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

Bortezomib for induction therapy in multiple myeloma before high dose chemotherapy and autologous stem cell transplantation

Produced by Southampton Health Technology Assessments Centre

Authors Keith Cooper, Senior Research Fellow, SHTAC
Debbie Hartwell, Senior Research Fellow, SHTAC
Vicky Copley, Research Fellow, SHTAC
Karen Pickett, Research Fellow, SHTAC
Jackie Bryant, Principal Research Fellow, SHTAC

Correspondence to Debbie Hartwell
Southampton Health Technology Assessments Centre
University of Southampton
First Floor, Epsilon House
Enterprise Road, Southampton Science Park
Southampton SO16 7NS

Date completed 22nd May 2013

Source of funding: This report was commissioned by the NIHR HTA Programme as project number 12/41/01.

Declared competing interests of the authors

None.

Acknowledgements

We are very grateful to the following expert who offered clinical advice and comments on the draft report: Dr Matthew Jenner, Consultant Haematologist and Haematology Clinical Lead, Southampton University Hospitals NHS Trust, Southampton General Hospital, Tremona Road, Southampton, Hampshire SO16 6YD.

We also thank: Karen Welch, Information Specialist, SHTAC, for commenting on the manufacturer's search strategy, and Jonathan Shepherd, Principal Research Fellow, SHTAC, for acting as internal editor for the ERG report.

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, Hartwell D, Copley V, Pickett K, Bryant J. Bortezomib for induction therapy in multiple myeloma before high dose chemotherapy and autologous stem cell transplantation: A Single Technology Appraisal. SHTAC. 2013.

Contributions of authors

K Cooper (Senior Research Fellow) critically appraised the health economic systematic review and the economic evaluation and drafted the report; D Hartwell (Senior Research Fellow) critically appraised the clinical effectiveness systematic review, drafted the report and project managed the review; V Copley (Research Fellow) critically appraised the health economic systematic review and the economic evaluation and drafted the report; K Pickett (Research Fellow) critically appraised the clinical effectiveness systematic review; J Bryant (Principal Research Fellow) critically appraised the clinical effectiveness systematic review and drafted the report.

TABLE OF CONTENTS

1 Introduction to ERG Report ... 10
2 BACKGROUND ... 10
2.1 Critique of manufacturer's description of underlying health problem ... 10
2.2 Critique of manufacturer's overview of current service provision ... 10
2.3 Critique of manufacturer's definition of decision problem ... 10
3 CLINICAL EFFECTIVENESS ... 12
3.1 Critique of manufacturer's approach to systematic review ... 12
3.2 Summary statement of manufacturer's approach ... 23
3.3 Summary of submitted evidence ... 24
3.4 Summary ... 30
4 ECONOMIC EVALUATION ... 32
4.1 Overview of manufacturer's economic evaluation ... 32
4.2 Critical appraisal of the manufacturer's submitted economic evaluation ... 34
4.3 Additional work undertaken by the ERG ... 55
4.4 Summary of uncertainties and issues ... 58
5 End of life ... 59
6 Innovation ... 59
7 DISCUSSION ... 59
7.1 Summary of clinical effectiveness issues ... 59
7.2 Summary of cost-effectiveness issues ... 59
8 REFERENCES ... 61
9 APPENDIX 1 ... 64
9.1 Clinical effectiveness critique of the Hovon, IFM and MRC MMIX trials ... 64
9.2 Economic analysis ... 73

LIST OF TABLES

Table 1: List of trials included in the MS ... 14
Table 2: Manufacturer and ERG assessment of trial quality ... 17
Table 3: ERG appraisal of MTC approach ... 22
Table 4: Quality assessment (CRD criteria) of MS review ... 23
Table 5: Response rates post-induction and post-transplant ... 26
Table 6: Median PFS (months) and HR of PFS (months) ... 27
Table 7: Median TTP (months) and HR of TTP (months) ... 27
Table 8: Overall survival HR of death ... 27
Table 9: Adverse events ... 29
Table 10: Withdrawals from treatment during induction ... 30
Table 11: Base case cost-effectiveness results ... 34
Table 12: Critical appraisal checklist of economic evaluation ... 34
Table 13: NICE reference case requirements ... 35
Table 14: Post-induction response rates (Pethema trial) ... 40
Table 15: Total SCT proportions by treatment arm (Pethema trial) ... 40
Table 16: SCT rate by post-induction response category (Pethema trial) ... 41
Table 17: Mortality rate during induction period by treatment arm (Pethema trial) ... 41
Table 18: Overall survival by maximal response to treatment category ... 44
Table 19: Summary of utility values for cost-effectiveness analysis ... 46
Table 20: Unit costs associated with the 1st line induction therapies VTD and TD: drugs, prophylaxis, administration and monitoring ... 48
Table 21: Base case cost-effectiveness results versus TD ... 55
Table 22: Base case cost-effectiveness results versus CTD ... 56
Table 23: ERG analysis of changes to median overall survival (in months) by post-induction response category, VTD vs. TD model ... 57

Table 24: Post-induction and post-SCT response achieved in Pethema trial, by treatment arm 58
Table 25: ERG scenario analysis using post-SCT response rates, VTD vs. TD model..... 58
Table 26: Manufacturer and ERG quality assessment of Hovon, IFM and MRC MMIX trials65
Table 27: Base case cost-effectiveness results for PAD vs. VAD and VD vs. VAD 74
Table 28: Post-induction response rates used in PAD vs. VAD and VD vs. VAD models 76
Table 29: Total SCT proportions by treatment arm for PAD, VAD, VD and VAD treatments 76
Table 30: Mortality rate during induction period by treatment arm 77
Table 31: Unit costs associated with the 1st line induction therapies: drugs, prophylaxis, administration and monitoring 78

LIST OF FIGURES

Figure 1: Schematic of the state transition model..... 36
Figure 2: Comparison of overall survival predicted by model and overall survival observed in Pethema trial, by treatment arm 51
Figure 3: Difference between OS observed in the Pethema trial and OS predicted by the model for the VTD and TD treatment arms..... 51

LIST OF ABBREVIATIONS

Abbreviation	Definition
AE	Adverse event
ASCT	Autologous stem cell transplant
BNF	British National Formulary
CI	Confidence interval
CR	Complete response
CrIs	Credible intervals
CSR	Clinical study report
CTD	Cyclophosphamide, thalidomide and dexamethasone
CVAD	Cyclophosphamide, vincristine, adriamycin (doxorubicin) and dexamethasone
DCEP	Dexamethasone combined with cyclophosphamide, etoposide or etoposide phosphate, and cisplatinum
ECOG	Eastern Cooperative Oncology Group
ERG	Evidence Review Group
EQ5D	Euro-QoL 5D
HDC	High-dose chemotherapy
HR	Hazard ratio
HRQoL	Health related quality of life
ICER	Incremental cost-effectiveness ratio
IFM	Intergroupe Francophone du Myelome
ISS	International Staging System
ITT	Intention-to-treat
LYG	Life years gained
MM	Multiple myeloma
MRC MMIX	Medical Research Council Multiple Myeloma IX
MTC	Mixed treatment comparison
n	Sample size
N	Group size
N/A	Not applicable
nCR	Near complete response
NICE	National Institute for Health and Clinical Excellence
NR	Non-responders
N/R	Not reported
ORR	Overall response rate
OS	Overall survival
PAD	Bortezomib, adriamycin (doxorubicin) and dexamethasone
PFS	Progression-free survival
PN	Peripheral neuropathy
PR	Partial response
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QALY	Quality adjusted life year
RCT	Randomised controlled trial
RR	Relative risk
SCT	Stem cell transplant
SD	Standard deviation
TD	Thalidomide and dexamethasone
TTP	Time to progression
VAD	Vincristine, Adriamycin (doxorubicin) and dexamethasone
VAT	Value added tax
VD	Bortezomib and dexamethasone
VGPR	Very good partial response
VTD	Bortezomib, thalidomide and dexamethasone

SUMMARY

Scope of the manufacturer submission

The scope of the manufacturer's submission (MS) is the clinical effectiveness and cost-effectiveness of bortezomib for induction therapy in multiple myeloma (MM) before high dose chemotherapy and autologous stem cell transplantation. The decision problem specified in the MS generally reflects the scope of the appraisal issued by the National Institute for Health and Clinical Excellence (NICE) except that comparators are not limited to chemotherapy regimens containing thalidomide and is therefore wider than the NICE scope.

Summary of submitted clinical effectiveness evidence

The clinical effectiveness evidence in the MS which meets the NICE scope comes from two open label RCTs (the Pethema and Gimema trials) both of which compared VTD (bortezomib, thalidomide and dexamethasone) with TD (thalidomide and dexamethasone) using bortezomib at a dose of 1.3 mg/m^2 in a population of newly diagnosed, previously untreated, multiple myeloma patients. Treatment pathways and subsequent maintenance/consolidation therapy were different in the two trials.

The primary outcomes were post-induction and post-transplant response rates. A statistically significantly greater number of patients treated with VTD compared with TD achieved an overall response rate (ORR) post-induction (Pethema 84.6% vs 61.4%, p<0.001; Gimema 93.2% vs 78.6%, p<0.0001) and post- transplant (Pethema 77.7% vs 56.7%, p<0.001; Gimema 93.2% vs 84.5%, p=0.0025). There was also a statistically significantly greater number of patients who achieved complete response (CR) when treated with VTD compared with TD post-induction (Pethema 35.4% vs 13.4%, p<0.001; Gimema 18.6 % vs 4.6%, p<0.0001) and post- transplant (Pethema 46.9% vs 23.6%, p<0.001; Gimema 37.7% vs 22.7%, p=0.0004). Both the Pethema and Gimema trials reported that a statistically significantly lower proportion of patients experienced disease progression when treated with VTD compared with TD post-induction (Pethema 6.2% vs 23.6%, p=0.0004; Gimema 0% vs 5%, p=0.0005). The difference was maintained post-transplant for the Gimema trial (<1% vs 7%, p=0.0001) but not in the Pethema trial.

Secondary outcomes included progression free survival (PFS), time to progression (TTP) and overall survival (OS). Unadjusted PFS hazard ratios (HRs) showed a statistically significant longer PFS for VTD compared with TD (Pethema HR 0.65, 95% CI 0.45, 0.92, p=0.015; Gimema HR 0.63, 95% CI 0.45, 0.88, p=0.0061) with median follow-up of 35.9 months (Pethema) and 36 months (Gimema). The unadjusted TTP HR showed a statistically

significantly lower hazard of progression in patients treated with VTD compared with TD (Pethema HR 0.64 95% CI 0.44, 0.93, p=0.017; not reported for Gimema). There were no statistically significant differences between VTD and TD for OS. Data for the proportion of patients who underwent stem cell transplant (SCT) were not powered nor were statistical tests reported so results are uncertain. Adverse events were similar for both treatments except for any grade 3/4 adverse event in the Gimema trial where they were statistically significantly higher for VTD compared with TD (relative risk (RR) 1.69, 95%CI 1.36, 2.08) and any treatment-related adverse event in the Pethema trial where they were statistically significantly higher for VTD compared with TD (RR 1.42, 95% CI 1.17, 1.73). In addition, there was a greater incidence of peripheral neuropathy in patients receiving bortezomib (VTD) than TD (Pethema 6.2% vs 0, no p values; Gimema 10% vs 2%, p=0.0004).

Results from the three additional trials in the MS (Hovon, IFM, Medical Research Council Multiple Myeloma IX [MRC MMIX]) are only presented in an appendix here because they include comparators not specified in the NICE scope.

Summary of submitted cost-effectiveness evidence

The MS includes:

i) A review of published economic evaluations of the treatment of newly diagnosed MM;
ii) Three cost-effectiveness analyses of bortezomib-based regimens for patients with newly diagnosed MM: VTD compared to TD; bortezomib, doxorubicin and dexamethasone (PAD) compared to vincristine, doxorubicin and dexamethasone (VAD); and bortezomib and dexamethasone (VD) compared to VAD.

A systematic search of the literature was conducted by the manufacturer to identify economic evaluations of treatments of newly diagnosed MM. The review identified three studies which met the MS inclusion criteria, and the MS considered relevant to the decision problem, however none of the studies are within the NICE scope for this appraisal.

State-transition models for each of the analyses were developed with a similar structure. The model structure is based upon the clinical pathway of care for MM, including the distinct phases of treatment for induction, SCT, and subsequent lines of treatment after disease progression. The estimation of long term survival and progression free survival is based upon surrogate outcomes for post-induction response (CR, partial response [PR], non-responders [NR]). The model adopts a 30 year time horizon to capture long term costs and health outcomes, with a cycle length of one month.

Results are presented for costs and quality-adjusted life years (QALYs) and incremental cost-effectiveness ratios (ICERs) for each of the analyses. For the base case analysis (VTD vs. TD), an ICER of £24,683 / QALY is presented. The other two analyses both had ICERs < £15,000 / QALY.

The model explores structural and parameter uncertainty in one-way and probabilistic sensitivity analyses (PSA). In the base case analysis (VTD vs. TD), the ICER is most sensitive to post induction CR mortality and VTD drug costs. The PSA estimates that there is a probability that VTD is cost-effective against TD at the £20,000 and £30,000 willingness to pay thresholds of 18.6% and 54.8% respectively.

Commentary on the robustness of submitted evidence

Strengths

- The MS contains systematic searches for the clinical and cost-effectiveness studies of bortezomib. It appears unlikely that these have missed any studies that would have met the inclusion criteria.
- The systematic review meets the Centre for Reviews and Dissemination (CRD) criteria for methodological quality.
- The model structure is consistent with the clinical pathway of care for multiple myeloma.
- The economic model has been presented in a clear and transparent format, and the Excel model is well presented and user-friendly.

Weaknesses and Areas of uncertainty

- Of the five trials included in the clinical effectiveness review, three trials do not meet the NICE scope as they do not contain a thalidomide comparator. The trials have different treatment pathways and it is unclear how these affect the results.
- There are a number of issues around the outcome measures: post-induction response rate is a surrogate outcome and it is not clear how good a predictor of long term outcomes it is. Furthermore, long-term outcomes (PFS, OS) may be confounded by post-induction consolidation and maintenance therapies which do not reflect current UK clinical practice. There is also uncertainty in the PFS and OS results due to the high censoring of data and the reporting of data unadjusted for maintenance therapy.

- There are key concerns over the mixed treatment comparison (MTC) analysis due to the assumptions made to develop a network of evidence in the absence of trial data, and heterogeneity across the trials.
- Of the three analyses submitted, two analyses do not meet the NICE scope (PAD vs. VAD and VD vs. VAD). Furthermore, for the other analysis (VTD vs. TD), neither treatment is currently used routinely in the NHS. A more appropriate comparator would be CTD, instead of TD, which is routinely used in the UK, but this has not been included in the MS economic analysis.
- The estimation of long term survival and progression free survival is based upon surrogate outcomes for post-induction response (CR, PR, NR). However, there is not a good fit between post-induction response and OS and time to progression compared to estimates from the Pethema trial, and the results presented are systematically biased in favour of VTD.

Summary of additional work undertaken by the ERG

The ERG conducted the following additional analyses:

a) Comparing all treatment analyses
b) Two alternative scenarios for post-induction mortality (VTD vs. TD model)
c) Post-SCT response rate from Pethema trial used instead of post-induction response rate (VTD vs. TD model)

The MS provides three pairwise analyses for bortezomib induction treatment: VTD vs. TD; PAD vs. VAD; VD vs. VAD. For illustrative purposes, the results of all relevant treatments have been compared. In addition, illustrative results have been shown for VD and PAD vs. CTD, by deriving the estimates for CTD using the response rates from the MRC MMIX trial.

The model results were sensitive to the parameters used for post-induction mortality in the VTD vs. TD model. The ICER ranged from £38,750 to £110,727 for the two alternatives. The results were also sensitive to using post-SCT response rate, rather than post-induction response rate in the VTD vs. TD model (ICER of £35,915 / QALY).

1 Introduction to ERG Report

This report is a critique of the manufacturer's submission (MS) to NICE from Janssen on the clinical effectiveness and cost-effectiveness of bortezomib (Velcade®) for induction therapy in multiple myeloma (MM) before high dose chemotherapy and autologous stem cell transplant (HDT-ASCT). It identifies the strengths and weakness of the MS. A clinical expert was consulted to advise the ERG and to help inform this review. The MS was received on 13th February 2013, and due to further work by the manufacturer, a re-submission was received on 12th March 2013.

Clarification on some aspects of the MS was requested from the manufacturer by the ERG via NICE on 22nd March 2013 (sent to the manufacturer on 27th March 2013). A response from the manufacturer via NICE was received by the ERG on 12th April 2013 and this can be seen in the NICE evaluation report for this appraisal.

2 BACKGROUND

2.1 Critique of manufacturer's description of underlying health problem

The MS provides a clear and accurate overview of MM.

2.2 Critique of manufacturer's overview of current service provision

The MS provides an accurate overview of current service provision. The MS notes that none of the available drug regimens are currently licensed for use in induction therapy of MM (MS p.28).

2.3 Critique of manufacturer's definition of decision problem

Population

The population described in the decision problem is appropriate for the NHS.

Intervention

Bortezomib has not yet been granted marketing authorisation [redacted]. It is anticipated to be indicated in combination with oral dexamethasone (VD), or with oral dexamethasone and oral thalidomide (VTD), for the induction treatment of adult patients with previously untreated MM eligible for high-dose chemotherapy with haemotopoietic stem cell

transplantation. [redacted] The recommended dose of bortezomib is 1.3 mg/m^2 administered subcutaneously as four injections per cycle for 3-6 cycles (depending on the combination).

In current UK clinical practice, cyclophosphamide, thalidomide, dexamethasone (CTD) is the most commonly used induction regimen. Bortezomib-based treatment is usually only given for induction therapy when patients are unsuitable for, or intolerant to, CTD or for those with renal impairment.

Comparators

The main comparator in the MS decision problem is CTD as 'it is the most widely used and for which there is the most UK clinical experience' (MS p.33). It is stated that due to the lack of head–to-head trials of bortezomib regimens and CTD, thalidomide and dexamethasone (TD) and vincristine, doxorubicin and dexamethasone (VAD) are also included as comparators. However, it should be noted that VAD is outside the scope of the appraisal as it does not contain thalidomide.

Outcomes

The outcomes included in the MS are appropriate and clinically meaningful to patients.

Economic analysis

The economic evaluation in the MS decision problem appears to be appropriate, being a cost-utility analysis from the NHS and Personal Social Services (PSS) perspective.

Other relevant factors

Subgroups reported in the MS include analysis by cytogenetic risk (in main report) and also International Staging System (ISS) and creatinine clearance (in appendix only).

The MS states that issues relating to equity or equality are not applicable and this is in line with the NICE scope.

3 CLINICAL EFFECTIVENESS

3.1 Critique of manufacturer's approach to systematic review

3.1.1 Description of manufacturer's search strategy

The search strategies and database selection, covering clinical, cost-effectiveness and quality of life are well documented and considered fit for purpose. A mix of index terms and free text have been applied and appropriately combined into sets, and suitable search filters were employed. All searches are reproducible and although the return of numbers on each search line is not documented, the total return is summarised in a flow chart. It is noted that exact replication of the clinical searches by the ERG would not be feasible on account of the use of different database hosts, however the strategy and syntax based on ERG searching expertise appears adequate. The re-submission of the MS did not affect the content of the cost search strategies as they were not drug specific.

It was not considered necessary to replicate all the searches as they appeared to be sensitive and designed for maximum recall. The ERG undertook update searches for 2012/2013, as the search undertaken by the manufacturer was September 2012 with the submission being received in February 2013. These results were screened by an ERG reviewer and no additional relevant trials were identified.

A bibliographic search of identified references has been undertaken and in-house manufacturer clinical study reports (CSRs) have been used in the submission. Key conferences relevant to the therapeutic area are recorded as having been searched, although the American Society of Haematology (ASH) 2012 conference was reported as not available at the time of the manufacturer's submission. This was searched by the ERG; no relevant abstracts were identified.

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

The MS clearly states the inclusion and exclusion criteria in MS Tables 10 and 11 (p.43-4) of the submission. The criteria deviate from the decision problem with regard to patient population, intervention and comparator. The criteria state that the patient population should include patients with MM, including symptomatic MM (MS Table 10, p.43) but do not stipulate that these should be newly diagnosed, treatment naïve patients eligible for HDT-ASCT. However, they have commented that the patient population was restricted to that stated in the decision problem which is in line with the final scope.

