

## Consensus agreement of the blinded review and interpretation of the results of the patient-reported outcomes from the CARFI trial

### Writing committee

Henrik Rode Eshoj<sup>1,5</sup>, Madeleine T King<sup>1, 2</sup>, Niels Abildgaard<sup>1</sup>, Lene Kongsgaard Nielsen<sup>1,3</sup>  
Henrik Gregersen<sup>4</sup>

### Affiliations:

<sup>1</sup>Quality of Life Research Center, Department of Haematology, Odense University Hospital, Denmark

<sup>2</sup>School of Psychology, University of Sydney, Australia

<sup>3</sup>Research Unit for Multimorbidity, Department of Internal Medicine and Cardiology, Viborg Regional Hospital, Denmark

<sup>4</sup>Department of Haematology, Aalborg University Hospital, Denmark

<sup>5</sup>OPEN - Open Patient data Explorative Network, Odense University Hospital and University of Southern Denmark, Denmark

## Consensus agreement: Blinded analyses, review and interpretation of the results of the patient-reported outcomes from the CARFI trial

The CARFI trial: Phase II study of carfilzomib-cyclophosphamide-dexamethasone and high-dose melphalan followed by randomization between observation or maintenance with carfilzomib and dexamethasone in patients with relapsed multiple myeloma after high-dose melphalan with autologous stem cell support

### Introduction

The result of the primary endpoint in the CARFI trial (time to progression) was known at the time of analyzing the secondary patient-reported outcome (PRO) endpoints. Unblinded interpretations of the PRO endpoints may be vulnerable to biases influenced by prior convictions, wishful thinking and conflict of interest (Gotzsche 1996). At the time of this agreement, the CARFI investigators who form the writing team for the CARFI PRO paper had not yet seen the unblinded results of PRO endpoints of the CARFI trial.

This document presents the results of the primary PRO, the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 Summary Score (EORTC QLQ-C30 Sum Sc), of the CARFI trial, while all the authors, including the independent statistician conducting the analyses, were still blinded as to the intervention and control group identification. Additionally, blinded review and interpretation of the secondary PRO endpoints was also conducted.

This document, including a presentation of the blinded results, was presented to the writing committee (undersigned below) of the PRO data from the CARFI trial. The writing committee developed two blinded interpretations of the PRO endpoint results prior to unblinding. One interpretation assumed that A was the experimental intervention (carfilzomib-dexamethasone (Kd) therapy) and another assumed that A was the control (observation alone). After agreeing on the interpretations, as stated in this document, the writing committee signed the resulting document. Subsequently, the randomisation code was broken, the correct interpretation chosen, and the manuscript finalized, as recommended by Jarvinen et al (2014).

Of note, to keep the blinding of group allocation, the number of patients still on study at each assessment time point was not revealed to the author group. The reason for this was that we knew that a higher number of patients allocated to observation experienced progressive disease earlier than patients allocated to Kd maintenance (Gregersen H 2021). Further, PRO questionnaire compliance rates (both the proportion of randomised patients with completed questionnaires and the proportion of patients expected to complete questionnaires (alive and still on study)) were also not revealed to keep the blinding. However, as these are important considerations when interpreting PROs (White et al. 2011; Bell and Fairclough 2014), they will be taken into consideration for the overall interpretation of the study results after the blinding is broken.

### Primary objective

To compare the effects of Kd maintenance therapy, relative to observation only, on changes in health-related quality of life (HRQL) as assessed by the QLQ-C30 Sum Sc from randomisation to eight months follow-up.

**Hypothesis:** *There will be no difference between the two groups in change from randomisation (baseline) to eight months follow-up in the QLQ-C30 Sum Sc*

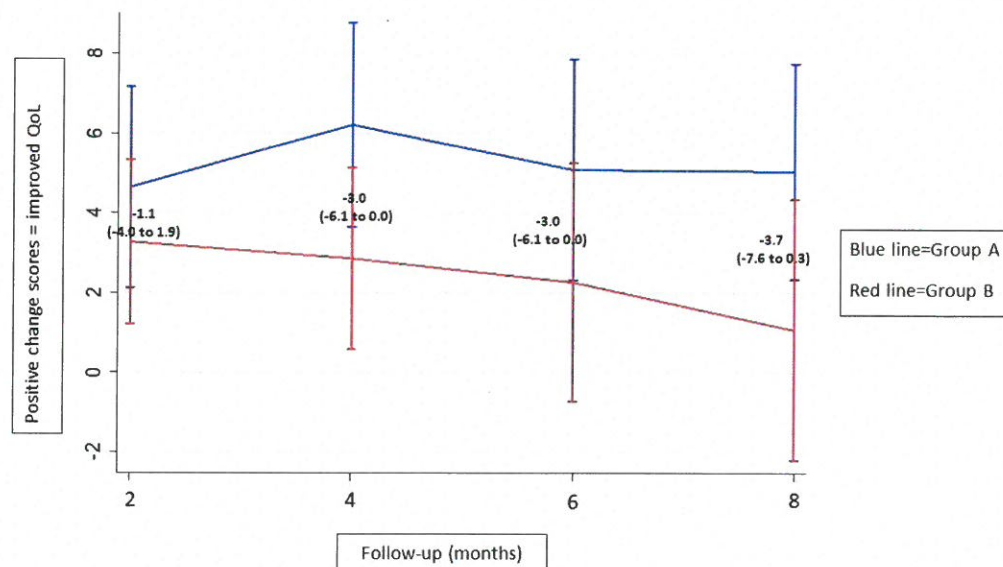
## Results from the intention-to-treat analysis of the primary analysis of the primary PRO endpoint

### Between-group differences

The primary PRO endpoint was mean change in the QLQ-C30 Sum Sc from randomisation to eight months follow-up (i.e. the primary data assessment time point). The mean difference between groups by linear mixed effects model for repeated measures, utilising data from all time points, was 3.7 units (95%CI -7.6 to 0.3,  $p=0.068$ ). Thus, the health-related quality of life (HRQL) experienced by Group A was estimated to be 3.7 points (95%CI -7.6 to 0.3) better than that experienced by Group B over the period from randomisation to eight months follow-up, as assessed by the QLQ-C30 Summary score (**Figure 1**). This mean difference did not reach the 10-point threshold for minimal important difference (MID) stated *a priori* in the published statistical analyses plan (Eshoj HR 2020), nor did its *p*-value achieve the *a priori* statistical significance threshold ( $p<0.05$ ).

