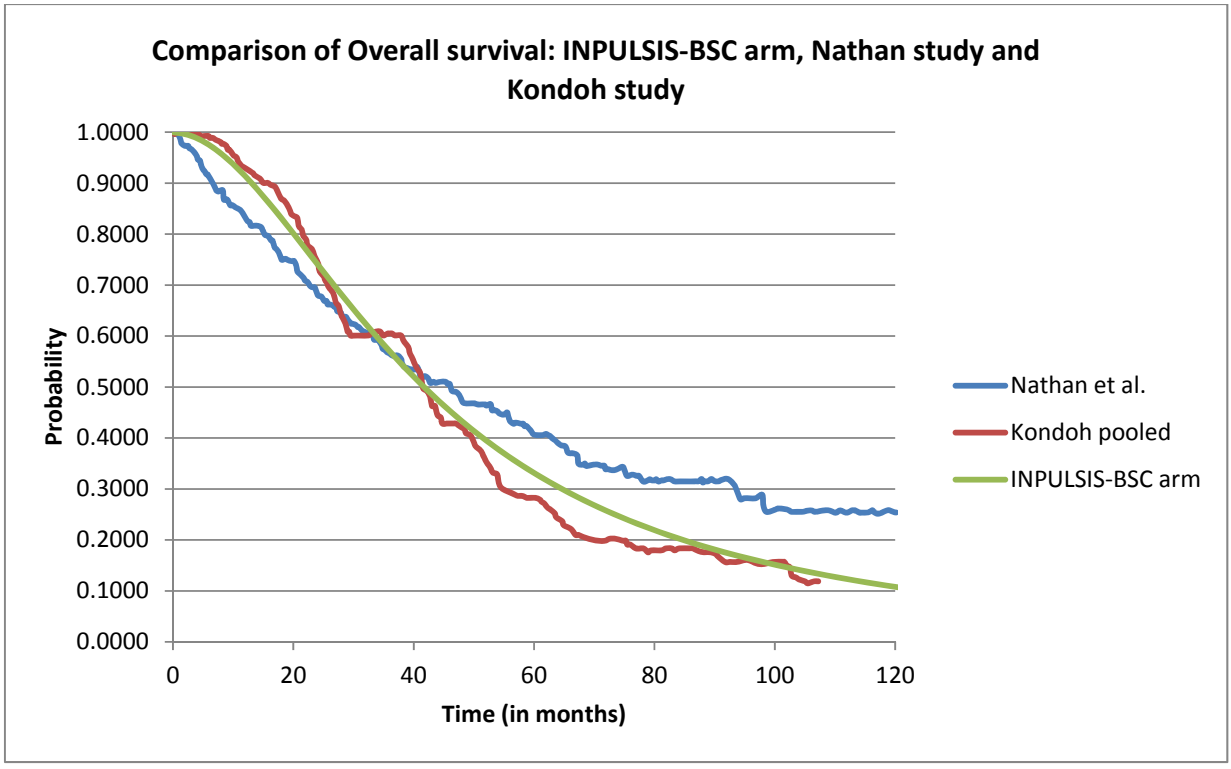

Figure 3 Comparison of overall survival of the INPULSIS-BSC arm against Nathan and colleagues and Kondoh and colleagues

As shown in the figure, patients' survival trajectory in the INPULSIS trial followed a relatively similar pattern to the pooled survival data obtained from the Kondoh study.[39] Patients in the Nathan study[40] had better survival compared to those in the INPULSIS trials in the long run.

Based on the above observations, the ERG felt that the selections of Kondoh study for validation and that of the log-logistic curve was appropriate to use in the company's analysis.

ii. Acute IPF exacerbation

The risk of acute exacerbation was incorporated in the model as time to first acute exacerbation and recurrent exacerbation. Time to first acute exacerbation was recorded in two ways based on: i) investigator-reported adverse events which was in line with the selection criteria as described in trial protocol and ii) adjudication committee classification of acute IPF exacerbation as "confirmed", "suspected", or "not" based on the cases that met all the criteria for the definition of acute IPF exacerbation. The company used the investigator reported approach for their base

case analysis which concurred with the suggestion of the ERG's clinical advisor that this approach was likely to represent current clinical practice.

The company fitted parametric models to extrapolate data for time to first acute exacerbation from the post hoc analysis of the INPULSIS I and II data, based on both the investigator reported and adjudication committee definitions. In both cases, the company assumed a constant hazard and fitted exponential models. The estimated risk of exacerbation per cycle applied for the placebo arm varied with a risk of 1.95% for the investigator-reported and 1.47% for adjudication committee definition respectively. The risks of exacerbation for nintedanib (0.56; 95% CrI: 0.35 to 0.89) and pirfenidone (1.01; 95% CrI: 0.22 to 4.50) were estimated by applying OR values obtained from the NMA scenario (Scenario 3 in CS Appendix B) that excluded Azuma and colleagues,[22] Taniguchi and colleagues,[23] and one of the NAC studies (Homma and colleagues[24]) to the baseline placebo risk.

For recurrent exacerbation, the model assumed that patients who experienced at least one exacerbation were at risk of recurrent exacerbation. This risk was assumed to be the same as for those patients who had not had any exacerbation. A range of one-way sensitivity analyses and PSA were conducted around these estimates as outlined in CS Tables 175 and 181 (p.279, 287).

To check for consistency and validity, the ERG compared the OR values applied in the model for nintedanib and pirfenidone with those obtained by Loveman and colleagues.[25] The ERG observed that ORs for nintedanib vs placebo estimated by the company (0.56, 95%CrI: 0.35 to 0.89) were close to the estimates obtained by Loveman and colleagues[25] (0.50, 95%CrI: 0.31 to 0.79) but the values differed significantly for pirfenidone vs placebo (CS: 1.01, 95%CrI: 0.22 to 4.50; Loveman and colleagues: 0.43, 95%CrI: 0.14 to 1.26). The ERG considers that these differences could be explained by the difference in studies included in the two analyses. For instance, whilst Loveman and colleagues[25] included studies with Japanese patients, the company excluded Japanese trials. The ERG also observes that there are differences in the definition of acute exacerbation in the studies included in the NMA. Further ERG critique of the NMA for acute exacerbation can be found in section 3.1.7. Secondly, the assumptions adopted to estimate the risk of acute exacerbation may be inappropriate as the ERG's clinical advisor suggested that the risk of exacerbation increases with IPF severity. Whilst the ERG's clinical advisor acknowledged the lack of evidence, the advisor was also of the opinion that patients

who have had one exacerbation were more likely to be at higher risk of recurrent exacerbation compared to those who have not had any.

iii. Loss of lung function

The company defined loss of lung function as a 10-point drop in FVC% predicted. Patients entered the model at different FVC% predicted health states to reflect the INPULSIS clinical trial as shown in CS Table 108 (p.174). Lung function declines with and without exacerbation were incorporated using a logistic model derived from a logistic regression of the phase III clinical trial data.[2] In both the scenarios (i.e., with and without exacerbation), there was a diminishing effect in progression with loss of lung function. However, the absolute risk of progression was significantly higher when there was an exacerbation. This is graphically presented in CS Figure 47 (p.178).