The inclusion criteria state that the intervention may be given as monotherapy and that inclusion was not restricted to the licensed dose (MS Table 10, p.43). These are not in line with the decision problem, the final scope nor the anticipated license which is for bortezomib combination therapy at a specific dose (1.3 mg/m^2) and for a specific number of cycles depending on the combination regimen (MS Appendix 1, draft SPC). The MS specified there was no exclusion on the basis of comparator and that the main comparator was CTD (current standard treatment in the UK) but also included studies that involved induction regimens not containing thalidomide in order to contribute to a mixed treatment comparison (MTC) analysis. This does not reflect the final scope which stipulates combination regimens containing thalidomide as the comparator. The inclusion criteria for outcomes were also broader than the decision problem and final scope in that no outcomes were specified.

Study quality and setting were not stated as inclusion or exclusion criteria, and this reflects the final scope. No limits were placed on the quality of randomised controlled trials (RCTs) and it is stated that RCTs were included regardless of blinding. Non-RCTs were included in the event that an insufficient number of relevant RCTs were found. In the non-RCT inclusion criteria, study design limitations were that non-RCTs reported as conference abstracts with a sample size ≤30, or that did not assess safety or efficacy, were excluded. Retrospective studies, case reports, case series, hospital records/database analyses, pharmacokinetic studies and phase 1 studies were also excluded (the MS reports that these studies are at higher risk of bias compared to other study designs, MS Table 11, p.44), and the ERG agrees that it is reasonable to exclude these studies.

The MS includes a flow diagram that shows the number of publications identified through searches and the number of publications included and excluded at each stage of the review process (MS Figure 3, p.45). Reasons (and corresponding numbers) for excluding studies at the abstract and full publication review stages, are given in the diagram. In the last box in the diagram, the MS reports a total of 53 studies including 15 RCTs. Clarification requested from the manufacturer by the ERG confirmed that the remaining 38 studies were excluded as they were non-RCTs. A list of the non-RCTs identified were reported in an appendix (MS Appendix 6).

The MS excluded non-comparative studies at screening due to issues with bias (see previous comment above). A critical appraisal of the included studies was presented in Section 6.4 of the MS (p.78) and in a summary table (MS Table 23, p.80), with further details in the separate appendices document (MS Appendix 3).

3.1.3 Identified studies

The MS identified 15 RCTs, of which a further six were excluded for not containing bortezomib (MS p.45), along with another four where both treatment arms contained bortezomib (MS Table 15, p.48). Five RCTs were included (MS Table 13, p.47) that the MS states are relevant to the decision problem. However, the ERG note that only two of these (Pethema and Gimema) compare a bortezomib regimen versus a thalidomide regimen as per the decision problem and NICE final scope. For this reason, we have restricted our review of the evidence submitted by the manufacturer to these two trials[1;2] in the main part of the ERG report. For information however, an overview of the other three trials included by the manufacturer (Hovon,[3;4] IFM[5;6] and MRC MMIX[7]) is provided in an appendix (see 9.1) as they provide data for the MTC and economic evaluation.

Table 1: List of trials included in the MS

Trial	Intervention	Comparator
PETHEMA[1;8]	Bortezomib, Thalidomide, Dexamethasone (**VTD**)	Thalidomide, Dexamethasone (**TD**)[a]
GIMEMA[2]	Bortezomib, Thalidomide, Dexamethasone (**VTD**)	Thalidomide, Dexamethasone (**TD**)
Hovon[3;4]	Bortezomib, Doxorubicin (Adriamycin), Dexamethasone (**PAD**)	Vincristine, Doxorubicin (Adriamycin), Dexamethasone (**VAD**)
IFM[5;6]	Bortezomib, Dexamethasone (**VD**)	Vincristine, Doxorubicin (Adriamycin), Dexamethasone (**VAD**)
MRC MMIX[7]	Cyclophosphamide, Thalidomide, Dexamethasone (**CTD**)	Cyclophosphamide, Vincristine, Doxorubicin (Adriamycin), Dexamethasone (**CVAD**)

[a]Comprised a second comparator arm of VBMCP/VBAD/bortezomib (vincristine, BCNU, melphalan, cyclophosphamide, prednisone/vincristine, BCNU, doxorubicin (Adriamycin), dexamethasone/bortezomib) which was not included in the MS.

The included RCT publications[1-3;5;7] and CSRs[4;6;8] were provided electronically by the manufacturer. The MS states that the trials were independently conducted and not sponsored by the manufacturer (MS p.48). However, the Pethema trial publication[1] states that the trial was sponsored by the Spanish Pethema Foundation and that Janssen-Cilag and Pharmion supported the study costs through two grants to Pethema. The Gimema publication[2] states that the trial was sponsored by Seragnoli Institute of Haematology with Janssen-Cilag providing bortezomib free of charge. The MS appears to have included all relevant RCTs. The ERG searches did not identify any other relevant studies.

The MS presents summary details for the included RCTs of trial design, intervention and comparators, treatment regimens, number of patients randomised and randomisation method, outcomes, time points for measurement of response and follow-up (MS Table 16,

p.51-52). Further details of study design and treatments for Pethema and Gimema were provided in MS Figures 6 & 7 (p.55-56). According to the ERG clinical expert, the Pethema treatment pathway most closely resembles that of UK current practice which is 4-6 cycles of induction therapy + ASCT + maintenance. The Gimema trial included a second consecutive ASCT which is not reflective of UK practice (although some patients will have a second ASCT, this will be held back until after relapse). Further information on the population was provided in MS Section 6.3.3 and Table 17 (p.59-62) – the ERG notes that there is less reported in the Pethema trial publication[1] (but is available in the CSR[8]). The number of patients randomised and allocated to each trial arm is shown in flow diagrams in MS Figures 11 & 12 (p.76-77). The number of patients screened for eligibility is only reported for one of the trials (Gimema[2]), although the publication for the Pethema trial[1] states that 4 of the 390 randomised patients were not eligible and thus 386 were randomised (as per the MS). There are some data discrepancies in patient numbers between the Pethema trial publication[1] and the MS Figure 11 (p.76)/CSR.[8] The manufacturer acknowledged in the MS (p.42) and in the response to clarification questions that there are a number of discrepancies between the Pethema trial publication[1] and CSR,[8] and that where there are discrepancies, data was taken from the CSR. Further details on study outcomes are presented in MS Table 20 (p.66). A summary of the statistical methods, sample size/power calculation and data management is presented in MS Table 21 (p.69-73).

The MS presents baseline data for age, percentage male, International Staging System (ISS) stage, performance status, immunoglobulin type & cytogenetics (MS Tables 18 and 19, p.64). Performance status is not reported in either trial publication;[1;2] it is available in the Pethema CSR[8] but the Gimema data reported in the MS cannot be checked. The Pethema publication[1] only reported ISS stage for all patients, not by treatment arm (although this is available in the Pethema CSR[8]). The MS does not comment specifically on baseline characteristics between treatment arms within the trials but both trial publications[1;2] report that treatment groups were well-balanced, though no statistical comparisons were presented. The ERG would agree on the whole, although some differences are noted. In the Pethema trial, patients in the TD arm had a slightly worse Eastern Co-operative Oncology Group (ECOG) performance status than patients in the VTD arm - the TD arm had a higher proportion of patients with an ECOG performance status of 1 (55% TD vs 44% VTD) and a lower proportion of patients with an ECOG performance status of 0 (32% TD vs 44% VTD). The VTD arm also had a higher proportion of patients with immunoglobulin-G type compared to the TD arm (66.2 vs 58.3% in MS but 65 vs 55% in publication).

The MS reports that overall, baseline characteristics of patients are similar across trials (MS p.62), pointing out only 'minor differences' between the MRC MMIX trial and the other four trials (see APPENDIX 1). The ERG would agree on the whole but notes some differences in ISS stage between the Pethema and Gimema trials in that the Pethema trial has a slightly lower proportion of ISS stage I and slightly higher proportion of ISS stage III patients compared to the Gimema trial. In the opinion of the ERG's clinical expert, the trials are fairly representative of UK patients, with the exception of ISS stage. The proportion of patients with ISS stage III in both trials (16-25%) is lower than would be seen in clinical practice where around one third of patients have ISS stage III. The MS does not report baseline ECOG performance status for the Gimema trial (and neither trial publication[1;2] reports this), hence the ERG cannot comment on the similarities/differences for this characteristic. It should also be noted that the Pethema and Gimema trials excluded patients >65 years which is not reflective of UK clinical practice where there is not generally an absolute age exclusion for ASCT.

The MS did not report whether they searched for on-going trials and no specific search was recorded. This was queried with the manufacturer in our questions for clarification and the manufacturer subsequently provided the relevant details of a search on one database (clinicaltrials.gov). The manufacturer reported that eight relevant trials were identified but none reported study results. Searches were undertaken by the ERG within the following on-going trials databases: UKCRN, WHO ICTRP, controlled-trials.gov and controlled-trials.com. Results were checked by an ERG reviewer. No relevant studies were identified.

3.1.4 Description and critique of the approach to validity assessment

The MS provides a summary of the quality assessment of each of the five included trials in Section 6.4 and Table 23 (MS p. 80) with a more detailed assessment in MS Appendix 3. The quality assessment in the MS follows the NICE criteria and is appropriate. The ERG carried out an independent quality assessment of the five trials included in the review. We present the quality assessment for the Pethema and Gimema trials in Table 2 as these trials matched the NICE scope, whilst the assessment for the Hovon, IFM and MRC MMIX trials is presented in Appendix 1 (Section 9.1). As Table 2 shows, the ERG and the manufacturer's quality assessments of the Pethema and Gimema trials agree in part. The ERG assessment differed to that of the manufacturer on the criteria of adequate allocation concealment, group similarity at baseline and whether adequate intention-to-treat (ITT) analyses had been used.

Table 2: Manufacturer and ERG assessment of trial quality

NICE QA criteria for RCTs		Pethema	Gimema
1. Was the method used to generate random allocations adequate?	MS:	Low risk	Low risk
	ERG:	Low risk	Low risk
Comment:			
2. Was the allocation adequately concealed?	MS:	Low risk	High risk
	ERG:	Low risk	Low risk
Comment: Gimema trial – the manufacturer has marked this as 'high risk' as "patients and investigators were not masked to the allocation of treatment" (MS p. 79). However, this question refers to whether the treatment allocation could be foreseen by patients and investigators prior to randomisation, rather than blinding, and the ERG notes that allocation concealment was adequate as the trial used a central, web-based allocation system.			
3. Were the groups similar at the outset of the study in terms of prognostic factors, e.g. severity of disease?	MS:	Low risk	Low risk
	ERG:	High risk	Low risk
Comment: Pethema trial – patients in the TD arm had slightly worse performance scores at baseline than patients in the VTD arm, as measured by the ECOG/WHO score.			
4. Were the care providers, participants and outcome assessors blind to treatment allocation? If any of these people were not blinded, what might be the likely impact on the risk of bias (for each outcome)?	MS:	High risk	High risk
	ERG:	High risk	High risk
Comment: The Pethema trial was open-label. Response rates were assessed locally and then re-assessed centrally by the principal investigator.[1;8] There is some risk of bias in the assessment of response rates and the classification of adverse events. Gimema trial – the MS states that "response assessors were blinded" (MS Table 16, p. 51), but it is unclear from the trial paper if this was the case.[2]			
5. Were there any unexpected imbalances in drop-outs between groups? If so, were they explained or adjusted for?	MS:	Low risk	Low risk
	ERG:	Low risk	Low risk
Comment: Pethema trial – MS p.103 and Table 12 in Appendix 8 states a significantly lower proportion of patients in the VTD arm withdrew compared to the TD arm; reasons for withdrawals were provided. The same trend was seen in the Gimema trial, but the differences were not significant.			
6. Is there any evidence to suggest that the authors measured more outcomes than they reported?	MS:	Low risk	Low risk
	ERG:	Low risk	Low risk
Comment:			
7. Did the analysis include an intention to treat analysis? If so, was this appropriate and were appropriate methods used to account for missing data?	MS:	Low risk	Low risk
	ERG:	Unclear risk	Unclear risk
Comment: ITT analyses were used in the Pethema and Gimema trials, but it is unclear how missing data were imputed for the response rate outcomes. The ITT analysis in the Gimema trial is not strictly an ITT analysis as it includes only patients who received induction therapy, but as the number of patients not included is small (5 and 1 in the VTD and TD arms, respectively), this is unlikely to have affected the results.			

Note: These questions are usually answered with 'yes', 'no' or 'unclear'. However, in the MS the manufacturer has answered these using 'low risk', 'high risk' and 'unclear risk', so the ERG has followed this approach for ease of comparison. 'Low risk' = 'yes' and 'high risk' = 'no' (except for question 6).

3.1.5 Description and critique of manufacturer's outcome selection

Treatment response rate was the primary outcome reported in the Pethema and Gimema trials and is reported in the MS. The MS reports the following types of response rate for both post-induction and post-transplant in each study (post first transplant for Gimema):

- Complete response (CR)
- Near CR (nCR)
- Very Good Partial Response (VGPR) (not for Pethema trial)
- Partial response (PR)
- Progressive disease
- Overall response rate (ORR) defined as CR+nCR+VGPR+PR.

There are discrepancies between the MS, CSR[8] and Pethema publication[1] as to what is the primary outcome. MS Tables 16 & 20 report CR+nCR+PR and CR/nCR as primary outcomes, the CSR[8] (p.37) reports CR, nCR or PR and CR/nCR, whereas the publication[1] just states CR. In response to clarification questions, the manufacturer stated that it was unclear why the trial publication authors had reported the primary outcome differently.

ORR results for the Pethema and Gimema trials were not reported in the trial publications.[1;2] Data presented in the MS (Table 24, p.83) for the Pethema trial do not correspond to the sum of the individual response rates as per the definition of ORR (stated on MS p.67). Clarification requested from the manufacturer stated that ORR was comprised of different response categories in the two trials, defined as CR+nCR+VGPR+PR in the Gimema trial and CR+nCR+PR in the Pethema trial (VGPR was assessed in a post-hoc analysis in the trial paper[1] and thus not reported in the CSR[8]). The ORR results across the two trials therefore cannot be directly compared.

The Pethema trial publication[1] reported CR, VGPR and PR (but the PR data differs from the CSR[8]) whilst data in the MS for nCR and progressive disease were derived from the CSR[8]. Data were only reported for CR and progressive disease for the Gimema trial in the MS (derived from the publication[2]) as other response rates were reported differently (the publication reported VGPR or better and PR or better). The data for nCR, VGPR and PR for the Gimema trial were not reported in the MS but were provided in response to the ERG's questions for clarification.

Secondary outcomes reported in the MS are:

- Progression-free survival (PFS)
- Overall survival (OS)
- Time to progression (TTP)
- Proportion of patients who underwent SCT
- Adverse events (AEs) – reported in the AE section of the MS

The Pethema trial publication[1] reported safety as a secondary outcome but the MS and CSR[8] report safety as 'other endpoint' and not as a secondary outcome. The MS presents data for AE in the post-induction period only, stating that AE post-transplant and AE across the whole treatment protocol were not relevant to the induction therapy under review, nor the decision problem, and were therefore not reported. The ERG would agree with this approach.

The MS report any AE, any grade 3/4 AE, any serious AE and any treatment-related AE for both trials (MS Table 42, p.104), as well as the incidence of the 10 most frequently occurring drug-related grade ≥3 AEs and AEs of special interest to bortezomib-based therapy (MS Table 45, p.106 Pethema trial only). It is not clear why AEs of all grades are not reported. The withdrawal rate during induction treatment is reported in MS Appendix 8.

The outcomes selected by the manufacturer from both trials are appropriate and match the scope/decision problem, with the exception of TTP, which was not specified in the scope. Health-related quality of life (HRQoL) was specified as an outcome in the scope, but the manufacturer has not included this in the MS as the trials did not measure this. The MS reports all relevant outcomes from the trials.

The manufacturer highlights that response rate is a critical endpoint, as evidence shows that patients who achieve a "robust response" (MS p.66), particularly a CR, to treatment have better OS than patients who experience less response. The ERG concurs that this is one prognostic factor, but other factors can also influence prognosis (e.g. age, ISS stage, type of cytogenetic abnormality).[9] The clinical expert consulted by the ERG suggests that although achieving a good post-induction response rate is beneficial to the patient, PFS and OS are more important endpoints as these offer insight into longer-term outcomes post-treatment, which are more meaningful to patients.

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

The MS reports the Pethema and Gimema trial results for all outcome measures relevant to the scope. Response rates are presented as n and %, and the associated p-values are provided for some response outcomes. ORR values reported in the MS do not correspond to the sum of the individual response rates (as noted in Section 3.1.5 above). The proportion of patients who underwent SCT is presented as n and %; no tests of statistically significant

differences were performed. AEs are reported as n, % and RR with 95% CIs but no absolute differences were reported. The incidence of grade ≥3 AEs are only reported as percentages of patients experiencing each event. The MS does not provide p-values nor RR, risk difference or associated 95% CIs statistics for these analyses, so it is not possible to tell whether the differences reported are statistically significant. The manufacturer has reported the number of patients included in each analysis. PFS and OS are reported in median months to the event, with the associated unadjusted hazard ratios (HRs), 95% confidence intervals (CIs) and p-values provided where available. Some PFS and OS data are not reported and stated to be 'not reached'; in response to the ERG's clarification questions the manufacturer stated that data was not available because the duration of follow-up was not long enough to provide the information. The MS states that the length of follow-up used in the studies means that the PFS and OS data presented are currently immature, as patients with newly diagnosed multiple myeloma have a relatively long post-transplant survival rate. It would appear from the Kaplan-Meier curves presented (MS Figures 14 and 16) that median PFS and OS had not been reached at the chosen follow-up points (35.9 and 36 months in the Pethema and Gimema trials, respectively), however the full follow-up period is five years and four years in the Pethema and Gimema trials, respectively, and sufficient data is available for calculation of HRs and p values. The two trial publications do not report OS at the full follow-up period but do report at four[1] and three[2] years respectively.

Comparisons of response rates between trial arms were conducted using the Cochran-Mantel-Haenszel test in the Pethema trial and the Chi^2 test in the Gimema trial. In both trials, PFS and OS were estimated using the Kaplan-Meier method and compared between trials arms using log-rank tests. The MS states that patients with missing data were censored for OS and PFS (MS p.68). The Pethema trial paper[1] states that the trial also censored patients who withdrew from the study due to AE in the induction phase and started on a different treatment. The MS does not report the number of patients who were censored. The CSR[8] for the Pethema trial shows that a high proportion of patients in both the VTD and TD arms were censored: 57.7% and 44.9% respectively in the PFS analysis, and 80.0% and 74.8% respectively in the OS analysis (Table 38 on p.72 and Table 44 on p.76 of the CSR[8]). Given this censoring, the ERG suggests that there is uncertainty about the robustness of the results. The proportion of patients censored in the PFS and OS analyses in the Gimema trial are not reported in the original trial paper.[2]

The MS states (p.68) that efficacy data in the Pethema and Gimema trials were analysed using the ITT population but no definitions of ITT were provided. In response to ERG clarification questions, the manufacturer confirmed that the ITT population included all

randomised patients. The ERG notes, though, that in the Gimema trial these were not strictly ITT analyses (except for the analysis of the proportion of patients who underwent SCT presented in MS Table 28, p.93) as they did not include all randomised patients, but rather only those who received induction treatment. As the number of patients not included in these analyses is small, this is unlikely to have affected the results. In both trials, the safety analyses were based on the 'safety analysis set', which consisted of patients who had received at least one dose of the study drug during induction (MS p.68). The MS reports that all subgroup analyses were pre-specified (MS p.74) but no further methodological details are provided and the publications do not report subgroup analyses in the methods. It should also be noted that some subgroup numbers are small and are likely under powered.

Overall, the manufacturer's approach to the trial statistics is appropriate and reasonably well reported. However, different definitions of ORR between the Pethema and Gimema trials means that results cannot be directly compared and should be interpreted with caution. In addition, the MS did not comment on the high censoring rate in the PFS and OS analyses, and the PFS and OS data should also be interpreted with caution.

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

The MS provides a narrative synthesis of the findings of Pethema and Gimema (the two trials that meet the scope of the appraisal) and also three studies outside the scope (Hovon, IFM, MRC MMIX), one of which does not include bortezomib (MRC MMIX).

A meta-analysis of the four bortezomib-based trials (Pethema, Gimema, Hovon and IFM) is not provided. The MS states that this is because the trials are not comparable in terms of intervention regimens, the variable duration of induction, comparator arms and study design. The ERG agrees with this decision. This also holds for the two studies that meet the scope of the review (Pethema and Gimema).