Considering the differences between Group A and Group B in change at each of the intervening time points (two, four, and six months follow-up), none reached the threshold for clinical relevance, and none was statistically significantly different at the 1% significance level (**Table 1**).

**Figure 1.** Change scores from randomisation to two, four, six and eight months follow-up in the primary endpoint the QLQ-C30 Summary Score, and indicates the between-group difference in change estimated by linear mixed effects model



This figure shows the effect of CARFI and observation only following salvage ASCT on the EORTC Quality of life (QLQ)-Core30 (QLQ-C30) Summary score over all time points until eight months follow-up. Estimates for the between group comparison for each time point and for the entire profile were obtained. The estimates are presented with the corresponding 95% confidence intervals in parentheses. Higher scores represent better quality of life. The corresponding mixed-model repeated-measures analysis for the QLQ-C30 Summary score for the mean difference over time between groups showed a difference of 3.7 (95%CI -7.6 to 0.3,  $p=0.068$ ) at the primary endpoint at eight months follow-up. Threshold for significance at the primary assessment time point at eight months follow-up was  $p<0.05$ .

### Within-group differences

Both groups improved in QLQ-C30 Sum Sc. relative to baseline at all time points (two, four, six and eight months from randomisation). For Group A, improvements were statistical significant ( $p < 0.001$ ) at each time point, whereas for Group B, the degree of improvement was statistical significant for only the two first time points ( $p < 0.001$ ,  $p < 0.007$ ). Notably, none of these improvements reached the threshold for clinical relevance (**Table 1**).

**Table 1. Mean QLQ-C30 Summary Scores at randomisation, two, four, six and eight months follow-up for each group, and linear mixed effects estimates of within-group change from randomisation.**

Assessment time point (months)	Group A			Group B			Between groups	
	Mean score per visit	Change (95%CI)	p-value	Mean score per visit	Change (95%CI)	p-value	Mean difference (95%CI)	p-value
Baseline	78.05	-	-	79.28	-	-	-	-
2	82.75	4.5 (2.1 to 6.9)	<0.001	81.90	3.4 (1.4 to 5.5)	0.001	-1.1 (-4.0 to 1.9)	0.483
4	84.54	6.1 (3.7 to 8.5)	<0.001	82.18	3.0 (0.8 to 5.2)	0.007	-3.0 (-6.1 to 0.0)	0.050
6	83.37	4.9 (2.3 to 7.6)	<0.001	82.68	2.4 (-0.5 to 5.3)	0.100	-2.5 (-6.2 to 1.2)	0.188
8	83.17	4.9 (2.3 to 7.5)	<0.001	81.60	1.2 (-1.9 to 4.5)	0.451	-3.7 (-7.6 to 0.3)	0.068

QLQ-C30 quality of life questionnaire-core 30; CI Confidence interval. A 10 points difference was set a priori as the minimal clinically meaningful between-group difference. Threshold for significance was  $p < 0.01$ .

## Results from the intention-to-treat analysis of the supportive secondary analysis of the primary PRO endpoint

### Summary measure approach for the QLQ-C30 Summary score

The supportive secondary analysis of the primary PRO endpoint used a summary measure approach to estimate the between-group difference in HRQL (as captured by QLQ-C30 Sum Sc). The summary measure was per-patient average QLQ-C30 Sum Sc (range 0-100), calculated using all available data for each patient from randomisation to progressive disease/death/drug discontinuation/end of study. The average per-patient mean QLQ-C30 Sum Sc was 82.2 (SD 12.3) for Group A and 81.2 (SD 12.9) for Group B. These means did not differ significantly between the groups (t-test,  $p = 0.623$ ).

## Results from the intention-to-treat analysis of the secondary PRO endpoints

### *Individual domains of the QLQ-C30, QLQ-MY20 and FACT/GOG-ntx subscale (SAP analysis 6.3.3.1)*

Of note, the FACT/GOG-ntx subscale could not be calculated due to high rates of missing questionnaires, and so is not reported below. Unlike the QLQ-C30 and QLQ-MY20, which were assessed electronically, the FACT/GOG-ntx questionnaires were assessed in hard copy. However, due to site procedural errors, FACT/GOG-ntx data were not collected at all at most study sites.

#### *Between-group differences*

There was a clinically relevant between-group difference at the eight months follow-up in favour of Group A in the following QLQ-C30 domains: social functioning, cognitive functioning, appetite loss, diarrhoea, fatigue, nausea and vomiting and insomnia. For cognitive functioning and insomnia, clinically important differences between the groups were achieved despite change within group failing to achieve the clinically meaningful change thresholds because Group A improved (by a small amount) while Group B worsened (by a small amount). However, none of the domains that reached clinically important between-group differences reached the statistical significance threshold ( $p < 0.01$ ) (**Table 2**). Considering between-group differences for the QLQ-MY20 domains, none reached the threshold for clinical importance and none was statistically significant (**Table 2**).

#### *Within-group changes*

Group A showed clinically meaningful change scores in the direction of improvements in three QLQ-C30 functioning domains (physical, role, social) and five QLQ-C30 symptom scales (appetite loss, diarrhoea, fatigue, nausea and vomiting and dyspnoea). However, only the improvements in the functioning domains and the symptom scales of appetite loss and fatigue reached the threshold for statistical significance ( $p < 0.01$ ) (**Table 2**). Considering change scores for the QLQ-MY20 domains for Group A, three domains (Body image, Future perspectives and Side effects) reached the threshold for statistical significance ( $p < 0.01$ ) (in the direction of improvement), however none of these changes were clinically meaningful (**Table 2**).

Group B showed clinically meaningful change scores in the direction of improvements in two QLQ-C30 functioning domains (role, social) and one QLQ-C30 symptom scale (dyspnoea). None of these, however, reached the threshold for statistical significance ( $p < 0.01$ ) (**Table 2**). Considering change scores for the QLQ-MY20 domains for Group B, two domains (Future perspectives and Side effects) reached the threshold for statistical significance ( $p < 0.01$ ) (in the direction of improvement), however neither of these changes were clinically meaningful (**Table 2**).