As in the cases of overall survival and acute exacerbation, the risks associated with loss of lung function for nintedanib and pirfenidone were obtained by applying ORs from a NMA scenario (Scenario 2 CS Appendix B) to baseline risk from the INPULSIS trials[2] assuming a constant hazard over time. This scenario excluded King and colleagues.[3] The ERG critiques the loss of lung function NMA in section 3.1.7. The OR estimates for nintedanib vs placebo were 0.54 (95%CrI: 0.42 to 0.69) and 0.69 (95%CrI: 0.47 to 1.00) for pirfenidone vs placebo respectively. For validation, the company compared the model projections for the distributions of patients in FVC% predicted health states after 1 year against the clinical trial results for the placebo and nintedanib arms, presented in CS Figures 48 and 49 (p.179, 180). The results of the predicted model agree with the clinical trial results.

The ERG considers that the methodologies adopted by the company to predict loss of lung function were reasonable. The OR values obtained by Loveman and colleagues[25] were 0.41 (95%CrI: 0.34 to 0.51) for nintedanib vs placebo and 0.62 (95%CrI: 0.52 to 0.74) for pirfenidone vs placebo. These values differed from those obtained by the company which could be due to the inclusion of different studies in the two analyses. Loveman and colleagues[25] included the ASCEND trial by King and colleagues[3] whereas the company excluded this study due to difference in patient characteristics. On closer inspection, the ERG found that whilst similar studies were included for nintedanib vs placebo in both the analyses, it was unclear as to why the results obtained were different.

Adverse events

The CS model only included AE which had a substantial impact on costs and QALYs, had an incidence of more than 5% or an incidence 1.5 times greater than the comparator arm. Serious cardiac events and serious GI events were included in the analysis. Gastrointestinal perforation (for nintedanib) and photosensitivity and rash (for pirfenidone) were also included based on their clinical importance. Liver enzyme elevations were excluded.

The incidences of each of the serious AEs were estimated from the placebo arm and their associated risks for nintedanib and pirfenidone were measured using OR values from the NMA presented in CS Table 117 (p. 181) and shown below in Table 36. Whilst for serious cardiac events the company used the NMA scenario that excluded the study by Richeldi and colleagues (Scenario 2 CS Appendix B),[1] for serious GI events the OR values obtained from the all evidence scenario of the NMA were used (Scenario 1). The incidences of other clinically important AEs were presented in CS Table 118 and Table 119 (p.182). A range of sensitivity analyses were conducted around these parameters (CS Table 175 and 183, p.279 and 287).

It was observed that although diarrhoea was a common adverse outcome in IPF patients occurring in 60% of the patients in the INPULSIS trials, the condition predominantly ranged from mild to moderate severity. In the trials, less than 5% of patients discontinued because of the condition (CS section 4.12, p.142). As a result, the ERG considered it appropriate to exclude diarrhoea from the economic analyses.

Overall, the ERG considers the company's approach to the inclusion of AEs in the economic model as reasonable and justified.

Discontinuation

The company estimated overall discontinuation risk for the baseline placebo arm by fitting parametric models to extrapolate the phase II and phase III clinical trial data.[2] Among five different types of distributions fitted, the exponential model was chosen based on smaller AIC values. The overall risk of discontinuation for the placebo arm was estimated to be 5.4% per month and the associated risk for nintedanib (OR 1.42; 95% CrI: 1.08 to 1.87) and pirfenidone (OR 1.34; 95% CrI: 1.04 to 1.73) were calculated by applying ORs obtained from the all evidence scenario of the NMA to the baseline risk (CS Table 122, p.183). The company assumed that patients would not discontinue from the placebo (BSC) arm, but they used this

discontinuation risk to estimate the relative discontinuation risks in patients receiving nintedanib and pirfenidone. The model also incorporated stopping rules for both nintedanib and pirfenidone for the proportion of the cohort that dropped below a certain FVC% predicted level (below FVC50% predicted; FVC60% predicted and FVC70% predicted) and discontinued treatment when patients experienced a fall of FVC10% predicted or more.

The OR values obtained from the NMA used in the economic model are summarised below in Table 36.

Table 36 OR values obtained from the NMA as used in the company's economic model

Comparison	OR median value (95% CrI[1])	Evidence source for the NMA
Overall Survival		
Nintedanib vs Placebo	0.70 (0.45 to 1.10)	Scenario 1, all evidence scenario
Pirfenidone vs Placebo	0.70 (0.46 to 1.05)	
Acute exacerbations		
Nintedanib vs Placebo	0.56 (0.35 to 0.89)	Scenario 3, excluding Azuma et al.,[22] Taniguchi et al.[23] and an NAC study (Homma et al.[24])
Pirfenidone vs Placebo	1.01 (0.22 to 4.50)	
Loss of lung function		
Nintedanib vs Placebo	0.54 (0.42 to 0.69)	Scenario 2, excluding King et al.[3]
Pirfenidone vs Placebo	0.69 (0.47 to 1.00)	
Serious adverse events		
Serious cardiac events		
Nintedanib vs Placebo	0.92 (0.53 to 1.63)	Scenario 2, excluding Richeldi et al.[1]
Pirfenidone vs Placebo	1.27 (0.66 to 2.49)	
Serious GI events		
Nintedanib vs Placebo	2.35 (1.05 to 5.88)	Scenario 1, all evidence scenario
Pirfenidone vs Placebo	0.60 (0.23 to 1.45)	
Discontinuation		
Nintedanib vs Placebo	1.42 (1.08 to 1.87)	Scenario 1, all evidence scenario
Pirfenidone vs Placebo	1.34 (1.04 to 1.73)	

NMA: Network Meta-Analysis; GI: Gastrointestinal;

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 Table 36. 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 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 4.3.

4.2.5 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.

SUPERSEDED - SEE ERRATUM

HRQoL is incorporated in the model using utility estimates applied to the model health states, in terms of FVC% predicted. A disutility is applied for acute exacerbation and serious adverse events.

The utility values used in the model are shown in Table 37 (CS Table 129, p. 190). These EQ-5D values are taken from the company's own trial data for the INPULSIS I and II[2] trials (unpublished data). The company supplied additional information on these data upon request from the ERG. The company reported that [redacted]

The CS states that the two HRQoL studies identified in their review (King and colleagues[41] and Zisman and colleagues[42]) reported EQ-5D scores broadly consistent with the values in the nintedanib clinical trials. The ERG concurs with this view and also notes that the utility values are also consistent with those used in a previous analysis by Loveman and colleagues.[27] The ERG also notes the scarcity of good quality HRQoL data in this population and have not identified any alternative relevant sources of HRQoL utility values.

Table 37 Summary of quality of life values used in the company's cost effectiveness analysis

FVC%pred	Mean EQ-5D utility	SD	Number of observations
90 and above	0.8380	0.1782	458
80-89.9	0.8105	0.2051	684
70-79.9	0.7800	0.2244	788
60-69.9	0.7657	0.2380	809
50-59.9	0.7387	0.2317	490
40-49.9	0.6634	0.2552	98

The utility decrements for acute exacerbation were also taken from the INPULSIS I and II trial data. The company considered two acute exacerbation definitions: investigator reported and adjudication committee exacerbations (CS Table 130). The model used the investigator reported exacerbation as base case and explored the effect of the adjudicated committee

exacerbation in a sensitivity analysis. The decrement was assumed to apply across all health states and to be more severe in the first month (disutility of -0.14), followed by a smaller decrement in the subsequent months (disutility of -0.078). The ERG was unable to find any alternative sources of disutility for acute exacerbations and note that the values used by Loveman and colleagues[27] were from patients with a different condition.