As no trials comparing bortezomib-based regimens with CTD (the current UK standard) were identified, the MS presents an MTC in order to rank all bortezomib-based regimens and CTD. The MTC is reported in Sections 6.7.3 to 6.7.9 of the MS (MS p.96 to p.102) and was conducted using the guidance outlined in the NICE DSU Technical Support Document 2.[10]

The justification for conducting an MTC is given which is appropriate (i.e. no head-to-head trials of bortezomib-containing regimens against CTD, the regimen most commonly used in

the UK). However, to be included in an MTC, trials are required to be homogeneous enough to allow pooling which is the same assumption as required for a standard pairwise meta-analysis. Therefore there is inconsistency in the MS as no standard pairwise meta-analysis is presented for reasons of heterogeneity between trials. The ERG feels that the similarity assumption for an MTC is not met due to the differences in trial designs and effect modifiers (such as post-induction treatment and follow-up) on the time-to-event outcomes chosen for the MTC. As such the ERG has limited its appraisal of the methodological quality of the MTC here to a checklist (Table 3) and brief summary. The checklist shows that some criteria are not met or partially met. Further assessment of the appropriateness of the methods used and of the results and conclusions presented are provided in Appendix 1.

Table 3: ERG appraisal of MTC approach

Appraisal criteria	Criteria met (YES / NO / UNCLEAR / NOT APPLICABLE)
A. CONCEPTUAL BASIS	
1. Is a justification given for conducting an MTC?	Yes (however, it may not be valid)
B. SYSTEMATIC PROCESSES	
2. Is a comprehensive and transparent search strategy reported?	Yes (though not specifically for MTC)
3. Are inclusion / exclusion criteria adequately reported?	No (no details)
4. Is the number of included /excluded studies from the MTC reported, with reasons for exclusions?	No (no details)
5. Is a visual representation of the data networks provided?	Yes
6. Are the data from included studies extracted and tabulated?	Yes
7. Is the quality of the included studies assessed?	Yes
C. STATISTICAL ANALYSIS	
8. Are the statistical procedures adequately described and executed?	Partial
9. Is there a sufficient discussion of heterogeneity?	No
10. Is the type of model used (i.e. fixed or random effects) reported and justified?	Partial (not fully justified)
11. Was sensitivity analysis conducted?	Partial (by doing a random effects model)
12. Is any of the programming code used in the statistical programme provided (for potential verification?)	Yes
D. PRESENTATION AND INTERPRETATION OF THE EVIDENCE	
13. Is there a tabulation/ illustration of results for each intervention and for each outcome?	No (only 2 outcomes, choice not justified)
14. Is there a narrative commentary on the results?	Partial (very limited)
15. Does the discussion of the results reflect the data presented?	No (no discussion of results)
16. Have the authors commented on how their results compare with other published studies (e.g. MTCs), and offer any explanation for discrepancies?	No
17. Have the authors discussed whether or not there are any differences in effects between the direct and indirect evidence?	N/A

The MTC uses the four bortezomib-based trials and the MRC MMIX trial which are presented in a network diagram (MS Figure 17, p.97). This shows that there is no closed loop of evidence and as such should not strictly be referred to as an MTC.[11] The creation of a network relies on assumptions (specifically that cyclophosphamide, vincristine, doxorubicin and dexamethasone (CVAD) and VAD are clinically equivalent, and TD and CTD are clinically equivalent) rather than direct evidence through any common comparator. The ERG clinical expert agrees with the assumption, acknowledging the absence of randomised data. As stated in the MS, this, combined with the heterogeneity in the trial designs of bortezomib-based regimens, means that the results of the MTC should be treated with 'utmost caution' (MS Section 6.7, p.98). The manufacturer recognises the limitations of the MTC and results are not used to inform the economic model. The ERG considers the MTC is flawed because: (1) the network is not supported by evidence from trials; (2) it may not be meaningful to generalise over the set of included studies as they may not be sufficiently similar. Therefore, the results may not be reliable. In addition, the limited data available in terms of the number of trials and missing outcomes adds to the unreliability of the results. The ERG agrees with the manufacturer's decision not to use the results of the MTC in the economic model.

3.2 Summary statement of manufacturer's approach

The ERG's assessment of the quality of the systematic review included in the MS, based on the CRD criteria,[12] is provided in Table 4.

Table 4: Quality assessment (CRD criteria) of MS 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. The inclusion and exclusion criteria are reported in MS Tables 10 and 11 (MS p.43 and p.44). However, the inclusion criteria are broader than the scope and decision problem in terms of patient population, intervention and comparator (as detailed in Section 3.1.2 of this report). The manufacturer also retrospectively excluded four identified RCTs from the review, but has provided reasons for this.
2. Is there evidence of a substantial effort to search for all relevant research? i.e. all studies identified	Yes. The manufacturer searched all the databases specified by NICE; conference abstracts and the reference lists of studies were included in the review. They also obtained CSRs where available.
3. Is the validity of included studies adequately assessed?	Yes. A quality assessment of each RCT that follows the CRD criteria is provided in MS Section 6.4, Table 23 (p.80) and in Appendix 3 of the MS. Some narrative discussion is provided.
4. Is sufficient detail of the individual studies presented?	Yes. Summary details of the included RCTs are provided in several tables, including methodology, participants and

	approach to statistical analysis.
5. Are the primary studies summarised appropriately?	Uncertain. The RCTs have been summarised in a narrative review and supporting data has been provided for all outcomes. The narrative review is mostly appropriate, but ORR definitions were not consistent between trials and the manufacturer did not report the high censoring rate for the PFS and OS analyses in the Pethema trial. An MTC is presented which is not appropriate due to the assumptions made regarding the evidence network and heterogeneity.

The systematic review is, on the whole, of a good quality according to the CRD criteria,[12] but the ERG has a few concerns about how the results of the included RCTs were summarised and presented.

Publications were screened for inclusion by two reviewers independently at the initial screening (on title and abstract) and full text screening stages, which is considered a desirable approach when conducting systematic reviews.[12] Data were extracted by two reviewers using a data extraction grid, although it is unclear in the MS whether or not they did this independently. It is also unclear whether the quality assessment was performed by one or more reviewers.

As discussed in Section 3.1.2 of this report, the included five RCTs reflect the decision problem as set out in the MS, but only two trials match the NICE scope.

Overall, there is a low chance of systematic error in the systematic review based on the methods used by the manufacturer.

3.3 Summary of submitted evidence

In this section of the report, the ERG provides a summary of the clinical effectiveness evidence reported in the MS. Results are only presented for the Pethema[1] and Gimema[2] trials which compare a bortezomib regimen with a thalidomide regimen, i.e. are relevant to the scope. Results for the IFM,[5] Hovon[3] and MRC MMIX[7] trials are briefly summarised for information in Appendix 1 (Section 9.1). Data have been checked by the ERG against the original Pethema and Gimema trial papers[1;2] and Pethema CSR[8] where possible. Results are summarised for the primary outcome and key secondary outcomes. Some points of clarification were requested from the manufacturer and these are noted where relevant.

Summary of results for response (primary outcome)

Results for the different categories of response are shown in Table 5. It should be noted that patients in the Gimema trial received two consecutive ASCTs compared to one ASCT in the Pethema trial which may have had an impact on the post-transplant response rates and thus makes comparisons between the studies difficult. In addition, ORR was defined differently and comprised of different response categories in the Pethema and Gimema trials, and therefore results cannot be directly compared (see Section 3.1.5 for further details).

ORR post-induction was achieved in a significantly greater number of patients receiving a bortezomib regimen (VTD) compared to a thalidomide regimen (TD) in both the Pethema (84.6% vs 61.4%, p<0.001) and Gimema (93.2% vs 78.6%, p<0.0001) trials. This significant difference in treatment effect on ORR was maintained post-transplant. Similarly, patients receiving bortezomib (VTD) achieved a significantly higher CR post-induction compared to those receiving TD for both the Pethema (35.4% vs 13.4%, p<0.001) and Gimema (18.6% vs 4.6%, p<0.0001) trials, with the significantly favourable effect of the bortezomib regimen (VTD) being maintained in the post-transplant period.

nCR and VGPR post-induction were higher in the VTD arm compared to the TD arm for both trials but the differences were not statistically significant, and there were no differences between treatment arms for these outcomes post-transplant. In contrast, a higher proportion of patients receiving TD achieved a PR post-induction (both trials) and post-transplant (Gimema trial) but there were no significant differences between treatment arms. The Pethema trial publication reports lower PR rates (25% and 33% for VTD and TD respectively) than that reported in the MS (35.4% and 44.1% for VTD and TD respectively) and this was queried with the manufacturer in the ERG questions for clarification. In response, the manufacturer reported that the CSR data reported in the MS are for all patients (ITT analysis) whilst in the trial publication the PR for some patients is missing.

A significantly lower number of patients receiving bortezomib treatment (VTD) experienced disease progression post-induction for both the Pethema and Gimema trials (p=0.0004 and p=0.0005 respectively). The difference was maintained post-transplant for the Gimema trial (p<0.0001[2]) but not in the Pethema trial.

Table 5: Response rates post-induction and post-transplant

Study	Induction treatment	% (n/N)	p value	% (n/N)	p value
		ORR[a] post-induction		**ORR post-transplant**	
Pethema	VTD	84.6 (110/130)	<0.001	77.7 (101/130)	<0.001
	TD	61.4 (78/127)		56.7 (72/127)	
Gimema	VTD	93.2 (220/236)	<0.0001	93.2 (220/236)	0.0025
	TD	78.6 (187/238)		84.5 (201/238)	
		CR post-induction		**CR post-transplant**	
Pethema	VTD	35.4 (46/130)	<0.001	46.9 (61/130)	<0.001
	TD	13.4 (17/127)		23.6 (30/127)	
Gimema	VTD	18.6 (44/236)	<0.0001	37.7 (89/236)	0.0004
	TD	4.6 (11/238)		22.7 (54/238)	
		nCR post-induction		**nCR post-transplant**	
Pethema	VTD	13.8 (18/130)	N/R	8.8 (11/130)	N/R
	TD	3.9 (5/127)		11.0 (14/127)	
Gimema	VTD	12 (29/236)	N/R	14 (34/236)	N/R
	TD	6 (16/238)		8 (20/238)	
		VGPR post-induction		**VGPR post-transplant**	
Pethema	VTD	25 (33/130)	N/R	NR	N/R
	TD	15 (19/127)		NR	
Gimema	VTD	31 (73/236)	N/R	27 (63/236)	N/R
	TD	17 (39/238)		27 (63/238)	
		PR post-induction		**PR post-transplant**	
Pethema	VTD	35.4[b] (46/130)	N/R	22.3 (29/130)	N/R
	TD	44.1[c] (56/127)		22.0 (28/127)	
Gimema	VTD	31 (74/236)	N/R	14 (34/236)	N/R
	TD	51 (121/238)		26 (64/238)	
		PD post-induction		**PD post-transplant**	
Pethema	VTD	6.2 (8/130)	0.0004	1.5 (2/130)	N/R
	TD	23.6 (30/127)		0.8 (1/127)	
Gimema	VTD	0 (0/236)	0.0005	<1 (1/236)	0.0001
	TD	5.0 (12/238)		7 (17/238)	

N/R, not reported.
[a]ORR for Pethema defined as CR+nCR+PR, ORR for Gimema defined as CR+nCR+VGPR+PR.
[b]25% and [c]33% reported in the trial publication.

Summary of results for disease progression and survival

For the longer-term outcomes of PFS, TTP and OS, comparisons between trials are difficult due to the different treatment pathways employed by the trials. The MS states that the consolidation treatment given in the Gimema trial (which is not standard practice in the UK according to the ERG clinical expert) may confound the PFS and OS (MS p.56). The ERG would agree with this and notes that these results should therefore be interpreted with caution. Results reported by the MS are shown in Tables 6-8.

The MS reports PFS in Table 25 (p.85) and in Kaplan-Meier plots (MS Figure 14, p.86-87). The MS notes that results are without SCT censoring and hazard ratios are unadjusted for maintenance therapy. The median follow-up of the trials was similar (35.9 months Pethema and 36 months Gimema). PFS was similar in both trials and was maintained for significantly

longer in the bortezomib (VTD) arm compared to the TD arm (Pethema HR 0.65 95% CI 0.45, 0.92, p=0.015; Gimema HR 0.63 95% CI 0.45, 0.88, p=0.0061).

Table 6: Median PFS (months) and HR of PFS (months)

Study	Induction treatment	N	Median (95% CI)	HR (95% CI; p value)
Pethema	VTD	130	55.5 (31.2, Not reached)	0.65 (0.45, 0.92; p= 0.015)
	TD	127	27.9 (19.8, 34.6)	
Gimema	VTD	236	Not reached	0.63 (0.45, 0.88; p=0.0061)
	TD	238	42 (Not reached, Not reached)	

HR, hazard ratio

TTP was reported in the MS (MS Table 26, p.88 and MS Figure 15, p.89-90) for the Pethema trial only (data derived from the CSR[8]), and hazard ratios are unadjusted for maintenance therapy. There was no statistically significant difference in median TTP (median TTP follow-up of 35.9 months), but there was a statistically significantly lower hazard of progression in bortezomib-treated patients (VTD) compared with the TD arm (HR 0.64 95% CI 0.44, 0.93, p=0.017).

Table 7: Median TTP (months) and HR of TTP (months)

Study	Induction treatment	N	Median (95% CI)	HR (95% CI; p value)
Pethema	VTD	130	Not reached (31.9, Not reached)	0.64 (0.44, 0.93; p= 0.017)
	TD	127	29.0 (23.3, 45.9)	

HR, hazard ratio

The unadjusted OS HR was presented in MS Table 27 (p.91) and MS Figure 16 (p.92) for the Pethema trial only. Median survival was not reached and there was no statistically significant difference in OS between induction treatment arms. The MS reports that the study was not powered to detect a difference in OS and that the trial duration was too short to allow a sufficient difference in OS to be measured (MS p.13 & 91).

Table 8: Overall survival HR of death

Study	Induction treatment	N	Median (95% CI)	HR (95% CI; p value)
Pethema	VTD	130	55.5 (55.5, Not reached)	0.80 (0.48, 1.34; p= 0.393)
	TD	127	Not reached (50.6, Not reached)	

HR, hazard ratio

Summary of results for proportion of people undergoing SCT

The MS reports the proportion of patients who underwent SCT (MS Table 28, p.93) but states that the studies were not powered for this endpoint (MS p.93). From observation of the Pethema trial data, more patients in the VTD arm than the TD arm underwent SCT (80.8% vs 61.4% respectively). However, no statistical tests were reported so it is unclear whether there is a significant difference. The Gimema trial data show that similar proportions of patients in the VTD and TD arms underwent SCT (88.0% vs 82.0% respectively).

Summary of Health-related quality of life

Health-related quality of life was not reported in the MS as this was not measured in the Pethema or Gimema trials.

Summary of sub-group analyses results

Cytogenetic risk subgroup

The MS reported response rates for patients with high and standard cytogenetic risk for the Pethema trial (MS Table 29, p.94-95). The MS reports CR/nCR (data derived from the CSR) whilst the publication[1] reports CR (as well as other response outcomes). In patients with both high risk and standard risk cytogenetics, the CR/nCR rate post-induction and post-transplant was higher in the bortezomib (VTD) arm compared with the TD arm, but no statistical comparison was reported so it is not clear whether these results were statistically significant.

The MS reported PFS, TTP and OS for patients with high and standard cytogenetic risk in MS Table 30 (p.94-95). PFS data were available for the Pethema and Gimema (high risk group only) trials, and TTP and OS for the Pethema trial. There were no statistically significant differences between patients treated with VTD or TD for PFS, TTP or OS, with the exception of the high risk group in the Gimema trial where PFS was significantly longer in the VTD group than in the TD group (HR 0.51 95% CI 0.29, 0.88, p=0.0174).

Other subgroups

The MS states (Section 6.5.3.4, p.93) that subgroup analysis data for response rate (CR/nCR post-induction and post-transplant), PFS, TTP and OS for the subgroups of age, ISS staging and creatinine clearance were provided in Appendix 17; however there were only data for the latter two subgroups, not for age (MS Tables 21 and 22 in Appendix 16). Very minimal data were reported for the Gimema trial (only CR/nCR post-induction and PFS for ISS stage III). For the Pethema trial, there appeared to be a higher CR/nCR response

post-induction and post-transplant in bortezomib-treated patients (VTD) compared to TD patients across subgroups (though differences between groups for ISS stage I post-transplant response were smaller). It should be noted that some of the subgroups were small and no statistical tests were reported. For PFS, TTP and OS, results were inconsistent across subgroups which is in disagreement with the MS which states that 'treatment effects associated with bortezomib-based regimens were consistent across all subgroups' (MS p.93).

Mixed Treatment Comparison results

As stated in Section 3.1.7, due to the limitations and unreliability of the MTC, results are confined to Appendix 1 (Section 9.1).

Summary of adverse events

The MS provides a summary and results table for adverse events (AE) for the 5 included trials (MS Section 6.9, p.102-106). Results for the Pethema and Gimema trials are shown in Table 9 (below), whilst a summary of AE findings for the Hovon, IFM and MRC MMIX trials are available in Appendix 1 (Section 9.1) of this report. The MS presents data for AE in the post-induction period only.

For most AE data, a similar proportion of patients in the VTD and TD treatment groups reported any AE, any grade 3/4 AE, any serious AE and any treatment-related AE across both trials. However, in the Gimema trial, a significantly greater proportion of patients receiving a bortezomib regimen (VTD) experienced any grade 3/4 AE compared to those receiving TD (55.9% vs 33.2% respectively, RR 1.69 95% CI 1.36, 2.08), and in the Pethema trial, a greater proportion of patients in the bortezomib (VTD) arm experienced any treatment-related AE compared to the TD arm (74.6% vs 52.4% respectively, RR 1.42 95% CI 1.17, 1.73).

Table 9: Adverse events

Induction regimen	Pethema			Gimema		
	VTD n (%)	**TD n (%)**	**RR (95% CI)**	**VTD n (%)**	**TD n (%)**	**RR (95% CI)**
Safety population	130	126	-	236	238	-
Any AE	110 (84.62)	102 (80.95)	1.05 (0.93, 1.17)	N/R	N/R	N/R
Any grade 3/4 AE	52 (40)	47 (37.3)	1.07 (0.79, 1.46)	132 (55.93)	79 (33.19)	1.69 (1.36, 2.08)
Any serious AE	34 (26.15)	42 (33.33)	0.78 (0.54, 1.15)	31 (13.14)	30 (12.61)	1.04 (0.65, 1.66)

Any treatment-related AE	97 (74.62)	66 (52.38)	1.42 (1.17, 1.73)	N/R	N/R	N/R

N/R, not reported; RR, relative risk

The MS reports frequently-occurring and treatment-related grade ≥3 AEs for the Pethema trial in MS Table 45 (p.106). Observation of the data shows no apparent differences with two exceptions. A greater proportion of patients treated with bortezomib (VTD) compared to TD experienced peripheral neuropathy (6.2% vs 0 respectively) and pneumonia (7.7% vs 4.0% respectively), although no statistical tests are reported so it is unclear whether this difference is statistically significant. The MS does not present data for the Gimema trial. However, the trial publication[2] reports 8 of the most common grade 3 or 4 AEs reported in at least 2% of patients. A significantly higher proportion of patients receiving VTD compared with TD experienced peripheral neuropathy (10% vs 2%, p=0.0004) and skin rash (10% vs 2%, p=0.0001).

As shown in Table 10, total withdrawals and withdrawals due to disease progression were statistically significantly less in the bortezomib (VTD) arm compared to the TD arm in the Pethema trial (MS Appendix 8 Table 12). The same trend was observed in the Gimema trial but the differences did not reach statistical significance.