Table 2. Mean change from randomisation to eight months follow-up in the individual domains of the QLQ-C30 and QLQ-MY20 questionnaires for each group, and linear mixed effects model estimates of the within-group change from randomisation and the between-group differences across time.

QLQ-C30 Functioning domains	Group A			Group B			Between groups		
	Mean score change (95%CI)	p-value	Mean score change (95%CI)	p-value	Mean score difference (95%CI)	p-value	Mean score difference (95%CI)	p-value	
	Positive values indicate better functioning			Negative values favor Group A			Negative values favor Group A		
-Physical	5.5 <sup>a</sup> (1.9 to 9.1)	0.002	1.1 (-3.0 to 5.2)	0.596	-4.4 (-9.5 to 0.8)	0.095			
-Role	10.2 <sup>a</sup> (3.7 to 16.7)	0.002	6.9 <sup>a</sup> (-0.5 to 14.3)	0.067	-3.3 (-12.6 to 5.9)	0.483			
-Social	8.1 <sup>a</sup> (3.3 to 12.8)	0.001	3.0 <sup>a</sup> (-2.3 to 8.3)	0.266	<b>-5.1</b> (-12.0 to 1.9)	0.157			
-Emotional	0.1 (-2.9 to 3.1)	0.933	0.3 (-3.4 to 3.9)	0.895	0.1 (-4.6 to 4.8)	0.960			
-Cognitive	1.6 (-2.2 to 5.4)	0.402	-1.5 (-5.1 to 2.1)	0.415	<b>-3.1</b> (-8.2 to 2.0)	0.234			
-Global Health State/Quality of Life	0.29 (-5.60 to 6.19)	0.923	0.31 (-6.02 to 6.65)	0.923	0.02 (-8.45 to 8.49)	0.996			
<b>QLQ-C30 Symptom scales</b>	Negative values indicate improvements			Positive values favor Group A					
-Appetite loss	-9.9 <sup>a</sup> (-15.7 to -4.1)	0.001	-2.6 (-8.6 to 3.5)	0.409	<b>7.4</b> (0.1 to 14.6)	0.048			
-Constipation	-1.5 (-5.8 to 2.9)	0.512	0.76 (-3.9 to 5.5)	0.753	2.2 (-3.8 to 8.3)	0.471			
-Diarrhoea	-4.6 <sup>a</sup> (-10.6 to 1.5)	0.137	0.11 (-6.1 to 6.3)	0.973	<b>4.7</b> (-2.9 to 12.2)	0.225			
-Dyspnoea	-5.7 <sup>a</sup> (-11.3 to -0.2)	0.041	-2.1 <sup>a</sup> (-8.4 to 4.2)	0.518	3.7 (-3.9 to 11.2)	0.343			
-Fatigue	-7.9 <sup>a</sup> (-12.6 to -3.2)	0.001	-2.3 (-7.7 to 3.1)	0.409	<b>5.6</b> (-1.4 to 12.6)	0.118			
-Pain	0.35 (-5.3 to 5.9)	0.903	0.9 (-4.5 to 6.5)	0.726	0.6 (-6.8 to 8.1)	0.867			
-Insomnia	-2.1 (-7.1 to 2.9)	0.402	2.7 (-3.3 to 8.6)	0.385	<b>4.9</b> (-2.6 to 12.2)	0.205			
-Nausea and vomiting	-4.7 <sup>a</sup> (-8.9 to -0.6)	0.025	-0.0 (-4.2 to 4.1)	0.983	<b>4.7</b> (-0.2 to 9.5)	0.058			
<b>QLQ-MY20 domains</b>	Positive values indicate improvements			Negative values favor Group A					
-Body image	12.33 (5.34-19.31)	0.001	6.46 (-0.11-13.02)	0.054	-5.87 (-15.06 to 3.32)	0.210			
-Future perspectives	6.95 (2.53 to 11.37)	0.002	6.83 (2.64 to 11.02)	0.001	-0.12 (-5.93 to 5.68)	0.967			
	Negative values indicate improvements			Positive values favor Group A					
-Disease symptoms	1.41 (-1.62 to 4.45)	0.362	2.75 (-0.27 to 5.78)	0.074	1.34 (-2.77 to 5.45)	0.523			
-Side effects	-6.35 (-9.11 to -3.59)	0.001	-5.13 (-8.24 to -2.02)	0.001	1.22 (-2.85 to 5.28)	0.557			

**QLQ-C30** quality of life questionnaire-core 30; **QLQ-MY20** quality of life questionnaire-multiple myeloma module, **CI** Confidence interval, <sup>a</sup> Within group minimal important change score achieved. The point estimates marked with bold are the ones reaching the threshold for between group small minimal important differences (MID). The point estimates both bold and italic are at the lower range of small MID, thus close to trivial effects. Threshold for significance was p<0.01.

**The proportion of patients that improved, remained stable or worsened in the EORTC QLQ-C30 functional domains (except Cognitive domain) and Global health scale/Quality of life and the Body image domain of QLQ-MY20 from randomisation to eight months follow-up (SAP analysis 6.3.3.2)**

There was no statistical significant between-group difference in the proportion of patients that improved, remained stable or worsened from randomisation to the eight months follow-up, nor in any of the other assessment time points (two, four and six months follow-up) (Table 3).

**Table 3. The proportion of patients who improved, remained stable or worsened in the QLQ-C30 Global health scale/Quality of life and functional domains (except Cognitive functioning) and in the QLQ-MY20 Body image domain from randomisation to two, four, six and eight months follow-up.**