The model includes utility decrements for serious cardiac events (-0.198), serious GI events (-0.068), skin disorders (-0.082) and GI perforation (-0.118), CS Table 133 page 224). These values are based on a study by Ara and Brazier[28] for serious cardiac events, skin disorders and GI perforation. Ara and Brazier[28] analysed data from four consecutive Health Surveys for England which included self-reported health status and EQ-5D values. They reported values for groups of patients with and without specific health status. The disutility values for serious GI events are taken from the INPULSIS trial data. It is assumed that the proportion of patients with adverse events remains constant over time and the disutility is applied for one cycle.

The ERG notes that the duration used for adverse event disutility is for one year and considers that the duration of the adverse event would be significantly less than this for GI events and skin disorders. For example, Costabel and colleagues[43] state that GI AEs for pirfenidone were mostly transient in nature, with the exception of dyspepsia which was present for a median duration of 168 days. Likewise, rash and photosensitivity reaction in most cases were resolved within 15 days through pirfenidone dose reduction.

The CS reports that the patients who had a serious GI event had a drop in HRQoL of -0.068 points and then recovered but it does not report the duration of the serious GI event. The CS reports an annual GI disutility for a study by Leontiadis and colleagues[44] as -0.025, and yet uses a disutility of -0.118. It is unclear whether the categories used for skin complaints and GI perforation are of the same definition and severity in the Health Survey for England as seen in the INPULSIS trials.

Many cases of photosensitivity and rash may now be avoided through patient advice to avoid sun exposure.[43] The RECAP study[45] was a long-term open label extension study of the CAPACITY trials. Rash was less prominent in RECAP (18%) than in CAPACITY (31%) ;[4] rates of photosensitivity were similar between RECAP and CAPACITY (11.9% vs. 11.8%).[4]

As stated above, the ERG considers that the disutilities have been overestimated in the company model. The ERG conducts a scenario analysis with changes to the disutility of adverse events for rash and photosensitivity in section 4.3. Changes to the disutilities for serious GI events and GI perforation should have minimal impact on the model results as both nintedanib and pirfenidone have these events, while rash and photosensitivity only occur in the pirfenidone arm of the model.

Overall, the health benefits have been measured and valued as per the NICE reference case. The utility estimates appear to be based upon a large sample with a directly relevant population group, however the ERG is not able to check or verify the estimates and they have not been published in full. The ERG considered the disutility for adverse events to be overestimated.

4.2.6 Resource use

The categories of resource use included by the company were treatment (including drug acquisition, and patient monitoring), health state resources and resources for treating acute exacerbations and adverse events.

The nintedanib dosing schedule is stated in CS Table 5 page 25. The recommended daily dose of nintedanib for patients with IPF is two doses of 150 mg oral capsules. This dosage is consistent with that used in the INPULSIS I and II trials.[2] The pirfenidone dosage was assumed to be 2403 mg/day from the Electronic Medicines Compendium (eMC 2014).[46] The ERG notes that the listed dosage for pirfenidone is increased over the first two weeks of treatment until the target dose is reached and that patients are recommended to receive pirfenidone 801 mg/day for the first week and 1602 mg/day for the second week.

The company conducted a search to identify existing studies reporting resource use and/or unit costs for nintedanib or its comparators in adults with IPF (CS section 5.5.2 page 225). One abstract was identified (Parfrey and colleagues 2013).[47] This study reported hospital resource unit data collected over a nine-month observation period of a multi-centre, retrospective, cohort review undertaken across four NHS trusts. The study reported on 100 patients treated with pirfenidone for six months.

Resource use data in the economic model was based on the resources in the INPULSIS trial. The CS refers to these resource data as the Health Care Research Unit (HCRU) data. These data were analysed and adjusted to match the model states, i.e. by FVC% predicted category. A per cycle probability (3 months) of incurring the resource use was calculated. The number of observations for each FVC% predicted group is presented in CS Table 140, page 232. The company provided more information on the collection of the resource data in response to NICE and the ERG's clarification letter (B1). The company reported that [redacted] The CS reports that there was discussion with two clinicians about the resource use.

The resources for patient monitoring consist of hospitalisation, A&E, GP visits, specialist visits, physiotherapist visits, chest HRCT, chest X-ray and oxygen requirement, bronchoalveolar, CT pulmonary angiogram, right heart catheterization procedure, and a general diagnostic procedure (for example bronchoscopy). These were derived from the HCRU data as a 3-month probability as reported in CS Tables 142 to 158. The description of the components of hospitalisation is reported in CS page 234. The CS reported that average number of hospitalisations per patient were 1.124 with an average duration of 8.72 days. A small proportion of hospitalisations included an Intensive Care Unit (ICU) stay, mechanical ventilation use, an overnight Emergency Room (ER) stay or use of an ambulance. The ERG's clinical expert considered that the frequency of hospitalisation and duration of stay appeared reasonable for IPF patients.

In addition to the health state resources described above, patients received oxygen supplementation if their FVC% predicted was lower than 80%. The CS stated that patients with FVC% predicted above 80% would be in relative good health and would not require oxygen supplementation.

The resources associated with an acute exacerbation were hospitalisations, ER visits, GP visits and specialist visits. The 3-month probability of patients with an acute exacerbation visiting a hospital and the other health professionals is shown in CS Table 157 page 244. The average number of hospitalisations was 1.3 and the average duration of each hospitalisation was 16.3 days.

Overall, the estimates used for the choice of IPF resources have been based upon a large sample collected from the clinical trials for IPF treatment. These data used in the modelling appear appropriate and relevant to the clinical pathway of IPF patients, however the ERG is not able to check or verify the estimates and they have not been published in full.

4.2.7 Costs

The main costs in the model are drug treatment costs, oxygen, liver function test costs, monitoring costs, hospitalisation costs and end of life costs. The CS states that NHS reference costs have been used for the cost of hospital procedures and interventions. The ERG confirms that this approach is appropriate and consistent with NICE guide to the methods of technology appraisal.[20] The costs used in the model are shown in the CS Table 159 (page 246-253).

