Methodological aspects of health-related quality of life measurement and analysis in patients with multiple myeloma

Kongsgaard Nielsen, Lene; Abildgaard, Niels; Jarden, Mary Ellen; Klausen, Tobias Wirenfeldt

Published in:
British Journal of Haematology

DOI:
10.1111/bjh.15759

Publication date:
2019

Document version
Accepted manuscript

Citation for published version (APA):

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:

• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 18. Apr. 2021
Methodological aspects of health-related quality of life measurement and analysis in patients with multiple myeloma

Lene Kongsgaard Nielsen1,2, Niels Abildgaard1,2, Mary Jarden3 and Tobias Wirenfld Klausen4

1Quality of Life Research Centre, Department of Haematology, Odense University Hospital, Denmark; 2The Academy of Geriatric Cancer Research (AgeCare), Odense University Hospital, Denmark; 3Department of Haematology, Copenhagen University Hospital, Copenhagen, Denmark; 4Department of Haematology, Herlev Hospital, Denmark

Corresponding author: Lene Kongsgaard Nielsen, Quality of Life Research Centre, Department of Haematology, Odense University Hospital, Klovverænget 10, 12. Floor, 5000 Odense, Denmark, email lene.kongsgaard.nielsen@rsyd.dk phone: +45 2324 1605.

Summary
Multiple myeloma (MM) is an incurable but treatment-sensitive cancer. For most patients, this means treatment with multiple lines of anti-myeloma therapy and a life with disease- and treatment-related symptoms and complications.
Health-related quality of life (HRQoL) issues play an important role in treatment decision-making. Methodological challenges in longitudinal HRQoL measurements and analyses have been identified, including non-responses (NR) to scheduled questionnaires. Publications were identified for inclusion in a systematic review of longitudinal HRQoL studies in MM, focussing on methodological aspects of HRQoL measurement and analysis. Diversity in timing of HRQoL data collection and applied statistical methods were noted. We observed a high rate of NR, but the impact of NR was investigated in only 8/23 studies. Thus, evidence-based knowledge of HRQoL in patients with MM is compromised. To improve quality of HRQoL results and their implementation in daily practice, future studies should follow established guidelines.

**Keywords** Health-related Quality of Life, Missing data, statistical analysis, review, multiple myeloma.

**Introduction**

Health-related quality of life (HRQoL) and other Patient Reported Outcomes (PRO) have become increasingly used as endpoints in clinical cancer studies to measure patient-experienced benefits and toxicities of treatments (Basch, *et al* 2016, Vodicka, *et al* 2015). PRO results are important in the approval of new drugs, as well as in shared decision-making in the daily care of patients (European Medicines Agency 2016, Speight and Barendse 2010).

Multiple myeloma (MM) is an incurable malignancy derived from plasma cells in the bone marrow. It is the second most common haematological cancer, and worldwide, it is estimated that 86,000 patients annually are diagnosed with MM (Becker 2011). The prognosis of MM has improved markedly over the past 20 years and it is expected to improve further in the coming years due to new treatment options (Kumar, *et al* 2014, Kumar, *et al* 2008). The median survival of MM patients under the age of 70 years at the time of diagnosis now exceeds 6-7 years (Kumar, *et al* 2014). MM is associated with severe morbidity caused by bone destruction/bone fractures, renal dysfunction, bone marrow failure, high infection rates and potential physical disability (Kyle and Rajkumar 2008, Rajkumar, *et al* 2014).

Current treatment of newly diagnosed MM patients involves induction treatment with repeated cycles of two or three drug combinations followed by stem cell harvest and high dose chemotherapy with stem cell support (HDT) in younger, eligible patients (Laubach, *et al* 2016, Lenhoff, *et al* 2000, Mateos, *et al* 2014). After HDT, patients are treated with consolidation and/or long-term maintenance treatment or enter a drug-free period (McCarthy, *et al* 2017, Richardson, *et al* 2018). Elderly and HDT-ineligible patients are treated with longer induction treatment for 8-12 months and in some patients, treatment is continued until relapse (Benboubker, *et al* 2014, Stewart, *et al* 2015a). However, eventually, the disease will progress or relapse, and initiation of a rescue treatment will be necessary (Laubach, *et al* 2016).

Chemotherapy as well as treatment with proteasome inhibitors (PI) and immune-regulatory agents (IMID) cause a high risk of acute adverse events, e.g. suppression of the bone marrow with risk of infections and hospitalization, as well as other side effects, e.g. peripheral neuropathy and fatigue (Boland, *et al* 2013, Mateos 2010, Molassiotis, *et al* 2011, Richardson, *et al* 2012). Treatment decision-making in MM is complex and involves factors such as disease stage, prognostic risk stratification, severity of myeloma symptoms and complications, expected progression-free survival and...