QLQ-C30 domain	Direction	Follow-up (months)	Group A (%)	Group B (%)	Odds ratio (95%CI)	P-value
<b>-Global health scale/Quality of life</b>						
	Improved	2	9.7	11.3	1.18 (0.40-3.44)	0.763
	Improved	4	8.6	11.4	1.38 (0.45-4.20)	0.573
	Improved	6	5.9	15.5	2.93 (0.89-9.71)	0.068
	Improved	8	14.1	11.9	0.83 (0.30-2.30)	0.718
	Remained stable	2	83.3	83.1	0.98 (0.41-2.36)	0.970
	Remained stable	4	82.9	85.7	1.24 (0.50-3.09)	0.642
	Remained stable	6	85.3	76.1	0.55 (0.23-1.30)	0.169
	Remained stable	8	71.9	71.6	0.99 (0.46-2.12)	0.976
	Worsened	2	6.9	5.6	0.80 (0.21-3.11)	0.747
	Worsened	4	8.6	2.9	0.31 (0.06-1.61)	0.145
	Worsened	6	8.8	8.5	0.95 (0.29-3.12)	0.938
	Worsened	8	14.1	16.4	1.20 (0.46-3.12)	0.708
<b>-Physical functioning</b>						
	Improved	2	6.7	11.3	1.78 (0.55-5.72)	0.329
	Improved	4	12.3	5.7	0.43 (0.13-1.47)	0.169
	Improved	6	11.3	7.0	0.60 (0.19-1.92)	0.383
	Improved	8	12.1	8.8	0.70 (0.23-2.14)	0.533
	Remained stable	2	89.3	88.7	0.94 (0.33-2.66)	0.908
	Remained stable	4	80.8	92.9	3.08 (1.05-9.09)	0.034
	Remained stable	6	81.7	90.1	2.05 (0.77-5.49)	0.148
	Remained stable	8	83.3	86.8	1.31 (0.50-3.41)	0.577
	Worsened	2	4.0	0.0	1.00 (-.)	0.089
	Worsened	4	6.8	1.4	0.20 (0.02-1.73)	0.106
	Worsened	6	7.0	2.8	0.38 (0.07-2.04)	0.245
	Worsened	8	4.5	4.4	0.97 (0.19-4.98)	0.970
<b>-Role functioning</b>						
	Improved	2	35.1	26.8	0.67 (0.33-1.37)	0.276
	Improved	4	32.9	28.6	0.82 (0.40-1.66)	0.577
	Improved	6	33.8	23.9	0.62 (0.30-1.28)	0.195
	Improved	8	30.3	25.0	0.77 (0.36-1.64)	0.492
	Remained stable	2	54.1	60.6	1.31 (0.67-2.53)	0.428
	Remained stable	4	61.6	68.6	1.36 (0.68-2.71)	0.385
	Remained stable	6	56.3	64.8	1.43 (0.73-2.80)	0.303
	Remained stable	8	60.6	63.2	1.12 (0.56-2.25)	0.754
	Worsened	2	10.8	12.7	1.20 (0.43-3.30)	0.727
	Worsened	4	5.5	2.9	0.51 (0.09-2.86)	0.434

**Table 3. The proportion of patients who improved, remained stable or worsened in the QLQ-C30 Global health scale/Quality of life and functional domains (except Cognitive functioning) and in the QLQ-MY20 Body image domain from randomisation to two, four, six and eight months follow-up.**

QLQ-C30 domain	Direction	Follow-up (months)	Group A (%)	Group B (%)	Odds ratio (95%CI)	P-value
-Role functioning	Worsened	6	9.9	11.3	1.16 (0.40-3.39)	0.785
	Worsened	8	9.1	11.8	1.33 (0.44-4.08)	0.613
-Emotional functioning	Improved	2	9.6	7.0	0.71 (0.22-2.37)	0.580
	Improved	4	9.7	5.7	0.56 (0.16-2.01)	0.372
	Improved	6	5.7	9.9	1.80 (0.50-6.46)	0.359
	Improved	8	4.6	13.4	3.21 (0.83-12.43)	0.078
	Remained stable	2	87.7	90.1	1.29 (0.45-3.66)	0.637
	Remained stable	4	87.5	87.1	0.97 (0.36-2.60)	0.949
	Remained stable	6	91.4	78.9	0.35 (0.13-0.96)	0.036
	Remained stable	8	87.7	79.1	0.53 (0.21-1.37)	0.186
	Worsened	2	2.7	2.8	1.03 (0.14-7.51)	0.978
	Worsened	4	2.8	7.1	2.69 (0.50-14.36)	0.230
	Worsened	6	2.9	11.3	4.32 (0.88-21.11)	0.052
	Worsened	8	7.7	7.5	0.97 (0.27-3.51)	0.960
-Social functioning	Improved	2	19.2	21.1	1.13 (0.50-2.55)	0.771
	Improved	4	20.8	15.7	0.71 (0.30-1.67)	0.430
	Improved	6	18.6	26.8	1.60 (0.72-3.56)	0.246
	Improved	8	20.0	20.9	1.06 (0.45-2.46)	0.899
	Remained stable	2	74.0	76.1	1.12 (0.53-2.38)	0.773
	Remained stable	4	76.4	75.7	0.96 (0.45-2.08)	0.925
	Remained stable	6	77.1	67.6	0.62 (0.29-1.31)	0.206
	Remained stable	8	75.4	65.7	0.62 (0.29-1.33)	0.221
	Worsened	2	6.8	2.8	0.39 (0.07-2.10)	0.261
	Worsened	4	2.8	8.6	3.28 (0.64-16.84)	0.134
	Worsened	6	4.3	5.6	1.33 (0.29-6.19)	0.713
	Worsened	8	4.6	13.4	3.21 (0.83-12.43)	0.078
<b>QLQ-MY20 domain</b>						
-Body image	Improved	2	29.6	34.8	1.27 (0.62-2.58)	0.510
	Improved	4	28.6	34.3	1.30 (0.64-2.67)	0.466
	Improved	6	32.4	30.3	0.91 (0.44-1.89)	0.798
	Improved	8	36.5	33.3	0.87 (0.42-1.79)	0.705
	Remained stable	2	66.2	55.1	0.63 (0.32-1.24)	0.178
	Remained stable	4	64.3	57.1	0.74 (0.38-1.46)	0.387
	Remained stable	6	61.8	62.1	1.02 (0.51-2.04)	0.966
	Remained stable	8	50.8	50.0	0.97 (0.49-1.93)	0.928
	Worsened	2	4.2	10.1	2.56 (0.63-10.33)	0.174
	Worsened	4	7.1	8.6	1.22 (0.35-4.19)	0.753
	Worsened	6	5.9	7.6	1.31 (0.34-5.11)	0.695
	Worsened	8	12.7	16.7	1.38 (0.51-3.68)	0.525

**QLQ-C30** quality of life questionnaire-core 30; **QLQ-MY20** quality of life questionnaire-multiple myeloma module, **CI** Confidence interval. Each time point (follow-up) were compared to baseline level. Threshold for significance was  $p < 0.01$ .