Table 10: Withdrawals from treatment during induction

Induction regimen	Pethema			Gimema		
	VTD n (%)	**TD n (%)**	**RR (95% CI)**	**VTD n (%)**	**TD n (%)**	**RR (95% CI)**
ITT N	130	127	-	241	239	-
Total withdrawals	25 (19.23)	48 (37.80)	0.51 (0.34, 0.77)	9 (3.73)	19 (7.95)	0.47 (0.22, 1.02)
Withdrawals due to death	3 (2.31)	6 (4.27)	0.49 (0.12, 1.91)	1 (0.42)	0	N/R
Withdrawals due to AE	8 (6.15)	9 (7.09)	0.87 (0.35, 2.18)	8 (3.32)	7 (2.93)	1.13 (0.42, 3.08)
Withdrawals due to disease progression	13 (10)	28 (22.05)	0.45 (0.25, 0.84)	0	8 (3.35)	N/R

ITT, intention to treat; N/R, not reported or not calculable; RR, relative risk

3.4 Summary

Results of the two RCTs that met the NICE scope (Pethema and Gimema) show that patients with newly-diagnosed MM, eligible for HDT and ASCT, who received bortezomib-

based induction therapy (VTD) had a statistically significantly higher ORR and CR post-induction and post-transplant compared to those receiving TD. The ERG clinical expert considers that CR post-transplant results are clinically meaningful to patients. There were no statistically significant differences in nCR, VGPR or PR for either trial. Disease progression was significantly lower in bortezomib-treated patients post-induction, though this was maintained post-transplant for the Gimema trial only.

For TTP, there was a statistically significantly lower hazard of progression in bortezomib-treated patients (VTD) compared with the TD arm (Pethema trial only), and PFS was maintained for significantly longer in bortezomib-treated patients (VTD) compared to TD (both trials). There were no statistically significant differences in median TTP, median OS (not reached) or OS.

A greater proportion of bortezomib-patients experienced any grade 3/4 AE (Gimema) and any treatment-related AE (Pethema), and also experienced a higher incidence of peripheral neuropathy.

The MS discusses the relevance of the evidence base to UK practice and its limitations. However some concerns/uncertainties include:

- Only two trials met the NICE scope, neither of which were blinded and therefore may be at risk of detection bias (although objective response outcomes minimise the risk).
- In one trial (Pethema) the patients in the bortezomib arm have a better baseline ECOG status, and a higher proportion with IgG type, and it is unclear what impact these may have on results.
- The patients in the Pethema and Gimema trials may not be representative of those in UK clinical practice in terms of ISS stage and age.
- There are uncertainties around the appropriateness of the primary outcome measure in these trials. Response rate is a surrogate outcome and it is not clear how good a predictor of long term outcomes it is; post-transplant response may be better than post-induction response. There is also a need for the whole treatment pathway to be considered in assessing treatment effectiveness.
- ORR is defined differently in the Pethema trial compared to the Gimema trial (and other three trials) making comparisons difficult.
- Long term outcomes (PFS, TPP, OS) may be confounded by consolidation/ maintenance therapy which does not reflect current UK practice, particularly for the Gimema trial (but also for Hovon and MRC MMIX); it is also unclear how two

consecutive ASCTs that patients in the Gimema, Hovon and IFM trials underwent would affect the results.

- There is uncertainty in the results due to the high censoring of data; results were also unadjusted for maintenance therapy.
- MTC results are uncertain (MS p.109) due to the assumption made to develop a network, heterogeneity across the trials and the limited amount of data available.

4 ECONOMIC EVALUATION

4.1 Overview of manufacturer's economic evaluation

The manufacturer's submission to NICE includes:

i) a review of published economic evaluations of the treatment of newly diagnosed MM.
ii) a report of an economic evaluation undertaken for the NICE STA process. The cost-effectiveness of three bortezomib-based regimens is evaluated for patients with newly diagnosed MM in three separate economic models: VTD compared to TD, PAD compared to VAD, and VD compared to VAD.

Here the ERG chiefly considers the VTD vs. TD model as it is the only model which meets the NICE scope for this submission. The PAD vs. VAD and VD vs. VAD models are discussed in more detail in an Appendix to this report (Section 9.2).

Manufacturer's review of published economic evaluations

A systematic search of the literature was conducted by the manufacturer to identify economic evaluations (and burden of illness studies) of treatments of newly diagnosed MM. The ERG critique of the search strategy used in the MS is in Section 3.1.1.

The inclusion and exclusion criteria for the systematic review are listed in Section 7.1.1 of the MS (MS Table 46, p.111). The inclusion criteria state that full economic evaluations, budget impact analyses and resource use studies would be included for treatment with bortezomib, thalidomide, vincristine, cyclosphoshamide and lenalidomide for first line induction therapy prior to SCT for patients with multiple myeloma. Studies were included for the time period from 2000 – November 2012 for full articles only. Only English language studies were included.

From 287 titles and abstracts screened, seventeen potential studies were identified for full paper screening: and 3 studies were included for full review (van Agthoven,[13] Gulbrandsen,[14]

Kouroukis[15]). Fourteen studies were excluded, mainly for the following reasons: the cost of treatment was not specified (n=5), the intervention was not relevant to this submission (n=2), or the study was not specific to patients who received transplant (n=1). The checklist suggested by NICE has been applied to the included references. The MS does not discuss the studies identified. The ERG notes that none of the studies identified are within the NICE scope for this appraisal.

CEA Methods

A cost-effectiveness model was submitted to estimate the cost-effectiveness of VTD vs. TD in patients with newly diagnosed MM. The model adopts a lifetime horizon, with monthly cycles. Costs and outcomes are discounted at 3.5% per annum and the model takes the perspective of the NHS England and Wales.

There are distinct phases of treatment and these are captured by the model, from induction prior to SCT, SCT, and post-SCT, and 2nd and 3rd line treatments. Patients progress to 2nd line treatment after disease progression. Patients are subdivided into groups relating to their response to induction (CR, PR and NR [non-responders]). Patients' progression to death or disease progression is dependent upon their response category.

The principal clinical-effectiveness measures were derived from the Pethema clinical trial[1] for post induction response rates (CR, PR and NR), induction mortality rates, SCT rates, and post induction progression.

Health-related quality of life (HRQoL) was included within the model using data from a study by van Agthoven *et al*.[13] of patients with newly diagnosed and untreated MM which reported patient EQ-5D at different time points for patients receiving SCT or no SCT. The model included a disutility for adverse events associated with induction therapy.

Drug costs were based upon the British National Formulary (BNF),[16] November 2012 edition, and the 2012-13 Chemotherapy Regimens List.[17] The costs relating to stem cell mobilisation, harvest and transplant and other outpatient visits and tests and those associated with adverse events were based upon the NHS reference costs.

The model explores parameter uncertainty in both one-way and probabilistic sensitivity analyses (MS Section 7.7.7 p.192 and MS Section 7.7.8 p.197). Several scenario analyses

are also performed. The MS reports clinical plausibility / external validity of the extrapolated portions of the model against long term survival data (MS p.183).

CEA Results

The results from the economic evaluation are presented in MS Table 93 (p.192) as incremental cost per QALY gained for VTD vs. TD. For the base case, an incremental cost per QALY gained of £24,683 is reported (see Table 11).

Table 11: Base case cost-effectiveness results

Technologies	Total costs (£)	Total LYG	Total QALYs	Incre-mental costs (£)	Incre-mental LYG	Incre-mental QALYs	ICER (£) incre-mental (QALYs)
VTD	£72,815	5.95	4.00	+£23,401	+1.38	+0.95	£24,683
TD	£49,414	4.57	3.06				

The PSA results show that the probability that VTD is a cost effective option over TD at £20,000 and £30,000 willingness-to-pay thresholds is estimated to be 19% and 55% respectively.

The MS states that bortezomib-based regimens offer an important licensed addition to the therapeutic interventions currently on offer, demonstrating higher post-induction response rates than non bortezomib-based regimens.

4.2 Critical appraisal of the manufacturer's submitted economic evaluation

The ERG has considered the methods applied in the economic evaluation in the context of the critical appraisal questions listed in Table 12 below, drawn from common checklists for economic evaluation methods (e.g. Drummond and colleagues[18]).

Table 12: Critical appraisal checklist of economic evaluation

Item	Critical Appraisal	Reviewer Comment
Is there a well-defined question?	Y	MS p.118. 'to evaluate the cost-effectiveness of bortezomib-based regimens for induction of newly diagnosed myeloma compared to alternative induction regimens.
Is there a clear description of alternatives?	Y	VTD vs. TD, PAD vs. VAD, VAD vs. VD.
Has the correct patient group / population of interest been clearly stated?	Y	Patients with newly diagnosed multiple myeloma
Is the correct comparator used?	N	The scope specifies bortezomib in combination with other chemotherapy regimens versus chemotherapy

		regimens containing thalidomide. The analyses PAD vs. VAD and VD vs. VAD are not within the scope because the comparators do not contain thalidomide. The analysis VTD vs. TD is within the scope but is not relevant to UK practice. *(Discussed in Section 4.2.34.2.3)*
Is the study type reasonable?	Y	
Is the perspective of the analysis clearly stated?	Y	
Is the perspective employed appropriate?	Y	NHS England and Wales
Is effectiveness of the intervention established?	N	See comments above for the comparator
Has a lifetime horizon been used for analysis (has a shorter horizon been justified)?	Y	
Are the costs and consequences consistent with the perspective employed?	Y	
Is differential timing considered?	Y	
Is incremental analysis performed?	Y	
Is sensitivity analysis undertaken and presented clearly?	Y	

NICE reference case

The NICE reference case requirements have also been considered for critical appraisal of the submitted economic evaluation in Table 13.

Table 13: NICE reference case requirements

NICE reference case requirements:	Included in submission	Comment
Decision problem: As per the scope developed by NICE	?	Two analyses submitted are outside of the NICE scope
Comparator: Alternative therapies routinely used in the UK NHS	N	Most relevant comparator not considered in the analysis (CTD)
Perspective on costs: NHS and PSS	Y	
Perspective on outcomes: All health effects on individuals	Y	
Type of economic evaluation: Cost-effectiveness analysis	Y	
Synthesis of evidence on outcomes: Based on a systematic review	Y	
Measure of health benefits: QALYs	Y	
Description of health states for QALY calculations: Use of a standardised and validated generic instrument	Y	

Method of preference elicitation for health state values: Choice based method (e.g. TTO, SG, not rating scale)	Y	
Source of preference data: Representative sample of the public	Y	
Discount rate: 3.5% pa for costs and health effects	Y	

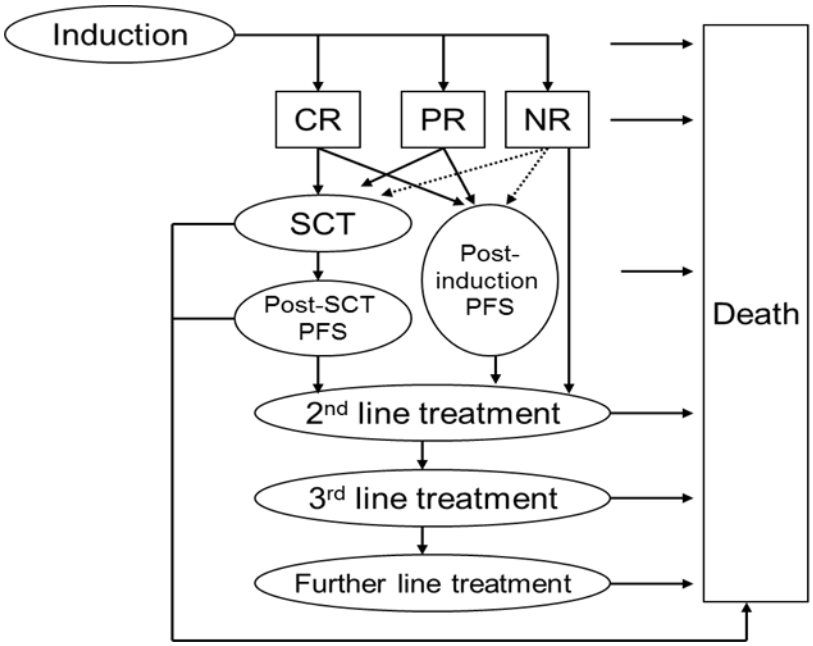
? = uncertain; N/A=not applicable

4.2.1 Modelling approach / Model Structure

A state-transition model was adopted as it allows the clinical pathway of care for transplant-eligible MM patients to be adequately represented. There are distinct phases of treatment and these are captured by the model, from induction prior to SCT, SCT, and post-SCT (MS Section 7.2.3, p.118). The model was developed in Microsoft Excel. A schematic is given in Figure 1 (reproduction of MS Figure 19, p.117).

The MS states that a number of potential model structures was considered but does not state by whom (MS p.118). Model structure and clinical assumptions were discussed at a meeting of the manufacturer's advisory board in October 2012 (MS Appendix 14).

Figure 1: Schematic of the state transition model



Patients enter the model at the start of induction therapy. Post-induction, patients enter one of three health states: complete response (CR), partial response (PR), or non-responders (NR). Some patients may then receive SCT and this is dependent upon their post-induction response (MS p.119). The post-induction response rate also defines the patient's PFS and

OS. Patients move from PFS to second line treatment, then third line treatment, then further line treatment. Patients may move to the death state at any stage. HRQoL varies by treatment state and in some cases also by the time spent in state.

The model has a lifetime horizon of 30 years in the base case. The model cycle length is one month which reflects the length of a course of treatment with VTD (28 days). Key clinical outcomes used by the model are also reported in months (MS Table 49, p.121). A half-cycle correction is not used as the cycle length is short relative to the model time horizon (MS Table 49, p.121).

The model captures the impact of the intervention and differential response to induction therapy with separate health states for CR, PR and NR post-induction, using data from the Pethema trial[1]. Time to progression (TTP) transition probabilities are derived from Pethema trial data[1] for each category of response (CR, PR and NR) and by treatment. Transition probabilities to 3rd and further lines of treatment are derived from the APEX trial data which compared bortezomib monotherapy with high dose dexamethasone in patients with relapsed multiple myeloma.[19] Parameter estimates obtained from median survival by response category in the MRC VII trial [20] are used to obtain OS probabilities by post-induction response.

The MS notes that the demonstration of a significant OS advantage for multiple myeloma interventions is difficult given the long duration of follow-up required, and that drug combination therapies such as VTD have been recommended by clinical experts based upon surrogate markers for OS such as response rates[21;22] (MS p.118). The Pethema trial was not powered to detect a statistically significant difference in OS and median overall survival had not been attained in this trial (MS p.118). Accordingly the MS considers that it is appropriate to use post-induction CR, PR and NR as surrogate markers for PFS and OS.

The MS states that post-induction all patients are assumed to incur the same survival benefit which is dependent only upon the response rate they achieve following the induction phase and is independent of the actual induction regimen that they received (MS p.119). The MS also states that given the data limitations associated with the available trials this is the optimal way to isolate the effect of VTD over TD (MS p.119).

The ERG considers that the structure of the model is consistent with the currently accepted theory of multiple myeloma. The model extrapolates level of response after induction therapy to long term survival and TTP based upon the MRC VII trial. The MRC VII trial is

reasonably old (it recruited patients between 1993 and 2000[20]) and its outcomes may be less good than those which would be achieved in the present day. The ERG clinical expert agrees that response rate at induction predicts PFS and OS. However other surrogate outcomes are available which may offer a better prediction of PFS and OS, for example post-SCT response rate. The ERG clinical expert states that maximum response to treatment (including post-SCT response) is probably the most predictive of long term outcome.

In contrast to the manufacturer's description in the MS, the ERG finds that the model implicitly assumes a continuing effect of induction treatment after induction finishes as separate TTP curves are used for each induction treatment arm. SCT mortality is also applied separately by treatment arm (Section 4.2.4). In addition the ERG finds that, contrary to statements in the MS, the probability of receiving an SCT is not dependent on post-induction response, but only on treatment received (see Section 4.2.4).

The ERG observes that whilst the model has separate states for those who receive an SCT and those who do not, the model attaches no explicit survival benefit to an SCT other than that achieved by delaying the transition to the post-induction/post-SCT PFS state for the duration of the SCT period (three months in the base case). Instead the effect of SCT is captured implicitly: complete responders have a better survival prognosis and this at least partly reflects a tacit assumption that post-induction complete response is associated with higher rates of SCT than partial or non-response. The ERG clinical expert states that SCT offers a survival benefit of 12-18 months compared with no transplant. The ERG considers that subject to data limitations it would have been more transparent to distinguish the separate effects of post-induction response and SCT on survival. Alternatively post-SCT response rate might have been considered for use in the model as this has been found elsewhere to be more significantly associated with OS than post-induction response rate.[23]

Overall the ERG considers that it would have been preferable for the economic model to have been based on OS and TTP Kaplan Meier curves or post-SCT response, rather than post-induction response, as the ERG considers that these would promote better external validity (Section 4.2.8). Several aspects of the economic model structure described in the MS are not implemented in the economic model itself but the overall impact of these differences on model outcomes is unclear.

4.2.2 Patient Group

The patient group included in the MS model is adult patients with previously untreated multiple myeloma, eligible for HDT-SCT. The characteristics of the modelled population are not specified. However as the main trial used for the model outcome was the Pethema[1] trial, the modelled cohort can be assumed to have these patient characteristics (MS Table 18, p.64). Our clinical expert considers that the clinical characteristics of the trial population are representative of clinical practice in the UK, with the exception of ISS Stage. Of the five trials in MS Table 18 (MS p.64), MRC MMIX is likely to be the most representative, especially in terms of age, cytogenic profile and ISS Stage.

4.2.3 Interventions and comparators

Based on the Pethema RCT, bortezomib is administered in combination with thalidomide and dexamethasone (VTD) for 6 cycles of 28 days each vs. thalidomide and dexamethasone (TD) for 6 cycles of 28 days.

The scope for this appraisal, developed by NICE, is for 'bortezomib in combination with other chemotherapy regimens for induction therapy' compared to 'chemotherapy regimens containing thalidomide'. The modelled analyses PAD vs. VAD, and VD vs. VAD are both therefore outside of the NICE scope. VTD and TD are not currently widely used in the UK NHS for first line treatment. The most common treatment for patients with this indication is CTD, and therefore this is the most appropriate comparator for this analysis. Therefore the ERG considers that the modelled intervention and comparator of VTD vs. TD are not wholly relevant to the UK NHS.

4.2.4 Clinical Effectiveness

The following clinical effectiveness parameters are used in the manufacturer's economic evaluation (MS Section 7.3): proportion of patients with post-induction CR, PR or NR; proportion of patients who receive SCT; mortality rate during induction period; mortality rate during transplant period; time to progression; time from 2nd to 3rd line treatment; time from 3rd to further lines of treatment; and overall survival post-induction. These are discussed below in turn.

The proportion of patients with post-induction CR, PR or NR by treatment arm was informed by the Pethema CSR.[4] These data are presented in Table 14 (extract of MS Table 50, p.123) and enter the economic model as baseline risks.

Table 14: Post-induction response rates (Pethema trial)

Trial	Treatment	Comparator
PETHEMA	**VTD** **N=130**	**TD** **N=127**
CR (CR+nCR+VGPR)	64 (49.2%)	22 (17.3%)
PR	46 (35.4%)	56 (44.1%)
NR (MR+SD+PD)	20 (15.4%)	49 (38.6%)

CR, complete response; NR, non-responders; MR, minimal response; PD, progressed disease; PR, partial response; SD, stable disease; VgPR, very good PR;

The MS indicates that the reason for using response rate after induction, rather than response rate after SCT, in the economic model is that it is less prone to confounding with other factors such as comorbidities which can influence the choice of treatment regimen post-induction, and the probability that a patient proceeds to transplant (MS p.119). The ERG considers that incidence of comorbidities and other patient characteristics may be assumed to be balanced between treatment arms in a properly randomised trial, and that on this basis it would be appropriate to use post-SCT response in the economic model.

Post-induction response is a surrogate outcome. Its relationship to the final model outcome, OS, is established using a series of data: TTP data from the Pethema trial; time to 3rd and further lines of treatment data from the APEX trial;[19] and data on OS by post-induction response category from the MRC VII trial.[20] No systematic searches for evidence to link post-induction response to OS are described in the MS. The MS does note a meta-analysis conducted by van de Velde *et al*. (2007)[24] (MS p.132) to assess the association between response and long-term outcomes but gives no justification why other studies included in this paper were not considered or used in the economic model.

The proportion of patients receiving SCT in the model only varies by treatment arm. These proportions are obtained from the Pethema CSR and are given in Table 15 (extract of MS Table 52 p.124).

Table 15: Total SCT proportions by treatment arm (Pethema trial)

	Total SCT
VTD (N=130)	105 (80.8%)
TD (N=127)	78 (61.4%)

Table 16 (adapted from Table 19 in the manufacturer's clarification letter) indicates the proportions receiving SCT by both post-induction response category and treatment in the Pethema trial. It is unclear why these more detailed figures were not applied in the economic model as they show appreciable variation across response categories.