Non-completion of an entire scheduled PRO questionnaire defines a non-response (NR) and can be subdivided into patterns of NR as monotone, intermittent or mixed (Fielding, et al 2009, Little, et al 2012). A monotone pattern of NR is a pattern of complete responses until NR occurs by e.g. drop-out; intermittent NR is a pattern of one or more NRs between completed questionnaires, and a mixed pattern of NR occurs when a patient first has an intermittent and later a monotone pattern. If a patient participating in a clinical trial or cohort study does not complete any scheduled questionnaires or is excluded from the PRO data analysis, the patient is defined as a complete non-responder.

Three different mechanisms for NR have been described (Rubin 1976), and each is exemplified here. “Missing completely at random” (MCAR) occurs, for example, if the questionnaires are not given to the patient. “Missing at random” (MAR) occurs, for example, if a specific subgroup of patients with similar outcomes, e.g. poorer PRO scores, has a higher proportion of NR. “Missing not at random” (MNAR) occurs, for example, if the patient does not complete the questionnaire due to experiencing adverse events or complications (Fielding, et al 2009, Palmer, et al 2018).

The objectives of this review were to investigate applied PRO measuring time points, statistical analysis methods and the magnitude and ways of handling NR in longitudinal PRO studies of patients with MM. Based on our findings we will discuss the quality of existing evidence of HRQoL in patients with MM and provide recommendations for future clinical trial investigators.

Material and methods

Publication selection
We used the earlier identified corpus of publications reported in Nielsen, et al (2017), which is based on a systematic literature search with the primary objective to identify longitudinal HRQoL studies in MM patients. The literature search and publication selection have been described in detail (Nielsen, et al 2017). Briefly, publications were eligible if the following criteria were met: patients were diagnosed with MM, and the study applied a longitudinal study design using the European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire (EORTC QLQ-C30) (Aaronson, et al 1993) instrument for HRQoL measurement of physical function, global QoL, fatigue and/or pain. Articles in languages other than English were excluded. There was no time limit set for the literature search. After the systematic literature search, separate publications with additional reporting of the HRQoL data from the ASPIRE trial and TOURMALINE-MM1 study were published (Leleu, et al 2018, Stewart, et al 2016). These two additional publications were included in the data extraction process for this review. When the PRO results from a clinical trial were presented in a separate publication, the first publication from the trial, including reporting of primary study endpoint, was identified and included in the data extraction process.

### Data extraction

Information extracted from the publications included: 1) whether the HRQoL data collection was a primary or secondary endpoint in the study or clinical trial; 2) scheduled timing of follow-up HRQoL assessment; 3) the statistical analysis method applied for between group differences and/or within group change estimation and the predefined statistical significance level; 4) a description of reasons for exclusion of patients from HRQoL analysis; and 5) the statistical handling of NRs. Numbers extracted from the publications were: 6) the number of patients included in the study at baseline; 7) the number of PRO assessments at baseline; 8) the number of participating patients from whom a completed PRO assessment was expected at each scheduled PRO assessment time point; 9) the number of completed PRO assessments at each scheduled PRO assessment time point; and 10) the number of PRO assessments at the last presented follow-up time point. If the last published PRO follow-up time point was an end of study/treatment discontinuation assessment, the number of available questionnaires at the former time point was used.

The total intermittent NR rate for each study was estimated by calculating the proportion of patients who did not complete scheduled PRO assessments of those from whom a completed PRO assessment was expected for each presented follow-up time point together. We calculated the magnitude of monotone NR by the proportion of NR at last presented follow-up time point compared to the number of patients enrolled in the study.

The statistical methods applied were divided into group A-D: A) Descriptive analyses; B) Non-parametric tests (Mann Whitney U-test, Wilcoxon signed rank test); C) Parametric tests, subdivided into C1) T-test, one-way analysis of variance (ANOVA) and C2) linear mixed model of repeated measures, generalized estimating equations; and D) Ordinal logistic regression, generalized mixed model with ordinal outcome.

The data extraction was done independently by three of the authors. Disagreements were discussed to achieve consensus.