Drug acquisition costs for nintedanib are £71.70 per day (CS p. 25) or £2151.10 per 30 days based on 150 mg capsules twice a day. The drug costs have not been published in the British National Formulary or MIMS at the time of writing. The company has provided a confidential PAS discount. Drug acquisition costs for pirfenidone are £71.70 per day (CS p. 229), based on three 267 mg capsules three times a day for a total of 2,403 mg/day (eMC 2014).[46] The manufacturer of pirfenidone provided a confidential PAS discount of as part of NICE Technology Appraisal 282.[7]

The costs for patient monitoring consist of a weighted average of the unit costs of the resources used and their 3-month probability for each FVC% predicted health state. The patient monitoring cost per 3-month cycle varies between £219.19 (FVC% predicted ≥ 110) and £649.17 (FVC% predicted 40-49.9). The largest component of the patient monitoring cost is hospitalisation costs. The total cost of hospitalisation consists of the hospitalisation stay cost, ICU cost, mechanical ventilation cost, ER cost and ambulance cost. The total hospitalisation cost is £3,044, as shown in CS Figure 54 page 234. The unit cost of hospitalisation per bed day is £303.73 and is taken from respiratory failure costs from NHS reference cost 2009/10 and inflated to 2012/13. The ERG is unclear why the company has not used a unit cost from NHS reference cost 2012/13. In response to NICE and the ERG's clarification letter (B14), the company supplied a unit cost of hospitalisation of £359.17 per day. The mechanical ventilation cost was taken from the unit cost of an outpatient procedure (£148). The ERG requested clarification on the use of this approach. In their letter of clarification, the company supplied an

alternative cost of £2830 per hospital stay. The ERG considered this a more appropriate cost estimate. However, the changes to these costs listed above, when tested by the ERG, had no significant impact on the model results.

The cost of oxygen supplementation was estimated at £418 per 3-month cycle (NHS reference cost 2009/10, inflated to 2012/13 costs). The values used in the cost are for an elective inpatient day in hospital receiving oxygen and the ERG considers this would be different to the 3-monthly cost of oxygen use. The ERG considers a more appropriate approach is that used by Loveman and colleagues They used a home oxygen costing tool from the Department of Health[48] and obtained a cost per year of £824.30 per patient. However, this alternative oxygen cost had no significant impact on the model results when tested by the ERG.

The model also includes end of life costs. The company justifies the inclusion of these costs on the basis that their clinical experts advised that palliative care is an important aspect of people's end of life care and that its inclusion affects the incremental cost effectiveness results. The end of life care costs were derived from a National Audit Office report by Hatziandreu and colleagues[49] which analysed end of life care costs for patients who suffered from cancer or organ failure (pulmonary and heart failure). The annual end of life costs consisted of £9,098 for home care and £8 for hospice care. These values were converted to a 3-month cycle and inflated to 2012/13 costs to give a cost of £3920.64 per cycle. The ERG notes that these costs have been incorrectly inflated and the correct inflated cost should be £2560.84, however this corrected cost had no significant impact on the model results when tested by the ERG.

The acute exacerbation cost consists of a synthesis of hospitalisation cost, ER visit, GP visit and specialist visit (CS Table 158 page 244). The acute exacerbation cost was £4133.59.

The costs of treating treatment-related adverse events are shown in CS Table 135. These are taken from NHS reference costs 2012/13.[30] The adverse event costs are for serious cardiac events (£2,054), serious GI events (£1749), skin disorders (£373) and GI perforation (£2353).

Overall the ERG considers that the approach for costing is appropriate. In general, the values used have been taken from standard sources, are indexed to the current price year and the estimates have been appropriately reported. The ERG identified a few cost values, which it

considered were not derived appropriately but changes to these costs had no significant impact on the model results.

4.2.8 Consistency/ model validation

There were no checklists explicitly listed in the CS for model validation. No evidence of model validation was provided, so whether coding and other mechanical checks were performed is unknown.

Internal consistency

The model developers had a senior modeller that was not involved with the nintedanib model development perform quality assurance checks on the model. The nature of these checks was not described, so it is unclear whether coding or other mechanical checks were performed. Additionally, the company performed basic input and output checks similar to those conducted by the ERG.

The ERG conducted a check of the model inputs and expected outputs by testing extreme input values for logical results and examining model code for appropriate mathematical and logical expressions. Setting quality of life to zero for the upper half of health states had a predictable reduction in quality of life, and did not change life years. This result is expected as patients do not die when their utility score is zero. Adjusting costs and treatment effectiveness parameters (ORs) produced consistently logical results. No input errors were detected and calculations appeared to function correctly. All Visual Basic (VBA) code was checked for errors and rerun, with expected outputs produced and no errors found. The model's logical components and cell reference structures worked as intended. The PSA was re-run using VBA, and a selection of DSA were rerun using built-in user defined cells. The results produced by the re-run analyses were consistent with those reported in the CS. Overall, the model was clear, relatively easy to work within and thorough.

External consistency

The model structure was checked for face validity by clinicians. The model was externally validated by the company as the model was developed by a consultancy. The methods are described in the section above. The conclusions followed logically from the inputs and made intuitive sense.

The company did not compare the model results to the results of the pirfenidone STA model,[36] one of two economic models in the disease area. The ERG notes the difficulty of assessing external validity when the inputs and results of the pirfenidone model are commercial in confidence, and considers the lack of comparison between the models reasonable and expected. The company did provide ae comparison of the CS model to the other economic model in the disease area, Loveman and Colleagues (2015).[27]

In the CS (Section 3.31, pp. 88-9), the NMA and the 2015 HTA model by Loveman and colleagues[27] were compared to the CS NMA and model.[35] The company found that a number of differences between the data inputs used in the model logically explained the differences between the company submission and the Loveman and colleagues (2015) NMA and model:[27]

- Due to the earlier date of the systematic searches in Loveman and colleagues,[27] INPULSIS 1 and 2, ASCEND, and PANTHER-IPF (which demonstrated the ineffectiveness of NAC) were not included.[2;3;38] Taniguchi and colleagues (2010) was also not included.[23] This resulted in all effectiveness data for nintedanib being derived from phase II evidence (TOMORROW).[1]
- The company had more current data (72 week follow-up for the CAPACITY trial instead of 52 week follow-up)[21]
- Loveman and colleagues excluded Azuma and colleagues (2005) and Taniguchi and colleagues (2010) from the NMA of mortality data[22;23]
- Exacerbation data was measured using different data and assumptions from the TOMORROW trial,[1] and no exacerbations were included from CAPACITY.[21]
- Azuma and colleagues (2005)[23] was included in the Loveman and colleagues (2015)[25] NMA based on an assumption of equivalence of vital capacity and FVC but this assumption was inconsistently applied, as Taniguchi and colleagues (2010)[22] was excluded from the NMA
- The price of nintedanib was incorrect in the Loveman and colleagues model

In general, OR between nintedanib and pirfenidone were more favourable to nintedanib in the Loveman and colleagues analysis due to the differences above.

The ERG was unable to compare the company model results with the pirfenidone STA model due to the almost total redaction of model results, inputs and even model structure. Only one

publically available source was available for the ICER of the pirfenidone model, the guidance issued by NICE, which indicated an ICER of £24,000/QALY for pirfenidone compared to BSC.[7] No data on total QALYs or total costs were available from the pirfenidone STA, making realistic, informative comparisons of the two models impossible. It should also be noted that the pirfenidone STA analysis included data from long-term pirfenidone follow-up in the open label extension study, RECAP,[4] whilst neither Loveman and colleagues nor the nintedanib submission included these data. Furthermore, the manufacturers of pirfenidone submitted two confidential PASs during the pirfenidone STA. The ERG has conducted an analysis with all PAS information in a separate confidential appendix as requested by NICE.