Table 16: SCT rate by post-induction response category (Pethema trial)

Post-induction response categories	Pethema	
	VTD % (n/N)	TD % (n/N)
CR category (CR+nCR)	96.9 (62/64)	95.5 (21/22)
PR category	82.6 (38/46)	89.3 (50/56)
NR category (MR + No change + PD + Death + not evaluable)	25.0 (5/20)	14.3 (7/49)
Total	80.8 (105/130)	61.4 (78/127)

CR: complete response; NR: non responders; MR: minimal response; PD: progressed disease; PR: partial response; SD: stable disease; VgPR: very good PR;

A result of this simplification is that the model makes some unrealistic assumptions, for example that 80.8% of non-responders (NR) received an SCT in the VTD treatment arm, when in fact only 25% of non-responders on VTD treatment received SCT; and similarly that 61.4% of non-responders on TD treatment received SCT, in contrast to the 14.3% observed (Table 16).

The ERG considers that as an SCT has little explicit impact on survival in the model (Section 4.2.1), the effect of this pooling on model outcome is likely to be small. Of greater importance to model outcomes are the survival differences between the post-induction response categories which tacitly reflect different SCT rates (see below).

Mortality rates by treatment arm during the induction phase were taken from the Pethema study and are given in Table 17. Mortality rates by treatment arm during the transplant period were also obtained from the Pethema study (MS Table 51, p.123).

Table 17: Mortality rate during induction period by treatment arm (Pethema trial)

	Mortality rate during induction (6 months)	Monthly probability of death during induction
VTD	3.8% (5/130)	0.7%
TD	4.8% (6/126)	0.8%

TTP is defined as the time from either SCT or the end of induction to the start of second-line therapies. The model implicitly assumes that TTP is affected by the interventions as TTP is modelled using separate parametric survival curves by treatment and response category. In the base case, TTP transition probabilities are derived from exponential curves fitted to the Pethema trial data. Weibull and log-logistic fits are explored by the manufacturer in scenario analyses as alternatives to the exponential fits, although the MS notes that the Weibull and log-logistic parametric fits lack face validity and clinical plausibility (MS p.140-141).

Treatment effects were calculated in parametric regression analyses and are used to modify the baseline TTP transition probabilities. The HRs used are not documented in the MS and are only supplied in the economic model. The HRs for the treatment effect for CR and PR are all non-significant, irrespective of functional form used ($p>0.05$), but HRs for the treatment effect for NR are significant for all functional forms ($p<0.05$).

The parameters of the TTP curves for each distribution are given in MS Table 56 (p.127). The ERG notes that the exponential distribution fitted to CR TTP data for VTD patients results in a shorter median survival time (approximately 61 months) than the exponential distribution fitted to CR TTP data for TD patients (median survival approximately 98 months), and that this contrasts with overall findings for PFS given in the trial publication where median PFS was significantly higher with VTD than with TD[1].

Transition probabilities to 3rd and further lines of treatment are derived from exponential fits to data from the APEX trial which compared bortezomib monotherapy with high dose dexamethasone in patients with relapsed multiple myeloma.[19] The MS states that the APEX trial represents the main trial supportive of the use of bortezomib as second-line therapy in MM patients, which is considered as the standard of care in this line of therapy in the UK (MS p.119). The ERG clinical expert notes that although bortezomib-based chemotherapy is standard of care in this line of therapy in the UK, bortezomib is not used as monotherapy but in a two or three drug combination. Given that the APEX trial concerns bortezomib monotherapy, it may have different survival outcomes to those seen with bortezomib combination therapy.

The APEX trial is reasonably old (conducted from June 2002 to October 2003[19]). However 68% of patients overall in the APEX trial had SCT or other high dose therapy[19] and this is similar to the 71% rate of SCT overall achieved in the Pethema trial (Table 15).

Estimates were obtained from the subgroups of patients in the APEX trial with one and two prior lines of treatment respectively (MS p.126). The same parameters are used for both treatments (i.e. VTD and TD) and response categories (MS Table 57, p.128). It is not possible to vary the choice of exponential distribution in scenario analysis. The MS notes that economic model results are not sensitive to the choice of distribution here (MS p.126).

Data from the MRC VII trial[20] were used to inform OS post-induction as it is the only long term UK-based study which provides mortality probabilities based on post-induction response (MS Section 7.3.8, p.143). The MRC VII trial was not powered to detect a difference in OS by post-induction response category, and no formal statistical tests were carried out on this outcome[20]. The trial is also rather old as it began recruiting patients in October 1993 and stopped recruiting in October 2000,[20] which means patients' survival rates for OS and PFS are likely to be lower than in current clinical practice.

The ERG notes that only 45% of the patients in the MRC VII trial received SCT,[20] in contrast to 71% of patients overall in the Pethema trial (Table 19 from manufacturer's clarification letter). The survival experience seen in the MRC VII trial is thus likely to be somewhat worse than that which has been, and will be, achieved in the Pethema trial, even for TD treatment where 61.4% of patients received SCT (Table 15). With its use of these data the model is likely to underestimate to some extent the survival that can be achieved in the present day. The ERG clinical expert considers that actual survival data will be much better today. A comparison of survival predicted by the economic model and survival observed in the Pethema trial is given in Section 4.2.8 (Figure 2). Two alternative scenarios for OS post-induction are considered by the ERG in Section 4.3.

A further difficulty with the model use of the MRC VII data is that they are not, as the MS states (MS Section 7.3.8, p.143) post-induction response data, but relate to maximal response to treatment.[20] The CR categorisation discussed in MRC VII trial publication thus encompasses not only those who achieved CR post-induction but also those who achieved CR post-SCT, and the resulting survival curves are consequently confounded to some extent with post-SCT response (when this was better than post-induction response).

The median five-year survival times from MRC VII used to calculate the survival probabilities in the economic model are presented in Table 18 (reproduced from MS Table 55, p125). The probabilities are calculated with the assumption that survival times are exponentially distributed. They are only differentiated by post-induction response rate, and not by treatment. Due to limited data availability the only parametric distribution that could be fitted

to the MRC VII data was the exponential distribution (MS p.140) and the MS notes that it was consequently not possible to explore alternative functional forms in scenario analysis.

Table 18: Overall survival by maximal response to treatment category

	5 year survival time			Monthly survival probability	Monthly probability of death
	Number of months	95%CI min	95% CI max		
CR	88.6	61.4	Not reported	99.2%	0.8%
PR	39.8	33.8	61.4	98.3%	1.7%
NR	25.6	7.0	31.3	97.3%	2.7%

Data from MRC VII trial[20]

Health effects of adverse events associated with bortezomib are included in the economic model as disutilities (Section 4.2.5) and have associated costs (Section 4.2.7).

In summary, the economic model makes a series of assumptions to extrapolate the post-induction response seen in the Pethema trial to an OS outcome. Extrapolations based on TTP data from this trial are in some cases counterintuitive. Key data are obtained from two other trials, APEX and MRC VII. However these trials were not conducted recently. A further issue is that the survival data from MRC VII are not categorised by post-induction response as indicated in the MS but by maximal response to treatment.

4.2.5 Patient outcomes

HRQoL changes over time according to the course of the disease, and stage of treatment. The utility values used in the model are shown in MS Table 65 (p.156) and Table 19 of this report.

A systematic search of the literature was conducted to identify publications that identified HRQoL information of relevance to the decision problem. The inclusion criteria for the HRQoL literature review are shown in MS Table 60 (p.146). Studies were included if they reported the utility or QoL of patients diagnosed with MM who underwent SCT as first line, had induction therapy, and used either the EQ-5D, SF-36 or EORTC-QLQ-C30 QoL instruments. Studies were excluded if they did not report results for first line induction therapy prior to SCT in adult patients with MM.

Five relevant studies were identified of which 3 reflected the current UK patient population, and current clinical practice (van Agthoven,[13] Gulbrandsen[14] and Uyl de Groot[25]). Of these

studies, Van Agthoven was considered the best data, because the utility values were obtained using the EQ-5D (using the UK tariff), and the HRQoL values obtained were the most extensive in terms of the frequency of measurement (pre-induction, post-induction and regularly post SCT/no SCT). Utility values from the trial were also reported in the Segeren thesis[26].

The study by van Agthoven *et al*.[13] compared chemotherapy (n=129) versus intensive chemotherapy followed by myeloablative chemotherapy with SCT (n=132) and total body irradiation treatment regimens in patients in the Netherland and Belgium under the age of 65 years with newly diagnosed and untreated MM. Patients received 3-4 cycles of VAD and two cycles of intermediate dose melphalan, where after they were randomised to either receive SCT and interferon maintenance, or interferon maintenance only.

The ERG notes that the van Agthoven *et al*.[13] study is larger than the study by Uyl-de-Groot *et al*.[25] The patient group in this study are largely representative of patients in this appraisal, although they are likely to be younger (age 54 years), are not from the UK[13] (based in Belgium and the Netherlands), and the treatments given in the trial differ from those in the current appraisal. The ERG clinical expert considered that total body irradiation is much more toxic conditioning than high dose melphalan used currently in the UK, and so the utility values from this study may not be representative of current patients.

The HRQoL associated with adverse events of induction therapy were included. A disutility of 0.02 was applied to each patient experiencing an adverse event with an induction therapy. A weighted average was then calculated to derive a disutility for the induction health state (MS Table 65). The aggregated disutility for the induction treatments are 0.007 and 0.005 for VTD and TD respectively.

The ERG considers that the disutility associated with induction therapy is captured in the HRQoL value for the induction period which is lower than for those patients who are no longer on treatment. Furthermore, assigning a similar decrement to all adverse events appears somewhat arbitrary. Nevertheless, the ERG notes that inclusion of the disutility associated with induction therapy has negligible effect on the model results.

Table 19: Summary of utility values for cost-effectiveness analysis

UTILITIES	Utility value	Confidence interval	Reference
1st line treatment			
From start treatment until post-induction response	0.57	0.34-0.78	Segeren [26]
From post-induction to post-SCT response	0.65	0.38-0.88	Segeren [26]
SCT patients	Up to 3 mos =0.59	0.35-0.81	Segeren [26]
	3-6 mos =0.65	0.38-0.88	Van Agthoven et al. [13]
	6-9 mos =0.68	0.39-0.91	Segeren [26]
	9-12 mos =0.62	0.37-0.84	Van Agthoven et al. [13]
	12-18 mos =0.69	0.39-0.92	
	18+ mos =0.75	0.41-0.97	
Non-SCT patients	CR =0.83	0.67-0.94	Beusterien et al.[27]
	PR =0.76	0.64-0.87	
	NR = 0.65	0.56-0.73	
2nd and 3rd line treatments	0.69	0.39-0.92	Van Agthoven et al. [13]
Further lines	0.644	0.38-0.87	
Disutility 1st line treatment	0.02	0.013-0.029	ScHARR HTA report[28]

4.2.6 Resource use

The resource categories included in the model were: drug acquisition and administration, on treatment monitoring, and resource use associated with SCT.

The MS conducted a systematic search to identify cost and resource inputs for the economic model using the same search criteria as for the cost-effectiveness review. Four trials were identified but none of these were used to provide costs input for the economic analysis as the results of the studies were not applicable to the UK.

The treatments of the induction regimens were based upon the Pethema trial[1] using the same dosages and durations of treatment. The dosage for bortezomib was based on the SPC (1.3 mg/m^2).[29] Four injections of bortezomib were administered on days 1, 4, 8 and 11 of each cycle as per SPC. The MS does not discuss assumptions concerned with unused vials. However, the model assumes that each person receives one 3.5mg vial, i.e. that there is no vial sharing. The ERG considers that this is the appropriate approach.

Six cycles were used for induction therapy from the draft SPC (MS Appendix 1), according to the duration in the Pethema trial. The dose for thalidomide was 50 mg daily (on days 1-14) and if well tolerated the dose was increased to 100 mg on days 15-28 and thereafter 200mg daily, as per SPC.[30] Dexamethasone was administered on days 1-4 and days 9-12 of each treatment cycle during cycles 1-2 and on days 1-4 during cycles 3-4. The dosage of dexamethasone was 40 mg.

In addition to the induction treatment, patients also received prophylaxis: herpes zoster, tumor lysis syndrome, anti-infective and gastroprotection (MS Table 68, p.163). Patients receive monitoring and laboratory testing and these are based upon the NICE submission for lenalidomide in the relapse setting[31] (MS Table 9, p.35). The economic model has different levels of monitoring for the induction period, for 2nd and 3rd line treatment and for post treatment period (Worksheets Monitoring 1-4).

The resources used for SCT are shown in MS Table 69 (p.165). The MS states that clinicians provided input on the drugs used for stem cell mobilization (i.e. cyclosphospomide 1.5g/m^2, lenograstim as G-CSF) and ablation (melphalan 200 mg/m^2 for 75% of patients and melphalan 140 mg/m^2 for 25% of patients).

The manufacturer assumes that 80% of patients would receive bortezomib and high dose dexamethasone as 2nd line therapy, 15% would receive CTD and 5% would receive high dose dexamethasone. Furthermore, for third line therapy, 75% of patients receive lenalidomide and high dose dexamethasone, 20% receive CTD and 5% high dose dexamethasone. Dosages and frequency are shown in MS Table 70 (p.166). The MS does not discuss the rationale for the choice of second and third line treatments.

The ERG considers that the 2nd line treatment would differ depending upon the induction treatment chosen. The ERG's clinical expert advised that for those given bortezomib as 1st line, bortezomib would not usually be given again as 2nd line. There is no clear UK consensus on what is given 2nd line, but would most likely include thalidomide or lenalidomide combinations. Conversely those patients receiving a thalidomide regime for their induction therapy would be unlikely to receive a thalidomide regimen for 2nd line therapy. The ERG considers that the MS assumptions around 3rd line therapy are reasonable.

The ERG notes that in the Pethema study patients received maintenance therapy for up to 3 years, or until disease progression, but this was not included in the manufacturer's economic model. The ERG asked for clarification from the manufacturer regarding maintenance therapy. In the manufacturers letter of clarification (p11), the manufacturer stated that maintenance was administered post SCT in the Pethema trial, and this may confound the long term outcomes in the trials such as OS, TTP and PFS. The trial data do not allow disentangling the induction treatment effect from the maintenance effect. They also acknowledge that maintenance could affect the total costs in the model.

4.2.7 Costs

For all treatments, the costs are from the BNF 64 for 2012 and (where appropriate) from the 2012-13 Chemotherapy Regimens List.[17] Administration of chemotherapy drugs, outpatient visits and tests as part of disease and treatment monitoring and the costs relating to SCT were taken from the 2011-12 National Schedule Reference costs.[32] The costs associated with treating adverse events were based upon inpatient outpatient or day case visit National Schedule Reference costs.[32]

The unit costs associated with each of the 1st line induction therapies, drugs, prophylaxis, administration and monitoring is shown in MS Table 68 (p.163), and summarised in Table 20. The average cost of a course of treatment for VTD is £24,840 compared to £8,720 for TD.

Table 20: Unit costs associated with the 1st line induction therapies VTD and TD: drugs, prophylaxis, administration and monitoring

	VTD	TD
Average cost of a course of treatment	£24,840	£8,720
Prophylaxis	£353.54	£298.97
Administration cost	£1,645.00	£828.00
Monitoring cost	£1,050.00	£1,050.00
TOTAL	£28,034	£8,865

The unit costs for SCT are shown in MS Table 69 (p.165). The total cost of SCT is £20,510.72, and this includes the cost for mobilisation (£547.68), harvest (£823), ablation (£451.50), transplant (£17,813) and post-transplant (£875.56).

The unit costs for 2nd and 3rd line treatments are shown in MS Table 70-71 (p.166-7). For 2nd line treatment, the weighted average treatment cost (for Velcade + HDD, CTD, HDD) is £24,440 for a mean duration of 9.8 months. For 3rd line treatment, the weighted average treatment cost (for RD, CTD, HDD) is £34,271.

The ERG has checked the costs used in the model with the referenced sources. All relevant costs have been considered and the manufacturer's approach is reasonable.

4.2.8 Consistency/ Model validation

The MS notes (p.118) that a number of potential model structures were considered in approaching the decision problem but it was felt that alternative approaches lacked both the

face and structural validity of the model which was eventually used. There was a previous economic model for this submission which had major flaws which were discovered shortly before the submission deadline (February 2013), and a new model was constructed within one week (MS Section 7.8.1 p.204). Thus although an earlier model structure and clinical assumptions were discussed at a meeting of the manufacturer's advisory board in October 2012 (MS Appendix 14), this discussion did not explicitly relate to the final model contained in the submission.

Internal consistency

The MS does not report if checklists were used for internal validation.

The ERG tested the predictive validity of the model by carrying out a number of sensitivity analyses to ensure model outputs varied in the expected direction. Results from this checking were all satisfactory.

External consistency

The MS notes in Section 7.8.1 (MS p.204) that due to the tight timescale for model construction the external validation of the model was to be completed in the weeks following the NICE submission. Elsewhere (MS Section 7.7.1, p.183) the MS describes a search of the literature to obtain long-term survival data for MM patients eligible for single SCT. These data are presented and compared to VTD vs. TD model results in MS Table 86 for two prospective trials (IFM90 and MRC VII) and one set of registry data (MS p.183). The MS states that the OS estimates calculated by the model are consistent with long-term OS data from the prospective trials (MS p.184) but that as the registry data only included patients that actually received a transplant these data overestimate survival.

The ERG considers that the manufacturer's conclusions relating to VTD vs. TD model validity against the prospective trials are appropriate, but that the prospective trials may not be the best comparators to use in this circumstance. Both of the sets of validation data[20;33] were used to populate the model to some extent and it would be surprising if the model did not show agreement with them. The trials are also rather old (MRC VII recruited between 1993 and 2000; IFM90 recruited between 1990 and 1993) and the good model agreement suggests that the model is underestimating the survival that may be achieved in the present day.

The model overestimates expected survival at 9 years for complete responders compared to the registry data (43% vs. 35%, MS Table 86, p.183), which is inconsistent with the

manufacturer's argument that the registry data overestimate survival as they only include SCT recipients (MS p.184). The model markedly underestimates 9-year survival of PR and NR patients compared to the registry data (15% vs 35% and 5% vs 23% respectively). Given that these differences are large, and inconsistent in direction for CR vs. PR and NR, the ERG suggests that to some extent they indicate poor external validity of the model as well as overestimation of survival in registry data.

The ERG considers that the OS data from the Pethema trial provide an appropriate contemporary validation dataset for the VTD vs. TD model. These data are not used to derive the OS estimates in the model and so they are a reasonably independent means of verification. Furthermore in-trial maintenance does not confound PFS or OS in Pethema as patients were re-randomised to maintenance treatment post-transplantation (MS p.119).

Given that data from the Pethema trial were used to inform post-induction response rates, the model OS curve should reflect the OS seen in the trial to some extent if the manufacturer has extrapolated post-induction response to OS correctly, and if post-induction response is a reasonable surrogate outcome for OS as is assumed. Accordingly the ERG has digitised the OS Kaplan-Meier curves presented in MS Figure 16C for the VTD and TD arms of the Pethema trial (MS p.92) and plotted these against OS predicted by the model for these treatment arms (Figure 2). Figure 2 indicates that OS predicted by the model is initially a reasonable fit to OS observed in the Pethema trial although survival for TD is somewhat overestimated. However after about one year the model values diverge from the observed values and thereafter the model consistently underestimates OS for both treatment arms. The underestimation of survival is worse for the TD treatment arm than the VTD treatment arm. This is shown more clearly in Figure 3 which plots the difference between OS observed in the trial and OS predicted by the model for the VTD and TD treatment arms.