### Results

Twenty-three longitudinal HRQoL datasets were identified for data extraction. The PRISMA flow diagram for the study selection and subdivision of publications into five treatment categories are presented in Nielsen, et al (2017). The
longitudinal PRO data from MM-015 studies were divided into two treatment categories: first line treatment without autologous stem cell transplantation and maintenance therapy, respectively, because the HRQoL data from those two treatments could be extracted separately from the studies (Dimopoulos, et al 2013, Dimopoulos, et al 2014). The included publications with references are presented in Table 1 with an additional reference to the publication reporting the primary study endpoint.

**Endpoint and timing of PRO data collection**

HRQoL was the secondary endpoint in 16 studies and the primary endpoint in seven studies. The most frequent time point for PRO assessment was at predefined calendar time points in ten studies, the second most common PRO assessment time point was at day 1 of a new treatment cycle in eight studies and at a predefined clinical time point in four studies.

**Statistical analyses method and evaluation strategy applied**

All 23 studies described the statistical methods used for analysing the longitudinal PRO data. The most frequently applied statistical method was C2) parametric statistical method of mixed model of repeated measures or generalized estimation equations of 11 studies, and the second most used was C1) parametric statistical methods of t-test or one-way ANOVA in eight studies. Non-parametric statistics were used in six studies, ordinal logistic regression in three studies and descriptive statistics in two studies. Adjustment of statistical significance level to avoid multiplicity testing and type I error was performed in eight of the studies. The PRO data was evaluated by inter-group differences in nine studies and intra-group change in eight studies. Both strategies (inter-group differences and intra-group changes) were used for evaluation of longitudinal PRO data in four studies, and in two studies, no strategy for evaluation was applied.

**Magnitude and handling of complete non-responders**

For 17 of the 23 longitudinal PRO study results, the number of PRO assessments at baseline was lower than the number of patients included in the clinical trial, leaving some patients as complete non-responders. The number of patients included in the clinical trial and number of PRO assessments at baseline are presented in Table I. The lowest proportion of all studied patients was included in the PRO analysis of the population-based PROFILES registry (Mols, et al 2012) of 51%, because the analysed cohort was limited to the patients with a completed one-year follow-up questionnaire. The second lowest studied cohort was the two randomized groups in the study of Gimsing, et al (2010), with 65% of the included patients treated with pamidronate 90 mg and 68% treated with pamidronate 30 mg. The analysed cohort in that study was the patients who returned questionnaires at the 12-month follow-up and who were still on study treatment. The third lowest proportion was the cohort of the SUMMIT study (Dubois, et al 2006), of 71%. Here the analysis cohort was limited to the patients with PRO information available and with a clinical response to bortezomib. For the remaining studies, the proportions of patients included in the PRO data analyses compared to the number of patients included in the clinical trial were between 81 and 96%.

Ten of the 17 publications described the selection strategy for the reduced number of baseline PRO assessments patients included in the PRO data analysis compared to the clinical trial (Table I). The most frequently used strategy was to include only participants with a non-missing baseline questionnaire and minimum one follow-up questionnaire. In three publications (Gulbrandsen, et al 2001; Wisloff, et al 1996a; Mols, et al 2012), characteristics of participants and non-participants were compared. In the Table 1 of the paper by Gulbrandsen, et al (2001) the baseline characteristics of the
participants and non-participants are presented. Wisloff, et al (1996a) found participants likely to be younger, female and to have longer survival than non-participants. In the PROFILES registry (Mols, et al 2012), the participants were diagnosed more recently and often treated with other regimens than chemotherapy only, compared to non-participants.

**Magnitude of intermittent non-responses**

In six of the studies, the number of completed PRO assessments together with the number of participating patients from whom a completed PRO assessment was expected at each scheduled PRO assessment time point was presented. The total number of participating patients and the total number of completed PRO assessments for the six studies are presented in Table II. The lowest presented total intermittent NR rate was 2% in the control group of the study by Gulbrandsen, et al (2001) and the highest was 22% in the study by Waage, et al (2004).

**Magnitude of monotone non-responses**

The number of completed questionnaires at last follow-up was presented in 16 studies, and in all of them, the number was lower than at baseline. The number of questionnaires is presented in Table I. The highest proportion of monotone NR was seen in the TOURMALINE study (Leleu, et al 2018): 99% in the lenalidomide-dexamethasone group and 98% in the ixazomib-lenalidomide-dexamethasone group. The PRO data results in that study were collected at a specified cut-off date and some patients are still in follow-up. The second highest proportion of monotone NR was in the MM-003 trial (Song, et al 2015, Weisel, et al 2015): 96% in the high dose dexamethasone group and 82% in the pomalidomide-dexamethasone group. The third highest proportion of monotone NR was in the APEX study (Lee, et al 2008), which was 88% in the bortezomib group and 83% in the dexamethasone group. For the remaining studies, the proportion of monotone NR was between 28% and 75%.