***Time to first recorded improvement in the functional domains and Global health scale/Quality of life scale of the QLQ-C30 and the QLQ-MY20 Body image domain (SAP analysis 6.3.3.3)***

There was no statistical significant between-group difference in time to first recorded improvement in any of the domains.

However, numerical trends in favour of Group A was seen for the QLQ-C30 physical, role, emotional and social functioning domains and the QLQ-MY20 body image domain with time to first clinically relevant improvement occurring faster for these domains (Hazard ratio (HR) (95%CI); range 0.72-0.90 (0.28-2.15)). Contrary, a numerical trend in favour of Group B was seen for the global health scale/quality of life domain with time to first clinically relevant improvement occurring faster for this domain (HR ratio (95%CI); 1.18 (0.57-2.43)) (Table 4).

**Table 4. Time to first recorded improvement in the functional domains and Global health scale/Quality of life of the EORTC QLQ-C30 and the EORTC QLQ-MY20 Body image domain.**

QLQ-C30 domains	Group A		Group B		Hazard ratio (95%CI)	P-value
	Patients included in analysis	Improved, n (%)	Patients included in analysis	Improved, n (%)		
-Global health scale/Quality of life	46	16 (35)	52	19 (37)	1.18 (0.57-2.43)	0.655
-Physical functioning	43	13 (30)	40	10 (25%)	0.72 (0.28-1.87)	0.502
-Role functioning	53	36 (68)	48	30 (63)	0.88 (0.52-1.50)	0.635
-Emotional functioning	22	9 (41)	32	11 (34)	0.82 (0.31-2.15)	0.685
-Social functioning	40	23 (58)	39	21 (54)	0.90 (0.47-1.69)	0.735
-Cognitive functioning	22	9 (41)	18	5 (28)	1.00 (0.33-3.06)	0.994
<b>QLQ-MY20 domain</b>						
-Body image	42	32 (76)	48	31 (65)	0.84 (0.50-1.43)	0.521

**QLQ-C30** quality of life questionnaire-core 30; **QLQ-MY20** quality of life questionnaire-multiple myeloma module, **CI** Confidence interval. Patients with no baseline score and patients with high functioning as well as good global health scale/Quality of life and body image score at randomization were excluded from this analysis since high/good baseline-scores leaves no room for improvement. Threshold for significance was  $p < 0.01$ .

***The proportion of patients with at least mild Kd-related symptoms and the proportion of patients with moderate to severe patient-reported symptomatic side effects (SAP analysis 6.3.3.4)***

There was no statistical significant between-group difference in proportion of patients with mild, nor moderate to severe, Kd-related symptoms at the eight months follow-up from randomisation.

Likewise, there was no statistical significant between-group difference in any of the Kd-related symptoms at the other assessment time points (2, 4 and 6 months from baseline), except for

symptoms of restlessness and agitation two months from randomisation, where 25% of patients in Group A reported to have "at least mild" symptoms of restlessness and agitation compared to 46% of patients in Group B (Odds ratio (95%CI) 0.12 (0.04-0.39), p= 0.006) (Table 5).

Generally, few patients (range 0-10%) in both groups experienced moderate to severe Kd-related symptoms throughout the four assessment time points (two, four, six and eight months from randomisation) (Table 5).

**Table 5. The proportion of patients with at least mild, respectively moderate to severe patient-reported Kd-related symptomatic side effects**

QLQ-C30	Severity of symptoms	Follow-up (months)	Group A (%)	Group B (%)	Odds ratio (95%CI)	P-value
<b>Item 8. Where you short of breath?</b>						
	At least mild	0	59	54	1.76 (0.64; 4.84)	0.529
	At least mild	2	47	48	0.88 (0.33; 2.40)	0.950
	At least mild	4	37	49	0.36 (0.13; 1.00)	0.132
	At least mild	6	42	49	0.56 (0.20; 1.55)	0.421
	At least mild	8	44	46	0.69 (0.24; 1.99)	0.747
	Moderate to severe	0	3	1	0.05 (0.00; 1.61)	0.575
	Moderate to severe	2	1	0	0.01 (0.00; 0.09)	0.332
	Moderate to severe	4	1	1	0.01 (0.00; 1.00)	0.962
	Moderate to severe	6	0	0	-	-
	Moderate to severe	8	1	3	0.01 (0.00; 0.45)	0.543
<b>Item 10. Did you need to rest?</b>						
	At least mild	0	83	82	5.47 (1.46; 20.55)	0.832
	At least mild	2	75	80	2.18 (0.67; 7.02)	0.440
	At least mild	4	71	71	2.31 (0.77; 6.92)	0.922
	At least mild	6	76	81	2.42 (0.72; 8.10)	0.476
	At least mild	8	73	74	2.59 (0.80; 8.43)	0.888
	Moderate to severe	0	5	4	0.07 (0.01; 0.74)	0.699
	Moderate to severe	2	1	3	0.01 (0.00; 0.37)	0.490
	Moderate to severe	4	1	3	0.01 (0.00; 0.39)	0.521
	Moderate to severe	6	3	5	0.01 (0.00; 0.25)	0.386
	Moderate to severe	8	1	4	0.00 (0.00; 0.28)	0.304
<b>Item 11. Have you had trouble sleeping?</b>						
	At least mild	0	54	53	1.23 (0.45; 3.32)	0.873
	At least mild	2	55	53	1.32 (0.48; 3.59)	0.834
	At least mild	4	47	60	0.54 (0.20; 1.47)	0.114
	At least mild	6	47	55	0.62 (0.22; 1.70)	0.286
	At least mild	8	46	54	0.64 (0.22; 1.81)	0.354
	Moderate to severe	0	3	3	0.03 (0.00; 0.61)	1.000
	Moderate to severe	2	3	7	0.01 (0.00; 0.18)	0.211
	Moderate to severe	4	4	5	0.03 (0.00; 0.34)	0.633
	Moderate to severe	6	5	1	0.23 (0.01; 3.89)	0.177
	Moderate to severe	8	1	3	0.01 (0.00; 0.46)	0.551

**Table 5. The proportion of patients with at least mild, respectively moderate to severe patient-reported Kd-related symptomatic side effects**