Figure 2: Comparison of overall survival predicted by model and overall survival observed in Pethema trial, by treatment arm

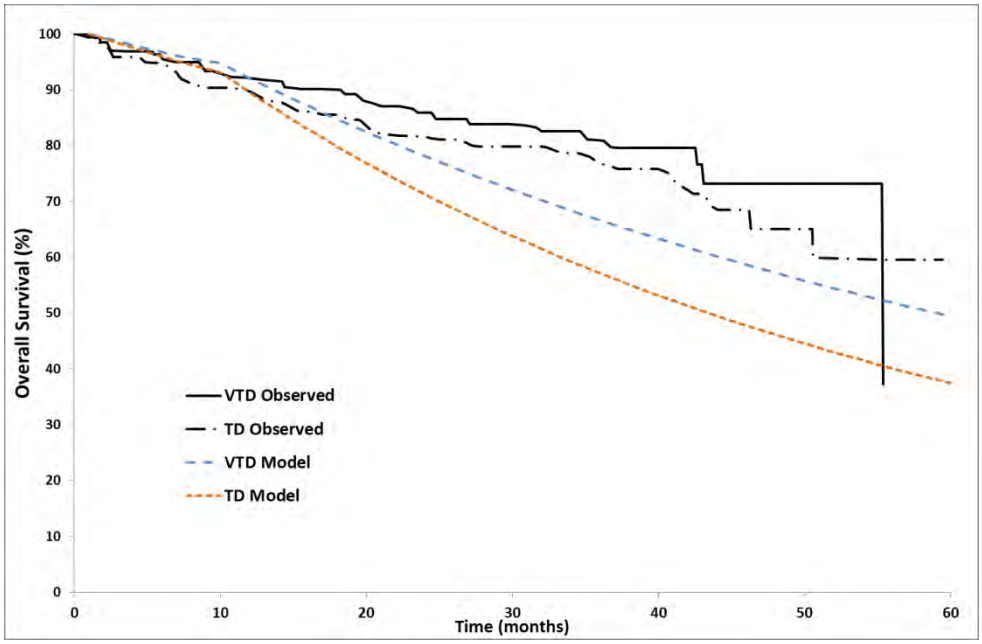
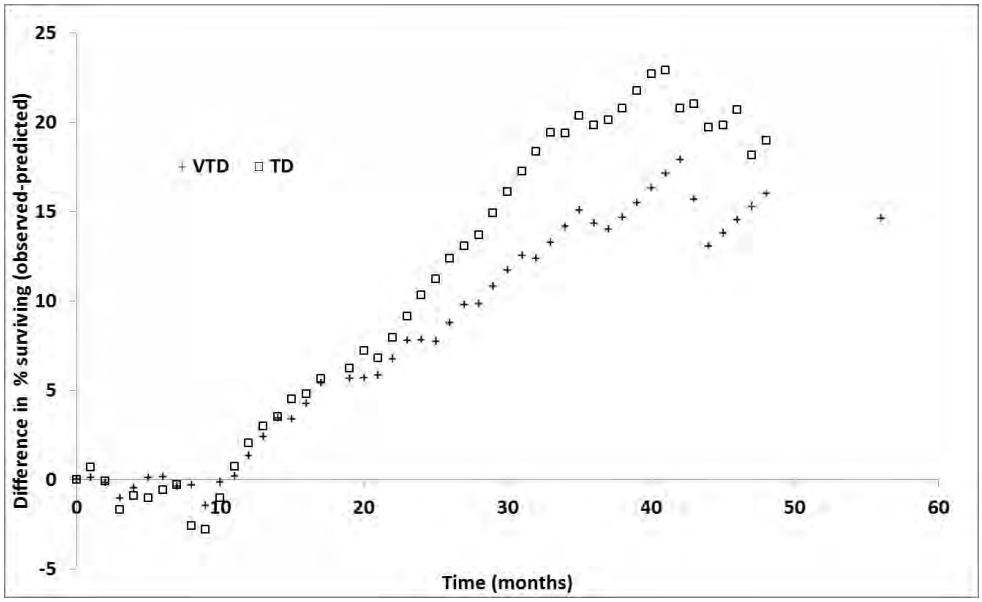


Figure 3: Difference between OS observed in the Pethema trial and OS predicted by the model for the VTD and TD treatment arms.



Uses data shown in Figure 2.

In summary the ERG does not consider that the manufacturer has provided satisfactory proof of the external validity of the model. Comparison of model OS with observed Pethema trial OS reveals that the model consistently underestimates OS, and that this underestimation is worse for TD than it is for VTD. Thus, in addition to external validity issues, the model also appears to be systematically biased in favour of VTD.

4.2.9 Assessment of Uncertainty

The manufacturer has addressed model methodological uncertainties by running alternative versions of the model with different assumptions. Discount rates are varied for costs and outcomes and alternative time horizons are examined. There is, however, no evidence that structural uncertainties have been addressed via sensitivity analysis. An economic analysis based upon subgroups was not carried out (MS p.205).

Deterministic sensitivity analysis, scenario analysis and PSA are all reported. Some scenario analyses use alternative published sources for key parameter values. The MS notes that sensitivity analyses were conducted to test several extreme scenarios using upper or lower 95% confidence interval limits for each of the post-induction response rate categories (MS p.171).

One-way sensitivity analyses

Some variables subject to one-way sensitivity analysis are given in MS Table 84 (p.179). These are: induction costs; SCT costs; 2nd and 3rd line costs; further line costs; end of life costs; AE-related costs during induction therapy; utility from start of treatment to post-induction assessment; utility from post-induction assessment to post-SCT response; utility over time by SCT/no SCT; utility for 2nd, 3rd and further lines of treatment; and AE related disutilities.

Other variables were also explored in one-way sensitivity analysis as shown in MS Figure 25 (p.194) and take distributions as noted in MS Table 85 (p.181) which are used to arrive at upper and lower 95% confidence intervals. The parameters of these distributions are not noted in MS Table 85, but are given in the economic model.

The percentage (+/-20%) by which a parameter is varied from the base-case analysis is clearly stated in MS Table 84 but the resulting upper and lower bounds are not supplied. 95% confidence intervals are provided in MS Table 84 for parameters varied over this range.

Results of the one-way sensitivity analyses indicate that the ICER is most sensitive to post-induction CR mortality and drug costs. A CR post-induction mortality of 1.1% per month (0.8% in base case) is associated with an ICER of £36,074. The high variation of VTD drug costs is associated with an ICER of £30,356. For all other considered parameters the ICER lies within the £20,00-£29,000 QALY range (MS Section 7.7.10, p.203).

Scenario Analysis

MS Tables 75 and 83 (p.172 and p.177) provide details of the scenario analyses undertaken and Table 95 (MS p.200) gives the results for these analyses. The following variables were included in scenario analyses: post-induction response rates; OS by post-induction response; TTP; number of cycles with VTD and TD; health state utility values; AE-related disutility; time horizon; discount rates. Justification for the selection of these variables and the alternative values examined is given in MS Section 7.6.1 (p.171-179).

The ERG considers that the selected variables are appropriate but that the alternative values examined may not fully test the uncertainty in the model. For example in the case of OS by post-induction response it might have been preferable to identify a more recent dataset to provide alternative values, rather than the IFM90 trial which first enrolled patients in 1990.[34] The ERG further explores uncertainty in OS by post-induction response rate in scenario analyses which are described in Section 4.3.

Results are presented for 24 scenarios in MS Table 95 (p.200). The ERG was not able to reproduce the exact results given in MS Table 95 for 5 of the 24 scenarios in the VTD vs. TD model but the differences in final ICER values were not substantial. ICERs generally remain below or close to £30,000/QALY with the exception of the following scenarios:

- 10 year time horizon (ICER £39,304/QALY)
- 2 VTD response rate variation scenarios (with CR<41%) which had ICERs of £41,226 and £39,272
- 2 TD response rate scenarios (with CR>24%) which had ICERs of £39,742 and £51,990)

The manufacturer concludes that the scenario analyses support the cost-effectiveness of bortezomib-based induction regimens as the ICERs generally remain below £30,000/QALY. Where they do not the manufacturer observes that quite extreme values were used i.e. those at either end of a 95% confidence interval (MS p.203).

The ERG considers that alternative values used by the manufacturer in scenario analysis may not fully explore the uncertainty in the model and may therefore not fully reflect the uncertainty in final ICER. Two alternative scenarios for OS by post-induction response are explored by the ERG in scenario analysis described in Section 4.3.

Probabilistic Sensitivity Analysis

The PSA uses 10,000 iterations and runs in approximately 6 minutes. The MS does not supply the final mean cost and final mean QALYs associated with the PSA runs but MS Table 96 gives the probabilities that VTD is cost-effective against TD at the £20,000 and £30,000 willingness-to-pay thresholds as 18.58% and 54.83% respectively (MS p.204). The manufacturer concludes that VTD is likely to be a cost-effective treatment option for the relevant patient population compared to TD (MS p.204).

Variables included in the PSA are reported in MS Table 85 (p.181). Base case values and assumed variability for some variables included in PSA are given in MS Table 84 (p.179). Assumed distributions are given in MS Table 85 (p.181). Parameters for these distributions are not provided in the MS but they are supplied in the economic model. MS Table 85 suggests that SCT rates depend upon post-induction response rate in the economic model, but they do not.

The ERG considers that the probability distributions are correctly applied and the methods of assessment of parameter uncertainty are appropriate. However parameter correlation is not addressed and this is a particular problem for the key clinical effectiveness measure, post-induction response rate. The CR, PR and NR proportions are drawn from independent Beta distributions and consequently the sum of transition probabilities may be more or less than 1 in PSA.

The ERG notes that the probabilistic and deterministic sensitivity analysis results are consistent. However although bortezomib is cost-effective at a willingness-to-pay (WTP) of £30,000/QALY in the great majority of deterministic analyses, the overall probability that VTD is cost-effective compared to TD at a WTP of £30,000/QALY is only 54.8%: there is a high probability that VTD is not cost effective when uncertainty in multiple parameters is considered together.

4.2.10 Comment on validity of results with reference to methodology used

The structure adopted for the economic evaluation reflects the clinical pathway for multiple myeloma. However, basing OS and TTP on the surrogate outcomes of treatment response has not been validated appropriately. Comparison of model OS with observed Pethema trial OS reveals that the model consistently underestimates OS, and this underestimation is biased in favour of VTD.

The patient population used in the model is from the relevant trial, but the treatments included are not representative of those used in secondary care in the UK. The most relevant comparator for the UK is CTD but this has not been included in the economic evaluation. The MS includes three separate pairwise analyses (VTD vs. TD; PAD vs. VAD; VAD vs. VAD), comparing several different treatments and these are not compared with each other. The MS conducted an MTC but did not use these analyses in the economic evaluation, as they noted that there was considerable uncertainty underlying the MTC and there was relative immaturity of the OS data from the pivotal trials.

4.3 Additional work undertaken by the ERG

The ERG has conducted the following scenario analyses:

a) Comparing all treatment analyses
b) Two alternative scenarios for post-induction mortality (VTD vs. TD model)
c) Post-SCT response rate from Pethema trial used instead of post-induction response rate (VTD vs. TD model)

a) Comparing all treatment analyses

The MS provides three pairwise analyses for bortezomib induction treatment: VTD vs TD; PAD vs VAD; VD vs. VAD, according to the trial evidence. As noted elsewhere, heterogeneity between the key trials makes indirect comparison between treatments very difficult (Section 3.1.7). However in order to draw together outputs from the three economic models, and to begin to isolate the cost-effectiveness of bortezomib-based regimens compared to thalidomide-based regimens, we have compared the model results of all treatments containing bortezomib but not thalidomide (PAD and VD) to the treatment containing thalidomide but not bortezomib (TD) (Table 21). Table 21 simply takes the relevant economic model outputs from MS Table 3 (MS p.15) and calculates the ICERs for VD and PAD compared to treatment with TD. These results should be treated with utmost caution as they compare individual arms of separate trials, without adjusting for trial populations, and are presented for information purposes only.

Table 21: Base case cost-effectiveness results versus TD

Treatment option	Costs	QALYs	Incremental Costs	Incremental QALYs	ICER (£/QALY)
TD	£49,414	3.06	-	-	-
VD	£62,874	3.79	£13,460	0.73	£18,318
PAD	£59,632	3.84	£10,218	0.78	£13,026

In addition, we show illustrative results for VD and PAD vs. CTD. CTD is a more relevant comparator than TD in a UK population. We have derived the cost and QALY estimates for CTD by applying the response rates achieved in the MRC MMIX trial to the TD arm of the TD vs. VTD model, and added in the additional costs for cyclophosphamide. Table 22 shows that CTD dominates VD and PAD, i.e. it is cheaper and more effective. This table is subject to the same limitations as Table 21, i.e. it compares heterogeneous trials, and consequently should also be treated with caution.

Table 22: Base case cost-effectiveness results versus CTD

Treatment option	Costs	QALYs	Incremental Costs	Incremental QALYs	ICER (£/QALY)
CTD	£48,237	3.90	-	-	-
VD	£62,874	3.79	£14,637	-0.11	Dominated
PAD	£59,632	3.84	£11,396	-0.06	Dominated

b) Two alternative scenarios for post-induction mortality (VTD vs. TD model)

The ERG considers that the MRC VII trial[20] is reasonably old and that the survival of patients in this trial may be poorer than would be achieved by similar patients today. The manufacturer examines uncertainty around survival probabilities by response category in sensitivity and scenario analyses but the ERG does not consider that the uncertainty is fully explored. In particular the manufacturer's scenario analysis uses data from the IFM90 trial which is older than the MRC VII trial and furthermore shows that no patient with less than partial response was alive at five years post-transplant. The model ICER is shown in the MS to be reasonably sensitive to variation in NR mortality (MS Figure 25, p.194) and the ERG is interested in the effect of longer survival for non-responders on model outcomes.

The ERG used data obtained from the meta-analysis of van de Velde et al (2007)[24] to inform two further scenario analyses for the VTD vs. TD model. These data were from the NMSG 5/94 study (van de Velde *et al*. Table 1) and a study by Alvares *et al*. (van de Velde *et al*. Table 2). The NMSG 5/94 study was a prospective study in Denmark, Norway and Sweden with 247 patients which recruited between 1994-1997 and is therefore more recent than IFM90.[5] The Alvares *et al.* study had a retrospective design and relates to 383 patients in England diagnosed with MM between 1985 and 2004.[35] Results for these alternative scenarios are given in Table 23.

Table 23: ERG analysis of changes to median overall survival (in months) by post-induction response category, VTD vs. TD model

Scenario	Treatment	Total costs, £	Total QALYs	ICER (£/QALY gained)
Base case MRC VII CR 88.6 mos. PR 39.8 mos. NR 25.6 mos.	TD	49,414	3.06	-
	VTD	72,815	4.00	-
	Incremental	23,401	0.95	24,683
NMSG 5/94 CR 71 mos. PR 64 mos. NR 64 mos.	TD	55,529	4.21	-
	VTD	75,552	4.39	-
	Incremental	20,023	0.18	110,727
Alvares et al CR 88.8 mos. PR 63.6 mos. NR 49.2 mos.	TD	55,076	4.07	-
	VTD	76,605	4.62	-
	Incremental	21,529	0.56	38,750

The NMSG 5/94 study shows less difference in median survival between the response categories than is seen in the base case MRC VII data. This leads to a much higher ICER than the VTD vs. TD base case, of £110,727 per QALY gained. The Alvares *et al*. study[35] has median OS for complete responders which is similar to the MRC VII study (88.8 months compared to 88.6 months respectively). However overall median survival for partial and non-responders in this study is much better than MRC VII and this leads to an increase in the VTD vs. TD ICER to £38,750 per QALY gained. The ERG considers that the Alvares study data provide a better fit to the Pethema OS data than either the NMSG or MRC VII data.

c) Post-SCT (maximal) response rate from Pethema trial (VTD vs. TD model)

In the VTD vs. TD model, post-induction response rates from the Pethema trial are extrapolated to OS using data from the MRC VII trial. However the ERG observes that the MRC VII trial survival data are categorised by maximal response to treatment, which is arguably more similar to post-SCT response than post-induction response, and so use of post-SCT response from Pethema would provide a more consistent fit to MRC VII data. Post-SCT response also appears to have more significant associations with OS, and be more predictive of OS, than post-induction response.[24] For these reasons the ERG conducted a scenario analysis using post-SCT response rates from the Pethema trial in the VTD vs. TD model, instead of post-induction response rates. Post-SCT response was a stated primary outcome of the Pethema trial.[1]

The Pethema CSR[8] notes that since approximately 20% fewer patients in the TD group continued on to receive an SCT transplant, a higher percentage were inevaluable or had

unknown response outcomes post-SCT (TD - 40.2%; VTD - 19.2%).[8] This group is incorporated in the non-response category in the economic model. Table 24 compares the percentages in each response category post-induction and post-SCT for the two treatment arms.[8]

Table 24: Post-induction and post-SCT response achieved in Pethema trial, by treatment arm

	Post-induction response %	Post-SCT response %
TD		
CR	17.3	34.6
PR	44.1	22.0
NR	38.6	43.4
VTD		
CR	49.2	55.4
PR	35.4	22.3
NR	15.4	22.3

Results for the post-SCT response rate scenario in the VTD vs. TD model are given in Table 25. An ICER of £35,915 per QALY gained is achieved. The ERG observes that the attribution of patients with inevaluable or unknown outcome after SCT to the NR category is a non-conservative assumption and that if some of these patients achieved PR or better the ICER would be higher than £35,915/QALY, i.e. VTD would become less cost-effective compared to TD.

Table 25: ERG scenario analysis using post-SCT response rates, VTD vs. TD model

Scenario	Treatment	Total costs, £	Total QALYs	ICER (£/QALY gained)
Base case	TD	49,414	3.06	-
Pethema post-induction response rates	VTD	72,815	4.00	-
	Incremental	23,401	0.95	24,683
Pethema post-SCT response rates	TD	50,378	3.43	-
	VTD	73,716	4.08	-
	Incremental	23,338	0.65	35,915

4.4 Summary of uncertainties and issues

Of the three analyses submitted, two analyses do not meet the NICE scope (PAD vs. VAD, and VD vs. VAD). Furthermore, the VTD vs. TD analysis is not wholly relevant to UK practice as TD is not currently routinely used in the NHS. A more appropriate comparator would be

CTD, which is routinely used in the UK, but this has not been included in the MS economic analysis.

The estimation of long term survival and progression free survival is based upon surrogate outcomes for post-induction response (CR, PR, NR). However, there is not a good fit between post-induction response and OS and time to progression compared to estimates from the PETHEMA trial, and the results presented are systematically biased in favour of the intervention.

5 End of life

NICE end of life treatment criteria were not applicable and not included in the MS.

6 Innovation

The manufacturer did not consider the treatment to be innovative and this was not included in the MS.

7 DISCUSSION

7.1 Summary of clinical effectiveness issues

The MS includes evidence on the clinical effectiveness of bortezomib for induction therapy in multiple myeloma before high dose chemotherapy and autologous SCTation, though only two of the five included trials are relevant to the NICE scope. Results presented in the MS suggest that VTD is superior to TD for ORR, CR and PFS. No differences were found between treatments for OS so it is unclear how well surrogate short-term response outcomes correlate with long-term survival. Other issues around the long-term outcomes, such as high censoring of data and confounding of post-induction consolidation/maintenance treatments, raise concerns over the reliability of the data.

7.2 Summary of cost-effectiveness issues

The MS includes evidence on the cost-effectiveness of bortezomib-based induction regimens through the submission of three analyses: VTD compared to TD, PAD compared to VAD, and VD compared to VAD. The MS considers the analysis of VTD vs. TD to be the most relevant because it contains a comparator relevant to the scope, i.e. a thalidomide-based regimen. The other two analyses do not meet the NICE scope. Furthermore, the analysis included is not wholly relevant to UK practice as TD is not currently routinely used in

the NHS. A more appropriate comparator would be CTD, which is routinely used in the UK, but this has not been included in the MS economic analysis.

State transition models for each of the analyses were developed with a similar structure. The model structure is consistent with the clinical pathway of care for multiple myeloma, including the distinct phases of treatment for induction, SCT, and subsequent lines of treatment after disease progression. The estimation of long term survival and progression free survival is based upon surrogate outcomes for post-induction response (CR, PR, NR). However, there is not a good fit between post-induction response and OS and time to progression compared to estimates from the PETHEMA trial, and the results presented are systematically biased in favour of the intervention.

The model results suggest that a bortezomib-based therapy is a cost effective option for a willingness to pay threshold of £30,000 per QALY, The results should be treated with caution, due to the issues noted above.

8 REFERENCES

1. Rosinol L, Oriol A, Teruel AI, Hernandez D, Lopez-Jimenez J, de la Rubia J *et al*. Superiority of bortezomib, thalidomide, and dexamethasone (VTD) as induction pretransplantation therapy in multiple myeloma: a randomized phase 3 PETHEMA/GEM study. *Blood* 2012;**120**:1589-96.

2. Cavo M, Tacchetti P, Patriarca F, Petrucci MT, Pantani L, Galli M *et al*. Bortezomib with thalidomide plus dexamethasone compared with thalidomide plus dexamethasone as induction therapy before, and consolidation therapy after, double autologous stem-cell transplantation in newly diagnosed multiple myeloma: a randomised phase 3 study. *Lancet* 2010;**376**:2075-85.

3. Sonneveld P, Schmidt-Wolf IG, van der Holt B, El JL, Bertsch U, Salwender H *et al*. Bortezomib induction and maintenance treatment in patients with newly diagnosed multiple myeloma: results of the randomized phase III HOVON-65/ GMMG-HD4 trial. *Journal of Clinical Oncology* 2012;**30**:2946-55.