**Statistical handling of intermittent and monotone non-responses**

Statistical methods for handling intermittent and monotone NR or methods to investigate the impact of NRs were used in eight of the 23 studies. The methods applied are presented in Table I. In the ASPIRE and TOURMALINE studies, a graphical approach was used to explore patterns of monotone NR (Leleu, et al 2018, Stewart, et al 2016). Multiple imputation was used to test the robustness of the PRO results in three studies (Gimsing, et al 2010; Lee, et al 2008; Waage, et al 2010). Other methods of exploring the missing data mechanisms or robustness of PRO data results were: confirming the results by standardized area under the curve, mixed method of repeated measures or comparing mean scores for patients with available questionnaires at follow-up to patients with early study discontinuation (Delforge, et al 2015, Stewart, et al 2016, Waage, et al 2004).

**Discussion**

In this review, we have investigated methodological aspects of PRO data measurements and analyses in 23 published longitudinal PRO studies of patients with MM, identified in a previously published systematic review (Nielsen, et al 2017). We observed diversity in the timing of PRO data collection and statistical methods for analysing the longitudinal PRO data among the studies. In 17/23 studies the number of PRO assessments at baseline was lower than the number of patients included in the clinical trial or study with proportions of complete non-responders being up to 51%. Reporting
of an intermittent NR rate was, in general, lacking, but was up to 22% when reported. For studies where every patient had reached the cut-off date for the analyses, we found proportions of monotone NR between 28% and 96%. Despite the high proportions of complete, intermittent and monotone NR, only 8/23 studies investigated the impact of NRs. Diversity in applied PRO methodologies and poor quality of PRO reporting from clinical trials has previously been reported (Brundage, et al. 2011a, Bylicki, et al 2015, Dirven, et al 2014, Efficace, et al 2015, Efficace, et al 2014, Hamel, et al 2017, Lemieux, et al 2011). However, to our knowledge, this is the first review to investigate methodological aspects in PRO measurements and analyses in studies of patients with MM.

In 2013, the CONSORT PRO guideline for reporting PRO from randomized clinical trials was published to improve the accuracy and validity of PRO data reporting (Calvert, et al 2013). Recently, in 2018, a guideline for inclusion of PROs in clinical trial protocols became available (Calvert, et al 2018). This guideline, together with international standards for analysing PRO data, have supported the increased application of PRO data results from clinical trials to daily clinical practice (Bottomley, et al 2016, Bottomley, et al 2018, Brundage, et al 2011b).

Clinical cancer trials are often designed with termination of PRO data collection if the patient drops out. Due to the nature of MM with risk of treatment failure, unacceptable adverse events to treatment and shortness of the patients’ life expectancy, PRO data collection in MM studies is at high risk of monotone NR. Also, patients with MM often experience disease complications and significant adverse events during treatment, particularly during HDT, which increases the risk of intermittent NR.

Missing PRO data can lead to a variety of problems, including loss of study power and precision (Bell and Fairclough 2014, Fairclough 2010). Participants who drop out early may have a poor HRQoL (Bell and Fairclough 2014, Mercieca-Bebber, et al 2017). Specific strategies to minimize missing PRO data should be implemented in the study design and data collection procedure (Calvert, et al 2018, Little, et al 2012, Mercieca-Bebber, et al 2016) and transparent reporting of the number of completed questionnaires at baseline and at subsequent time points are recommended (Calvert, et al 2013). NR to questionnaires might cause biased results if appropriate statistical handling guided by the missing data mechanisms is not performed (Bell and Fairclough 2014, Bell, et al 2013, Fairclough 2010). Handling of NR by simple imputation of “last observation carried forward” is not recommended for longitudinal data (Lavori, et al 2008). In our review, the most frequently used statistical analysis method was linear mixed model of repeated measures or generalized estimation equations including multiple imputation, where NR are handled as MAR. Multiple imputation was used as the statistical analysis method in three studies (Gimsing, et al 2010, Lee, et al 2008, Waage, et al 2010), and the authors found that the MAR assumption was not the correct missing data mechanism in all cases.