QLQ-C30	Severity of symptoms	Follow-up (months)	Group A (%)	Group B (%)	Odds ratio (95%CI)	P-value
<b>Item 12. Have you felt week?</b>						
	At least mild	0	74	71	3.40 (1.10; 10.50)	0.625
	At least mild	2	61	56	1.86 (0.67; 5.14)	0.583
	At least mild	4	51	53	0.97 (0.36; 2.63)	0.792
	At least mild	6	53	50	1.23 (0.45; 3.39)	0.747
	At least mild	8	56	54	1.40 (0.49; 4.00)	0.806
	Moderate to severe	0	3	8	0.01 (0.00; 0.16)	0.152
	Moderate to severe	2	3	0	0.03 (0.01; 0.11)	0.177
	Moderate to severe	4	1	0	0.01 (0.00; 0.09)	0.332
	Moderate to severe	6	4	7	0.02 (0.00; 0.28)	0.444
	Moderate to severe	8	1	4	0.00 (0.00; 0.28)	0.304
<b>Item 14. Have you felt nauseated?</b>						
	At least mild	0	24	32	0.22 (0.07; 0.70)	0.318
	At least mild	2	16	21	0.15 (0.04; 0.55)	0.489
	At least mild	4	12	18	0.08 (0.02; 0.36)	0.275
	At least mild	6	20	25	0.18 (0.05; 0.64)	0.470
	At least mild	8	20	23	0.20 (0.05; 0.73)	0.617
	Moderate to severe	0	3	0	0.03 (0.01; 0.11)	0.160
	Moderate to severe	2	0	4	0.04 (0.01; 0.14)	0.067
	Moderate to severe	4	0	3	0.03 (0.01; 0.11)	0.141
	Moderate to severe	6	0	1	0.01 (0.00; 0.10)	0.306
	Moderate to severe	8	0	0	-	-
<b>Item 15. Have you vomited?</b>						
	At least mild	0	12	8	0.20 (0.04; 1.03)	0.415
	At least mild	2	8	10	0.06 (0.01; 0.41)	0.659
	At least mild	4	8	5	0.12 (0.02; 0.84)	0.585
	At least mild	6	5	7	0.04 (0.00; 0.39)	0.700
	At least mild	8	6	13	0.02 (0.00; 0.20)	0.138
	Moderate to severe	0	1	0	0.01 (0.00; 0.09)	0.316
	Moderate to severe	2	0	3	0.03 (0.01; 0.12)	0.138
	Moderate to severe	4	0	0	-	-
	Moderate to severe	6	0	0	-	-
	Moderate to severe	8	0	1	0.01 (0.00; 0.10)	0.130
<b>Item 17. Have you had diarrhoea?</b>						
	At least mild	0	41	42	0.65 (0.23; 1.79)	0.848
	At least mild	2	36	30	0.73 (0.25; 2.08)	0.452
	At least mild	4	32	31	0.51 (0.17; 1.48)	0.844
	At least mild	6	33	22	0.87 (0.29; 2.63)	0.121
	At least mild	8	37	36	0.60 (0.20; 1.77)	0.911
	Moderate to severe	0	1	3	0.01 (0.00; 0.45)	0.575
	Moderate to severe	2	0	0	-	-
	Moderate to severe	4	0	0	-	-
	Moderate to severe	6	1	0	0.01 (0.00; 0.10)	0.322
	Moderate to severe	8	0	3	0.03 (0.01; 0.12)	0.151

**Table 5. The proportion of patients with at least mild, respectively moderate to severe patient-reported Kd-related symptomatic side effects**

QLQ-C30	Severity of symptoms	Follow-up (months)	Group A (%)	Group B (%)	Odds ratio (95%CI)	P-value
<b>Item 18. Where you tired?</b>						
	At least mild	0	76	78	2.89 (0.88; 9.48)	0.780
	At least mild	2	71	77	1.90 (0.61; 5.86)	0.461
	At least mild	4	64	71	1.29 (0.45; 3.70)	0.350
	At least mild	6	68	73	1.74 (0.58; 5.20)	0.540
	At least mild	8	72	75	2.13 (0.66; 6.85)	0.636
	Moderate to severe	0	4	9	0.02 (0.00; 0.20)	0.206
	Moderate to severe	2	1	5	0.00 (0.00; 0.18)	0.154
	Moderate to severe	4	1	4	0.00 (0.00; 0.24)	0.280
	Moderate to severe	6	3	7	0.01 (0.00; 0.19)	0.231
	Moderate to severe	8	4	10	0.02 (0.00; 0.20)	0.174
<b>QLQ-MY20</b>						
<b>Item 44. Did you feel restless or agitated?</b>						
	At least mild	0	28%	41%	0.22 (0.07; 0.68)	0.098
	At least mild	2	25%	46%	0.12 (0.04; 0.39)	<b>0.006</b>
	At least mild	4	27%	36%	0.24 (0.07; 0.75)	0.229
	At least mild	6	29%	40%	0.26 (0.08; 0.79)	0.189
	At least mild	8	28%	37%	0.27 (0.08; 0.87)	0.294
	Moderate to severe	0	1%	1%	0.01 (0.00; 1.16)	0.992
	Moderate to severe	2	0%	1%	0.01 (0.00; 0.11)	0.289
	Moderate to severe	4	0%	1%	0.01 (0.00; 0.10)	0.319
	Moderate to severe	6	0%	0%	-	-
	Moderate to severe	8	0%	2%	0.02 (0.00; 0.11)	0.308

**QLQ-C30** quality of life questionnaire-core 30; **QLQ-MY20** quality of life questionnaire-multiple myeloma module, **CI** Confidence interval. The QLQ-C30 and QLQ-MY20 scale score ranges from 0-100 with the response options being “Not at all” (<33 points), “A little” (≥33 points) and “Quite a bit” to “Very much” (≥67 points). The proportion of patients reporting any Kd-related symptoms corresponds to “at least a little”, thus a score ranging between ≥33 to 100 points, indicated as “at least mild” severity of symptoms. The proportion of patients reporting moderate or severe Kd-related symptoms corresponds to answers of “quite a bit” and “very much”, thus a score of ≥67 on a scale of 0-100 points, indicated as “moderate to severe” severity of symptoms. Threshold for significance was p<0.01.