4.2.9 Assessment of uncertainty

The company conducted a range of sensitivity analyses, including: one way sensitivity analyses, scenario analyses, and PSA. No methodological assumptions were tested in sensitivity analyses. No subgroups were identified by the CS. However, a scenario analysis did analyse an 'ASCEND-like population' that functions in a similar manner to a subgroup analysis with additional structure and parameter modifications to the model.

The company asserted that because pirfenidone is currently the only treatment accepted for treatment of IPF, that it is the correct comparator for cost-effectiveness analysis. Nintedanib dominated pirfenidone in the base case deterministic analysis, most of the one-way sensitivity analyses, and in PSA.

The company found that the choice of survival model for patient mortality was the most influential cost-effectiveness factor. However, the model assumes proportional hazards, so any changes in the survival model for the best supportive care arm have no effect on the ranking of interventions, they only increase the magnitude of the ICERs for nintedanib and pirfenidone compared to best supportive care.

One-way sensitivity analyses

The company conducted 46 one way sensitivity analyses. CS Tables 179-185 (p.286) list 45 of these one-way sensitivity analyses, each with a number 1-45.

- Fourteen one way sensitivity analyses tested 95% CI for all model parameters (Table 179, p.286).

- Seven one way sensitivity analyses evaluated alternative overall survival assumptions (Table 180, p. 286).
- Seven one way sensitivity analyses evaluated alternative values for acute exacerbations by changing assumptions or by changing the studies included in the NMA (Table 181, p. 287).
- Five analyses evaluated alternative values for loss of lung function by including an exacerbation coefficient in the survival regression or by changing the studies included in the NMA (Table 182, p. 287).
- Five analyses evaluated alternative values relating to drug safety (Table 183, p. 287).
- Five analyses evaluated alternative values for discontinuation using by using values from a Canadian Registry Study[50] or by changing the studies included in the NMA (Table 184, p. 287).
- Two analyses tested alternative values for starting FVC% predicted by using the top or bottom of the respective decile value ranges (i.e. 50 or 59.9 instead of the centre of the decile) (Table 185, p. 288).
- The 46th one-way sensitivity analysis evaluated the impact of applying hypothetical PASs to the price of pirfenidone with nintedanib at list price (Table 189, p. 294).

The model was robust to parameter uncertainty, with nintedanib remaining dominant compared to pirfenidone in analyses 1-45. This is due to all the analyses being applied to both comparator arms simultaneously due to most analyses adjusting baseline rates for BSC that are shared by both the nintedanib and pirfenidone arms. Adjusting mortality probabilities for BSC does not change the odds ratios applied in the model to pirfenidone and nintedanib. A similar logic is applicable to all analyses where a value is changed with no corresponding changes to ratios. The only analyses where there was the potential for changes in odds ratios were those that used alternative studies in the NMA, but these are not guaranteed to change the ranking of the odds ratios and the company did not report what the alternative odds ratio values were in Table 187 (p. 291-3). Partly due to most NMA scenarios not changing odds ratio rankings, multiple simultaneous changes are required to affect cost-effectiveness conclusions; the ERG explores this in section 3.1.7 In addition to the alternative values used in the analyses being absent in the results section (they are present in CS Appendix B), the actual ICER values and cost-effectiveness plane quadrant for the analyses were not presented, so the magnitude of the effects of the analyses was not transparent.

The one way sensitivity analysis that varied the cost of pirfenidone (Table 189, p. 294) showed that if the cost of pirfenidone was 5% lower, nintedanib no longer dominated pirfenidone, instead having an ICER of £13,663/QALY. The results from this analysis are presented in Table 38, below.

Structural uncertainty was addressed using alternative survival distributions (Weibull and Gompertz instead of log-logistic), and by allowing treatment discontinuation with or without maintenance of treatment effect.

Table 38 Impact of pirfenidone discount rate on the ICER, nintedanib at list price (CS Table 189, p.294)

	Discount Applied to Drug Cost						
Pirfenidone	**Base-case (0% discount)**	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**
Cost per pack	£2,151.10	£2043.55	£1935.99	£1828.44	£1720.88	£1613.33	£1505.77
ICER	Dominates	£13,663.45	£78,108.24	£142,553.03	£206,997.81	£271,442.60	£335,887.39
Pirfenidone	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**
Cost per pack	£1398.22	£1290.66	£1183.11	£1075.55	£968.00	£860.44	£752.89
ICER	£400,332.17	£464,776.96	£529,221.75	£593,666.54	£658,111.32	£722,556.11	£787,000.90
Pirfenidone	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**	
Cost per pack	£645.33	£537.78	£430.22	£322.67	£215.11	£107.56	
ICER	£851,445.68	£915,890.47	£980,335.26	£1,044,780.04	£1,109,224.83	£1,173,669.62	

Scenario Analysis

Three scenario analyses were undertaken:

- Analysis 46 explored the effect of a stopping rule for patients who observed a decline of 10% FVC% predicted or more with a loss of treatment effect for pirfenidone patients.
- An analysis using the relative effectiveness of nintedanib from the clinical trials rather than the NMA (Analysis 47).
- An analysis that compared an ASCEND-like population to pirfenidone by restricting selection criteria to 50-90 FVC% predicted; altering regression equations and data inputs for mortality, time to exacerbation, and loss of lung function; only including SAEs that occurred in more than 10% of patients; using a hazard ratio instead of an odds ratio to measure treatment effect between pirfenidone and nintedanib; using a relative risk instead of odds ratio for measuring lack of lung function; and taking pirfenidone discontinuation directly from the ASCEND trial instead of the NMA.

CS Table 187 (p. 291) presents the results for sensitivity analyses one to 47. For analysis 46 nintedanib no longer dominated pirfenidone and had an ICER of £82,784/QALY. The results for analysis 47 were reported as N/A. CS Table 188 (p. 293) provided the results for the ASCEND-like population (Analysis 48).The results of the ASCEND-like population are presented in Table 39 for the comparison of nintedanib and pirfenidone.

Table 39 ASCEND-like population analysis results (CS Table 188, p. 293)

	Pirfenidone	**Nintedanib**	**Incremental**
Treatment costs	£58,803.29	£64,387.68	£5,584.39
Adverse event costs	£361.74	£256.00	-£105.75
Liver panel tests	£9.01	£9.87	£0.86
Patient monitoring and O2 use	£9,276.29	£10,770.37	£1,494.08
Acute exacerbation costs	£1,142.57	£929.36	-£213.20
End of life costs	£14,094.61	£13,871.76	-£222.85
Total costs	£83,687.52	£90,225.03	£6,537.52
Total QALYs	2.5024	2.9881	0.4857
ICER (per QALY)			£13,459.17

Probabilistic Sensitivity Analysis

The company conducted a full PSA. All variables that were included in the PSA were given in CS Table 175 (p.279). Where variance and standard errors were not reported, the distributions used to model the outcomes and the parameters for generating those distributions were reported. Otherwise, mean and standard error were reported along with the distribution used to model each model parameter. For some parameters, lower and upper confidence intervals were reported.

The ERG considers the distributions chosen for the model parameters appropriate for their respective data and the list of parameters included in the PSA was comprehensive.