When using descriptive statistics, non-parametric or parametric methods of t-test and one-way ANOVA, the missing data are handled as MCAR, which however, is rarely the case for the majority of missing PRO data (Bell and Fairclough 2014, Fairclough 2010). The assumption for MCAR missing data mechanism could not be confirmed in the included studies, where this aspect was investigated. In the two studies where a graphical approach was used (Leleu, et al 2018, Stewart, et al 2016), a slightly different pattern of change in global quality of life score over time was found for patients dropping out earlier compared to patients staying in the study for the whole period, but the differences were not statistically significant or clinically meaningful.

We observed variations in the timing of PRO data measurements among the included studies. The scheduled PRO data assessments time point should ensure capturing the patient experienced effect of the intervention aimed at PRO data
collection. Using *day 1 of a new treatment cycle* for PRO data measurement has a clear advantage of reduced risk of missing response, as the patient can complete the questionnaire in hospital with assistance from a study coordinator. Most PRO data instruments for clinical research have a 7-day recall period, and most anti-myeloma regimens are administered in 21- or 28-day cycles, with the last week being drug free. Also, a general principle is rescheduling the next treatment cycle if the patient experiences severe toxicity or complications, such as admission with neutropenic fever, too much fatigue, etc. Therefore, when PRO data collection is scheduled on *day 1 of a new treatment cycle*, the PRO data measurement during periods with complication is missed. In addition, the patients might have completed the questionnaires with reflection of a drug-free week. This might lead to overestimation of HRQoL and underestimation of toxicities (Giesinger, et al 2014).

A limitation in our review is the restriction in selection criteria for the systematic review that comprised the publications using the EORTC QLQ-C30 instrument only. The EORTC instruments are traditionally used in European clinical trials, and applied PRO data research methods might be different in other parts of the world. An important aspect of the generalisability of PRO data results is that most PRO data deviates from patients included in clinical trials. This is also the case for patients with MM because 22 of the 23 studies in the systematic review are clinical trials (Nielsen, et al 2017). Newly diagnosed patients with MM included in clinical trials are not representative for the general MM population and PRO data from clinical trials might not be generally applicable (Klausen, et al 2018). HRQoL data from population-based studies of patients with MM with high focus on minimization of NR are needed. In this review, we focus on how the EORTC QLQ-C30 has been used for HRQoL measurement in studies of patients with MM, because it is the most used HRQoL instrument in this population. Another relevant consideration is whether the EORTC QLQ-C30 questionnaire was the most suitable tool to measure what matters to MM patients in each identified study. We did not review the identified studies and protocols to investigate whether there was a specific research question and rationale for choosing the EORTC QLQ-C30 instrument to elucidate HRQoL in each study (Calvert, et al 2018).

In conclusion, we found diversity in the PRO data measurements and analyses applied in clinical studies of MM patients and observed a large element of NR. We found no transparent reporting of NR, and the missing data mechanisms were rarely investigated, which resulted in the use of statistical methods of PRO data analysis based on untested assumptions. Based on the publications investigated, these findings suggest that the evidence-based knowledge of HRQoL in patients with MM is compromised by significant rates of complete, intermittent and monotone NR. This threatens the generalizability of PRO data results in MM and their application to daily clinical practice. In order to improve quality of PRO data and translation of PRO data results in patients with MM, we recommend PRO data investigators to follow the SPIRIT-PRO Extension Checklist during clinical trial protocol writing (Calvert, et al 2018) and the CONSORT PRO Extension Checklist Item when reporting PRO results from randomized trials (Calvert, et al 2013). Strategies to reduce NR that are suitable for the investigated cohort should be integrated in the study design, PRO data collection and procedures. Linear mixed models of repeated measures have been found to be the most suitable for analysing longitudinal PRO data and multiple imputation is considered the best method for sensitivity analyses, but these are not general recommendations (Hamel, et al 2017, Rombach, et al 2018). International standards for analysing PRO data from clinical trials are currently being developed (Bottomley, et al 2016). Therefore, it is important to be aware of potential pitfalls in PRO methodology when international standards are not available.
Author contributions
LKN and NA planned the study, performed literature search, data extraction, data analyses, interpretation and wrote the first manuscript draft. TWK and MJ contributed to data extraction, data analysis and interpretation. All authors approved the submitted and final version.

Acknowledgements
We acknowledge Lykkegaard Andersen C. and Frederiksen H. for critical review of the manuscript. This work was supported by The Danish Cancer Society and University of Southern Denmark.