#### ***Time in quality adjusted progression free period (SAP analysis 6.3.3.5)***

There was no statistical significant between group difference in time in quality adjusted progression free period at the eight months follow-up from randomisation. The quality adjusted progression free period was approximately seven months (range 218 to 221 days) for both groups (mean difference (days); 95%CI for EORTC Summary Score: 2.4; -16.1 to 20.9, p-value=0.80; for QLUC10D: -0.5; -22.6 to 21.6, p=value 0.97) (**Table 6**).

The time in quality adjusted progression free period for the full study period (including all visits until progression, withdrawal, end of study) could not be presented here as part of this blinded review

and interpretation document. The reason was that the primary endpoint in the CARFI trial (time to progression) showed that Kd maintenance therapy post salvage ASCT significantly prolonged the time to progression by approximately eight months compared to observation only. As we found no between group differences in the EORTC Summary Score at the 8 months follow-up and that quality adjusted progression free period is calculated as the product of the EORTC Summary score and the time to progression there was a risk that the quality adjusted progression free period for the full study period would reveal the blinding of Group A and Group B. However, as with PRO questionnaire compliance, which were also not revealed to the author group to keep the blinding, the quality adjusted progression free period for the full study period will be calculated and taken into consideration for the overall interpretation of the study results after the blinding is broken.

**Table 6. Time in quality-adjusted progression free period up to eight months follow-up from randomisation.**

Quality-adjustment score	Quality Adjusted Progression Free Period			p-value
	Group A	Group B	Between groups	
	Mean score (days; 95%CI)		Mean score difference (days; 95%CI)	
-QLQ-C30 Summary Score	218.5 (204.8 to 232.1)	220.9 (212.3 to 229.5)	2.43 (-16.1 to 20.9)	0.797
-QLU-C10D Utility score	219.84 (203.4 to 236.3)	219.3 (210.1 to 228.3)	-0.50 (-22.6 to 21.6)	0.965

**QLQ-C30** quality of life questionnaire-core 30; **QLU-C10D** Quality of Life Utility Measure-Core 10 dimensions. Threshold for significance was  $p < 0.01$ .

## Interpretation

Two interpretations were conducted based on the assumption that Group A received Kd maintenance therapy and the other version assuming that Group A received observation alone.

### Interpretation 1: “Group A received Kd maintenance therapy”

The hypothesis was that Kd maintenance therapy following salvage autologous stem cell support (ASCT) would be well tolerated and not adversely impact HRQL compared to observation alone (in patients with relapsed MM).

The primary finding of the CARFI trial showed that the time to progression was significantly prolonged for the group receiving Kd maintenance (Gregersen H 2021). The PRO results provide an important complement to that, showing that the progression-free survival benefit conferred by Kd maintenance did not come at the cost of adverse impact of overall HRQL relative to observation alone, as evaluated by the HRQL summary score, EORTC QLQ-C30 Sum Sc. Further, results for the individual functioning and symptom scales of the QLQ-C30 showed that Kd maintenance results in clinically relevant improvements with regards to less symptoms of diarrhoea, nausea and vomiting, insomnia, fatigue and appetite loss as well as better cognitive and social functioning across the eight months period from randomisation compared to observation only. The main reason for these

between groups differences were that the group allocated to Kd maintenance experienced clinical meaningful improvements in most domains whereas the group allocated to observation alone remained stable in most domains.

With regards to the other secondary PRO endpoints, 1) proportion of patients that improved, remained stable or worsened in QLQ-C30 and QLQ-MY20 specific domains; 2) time to first recorded improvements in QLQ-C30 and QLQ-MY20 specific domains; 3) proportion of patients with at least mild or moderate to severe Kd-related symptoms and 4) quality-adjusted progression free period, none reached statistical significant between-group differences, except for the proportion of patients with at least mild Kd-related symptoms of restlessness and agitation at the two months follow-up from randomisation, in favour of the group receiving Kd maintenance.

This is the first study to investigate the PRO effects of Kd as maintenance following salvage ASCT on HRQL in patients with relapsed MM. The study provides evidence that Kd maintenance do not impair recovery from salvage ASCT in terms of HRQL. In fact, the group receiving Kd maintenance appears to experience greater relief of numerous symptoms (diarrhoea, nausea and vomiting, fatigue, dyspnoea and appetite loss) and improvement in various aspects of functioning (physical, role and social) throughout the period of recovery from salvage ASCT compared to observation alone. Further, a significantly smaller proportion of patients in the group receiving Kd maintenance experienced symptoms of restlessness and agitation at the two months follow-up after randomisation. However, this may have been a chance finding, given the many statistical comparisons conducted.

The current results showing clinically relevant improvements in appetite loss eight months after salvage ASCT are in line with a previous study investigating the effect of salvage ASCT in relapsed MM patients (Ahmedzai et al. 2019). Although that study did not include maintenance therapy as an intervention following salvage ASCT both studies (ours and the Myeloma X trial by Ahmedzai and colleagues, 2019) show that appetite loss recover following salvage ASCT, regardless of whether Kd maintenance is provided or not. The current finding of improvements in nausea and vomiting are identical to the findings of the ENDEAVOUR trial investigating Kd-based therapy in relapsed/refractory MM patients, not receiving salvage ASCT (Ludwig et al. 2019). That study also showed a statistical, however not clinically meaningful, improvement in nausea and vomiting from start of treatment. However, as nausea and vomiting is known to be one of the most frequent side effects to Kd-based regimens (Siegel et al. 2013), the reason for improvements in nausea and vomiting following ASCT and/or Kd-based therapy compared to observation alone, as in our study, remains to be investigated.

In conclusion, Kd maintenance does not impair recovery of HRQL after salvage ASCT and provides equal duration of good quality of life (quality adjusted progression free period) eight months from randomisation compared to observation alone. Further, Kd maintenance appears to provide greater symptom relief and improved functioning in certain aspects of HRQL. Thus, this trial supports the use of Kd-based maintenance therapy following salvage ASCT in relapsed MM patients. In conjunction with the significant prolonged time to progression for the group allocated to Kd maintenance (Gregersen H 2021), this treatment should thus be considered following salvage ASCT. Of note though, as Kd maintenance requires continued treatment visits at the hospital, individual situations

and clinical status of the patients should be taken into consideration when choosing treatment pathways after salvage ASCT in this patient group.