The PSA took 159 seconds to run for 1000 iterations. The results of the PSA were presented in CS Table 178 (p.283) and CS Figures 76 and 77 (p. 285), but neither total QALYs and costs nor ICERs were reported. Whilst the complete deterministic results were available in CS Table 165 (p. 266), the probabilistic results were not fully reported, so the probabilistic results in Table 40 were derived directly from the CS model by the ERG. The probabilistic and deterministic models produced nearly identical results.

Table 40 Base case deterministic and probabilistic results of the CS model (derived directly from the model)

	Total Costs	**Total QALY**	**ICER vs. BSC**	**Full incremental analysis**
Deterministic				
BSC	£25,359	3.27		
Pirfenidone	£87,479	3.62	£176,081	Dominated by nintedanib
Nintedanib	£85,087	3.67	£149,361	£149,361
Probabilistic				
BSC	£25,961	3.28		
Pirfenidone	£88,183	3.62	£181,248	Dominated by nintedanib
Nintedanib	£85,800	3.68	£146,630	£146,630

4.2.10 Comment on validity of results with reference to methodology used

The structure of the economic model adopted for the economic evaluation was appropriate, comprehensive and reflected the clinical pathway for patients with IPF. The economic model, developed in Microsoft Excel was transparent and easy to follow. The ERG did not find any errors in the coding of the model structure.

The methods chosen for the analysis were generally appropriate and conformed to NICE methodological guidelines. Similarly, the model parameters were generally appropriate. However, the ERG identified several areas where choice of parameter was not sufficiently justified or uncertainty was not insufficiently explored. Where these concerns were identified, the ERG has conducted additional analyses to address the uncertainty surrounding these parameters.

As identified in section 4.2.9 the company's sensitivity analyses did not adequately demonstrate the effect of varying a parameter for only one intervention at a time. The ERG conducted additional one way sensitivity analyses wherein only ORs for nintedanib were varied.

The ERG observed that the patient population in the company model may not reflect current clinical practice as a significant proportion of patients with milder IPF were included in the analysis, as discussed in 4.2.2 . In the pirfenidone STA,[7] clinical experts indicated that patients with an FVC greater than 80% predicted were unlikely to be treated in the UK. In the CS model 45.7% of the patients have FVC 80% of the predicted valueor above. To account for the disparity between the population in the model, and the population likely to present for treatment in the UK the ERG performed an additional analysis that restricted the starting model population to patients with FVC between 50% and 79.9% of the predicted value.

The ERG had reservations with regard to the company's choice of NMA scenarios informing the clinical effectiveness parameters within the CS model, as identified in sections 3.1.7 and 4.2.4. The company inconsistently switched which studies were included in the model and provided little justification for their NMA scenario choices. The company frequently chose scenarios that did not include all available evidence. Additionally, the company's choice of fixed or random effects models was inconsistent between the clinical effectiveness description of the NMA and the description of the model. In the model, no random effects NMA models were utilised. In the clinical effectiveness description of the NMA, some outcomes (e.g. acute exacerbation, serious

cardiac events) were identified as having a best fit with a random effects model. The ERG has conducted additional analyses to address this model inconsistency.

There are some areas of inconsistency between the company description of the model and the actual values used in the model for utility decrements. The utility decrement for new exacerbations does not match the utility decrement used in the model. Additionally, the model has assumed high proportions of patients experiencing utility decrements for the rash SAE in the pirfenidone arm of the model and has assumed a utility decrement equivalent to experiencing the SAE for a year rather than the time of less than one month stated by the ERG's clinical advisor.

The ERG undertook sensitivity analyses to assess the effects of alterations to the identified inconsistencies and poorly justified parameter choices. The methods used in these sensitivity analyses and results of these analyses are reported in section 4.3.

4.3 Additional work undertaken by the ERG

In order to investigate methodological, structural, and parameter uncertainty issues raised in their assessment, the ERG undertook a series of deterministic sensitivity analyses and scenario analyses. These analyses are conducted without PAS submissions from Boehringer-Ingelheim (nintedanib) and Intermune (pirfenidone). All analyses with a PAS for both nintedanib and pirfenidone are conducted in a separate confidential appendix, as requested by NICE.

The base case analysis is provided in Table 41 for quick reference to the effects of changes to the model.

Table 41 Base case analysis

Treatment	Total costs	Total QALYs	ICER vs. BSC	Incremental ICER (£/QALY)
BSC	£25,359	3.27		
Nintedanib	£85,087	3.67	£149,361	£149,361
Pirfenidone	£87,479	3.62	£176,081	dominated by nintedanib

The ERG raised concerns with regard to the deterministic sensitivity analyses in the company model. Many of the deterministic sensitivity analyses conducted in the company submission

adjust the BSC arm, whilst leaving the OR that determines effectiveness in nintedanib and pirfenidone fixed. The ERG performed one way sensitivity analyses on only nintedanib using the upper and lower bounds of 95% CI for the following ORs for nintedanib vs. BSC: overall survival, exacerbation, loss of lung function, serious cardiac events and serious gastrointestinal events. Table 42 presents the results of these one-way sensitivity analyses.

Table 42 One way sensitivity analyses using 95% CI of nintedanib efficacy OR

Scenario	Value in analysis	ICER vs. BSC	ICER vs. pirfenidone
Basecase	--	£149,361	Dominant
Overall survival			
Basecase	0.70		
Lower limit	0.447	£87,246	Dominant
Upper limit	1.095	Dominated	£27,030
Exacerbation			
Basecase	0.56		
Lower limit	0.350	£145,272	Dominant
Upper limit	0.889	£155,751	Dominant
Loss of lung function			
Basecase	0.54		
Lower limit	0.416	£143,279	Dominant
Upper limit	0.687	£158,035	Dominant
Serious cardiac events			
Basecase	0.92		
Lower limit	0.533	£148,220	Dominant
Upper limit	1.630	£151,436	Dominant
Serious gastrointestinal events			
Basecase	2.35		
Lower limit	1.052	£148,843	Dominant
Upper limit	5.875	£150,751	Dominant

Of these analyses, only setting the OR for overall survival compared to BSC to 1.095 changed the results from nintedanib dominating pirfenidone. For the other sensitivity analyses the ICERs vs. BSC for nintedanib varies between £143,279 and £155,751 per QALY in Table 42.

To further explore uncertainty in the model the ERG conducted a number of scenario analyses. Table 43 provides brief descriptions of these analyses with full descriptions in the paragraphs below. Table 44 provides the results of the scenario analyses.

Table 43 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 = 0.014
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

SUPERSEDED - SEE ERRATUM

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 0.014 to all new exacerbations. The company submission stated that new exacerbations have a utility decrement of 0.014 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 0.014 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.

SUPERSEDED – SEE ERRATUM

Table 44 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 0.014				
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 44, 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 photsensitivity and rash duration lowered perfinidone's ICER vs. BSC by £8,248 (Table 44). 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 Table 45.