References


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


Table I. Methodological and statistical aspects of measurement, analyses and interpretation of PRO data extracted from publications

<table>
<thead>
<tr>
<th>Trial name, study design and references</th>
<th>PRO data measurement</th>
<th>PRO data statistical analyses and handling of non-responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Endpoint</td>
<td>Follow-up measurement time points</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-line treatment studies including induction therapy and ASCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomized phase II study (Ludwig et al 2013)</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
</tr>
<tr>
<td>Phase II study (Etto et al 2011)</td>
<td>Primary</td>
<td>Clinical time points</td>
</tr>
<tr>
<td>Evaluation, phase II study (Gulbrandsen et al 2001; Lenhoff et al 2000)</td>
<td>Secondary</td>
<td>Calendar time points</td>
</tr>
<tr>
<td>First-line treatment studies without ASCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIRST trial Randomized phase III study (Benboubker et al 2014; Delforge et al 2015)</td>
<td>Secondary</td>
<td>Clinical time points</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved
<table>
<thead>
<tr>
<th>Study</th>
<th>Type</th>
<th>Secondary/Primary</th>
<th>Time Point</th>
<th>Analysis Method</th>
<th>Sample Size</th>
<th>Significant p-values</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM-015 study</td>
<td>Randomized phase III</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Paired t-test (p&lt;0.05, p&lt;0.01 and p&lt;0.001) and Mixed model repeated measures for within group change.</td>
<td>252 and 252</td>
<td>152 and 153</td>
<td>Only patients with a non-missing baseline and minimum one follow-up assessment were included. HRQoL observations at PD/DC, if occurred earlier than cycle 16 were carried forward to the next measurement time point.</td>
</tr>
<tr>
<td>VISTA trial</td>
<td>Randomized phase III</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>T-test (p&lt;0.05) for between group differences</td>
<td>344 and 338</td>
<td>344 and 338</td>
<td>Only patients with a valid baseline and at least one follow-up HRQoL assessment.</td>
</tr>
<tr>
<td>HOVON 49</td>
<td>Randomized phase III</td>
<td>Secondary</td>
<td>Clinical time points</td>
<td>Mixed model repeated measures for multi-item domains and ordinal logistic regression repeated measures for one-item domains between group differences</td>
<td>168 and 165</td>
<td>168 and 165</td>
<td>Participants and non-participants group comparison. Patients with at least one HRQoL assessment, either at baseline, during treatment or follow-up, were included in the analysis.</td>
</tr>
<tr>
<td>Randomized phase III study</td>
<td>Primary</td>
<td>Calendar time points</td>
<td>T-test for within group change. Generalised estimating equations for within group change.</td>
<td>252 and 252</td>
<td>164 and 171</td>
<td>Only patients who returned questionnaires at 12 months and who were still on study treatment were included in the HRQoL analyses. Multiple imputation.</td>
<td></td>
</tr>
<tr>
<td>Randomized phase III study</td>
<td>Secondary</td>
<td>Calendar time points</td>
<td>Generalised estimating equations confirmed by standardized area under the curve for within group change</td>
<td>182 and 175</td>
<td>182 and 175</td>
<td>Not reported</td>
<td>Analyses were based on patients who returned completed questionnaires. Multiple imputation.</td>
</tr>
<tr>
<td>NMSG 4/90</td>
<td>Cohort study</td>
<td>Secondary</td>
<td>Calendar time points</td>
<td>Descriptive</td>
<td>583</td>
<td>583</td>
<td>Only eligible patients completing the first questionnaire were included. Comparison of characteristics of participants and</td>
</tr>
<tr>
<td>Study</td>
<td>Primary</td>
<td>Secondary</td>
<td>Calendar/time points</td>
<td>Test statistic</td>
<td>Between/within differences</td>
<td>N</td>
<td>Comments</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-----------</td>
<td>---------------------</td>
<td>---------------</td>
<td>----------------------------</td>
<td>---</td>
<td>----------</td>
</tr>
<tr>
<td>NMSG 4/90, Randomized phase III study (The Nordic Myeloma Study Group 1996; Wisloff et al 1996b)</td>
<td>Secondary</td>
<td>Calendar time points</td>
<td>Mann Whitney U-test (p&lt;0.01) for between group differences</td>
<td>297 and 286</td>
<td>271 and 253</td>
<td>67 and 74</td>
<td>Only eligible patients completing the first questionnaire were included.</td>
</tr>
</tbody>
</table>