4. Janssen-Cilag Ltd. Clinical Study Report (HOVON): A randomized phase 3 study on the effect of bortezomib combined with adriamycin, dexamethasone (AD) for induction treatment, followed by high-dose melphalan, and bortezomib alone during maintenance in patients with multiple myeloma; 2012.

5. Harousseau JL, Attal M, Avet-Loiseau H, Marit G, Caillot D, Mohty M *et al*. Bortezomib plus dexamethasone is superior to vincristine plus doxorubicin plus dexamethasone as induction treatment prior to autologous stem-cell transplantation in newly diagnosed multiple myeloma: results of the IFM 2005-01 phase III trial. *Journal of Clinical Oncology* 2010;**28**:4621-9.

6. Janssen-Cilag Ltd. Clinical Study Report (IFM): Open-label, multicentre, randomised phase 3 study to compare the combination of Velcade and dexamethasone with VAD-type chemotherapy in the treatment of patients up to the age of 65 with newly diagnosed multiple myeloma; 2012.

7. Morgan GJ, Davies FE, Gregory WM, Bell SE, Szubert AJ, Navarro Coy N *et al*. Cyclophosphamide, thalidomide, and dexamethasone as induction therapy for newly diagnosed multiple myeloma patients destined for autologous stem-cell transplantation: MRC Myeloma IX randomized trial results. *Haematologica* 2012;**97**:442-50.

8. Janssen-Cilag Ltd. Clinical Study Report (PETHEMA): A phase III national, open label, multicenter, randomized, comparative study of VBMCP-VBAD/Velcade versus thalidomide/dexamethasone versus Velcade/thalidomide/dexamethasone as induction therapy, followed by high-dose chemotherapy with autologous hematopoietic transplantation and subsequent maintenance treatment with interferon alpha-2b versus thalidomide versus thalidomide/Velcade in patients with multiple myeloma; 2012.

9. Picot J, Cooper K, Bryant J, Clegg AJ. The clinical effectiveness and cost-effectiveness of bortezomib and thalidomide in combination regimens with an alkylating agent and a corticosteroid for the first-line treatment of multiple myeloma: a systematic review and economic evaluation. *Health Technology Assessment* 2011;**15**.

10. Dias S, Welton NJ, Sutton AJ and Ades AE. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials; 2011.

11. Jansen J, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N *et al*. Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1. *Value in Health* 2011;**14**:417-28.

12. Centre for Reviews and Dissemination. *Systematic reviews: CRD's guidance for undertaking reviews in health care* (3rd edition). York Publishing Services Ltd.: CRD; 2009.

13. van Agthoven M, Segeren CM, Buijt I, Uyl-de Groot CA, van der Holt B, Lokhorst HM *et al*. A cost-utility analysis comparing intensive chemotherapy alone to intensive chemotherapy followed by myeloablative chemotherapy with autologous stem-cell rescue in newly diagnosed patients with stage II/III multiple myeloma; a prospective randomised phase III study. *European Journal of Cancer* 2004;**40**:1159-69.

14. Gulbrandsen N, Wisloff F, Nord E, Lenhoff S, Hjorth M, Westin J. Cost-utility analysis of high-dose melphalan with autologous blood stem cell support vs. melphalan plus prednisone in patients younger than 60 years with multiple myeloma. *European Journal of Haematoology* 2001;**66**:328-36.

15. Kouroukis CT, O'Brien BJ, Benger A, Marcellus D, Foley R, Garner J *et al*. Cost-effectiveness of a transplantation strategy compared to melphalan and prednisone in younger patients with multiple myeloma. *Leukaemia and Lymphoma* 2003;**44**:29-37.

16. Joint Formulary Committee. British National Formulary 64. London: British Medical Association and Royal Pharmaceutical Society of Great Britain; 2012.

17. Department of Health/Payment by Results Team. *Chemotherapy Regimens List 2012-13* (version 2.0). Department of Health; 2012.

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

19. Richardson PG, Sonneveld P, Schuster MW, Irwin D, Stadtmauer EA, Facon T *et al*. Bortezomib or high-dose dexamethasone for relapsed multiple myeloma. *New England Journal of Medicine* 2005;**352**:2487-98.

20. Child JA, Morgan GJ, Davies FE, Owen RG, Bell SE, Hawkins K *et al*. High-dose chemotherapy with hematopoietic stem-cell rescue for multiple myeloma. *New England Journal of Medicine* 2003;**348**:1875-83.

21. Bird JM, Owen RG, D'Sa S, Snowden JA, Pratt G, Ashcroft J *et al*. Guidelines for the diagnosis and management of multiple myeloma 2011. *British Journal of Haematology* 2011;**154**:32-75.

22. Moreau P, Avet-Loiseau H, Harousseau JL, Attal M. Current trends in autologous stem-cell transplantation for myeloma in the era of novel therapies. *Journal of Clinical Oncology* 2011;**29**:1898-906.

23. Lahuerta JJ, Mateos MV, Martinez-Lopez J, Rosinol L, Sureda A, de la Rubia J *et al*. Influence of pre- and post-transplantation responses on outcome of patients with multiple myeloma: sequential improvement of response and achievement of complete response are associated with longer survival. *Journal of Clinical Oncology* 2008;**26**:5775-82.

24. van de Velde HJ, Liu X, Chen G, Cakana A, Deraedt W, Bayssas M. Complete response correlates with long-term survival and progression-free survival in high-dose therapy in multiple myeloma. *Haematologica* 2007;**92**:1399-406.

25. Uyl-de Groot CA, Buijt I, Gloudemans IJ, Ossenkoppele GJ, Berg HP, Huijgens PC. Health related quality of life in patients with multiple myeloma undergoing a double transplantation. *European Journal of Haematology* 2005;**74**:136-43.

26. Segeren, C. M. *Intensive therapy in multiple myeloma.* Rotterdam: Erasmus University; 2002. Available at: *http://repub.eur.nl/res/pub/31962/020913_Segeren,%20Christine%20Maria.pdf*.

27. Beusterien KM, Davies J, Leach M, Meiklejohn D, Grinspan JL, O'Toole A *et al*. Population preference values for treatment outcomes in chronic lymphocytic leukaemia: a cross-sectional utility study. *Health and Quality of Life Outcomes* 2010;**8**:50.

28. Stevenson, M, Gomersall, T, Lloyd Jones, M *et al*. Percutaneous vertebroplasty and percutaneous balloon kyphoplasty for the treatment of osteoporotic vertebral fractures. Available at: http://www.nice.org.uk/nicemedia/live/13445/60625/60625.pdf

29. Janssen-Cilag Ltd. *Velcade 3.5mg powder for solution for injection.Summary of Product Characteristics.* Electronic Medicines Compendium; 2012. Available at: *Available from: URL: http://www.medicines.org.uk/emc/*. Accessed

30. Celgene Ltd. *Thalidomide Celgene 50mg Hard Capsules. Summary of Product Characteristics.* Electronic Medicines Compendium 2013. Available at: *http://www.medicines.org.uk/emc/*.

31. National Institute for Health and Clinical Excellence (NICE). *Lenalidomide for the treatment of multiple myeloma in people who have received at least one prior therapy. Technology Appraisal No.171;* 2009. Available at: *http://www.nice.org.uk/nicemedia/live/11898/44812/44812.pdf*.

32. Department of Health Payment by Results Team. *National Schedule of Reference Costs 2011-12 for NHS trusts and NHS foundation trusts.* Department of Health; 2012. Available at: *http://www.dh.gov.uk/health/2012/11/2011-12-reference-costs/*

33. Barlogie B, Attal M, Crowley J, van Rhee F, Szymonifka J, Moreau P *et al*. Long-term follow-up of autotransplantation trials for multiple myeloma: update of protocols conducted by the intergroupe francophone du myelome, southwest oncology group, and university of arkansas for medical sciences. *Journal of Clinical Oncology* 2010;**28**:1209-14.

34. Attal M, Harousseau JL, Stoppa AM, Sotto JJ, Fuzibet JG, Rossi JF *et al*. A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. Intergroupe Francais du Myelome. *New England Journal of Medicine* 1996;**335**:91-7.

35. Alvares CL, Davies FE, Horton C, Patel G, Powles R, Sirohi B *et al*. Long-term outcomes of previously untreated myeloma patients: responses to induction chemotherapy and high-dose melphalan incorporated within a risk stratification model can help to direct the use of novel treatments. *British Journal of Haematology* 2005;**129**:607-14.

9 APPENDIX 1

9.1 Clinical effectiveness critique of the Hovon, IFM and MRC MMIX trials

9.1.1 Context and description of the Hovon, IFM and MRC MMIX trials

The three additional studies presented in the MS (without thalidomide as comparator and therefore outside the NICE scope) are:

- Hovon trial[3;4] which evaluates bortezomib, doxorubicin and dexamethasone (PAD) vs vincristine, doxorubicin and dexamethasone (VAD);
- IFM trial[5;6] which evaluates bortezomib and dexamethasone (VD) vs vincristine, doxorubicin and dexamethasone (VAD), with and without intensification therapy;
- MRC MMIX trial[7] which evaluates cyclophosphamide, thalidomide and dexamethasone (CTD) vs cyclophosphamide, vincristine, doxorubicin and dexamethasone (CVAD).

[redacted]

Summary details relating to trial design, methodology and patient characteristics for the three trials are reported in the MS in Section 6.3 (p.49-67). The trials differ in their study design in terms of the interventions, comparators and the treatment pathways, and differ also from the Pethema and Gimema trials. In both the Hovon and IFM trials, some patients had a second consecutive ASCT which, according to the ERG clinical expert, is not standard UK practice (some patients may have a second ASCT but this will be held back until after relapse and would not be given consecutively). In all three studies groups appear to be well-balanced with respect to patient baseline characteristics. For the Hovon and IFM trials, the data appear similar to the Pethema and Gimema trials on observation. For IFM, it is not possible to see the comparison of VAD and VD without intensification therapy (as these are not reported separately). The ISS stage of patients in the MRC MMIX trial appeared to differ from the other four trials in that there were a higher proportion of patients with ISS stage III and a lower proportion of patients with ISS stage I. The ERG clinical expert notes that the MRC MMIX trial is more reflective of UK patients in terms of ISS disease stage. In addition, the MRC MMIX trial included patients >65 years which again is reflective of UK practice.

9.1.2 Manufacturer and ERG assessment of trial quality

The MS provides a quality assessment of the included trials in Section 6.4 and Table 23 (MS p. 79-80) with a more detailed assessment in MS Appendix 3. The quality assessment in the

MS follows the NICE criteria and is appropriate. The ERG carried out an independent quality assessment and this is shown in Table 26**Error! Reference source not found.**.

Table 26: Manufacturer and ERG quality assessment of Hovon, IFM and MRC MMIX trials

NICE QA criteria for RCTs		Hovon	IFM	MRC MMIX
1. Was the method used to generate random allocations adequate?	MS:	Low risk	Low risk	Low risk
	ERG:	Low risk	Unclear risk	Low risk
IFM trial – patients were 'centrally randomised' to treatment arms, but it is unclear what method was used to generate the randomisation sequence.				
2. Was the allocation adequately concealed?	MS:	Low risk	Low risk	High risk
	ERG:	Low risk	Low risk	Unclear risk
Comment: Hovon trial – patients were randomised using a web-based application. IFM trial – patients were randomised centrally. MRC MIX trial – the method used to conceal allocation is not described; the trial paper[7] states that "randomisation was on a 1:1 basis and open-labeled" (p. 443), which may be why the manufacturer marked this as 'high risk', but the ERG suggests that this refers to the trial being unblinded rather than to allocation concealment.				
3. Were the groups similar at the outset of the study in terms of prognostic factors, e.g. severity of disease?	MS:	Low risk	Low risk	Low risk
	ERG:	Low risk	Unclear risk	Low risk
IFM trial – this is difficult to assess as data were only reported for the two VAD groups together and the two VD groups together.				
4. Were the care providers, participants and outcome assessors blind to treatment allocation? If any of these people were not blinded, what might be the likely impact on the risk of bias (for each outcome)?	MS:	High risk	Low risk	Low risk
	ERG:	High risk	High risk	High risk
Comment: The Hovon, IFM and MRC MIX trials were open-label trials. IFM trial – response rate outcomes were assessed by an independent review committee (which is why the manufacturer has marked this as 'low risk'), but it is unclear if they were blinded to patient treatment allocation.				
5. Were there any unexpected imbalances in drop-outs between groups? If so, were they explained or adjusted for?	MS:	Low risk	Low risk	Low risk
	ERG:	Low risk	Low risk	Low risk
6. Is there any evidence to suggest that the authors measured more outcomes than they reported?	MS:	Low risk	Low risk	Low risk
	ERG:	Low risk	Low risk	Low risk
Comment: IFM trial – it should be noted that results were not reported for VAD without intensification and VD without intensification.				
7. Did the analysis include an intention to treat analysis? If so, was this appropriate and were appropriate methods used to account for missing data?	MS:	Low risk	Low risk	Low risk
	ERG:	Low risk	Low risk	Low risk
Comment: ITT analyses were used in the Hovon and MRC MIX trials (for response rates and proportion of patients who underwent SCT only in MRC MIX). PFS and OS in MRC MMIX were analysed in the per protocol population, according to actual treatment received, including five patients who crossed over from CVAD to CTD.[7] As the number of randomised patients not included and who crossed over is small, this is unlikely to have affected the results. ITT analyses were used in the IFM trial for all the efficacy outcomes except response rates, which were assessed in the "evaluable population".[5] As the number of patients not included is similar across arms, this is unlikely to have affected the outcomes.				

Note. These questions are usually answered with 'yes', 'no' or 'unclear'. However, in the MS the manufacturer has answered these using 'low risk', 'high risk' and 'unclear risk', so the ERG has followed this approach for ease of comparison. 'Low risk' = 'yes' and 'high risk' = 'no' (except for question 6).

The ERG's quality assessment agreed with the manufacturer's on all criteria for the Hovon trial, but differed to the manufacturer's for the IFM and MRC MMIX trials on the criteria of randomisation, allocation concealment, similarity of patient baseline characteristics and blinding.

9.1.3 Key clinical effectiveness results

The manufacturer reported all relevant results from the three trials. For the IFM trial, the MS reported results which included patients receiving intensification therapy so it was not possible to compare VAD and VD groups without the effects of intensification therapy. In response to the ERG clarification questions, the manufacturer subsequently supplied the data for the VD and VAD arms without intensification therapy and these are presented here.

Primary outcome – Response rates

The MS presents results for response outcomes in MS Table 24 (p.83-84) and in the questions for clarification response (for the IFM trial). ORR results for the Hovon, IFM & MRC MMIX trials correspond to the sum of the individual response rates.

- ORR post-induction was achieved in a significantly greater number of patients receiving a bortezomib regimen compared to a non-bortezomib regimen (Hovon PAD 84.2% vs VAD 61.3%, p<0.001; IFM VD 77.5% vs 59.5%, p=0.0029) which was maintained post-transplant in the Hovon trial only.
- Significantly higher ORR post-induction was reported in the CTD arm compared to CVAD (MRC MMIX 82.5% vs 71.2%, p<0.0001) but rates were similar post-transplant.
- Significantly higher CR post-induction was achieved in patients receiving a bortezomib regimen compared to a non-bortezomib regimen in one trial (Hovon PAD 11% vs VAD 2.9%, p<0.001) which was maintained in the post-transplant period.
- CR was significantly higher in the CTD arm compared to CVAD post-induction and post-transplant (MRC MMIX CTD 13.0% vs CVAD 8.1%, p=0.0083; CTD 33.3% vs CVAD 25.4%, p=0.00052, respectively).
- Higher nCR was achieved in bortezomib regimens (PAD or VD) compared to VAD post-induction (Hovon and IFM trials, although only statistically significant in the IFM trial). No significant differences in either trial post-transplant.
- VGPR was higher in bortezomib regimens (PAD or VD) compared to VAD post-induction and post-transplant (Hovon and IFM, only statistically significant in IFM trial).
- VGPR post-induction was significantly higher in patients receiving CTD compared to those receiving CVAD (MRC MIX trial) but this was not maintained post-transplant.

- PR rates post-induction and post-transplant were lower in the bortezomib groups (PAD or VD) compared to VAD (only statistically significant post-induction in IFM).
- Post-induction PR was similar in the two groups in the MRC MMIX trial, but significantly higher in the CVAD arm compared to CTD arm post-transplant.
- No differences between treatment groups in patients experiencing disease progression in any of the three trials

Secondary outcomes – disease progression and survival

For the longer-term outcomes of PFS, TTP and OS, comparisons between trials are difficult due to the different treatment pathways employed by the trials. Additionally, for both the Hovon and IFM trials, the MS states (p.53-54) that the maintenance therapies following induction may confound the long-term outcomes, and neither PFS nor OS were adjusted for maintenance. The ERG would agree with this and notes that these results should therefore be interpreted with caution.

- PFS was significantly longer in the bortezomib group (PAD or VD) compared to the VAD group (Hovon and IFM trials); no differences in PFS between the CTD and CVAD groups (MRC MMIX trial).
- For TTP, there was a statistically significant lower HR in patients treated with bortezomib (PAD or VD) compared with VAD (Hovon and IFM trials).
- Median OS was not reached and there were no statistically significant differences in OS between treatment arms in any of the trials.

Proportion of patients undergoing SCT

The MS reports the proportion of patients who underwent ASCT (MS Table 28, p.93) and in the ERG questions for clarification response (for the IFM trial), but states that the studies were not powered for this endpoint (MS p.93). The trial data show that similar proportions of patients in the two treatment groups underwent ASCT for all three trials though no statistical tests were reported.

Cytogenetic risk subgroup

Subgroup results are only available for the Hovon trial as data from the IFM trial included patients receiving intensification therapy, and no subgroup results were reported for the MRC MMIX trial.

The MS reported response rates for patients with high and standard cytogenetic risk in MS Table 29 (p.94-95). In patients with both high risk and standard risk cytogenetics, the CR/nCR rate post-induction and post-transplant was higher in the bortezomib (PAD) arm compared with the VAD arm, but no statistical comparison was reported so it is not clear whether these results were statistically significant.

The MS reported PFS and OS for patients with high and standard cytogenetic risk in MS Table 30 (p.94-95). PFS and OS were significantly longer in the PAD arm compared to VAD in the high risk subgroup. No other differences were observed between treatment arms. TTP was not reported in this subgroup.

Adverse Events

The MS presents data for AE in the post-induction period in MS Section 6.9 (p.102-102), withdrawal rates in MS Appendix 8 and data presented in response to ERG clarification questions. Patients receiving bortezomib (PAD) experienced statistically significantly more grade 3/4 AE (Hovon) and serious AE (Hovon). In the MRC MMIX trial, there was a significantly higher incidence of any serious AE in patients receiving CVAD compared to those receiving CTD.

The most frequently-occurring grade ≥3 AEs and AEs of special interest to bortezomib treatment are presented in MS Tables 43 (p.105) for the Hovon trial and in response to ERG clarification questions for the IFM trial. On observation of the data it appears that peripheral neuropathy occurred more frequently in those receiving bortezomib (PAD or VD) in the Hovon and IFM trials. In addition, there was a higher incidence of thrombocytopenia and herpes zoster (Hovon) and lymphopenia (IFM) in those receiving bortezomib (PAD or VD) compared to VAD, but a lower incidence of mucosal inflammation (IFM). The MS does not present statistical tests for AE data so it is unclear whether any of these differences are statistically significant. There were no statistically significant differences in withdrawals for the Hovon trial (no withdrawal rates available for the IFM trial and the MRC MMIX trial).

Summary of results

Results of the three RCTs included by the MS that evaluated interventions/comparators outside the NICE scope (Hovon, IFM and MRC MMIX) were presented in this appendix. Patients with MM, eligible for HDT-ASCT, who received bortezomib (PAD or VD) had a statistically significantly higher ORR post-induction (Hovon and IFM trials) and post-transplant (Hovon trial only). For other response outcomes (CR, nCR, VGPR), there tended to be a favourable effect observed in the bortezomib arms (PAD or VD) but results were only

statistically significant for one or other trial (Hovon or IFM) and not always maintained post-transplant. In contrast, PR rates in the Hovon and IFM trials were lower in the bortezomib groups (PAD or VD) compared to VAD (only statistically significant post-induction in the IFM trial). In the MRC MMIX trial, CTD treatment was significantly more favourable compared to CVAD for ORR (post-induction only), CR (post-induction and post-transplant) and VGPR (post-induction only). There were no differences in disease progression for any trial.