### **Interpretation 2: “Group A received observation alone”**

The hypothesis was that Kd maintenance therapy following salvage ASCT would be well tolerated and not adversely impact HRQL compared to observation alone in patients with relapsed MM.

The primary finding of the CARFI trial showed that time to progression was significantly prolonged for the group receiving Kd maintenance (Gregersen H 2021). The PRO results provide an important complement to that, showing that the progression-free survival benefit conferred by Kd maintenance therapy did not come at the cost of adverse impact of overall HRQL relative to observation alone, as evaluated by the HRQL summary score, EORTC QLQ-C30 Sum Sc. However, our study shows that observation alone improves the recovery of certain symptoms (dyspnoea, nausea and vomiting, fatigue, diarrhoea and appetite loss) and aspects of functioning (physical, role and social) at clinical relevant levels compared with patients receiving Kd maintenance following salvage ASCT. Furthermore, a significantly smaller proportion of patients receiving observation alone experienced symptoms of restlessness and agitation compared to patients receiving Kd maintenance at the two months follow-up from randomisation. Thus, the results indicate that patients receiving Kd maintenance after salvage ASCT on average have an inferior recovery in certain aspects of symptom and functioning scales, but not in overall HRQL or quality adjusted progression free period eight months from randomisation.

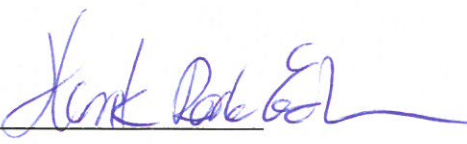
The current finding that functioning and symptom scales recover following salvage ASCT in relapsed MM receiving observation alone is in line with another study showing alone (Ahmedzai et al. 2019) that most aspects of HRQL, as assessed by the QLQ-C30 questionnaire, recovered to pre-salvage ASCT levels two to three months post salvage ASCT.

In conclusion, observation alone does not provide better overall HRQL and quality adjusted progression free period than Kd maintenance following salvage ASCT. However, as observation alone boosted the recovery of certain aspects of functioning and symptoms following salvage ASCT, and as observation alone may be less burdensome to patients due to fewer hospital visits for treatment administration, but in contrast leads to a significant shorter time to progression compared to Kd maintenance, our findings support the relevance of shared decision making for relapsed MM patients that have undergone salvage ASCT.

## Signatures

Henrik Rode Eshoj,

Date, signature: \_\_\_\_\_

11-9-2021 

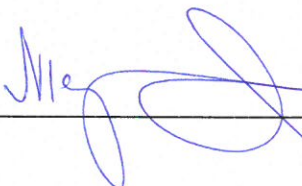
Madeleine T King,

Date, signature: \_\_\_\_\_

31 August 2021, Madeleine T King

Niels Abildgaard,

Date, signature: \_\_\_\_\_

1/9-2021 

Lene Kongsgaard Nielsen,

Date, signature: \_\_\_\_\_

31. August 2021, Lene Kongsgaard Nielsen

Henrik Gregersen,

Date, signature \_\_\_\_\_

31 August 2021, Henrik Gregersen



## References

- Ahmedzai, S. H., J. A. Snowden, A. J. Ashcroft, D. A. Cairns, C. Williams, A. Hockaday, J. D. Cavenagh, D. Ademokun, E. Tholouli, D. Allotey, V. Dhanapal, M. Jenner, K. Yong, J. Cavet, H. Hunter, J. M. Bird, G. Pratt, C. Parrish, J. M. Brown, T. C. M. Morris, G. Cook, and Group National Cancer Research Institute Haemato-Oncology Clinical Studies. 2019. 'Patient-Reported Outcome Results From the Open-Label, Randomized Phase III Myeloma X Trial Evaluating Salvage Autologous Stem-Cell Transplantation in Relapsed Multiple Myeloma', *J Clin Oncol*, 37: 1617-28.
- Bell, M. L., and D. L. Fairclough. 2014. 'Practical and statistical issues in missing data for longitudinal patient-reported outcomes', *Stat Methods Med Res*, 23: 440-59.
- Eshoj HR, King MT, Abildgaard N, Gregersen H, Nielsen LK, Möller S. 2020. 'Statistical analysis plan for patient-reported outcomes of the CARFI trial: a multi-centre randomized controlled trial of Carfilzomib-cyclophosphamide-dexamethasone and High-dose Melphalan (HDT) followed by Randomization between Observation or Maintenance with Carfilzomib and Dexamethasone in Patients with Relapsed Multiple Myeloma after HDT'.
- Gotzsche, P. C. 1996. 'Blinding during data analysis and writing of manuscripts', *Control Clin Trials*, 17: 285-90; discussion 90-3.
- Gregersen H, et al. 2021. "A Randomized Phase 2 Trial Comparing Carfilzomib-Dexamethasone Vs Observation As Maintenance after Induction with Carfilzomib-Cyclophosphamide-Dexamethasone in Salvage ASCT in Multiple Myeloma: A Trial By the Nordic Myeloma Study Group." In *ASCO*, 601-06. Blood.
- Jarvinen, T. L., R. Sihvonen, M. Bhandari, S. Sprague, A. Malmivaara, M. Paavola, H. J. Schunemann, and G. H. Guyatt. 2014. 'Blinded interpretation of study results can feasibly and effectively diminish interpretation bias', *J Clin Epidemiol*, 67: 769-72.
- Ludwig, H., P. Moreau, M. A. Dimopoulos, M. V. Mateos, M. Kaiser, R. Hajek, S. Feng, K. Cocks, J. Buchanan, and K. Weisel. 2019. 'Health-related quality of life in the ENDEAVOR study: carfilzomib-dexamethasone vs bortezomib-dexamethasone in relapsed/refractory multiple myeloma', *Blood Cancer J*, 9: 23.
- Siegel, D., T. Martin, A. Nooka, R. D. Harvey, R. Vij, R. Niesvizky, A. Z. Badros, S. Jagannath, L. McCulloch, K. Rajangam, and S. Lonial. 2013. 'Integrated safety profile of single-agent carfilzomib: experience from 526 patients enrolled in 4 phase II clinical studies', *Haematologica*, 98: 1753-61.
- White, I. R., N. J. Horton, J. Carpenter, and S. J. Pocock. 2011. 'Strategy for intention to treat analysis in randomised trials with missing outcome data', *BMJ*, 342: d40.