**Consolidation treatment studies**

<table>
<thead>
<tr>
<th>Study</th>
<th>Primary</th>
<th>Secondary</th>
<th>Calendar/time points</th>
<th>Test statistic</th>
<th>Between/within differences</th>
<th>N</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized phase II study (Melliqvist et al 2013)</td>
<td>Primary</td>
<td>Calendar time points</td>
<td>Mann Whitney U-test (p&lt;0.01) for between group differences</td>
<td>187 and 183</td>
<td>311 in total</td>
<td>Not reported</td>
<td>Only patients completing the baseline questionnaire were included in the analysis</td>
</tr>
<tr>
<td>Phase II study (Frodin et al 2011)</td>
<td>Primary</td>
<td>Clinical time points</td>
<td>Descriptive</td>
<td>56</td>
<td>56</td>
<td>25</td>
<td>Not described</td>
</tr>
<tr>
<td>Phase II study (Khalafallah et al 2011)</td>
<td>Primary</td>
<td>Calendar time points</td>
<td>Ordinal logistic regression repeated measures for within group change</td>
<td>18</td>
<td>18</td>
<td>Not reported</td>
<td>Not described</td>
</tr>
</tbody>
</table>

**Maintenance treatment studies**

<table>
<thead>
<tr>
<th>Study</th>
<th>Primary</th>
<th>Secondary</th>
<th>Calendar/time points</th>
<th>Test statistic</th>
<th>Between/within differences</th>
<th>N</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM-015 study, Randomized phase III study (Dimopoulos et al 2013; Dimopoulos et al 2014; Palumbo et al 2012)</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Paired t-test (p&lt;0.05, p&lt;0.01 and p&lt;0.001) and Mixed model repeated measures for within group change.</td>
<td>88 and 94 and 102</td>
<td>91 and 99 and 99$^{2,4}$</td>
<td>65 and 56 and 65$^2$</td>
<td>Only patients with a non-missing baseline and minimum one follow-up assessment were included. HRQoL observations at PD/DC, if occurred earlier than cycle 16 were carried forward to the next measurement time point.</td>
</tr>
<tr>
<td>Randomized phase II study</td>
<td>Primary</td>
<td>Calendar time points</td>
<td>Two-sample t-test Mixed model repeated measures for between</td>
<td>30 and 30</td>
<td>30 and 30</td>
<td>Not reported</td>
<td>90% completed all three questionnaires and all patients completed at least two.</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved
<table>
<thead>
<tr>
<th>Study</th>
<th>Type</th>
<th>Secondary</th>
<th>Day of Treatment</th>
<th>Methodology</th>
<th>n Group 1</th>
<th>n Group 2</th>
<th>n Response</th>
<th>Sensitivity Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOURMALINE-MM1 study (Leleu et al 2018; Moreau et al 2016)</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Mixed model repeated measures for between group difference and within group change (p&lt;0.05)</td>
<td>360 and 362</td>
<td>337 and 349</td>
<td>7 and 2$^2$</td>
<td>Graphical examination stratified by time of last assessment and pattern mixture model as sensitivity analysis.</td>
<td></td>
</tr>
<tr>
<td>ASPIRE trial (Stewart et al 2016; Stewart et al 2015b)</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Mixed model repeated measures for between group difference and within group change (p&lt;0.01 and p&lt;0.001)</td>
<td>396 and 396</td>
<td>348 and 348</td>
<td>227 and 148</td>
<td>Analysis includes patients with at least one HRQoL assessment. Graphical examination stratified by time of last assessment. Results were confirmed with standardized area under the curve and pattern mixture model as sensitivity analysis.</td>
<td></td>
</tr>
<tr>
<td>MM-003 study (Miguel et al 2013; Song et al 2015; Weisel et al 2015)</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Mixed model repeated measure. Paired t-test (p&lt;0.05) for within group change. Unpaired t-test (p&lt;0.05) for between group differences confirmed. Logistic regression analysis for analysis of responders</td>
<td>302 and 153</td>
<td>289 and 144</td>
<td>51 and 6$^4$</td>
<td>Analysis includes patients who received at least one study drug and had one HRQoL assessment. Missing data were categorically evaluated for all HRQoL assessments.</td>
<td></td>
</tr>
<tr>
<td>NMSG 17/07 study (Hjorth et al 2012)</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Mann Whitney U-test (p&lt;0.01) for between group differences</td>
<td>67 and 64</td>
<td>67 and 61</td>
<td>Not reported$^8$</td>
<td>Not described</td>
<td></td>
</tr>
<tr>
<td>APEX study (Lee et al 2008;)</td>
<td>Secondary</td>
<td>Calendar time points</td>
<td>Generalised estimating equations for between group difference (p&lt;0.05)</td>
<td>333 and 336</td>
<td>288 and 287</td>
<td>65 and 81</td>
<td>Analyses are based on patients with a valid HRQoL assessment and at least one post-baseline questionnaire. Multiple imputation taking deaths into account and other parameters as sensibility.</td>
<td></td>
</tr>
<tr>
<td>Study Type</td>
<td>Design</td>
<td>Time Points</td>
<td>Test Used</td>
<td>N Primary</td>
<td>N Follow-up</td>
<td>N Discontinued</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------</td>
<td>---------------------------------------</td>
<td>--------------------------------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>SUMMIT study</td>
<td>Secondary</td>
<td>Day 1 of treatment cycles</td>
<td>Wilcoxon signed rank test (p&lt;0.05) for within group change</td>
<td>202</td>
<td>144</td>
<td>144</td>
<td>Analysis includes 144 patients with both clinical response and PRO information available</td>
<td></td>
</tr>
<tr>
<td>Phase II study</td>
<td>Secondary</td>
<td>Calendar time points</td>
<td>Wilcoxon signed rank test (p&lt;0.01) for related samples for within group change</td>
<td>65</td>
<td>62</td>
<td>20</td>
<td>Comparing mean score for patients with available questionnaire at 24 weeks to patients with early study discontinuation</td>
<td></td>
</tr>
<tr>
<td>Non-interventional study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROFILES registry</td>
<td>Primary</td>
<td>Calendar time point</td>
<td>Paired t-test for within group change (p&lt;0.01)</td>
<td>156</td>
<td>156</td>
<td>80</td>
<td>Non-responders and responders were compared</td>
<td></td>
</tr>
</tbody>
</table>