SUPERSEDED – SEE ERRATUM

Table 45 Combined scenario analysis conducted by the ERG of analyses 1,2,4 and 5

Treatment	Total costs	Total QALYs	ICER vs. BSC	Incremental ICER (£/QALY)
BSC	£27,960	3.0441		
Nintedanib	£87,941	3.4365	£152,861	£152,861
Pirfenidone	£89,984	3.4443	£155,000	£263,051

The ERGs alternative base case further narrows the ICERs for nintedanib and pirfenidone vs. BSC. Additionally, with all the model changes in place, pirfenidone produces 0.008 more total QALYs than nintedanib. It seems clear that at list price there are no meaningful differences in cost-effectiveness between pirfenidone and nintedanib, and that they are likely interchangeable for the purpose of cost-effectiveness decisions.

4.4 Summary of uncertainties and issues

The CS reports that nintedanib dominates pirfenidone across a wide range of sensitivity analyses, i.e. nintedanib is more effective and less costly. This dominance is also apparent from the additional analyses conducted by the ERG. However, the results of the cost effectiveness analyses at nintedanib list price indicate that the base case results, including total costs, total life years and total QALYs for both nintedanib and pirfenidone are similar. The cost effectiveness results between the two treatments are largely driven by overall survival, which has been modelled to be equal for patients receiving the two drugs.

There remain some uncertainties with regard to the external consistency between the CS model in this STA (nintedanib), and the pirfenidone model in TA 282.[7] In the nintedanib model, neither nintedanib nor pirfenidone are cost effective, with average cost effectiveness estimates compared to best supportive care of over £149,000 per QALY for both treatments. The ERG notes that in most of the scenario analyses the ICER values obtained for nintedanib compared to BSC remain around £150,000 per QALY using the list price of nintedanib which would not be considered cost effective at the NICE willingness to pay threshold of £20,000 to £30,000 per QALY.

[redacted] In contrast, the pirfenidone model produces an ICER of £24,000/QALY with PAS included.[7] The disparity

between the model results highlights a need for careful examination of the differences between the two models, but this is not possible in the STA process due to confidential data.

5 End of life

The company does not apply the NICE end of life criteria in the submission. The NICE methods guide states the end of life criteria includes 'that treatment is indicated for patients with a short life expectancy, normally less than 24 months' and for a small population not exceeding a cumulative total of 7000.[20] The ERG notes that the CS states the life expectancy of IPF patients is approximately 2 to 5 years and the patient population is currently 15,000 and concludes that the submission does not meet NICE's end of life criteria.

6 Innovation

The company highlights the limited treatment options for adults with IPF. Only one treatment, pirfenidone, has been recommended by NICE[7] for patients with IPF whose FVC% predicted is between 50% and 80% (generally considered to be mild to moderate IPF) and pirfenidone treatment should be stopped if FVC falls by 10% or more in 12 months. Best supportive care is the only alternative option for patients whose FVC% predicted lies outside the 50-80% range. The ERG notes that the licensed indication for pirfenidone states mild to moderate IPF but does not provide a definition of this based on FVC% predicted.

In contrast to pirfenidone, nintedanib is licensed for adults with IPF of any severity. Therefore nintedanib could be an alternative treatment option for patients who are currently eligible for pirfenidone treatment but could also be a treatment option for those patients whose FVC% predicted lies outside the 50-80% range.

The company also points out the reduced 'pill burden' with nintedanib of one 150mg capsule twice daily in comparison to pirfenidone which has a recommended dose of three 267mg capsules three times a day.

7 DISCUSSION

7.1 Summary of clinical effectiveness issues

The company identified one phase II RCT and two replicate phase III RCTs that are relevant to the decision problem. The trials enrolled participants with an FVC that was 50% or more of the predicted value. There are no head-to-head trials comparing nintedanib to pirfenidone.

An NMA was conducted to provide indirect evidence for the nintedanib versus pirfenidone comparison via a common placebo comparator placebo. The CS presents NMA results for nine outcomes, six of which contribute to the economic model. For each outcome an 'all evidence' scenario which included all the available evidence was reported. For most outcomes one or more additional scenarios were reported in which a trial (or trials) was excluded from the NMA. For several outcomes 52 week data from nintedanib trials is compared to 72 week data from pirfenidone trials and the impact of these differing lengths of follow up which could potentially disadvantage pirfenidone is uncertain. For some NMA outputs contributing to the economic model scenario analyses were used instead of all of the evidence. The ERG therefore has some concerns regarding the potential for selection bias among the outputs from the NMA.

7.2 Summary of cost effectiveness issues

The CS includes evidence on the cost effectiveness of nintedanib compared to pirfenidone and BSC for IPF. The model structure and methods adopted for the economic evaluation are reasonable and are generally appropriate. The model structure and model parameter inputs are consistent with the clinical disease pathways and the available clinical trial evidence. The model results suggest that nintedanib has a cost effectiveness versus BSC of £149,361 per QALY gained using the list price of nintedanib and [redacted] using the nintedanib PAS. In the comparison between nintedanib and pirfenidone, the total costs and QALYs are similar but nintedanib dominates pirfenidone.

The company has used a population in the economic model than are milder than would be likely be seen in current UK practice, by including patients with FVC% predicted more than 80%.

The company did not fully investigate the uncertainty around the model results in their deterministic sensitivity analyses. Many of the deterministic sensitivity analyses conducted in

the company submission adjust the BSC arm, whilst leaving the ORs that determine effectiveness in nintedanib and pirfenidone fixed.

8 REFERENCES

(1) Richeldi L, Costabel U, Selman M, Kim DS, Hansell DM, Nicholson AG et al. Efficacy of a tyrosine kinase inhibitor in idiopathic pulmonary fibrosis. *N Engl J Med* 2011; 365(12):1079-1087.

(2) Richeldi L, du Bois RM, Raghu G, Azuma A, Brown KK, Costabel U et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* 2014; 370(22):2071-2082.

(3) King TE, Jr., Bradford WZ, Castro-Bernardini S, Fagan EA, Glaspole I, Glassberg MK et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* 2014; 370(22):2083-2092.

(4) Costabel U, Albera C, Bradford WZ, Hormel P, King TE, Jr., Noble PW et al. Analysis of lung function and survival in RECAP: An open-label extension study of pirfenidone in patients with idiopathic pulmonary fibrosis. *Sarcoidosis Vasculitis & Diffuse Lung Diseases* 2014; 31(3):198-205.

(5) National Institute for Health and Care Excellence. NICE pathway: Idiopathic pulmonary fibrosis. 2013. 8-6-0015.

(6) National Institute for Health and Care Excellence. Idiopathic pulmonary fibrosis: The diagnosis and management of suspected idiopathic pulmonary fibrosis (NICE clinical guideline 163). 2013. 8-6-2015.

(7) National Institute for Health and Care Excellence. Pirfenidone for treating idiopathic pulmonary fibrosis (TA282 NICE guidance). 2013. NICE. 8-6-2015.

(8) European Medicines Agency. Summary of product characteristics: Ofev. 2015. 8-6-2015.

(9) du Bois RM, Weycker D, Albera C, et al. Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference. *Am J Respir Crit Care Med* 2011; 184(12):1382-1389.

(10) Collard HR, Moore BB, Flaherty KR, et al. Acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respi Crit Care Med* 2007; 176(7):636-643.

(11) Kim HJ, Perlman D, Tomic R. Natural history of idiopathic pulmonary fibrosis. *Respiratory Medicine* 2015; 109(6):661-670.