PFS was significantly longer in the bortezomib group (PAD or VD) compared to VAD (Hovon and IFM); there were no differences between the CTD and CVAD groups (MRC MMIX). For TTP, there was a statistically significant lower hazard of progression in patients treated with bortezomib (PAD or VD) compared with VAD (Hovon and IFM). There were no statistically significant differences in OS or the proportion of patients undergoing SCT for all three trials.

Patients receiving bortezomib (PAD) experienced statistically significantly more grade 3/4 AE (Hovon) and serious AE (Hovon), whilst patients receiving CVAD experienced a significantly higher incidence of any serious AE compared to CTD (MRC MMIX).

9.1.4 Mixed treatment comparison

This Appendix provides further details and critique of the MTC within the MS. As stated in Section 3.1.7 of this report, the rationale for doing an MTC is given and is appropriate; however, some assumptions relating to the MTC may not be valid.

Methods for ascertainment of studies

Searches undertaken for the systematic review of clinical effectiveness included all potential comparators so no additional searches were performed for the MTC. This strategy, using the interventions bortezomib, thalidomide, vincristine and cyclophosphamide as monotherapy or in combination with any other intervention, is probably wide enough to have identified all trials relevant to an MTC. However, inclusion/exclusion criteria for the MTC are not adequately reported. It is stated in the MS Section 6.7.1 (p.96) that ‘all comparators which could potentially contribute to an MTC were included at the search stage, and only excluded at citation screening’. No further details are reported. No details are given on included/excluded studies for the MTC or reasons for exclusions; the only comment given is that ‘the MRC MMIX does not assess a bortezomib-based regimen but it is relevant to the MTC’ (MS Section 6.2.4, p.46). No QUOROM flow chart is presented for the MTC. Data from the five included RCTs is extracted and tabulated in the systematic review of clinical effectiveness (Baseline characteristics in MS Section 6.3, Results in MS Section 6.5). Data

for the MTC is presented in MS Table 32 (p.98). All trials used in the MTC were critically appraised in the systematic review of clinical effectiveness section (MS Section 6.4). Risk of bias for the five included studies is presented in Table 23 and Appendix 3 of the MS and briefly discussed in MS Section 6.4.3. No studies were judged to be of poor quality and no trials were excluded because of any potential risk of bias. In addition, the MTC was only performed on two outcomes (TTP and OS) and no justification was given for excluding other outcomes which could have been used and may not have been so heterogeneous.

Network of evidence

The MTC was based on the Pethema, Gimema, Hovon, IFM and MRC MMIX RCTs. A visual network diagram was provided for the MTC (MS Section 6.7.3, p.97). This shows no connected network of evidence; that is, no single network could be formed. In order to provide a network it was necessary to rely on a series of assumptions. An explanation is provided for the assumptions made (i.e. that CVAD and VAD are clinically equivalent and that TD and CTD are clinically equivalent based on clinical opinion) and it is acknowledged in the MS that they add uncertainty to the analysis. Two key bortezomib-based trials (Pethema and Gimema) connect VTD and TD, whilst the other two bortezomib-based trials connect VD and PAD with VAD (IFM and Hovon respectively). The assumption that CVAD is equivalent to CTD and that TD is equivalent CTD allows the two separate networks using the four bortezomib trials to be 'connected' via the MRC MMIX trial which connects CTD and CVAD. It is stated in MS Appendix 14 that clinicians on the Advisory Board were 'reasonably comfortable with the assumption that CVAD and VAD would be equivalent regimens' and 'felt that the assumption that TD and CTD are equivalent could be used.' The ERG clinical expert agrees that they are probably equivalent whilst acknowledging the absence of randomised data. It is the ERG's opinion that such assumptions are not fully justified (given the lack of trial evidence) and therefore may not be valid.

Statistical Analysis

Overall statistical procedures used for the MTC are reported but specific details of the analyses for the two outcomes (OS and TTP) are limited. The MTC for TPP and OS was conducted using time-to-event data (ITT) and where HR or CI data were missing or incomplete, these were derived from digitalised versions of the Kaplan-Meier survival curves from the clinical trial reports. OS data was not reported in the Gimema trial so this trial was not included in this analysis. Thus the VTD vs. TD OS comparison was based solely on data from Pethema making results less tenable. TTP data were not available from the Gimema and MRC MMIX trials so PFS data was used as a proxy.

A fixed effect (FE) model was performed as the base case which assumed no heterogeneity between RCTs. This is not fully justified and seems contradictory to the rationale for not doing a standard pairwise meta-analysis as the trials were deemed too heterogeneous. A random effects (RE) model was also conducted as a sensitivity analysis to take into account heterogeneity. The MS states for both OS and TTP that the deviance information criteria (DIC) assessment of model fit supported the use of the FE model (OS: FE DIC = 0.822 vs. RE DIC = 0.833; TTP: FE = -3.604 vs RE = -2.276; updated data after clarifications OS: FE DIC = 1.663 vs. RE DIC = 1.65; TTP: FE = -2.902 vs RE = -1.1 to 1-.5). That is, as the FE model has the lower DIC and is more parsimonious (fewer parameters) it is assumed to be the most appropriate. However, given the small difference between FE and RE DIC and the acknowledged heterogeneity across the included trials, the ERG considers that an RE model would be the most appropriate as it allows for variability between treatment effects estimated by individual studies even though there are not enough data in the network to robustly estimate such a model.

Bayesian MTC analyses were used to compare the different treatments. The models employed Markov chain Monte Carlo simulation, based upon 20,000 iterations after 10,000 burn-in iterations (to ensure the model had converged on the posterior distribution). A vague prior was assumed for the treatment effects (0, 10000) for the FE model. A vague prior was also assumed for the treatment effects (0, 10000) for the RE model and a weakly informative prior (0, 2) for the between-study standard deviation. It is stated in the MS (p101) that the prior distribution for the RE model was somewhat dominating the posterior distribution due to the limited data points available. When there is a limited number of trials it may be appropriate to use a more informative prior distribution for between-study standard deviation.[10] The ERG suggests that a sensitivity analysis using a more informative prior such as (0, 0.6) could have been performed which would still be reasonably uncertain and acknowledge the possibility of heterogeneity between studies.

Sampled values were used to estimate the posterior medians, 95% credible intervals (CrI) for the HRs and the probabilities for the HRs to be smaller than 1. Treatment efficacy was assessed according to the probability of each treatment having the largest beneficial effect, calculated as the proportion of simulations in which the treatment was ranked as most efficacious.

Although heterogeneity is recognised in the MS there are no numerical estimates of the degree of heterogeneity and no meta-regression using covariates to explore heterogeneity. However, given the limited amount of data available it may not be is possible to do this. Also,

no sensitivity analyses were undertaken on the trials included or on alternative prior distributions for model parameters. The MTC model was built in WINBUGS 1.4 and programming codes used in both the FE and RE models are provided (MS Appendix 17) which appear reasonable.

Results

Results are presented through a series of tabulations with no illustrations for the two selected outcomes (OS and TTP). TTP was not a primary outcome for any of the included trials and OS was a primary outcome for the MRC MMIX trial only. These two outcomes are likely to suffer confounding due to additional treatment post-SCT for patients in the Gimema, Hovon and MRC MMIX trials. The results are presented in terms of a simultaneous assessment ranking of superiority, pairwise comparisons of superiority (for three interventions and two comparators), and pairwise assessment of HRs for five interventions. The MTC should provide a full set of HRs for all ten comparisons, effectively combining all the direct evidence and indirect evidence for each comparison. However, only six pairwise comparisons are presented; the four omitted are those comparing bortezomib-containing regimens (PAD vs VD, VD vs VTD and PAD vs VTD) and CVAD/VAD vs CTD/TD. No tabulation of direct comparisons and multiple comparisons is provided. Direct data exist for only four comparisons and for three of these comparisons only one trial is available. When direct evidence is available, it agrees with the results of the MTC although one comparison is not presented for the MTC.

Comparisons of MTC results using both FE and RE models are reported (MS p.100-2) and are broadly similar, although RE CrIs are wider. (NB. Updated data after manufacturer clarifications produced similar results). Very limited narrative comments on the results are presented in the MS. For TTP it is stated that VTD had the highest probability of being the most effective treatment. Bortezomib-based regimens had probabilities close to 100% (FE model) or >50% (RE model) of being superior to CTD. Patients treated with VTD had significantly lower HR compared with CTD treated patients (FE model only). For OS VTD had the highest probability of being the most effective treatment followed by VD, PAD, CTD/TD and VAD/CVAD. No discussion of the results is presented. No mention is made of the fact that all 95% CrIs for HRs using the RE model and all CrIs for OS and most for TTP using the FE model exceed 1 which is indicative of an unstable model with not enough data. However, a statement is made that the limitations of the MTC due to the assumptions made and the heterogeneity in the trial designs means that results should be treated with ‘utmost caution’ (MS p.98). There is no comment in the MS on how results compare to other reviews, meta-analyses, studies or to routinely collected data. The direction and magnitude

of pairwise comparisons are stated to be consistent across the MTC analyses that were performed but not all results are presented (as mentioned above). The MS reports that a comparison of direct and indirect evidence comparing bortezomib-based induction therapy with CTD is not possible due to lack of head-to-head data.

Conclusion

Overall the methods and execution of the MTC appear adequate. However, there are two key concerns: firstly, the assumptions made for devising a network of evidence with the resulting network not forming a closed loop necessary for an MTC; and secondly, issues of heterogeneity, with too much heterogeneity for a FE model to be credible but too few data to fit a RE model.

9.2 Economic analysis

CEA Methods

The two additional models submitted by the manufacturer, for PAD vs. VAD and VD vs. VAD, have the same structure as the VTD vs. TD model considered in Section 4.2. They estimate the cost-effectiveness of PAD vs. VAD and VD vs. VAD in patients with newly diagnosed MM. As with the VTD vs. TD model, the models adopt a lifetime horizon, with monthly cycles. Costs and outcomes are discounted at 3.5% per annum and the models take the perspective of the NHS England and Wales.

The principal clinical-effectiveness measures were derived from the relevant clinical trials (Hovon,[3] IFM[5]), for post induction response rates (CR, PR and NR), induction mortality rates, SCT rates, and post induction progression.

The models use the same utilities as the VTD vs. TD model and costs are obtained from the same sources.

The models explore parameter uncertainty in both one-way and probabilistic sensitivity analyses (MS Section 7.7.7 p.192 and MS Section 7.7.8 p.197). Several scenario analyses are also performed. The MS does not report clinical plausibility / external validity of the extrapolated portions of these models against long term survival data (MS p.183). This is only done for the VTD vs. TD model.

CEA Results

The results from the economic evaluation are presented in MS Table 93 (p.192) as incremental cost per QALY gained for PAD vs. VAD and VD vs. VAD.

For the base case, an incremental cost per QALY gained of £11,041 is reported (see Table 27) for PAD vs. VAD, and £14,446 for VD vs. VAD.

Table 27: Base case cost-effectiveness results for PAD vs. VAD and VD vs. VAD

Technologies	Total costs (£)	Total LYG	Total QALYs	Incre-mental costs (£)	Incre-mental LYG	Incre-mental QALYs	ICER (£) incre-mental (QALYs)
VAD (Hovon)	£49,359	4.41	2.91	+£10,274	1.31	0.93	£11,041
PAD	£59,632	5.72	3.84				
VAD (IFM)	£50,163	4.42	2.91	+£12,710	1.22	0.88	£14,446
VD	£62,874	5.64	3.79				

The probability that PAD is a cost effective option over VAD at £20,000 and £30,000 willingness-to-pay thresholds is estimated to be 84% and 89% respectively. The probability that VD is a cost effective option over VAD at £20,000 and £30,000 thresholds is estimated to be 69% and 83% respectively.

9.2.1 Critical appraisal of the manufacturer's submitted economic evaluation

Please refer to Section 4.2 for a critical appraisal of these models.

9.2.2 Modelling approach / Model Structure

The structure of the PAD vs. VAD and VD vs. VAD models is identical to the structure of the VTD vs. TD model. Please see Section 4.2.1 for further details.

The model captures the impact of the intervention and differential response to induction therapy with separate health states for CR, PR and NR post-induction. Time to progression (TTP) transition probabilities are derived from Hovon and IFM trial data[3;5] for each category of response (CR, PR and NR) and by treatment. As with the VTD vs. TD model, transition probabilities to 3rd and further lines of treatment are derived from the APEX trial data which compared bortezomib monotherapy with high dose dexamethasone in patients with relapsed multiple myeloma.[19] Parameter estimates obtained from median survival by response category in the MRC VII trial [20] are used to obtain OS probabilities by post-induction response.

Given that the PAD vs. VAD and VD vs. VAD models have the same structure as the VTD vs. TD model, the ERG considers that they have the same limitations (Section 4.2.1).

9.2.3 Patient Group

The patient group included in the PAD vs. VAD and VD vs. VAD models is adult patients with previously untreated multiple myeloma, eligible for HDT-SCT. The characteristics of the modelled populations are not specified. However as the main trials used for the model outcomes of PAD vs. VAD and VD vs. VAD are the Hovon[3] and IFM[5] trials respectively, the modelled cohorts can be assumed to have these patient characteristics (MS Table 18, p.64).

Our clinical expert considers that the clinical characteristics of the trial populations are representative of clinical practice in the UK, with the exception of ISS Stage.

9.2.4 Interventions and comparators

For PAD vs. VAD, bortezomib is administered in combination with doxorubicin and dexamethasone (PAD) for 3 cycles of 28 days vs. vincristine, doxorubicin and dexamethasone (VAD) (based on the Hovon RCT[3]).

For VD vs. VAD, bortezomib is administered in combination with dexamethasone (VD) for 4 cycles of 21 days vs. vincristine, doxorubicin and dexamethasone (VAD) (based on the IFM RCT[5]).

The scope for this appraisal, developed by NICE, is for 'bortezomib in combination with other chemotherapy regimens for induction therapy' compared to 'chemotherapy regimens containing thalidomide'. The modelled analyses PAD vs. VAD, and VD vs. VAD are both therefore outside of the NICE scope.

9.2.5 Clinical Effectiveness

The clinical effectiveness parameters which are specific to the PAD vs. VAD and VD vs. VAD models are given below. All other model clinical parameters, and issues arising, are discussed in Section 4.2.4.

The proportion of patients with post-induction CR, PR or NR by treatment arm was informed by the Hovon and IFM CSRs.[6;8] These data are presented in Table 28 (extract of MS Table 50, p.123) and enter the economic model as baseline risks.

Table 28: Post-induction response rates used in PAD vs. VAD and VD vs. VAD models

Trial	Treatment	Comparator
Hovon	**PAD** **N=417**	**VAD** **N=416**
CR (CR+nCR+VGPR)	209 (50.1%)	81 (19.5%)
PR	142 (34.1%)	174 (41.8%)
NR (MR+SD+PD)	66 (15.8%)	161 (38.7%)
IFM 2005	**VD without DCEP** **N=121**	**VAD without DCEP** **N=121**
CR (CR+nCR+VGPR)	49 (40.8%)	19 (15.7%)
PR	44 (36.7%)	53 (43.8%)
NR (MR+SD+PD)	27 (22.5%)	49 (40.5%)

CR: complete response; NR: non responders; MR: minimal response; PD: progressed disease; PR: partial response; SD: stable disease; VgPR: very good PR;

The proportions of patients receiving SCT are obtained from the Hovon and IFM CSRs and are given in Table 29 (extract of MS Table 52 p.124).

Table 29: Total SCT proportions by treatment arm for PAD, VAD, VD and VAD treatments

Treatment	Total SCT
PAD (N=417)	354 (84.9%)
VAD (N=416)	348 (83.7%)
VD (N=121)	100 (82.6%)
VAD (N=121)	106 (87.6%)

Mortality rates by treatment arm during the induction phase were taken from the Hovon and IFM studies and are given in Table 30. Mortality rates by treatment arm during the transplant period were also obtained from these two studies (MS Table 51, p.123).

Table 30: Mortality rate during induction period by treatment arm

Treatment	Mortality rate during induction (6 months)	Monthly probability of death during induction
PAD	4.6% (19/410)	1.6%
VAD	5.6% (23/411)	1.9%
VD	0.7% (1/135)	0.2%
VAD	3.7% (5/136)	0.9%

TTP transition probabilities are derived from exponential curves fitted to the Hovon and IFM trial data. Weibull and log-logistic fits are explored by the manufacturer in scenario analyses as alternatives to the exponential fits, although the MS notes that the Weibull and log-logistic parametric fits lack face validity and clinical plausibility (MS p.140-141). Treatment effects were calculated in parametric regression analyses and are used to modify the baseline transition probabilities. The parameters of the TTP curves for each distribution are given in MS Table 56 (p.127).

9.2.6 Patient outcomes

Patient outcomes for the PAD vs. VAD and VD vs. VAD models are discussed in Section 4.2.5. They are identical to the outcomes used in the VTD vs. TD model.

9.2.7 Resource use

Resource use assumptions for the PAD vs. VAD and VD vs. VAD models are given below in instances where they differ to the assumptions in the VTD vs. TD model. For all other resource use assumptions, which are general to all three models, see Section 4.2.6.

The treatments of the induction regimens for each of the analyses were based upon those from the Hovon[3] and IFM[5] trials using the same dosages and durations of treatment.

For PAD vs VAD: 3 cycles were used for induction therapy based on the Hovon trial. Doxorubicin was administered on days 1-4 of the treatment cycle with a dosage of 9 mg/m². Dexamethasone was administered on days 1-4 and days 9-12 and 17-20 of each treatment cycle during each cycle. The dosage of dexamethasone was 40 mg.

For VD vs. VAD: 4 cycles were used for induction therapy based on the IFM trial. Doxorubicin was administered on days 1-4 of the treatment cycle with a dosage of 9 mg/m².

Dexamethasone was administered on days 1-4 for all cycles and days 9 to 12 for cycles 1 and 2. The dosage of dexamethasone was 40 mg.

9.2.8 Costs

Refer to Section 4.2.7 for general cost details for all models. The unit costs associated with each of the 1st line induction therapies, drugs, prophylaxis, administration and monitoring are shown in MS Table 68 (p.163), and summarised Table 20 for the PAD vs. VAD and VD vs. VAD models.

Table 31: Unit costs associated with the 1st line induction therapies: drugs, prophylaxis, administration and monitoring

	PAD	VAD
Average cost of a course of treatment	£9,692.09	£705.17
Prophylaxis	£80.65	£53.37
Administration	£781.00	£781.00
Monitoring cost	£520.06	£520.06
TOTAL	£11,073.80	£2,059.60
	VD	**VAD**
Average cost of a course of treatment	£12,260.91	£898.34
Total prophylaxis	£80.65	£71.15
Administration Cost	£1,069.00	£1,069.00
Monitoring cost	£693.42	£693.42
TOTAL	£14,103.98	£2,731.91

The ERG has checked the costs used in the model with the referenced sources. All relevant costs have been considered and the manufacturer's approach is reasonable.

9.2.9 Consistency/ Model validation

The MS does not report if checklists were used for internal validation of the PAD vs. VAD and VD vs. VAD models. No validation of outputs from these models against external data is reported in the MS.

9.2.10 Assessment of Uncertainty

Refer to Section 4.2.9 for a description of the work described in the MS to assess model uncertainty.

One-way sensitivity analyses

Results of the one-way sensitivity analyses for the PAD vs. VAD and VD vs. VAD models indicate that the ICER is most sensitive to post-induction CR mortality and drug costs. For the VD vs. VAD model the tornado diagram also shows sensitivity to the TTP hazard ratio for the PR group (MS Figure 25, p.195).

Scenario Analysis

Refer to Section 4.2.9 for general details.

Results are presented for 24 scenarios in MS Table 95 (p.200). The ERG was unable to reproduce the exact results in MS Table 95 for a small number of scenarios in the PAD vs. VAD and VD vs. VAD models but the differences in final ICER values were not substantial. In all analyses ICERs remain below or close to £30,000/QALY.

Probabilistic Sensitivity Analysis

Please refer to Section 4.2.9 for general details.

The probabilities that PAD is a cost-effective option compared to VAD at the £20,000 and £30,000 willingness-to-pay thresholds are estimated to be 84% and 89% respectively. The corresponding probabilities for VD vs. VAD are 69% and 83% (MS Table 96, p.204).

The ERG notes that the probabilistic and deterministic sensitivity analysis results are consistent for these two models.