1. In case of HRQoL measurement at study discontinuation, the number of available questionnaires at the former time point is presented unless another time point is specified.
2. The patients at follow-up are not all the same as at diagnosis.
3. Based on mean score of physical functioning.
4. Patients in each group at cycle 10, which was at start of maintenance.
5. The number of questionnaires used for later follow-up time point evaluation is not reported.
6. 50% of the patients in the melphalan-prednisone-thalidomide arm and 62% of the patients in the melphalan-prednisone arm. The exact numbers could not be extracted.
7. The results of the PRO data in the study were made at a specified cut-off date and some patients were still in follow-up after.
8. 29 vs. 29 patients were alive at the time of last follow-up. Study design with crossover at treatment failure.
9. The change in PRO over time was assessed by comparing the change in scores according to clinical response between baseline and best end point.
10. The non-parametric Mann Whitney U-test is used for related samples were used to compare the score at different time points, which is interpreted as a Wilcoxon signed rank test.

ANOVA: analysis of variance; ASCT: autologous stem cell transplantation; DC: discontinuation for other reasons; HRQoL: health-related quality of life; PD: Progressive disease; PRO: patient-reported outcomes.
Table II. The total number of participating patients and the total number of completed PRO assessments at each scheduled PRO assessment time point for the six studies with the numbers presented.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Study arm/cohort</th>
<th>Total sum for all presented PRO assessment time points</th>
<th>Intermittent non-responses rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sum of expected PRO assessments</td>
<td>Sum of completed PRO assessments</td>
</tr>
<tr>
<td>Gulbrandsen et al (2001)</td>
<td>Induction therapy and HDT</td>
<td>1076</td>
<td>966</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>541</td>
<td>528</td>
</tr>
<tr>
<td>Delforge et al (2015)</td>
<td>Lenalidomide-dexamethasone</td>
<td>5166</td>
<td>4743</td>
</tr>
<tr>
<td></td>
<td>Melphalan-prednisone-thalidomide</td>
<td>2492</td>
<td>2179</td>
</tr>
<tr>
<td>Wisloff et al (1996)</td>
<td>Cohort</td>
<td>2541</td>
<td>2055</td>
</tr>
<tr>
<td>Stewart et al (2016)</td>
<td>Carfilzomib-lenalidomide-dexamethasone</td>
<td>1706</td>
<td>1543</td>
</tr>
<tr>
<td></td>
<td>Lenalidomide-dexamethasone</td>
<td>1556</td>
<td>1351</td>
</tr>
<tr>
<td>Leleu et al (2018)</td>
<td>Ixazomib-lenalidomide-dexamethasone</td>
<td>3242</td>
<td>3007</td>
</tr>
<tr>
<td></td>
<td>Lenalidomide-dexamethasone</td>
<td>3209</td>
<td>2991</td>
</tr>
</tbody>
</table>

*Based on Figure 1 in Gulbrandsen et al (2001)

bCalculated on basis of Table S2 in Leleu et al (2018)

HDT: high dose chemotherapy with stem cell support; PRO: patient-reported outcomes.