(12) American College of Rheumatology. Six minute walk test (6MWT). 2011. 8-6-2015.

(13) Cooper K, Mendes D, Picot J, Loveman E. Pirfenidone for the treatment of idiopathic pulmonary fibrosis: A single technology appraisal (TA282 ERG Report). 2012. SHTAC.

(14) Taniguchi H, Xu Z, Azuma A, Inoue Y, Li H, Fujimoto T. Subgroup analysis of asian patients in the Inpulsis trials of nintedanib in idiopathic pulmonary fibrosis. *Respirology (Carlton, Vic)* 2014; 19:30.

(15) Ogura T, Taniguchi H, Azuma A, Inoue Y, Kondoh Y, Hasegawa Y et al. Safety and pharmacokinetics of nintedanib and pirfenidone in idiopathic pulmonary fibrosis. *European Respiratory Journal* 2015; 45(5):1382-1392.

(16) Inoue Y, Azuma A, Taniguchi H, Ogura T, Tadayasu Y, Fujimoto T et al. The Pharmacokinetics of Bibf 1120 Alone Or in Combination with Pirfenidone in Japanese Patients with Idiopathic Pulmonary Fibrosis (Ipf). *Respirology* 2011; 16:318.

(17) Jones PW, Quirk FH, Baveystock CM, et al. A self-complete measure of health status for chronic airflow limitation. *Am Rev Respir Dis* 1992; 145:1321-1327.

(18) Yorke J, Jones PW, Swigris JJ. Development and validity testing of an IPF-specific version of the St George's Respiratory Questionnaire. *Thorax* 2010; 65:921-926.

(19) Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med* 2001; 33(337):343.

(20) National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. *NICE, London* 2013.

(21) Noble PW, Albera C, Bradford WZ, et al. Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials. *Lancet* 2011; 377:1760-69.

(22) Azuma A, Nukiwa T, Tsuboi E, et al. Double-blind, placebo-controlled trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2005; 171(9):1040-47.

(23) Taniguchi H, Ebina M, Kondoh Y, et al. Pirfenidone in idiopathic pulmonary fibrosis. *European Respiratory Journal* 2010; 35:821-29.

(24) Homma S, Azuma A, Taniguchi H, Ogura T, Mochiduki Y, Sugiyama Y et al. Efficacy of inhaled N-acetylcysteine monotherapy in patients with early stage idiopathic pulmonary fibrosis. *Respirology* 2012; 17(3):467-477.

(25) Loveman E, Copley VR, Scott DA, Colquitt JL, Clegg AJ, O'Reilly KMA. Comparing new treatments for idiopathic pulmonary fibrosis- a network meta-analysis. *BMC Pulmonary Medicine* 2015; 15(37).

(26) Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. 2011. Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA.

(27) Loveman E, Copley VR, Colquitt J, Scott DA, Clegg A, Jones J et al. The clinical effectiveness and cost-effectiveness of treatments for idiopathic pulmonary fibrosis: a systematic review and economic evaluation. *Health Technol Assess* 2015; 19(20):i-336.

(28) Ara R, Brazier J. Using health state utility values from the general population to approximate baselines in decision analytic models when condition specific data are not available. HEDS discussion paper 10/11. [Last update 2010

(29) MIMS. Presciption drug database and drug prescribing guide. [Last update 2011

(30) Department of Health. NHS reference costs 2012-2013. [Last update 2014 , cited 2010 June 7];

(31) Curtis L. Unit Costs of Health and Social Care 2009. [Last update 2010

(32) Hagaman JT, Kinder BW, Eckman MH. Thiopurine S-Methyltransferase Testing in Idiopathic Pulmonary Fibrosis: A Pharmacogenetic Cost-Effectiveness Analysis. *Lung* 2010; 188:125-132.

(33) Drummond MF, O'Brien B, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes (3rd edition). *Oxford University Press* 2005.

(34) Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technology Assessment* 2004; 8(36):1-158.

(35) Boehringer Ingelheim. Nintedanib for the treatment of adults with idiopathic pulmonary fibrosis (Company Evidence Submission). 2015.

(36) Intermune UK and Ireland. Pirfenidone for the treatment of idiopathic pulmonary fibrosis (TA282 Company Submission). 1-12-2011. National Institute for Health and Care Excellence (NICE). Single Technology Appraisals.

(37) Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. *Health Econ* 2006; 15(12):1295-1310.

(38) Martinez FJ, de Andrade JA, Anstrom KJ, King TE, Jr., Raghu G. Randomized trial of acetylcysteine in idiopathic pulmonary fibrosis. *N Engl J Med* 2014; 370(22):2093-2101.

(39) Kondoh Y, Taniguchi H, Katsuta T, Kataoka K, Kimura T, Nishiyama O et al. Risk Factors of Acute Exacerbation of Idiopathic Pulmonary Fibrosis. *Sarcoidosis Vasculitis and Diffuse Lung Disease* 2010; 27:103-110.

(40) Nathan SD, Shlobin OA, Weir N, Ahmad S, Kaldjob JM, Battle E et al. Long term Course and Prognosis of Idiopathic Pulmonary Fibrosis in the New Millennium. *Chest* 2011; 140(1):221-229.

(41) King TE, Jr., Brown KK, Raghu G, du Bois RM, Lynch DA, Martinez F et al. BUILD-3: a randomized, controlled trial of bosentan in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011; 184(1):92-99.

(42) Zisman DA, Schwarz M, Anstrom KJ, Collard HR, Flaherty KR, Hunninghake GW. A controlled trial of sildenafil in advanced idiopathic pulmonary fibrosis. *N Engl J Med* 2010; 363(7):620-628.

(43) Costabel U, Bendstrup E, Cottin V, Dewint P, Egan JJ, Ferguson J et al. Pirfenidone in idiopathic pulmonary fibrosis: expert panel discussion on the management of drug-related adverse events. *Adv Ther* 2014; 31(4):375-391.

(44) Leontiadis GI, Sharma VK, Howden CW. Proton pump inhibitor therapy for peptic ulcer bleeding: Cochrane collaboration meta-analysis of randomized controlled trials. *Mayo Clin Proc* 2007; 82(3):286-296.

(45) Costabel U, Albera C, Fagan E, Bradford W, King T, Noble P et al. Long-term safety of pirfenidone in RECAP, an open-label extension study in patients with idiopathic pulmonary fibrosis, interim results. *European Respiratory Journal* 2014; Conference(var.pagings).

(46) Datapharm. Electronic medicines compenium (eMC). [Last update 2015

(47) Parfrey HL, Gibbons MA, Armstrong E, Harris E, Frank R, Sharp C et al. Healthcare utilisation by patients with idiopathic pulmonary fibrosis: observations form the Uk pirfenidone named patient programme. Thorax 68, A164-A165. 2013.

(48) Department of Health. Costing tool for home oxygen: Assessment and review. [Last update 2015

(49) Hatziandreu E, Archontakis F, Daly A. The potential cost savings of greater use of home and hospice-based end of life care in England. [Last update 2015

(50) MacQuarrie J, Lebel F. Inspiration: An assistance program for idiopathic pulmonary fibrosis patients on pirfenidone. *European Respiratory Journal* 2014; Conference(var.pagings).