Frequentist and Bayesian meta-regression of health state utilities for multiple myeloma incorporating systematic review and analysis of individual patient data

Running title: Health state utilities in multiple myeloma

Abstract

This analysis presents the results of a systematic review for health state utilities in multiple myeloma, as well as analysis of over 9000 observations taken from registry and trial data. The 27 values identified from 13 papers are then synthesized in a frequentist non-parametric bootstrap model, and a Bayesian meta-regression. Results were similar between the frequentist and Bayesian models with low utility on disease diagnosis (approximately 0.55), raising to approximately 0.65 on first line treatment, and declining slightly with each subsequent line. Stem Cell Transplant was also found to be a significant predictor of health related quality of life in both individual patient data and meta-regression, with an increased utility of approximately 0.06 across different models. The work presented demonstrates the feasibility of Bayesian methods for utility meta-regression, whilst also presenting an internally consistent set of data from the analysis of registry data. To facilitate easy updating of the data and model, data extraction tables and model code are provided as supplementary materials. The main limitations of the model relate to the low number of studies available, particularly in highly pre-treated patients.

1 Introduction

Multiple myeloma (MM) is a haematological malignancy characterised by clonal proliferation of immunoglobulin-secreting plasma cells. This can lead to reduced haemopoiesis, renal failure and bone lesions. While the disease is incurable with conventional therapy, there have been dramatic improvements in treatments over the past 20 years, with multiple classes of therapy becoming available. These include proteasome inhibitors (PIs, such as bortezomib and carfilzomib) and immunomodulatory agents (IMIDs, such as lenalidomide and pomalidomide), as well as novel agents. Patients are treated with sequential lines of therapy, which can include stem cell transplant (SCT).

In the UK, the majority of MM treatments have been reviewed by the National Institute for Health and Care Excellence (NICE), including bortezomib, lenalidomide and newer treatments such as pomalidomide or panobinostat (NICE, 2009, 2014, 2017). As a part of the economic modelling in each appraisal, health state utilities were taken from the trials for each treatment. In each appraisal, utilities from the relevant clinical trial(s) were used, and utility values from previous trial(s) were then used in sensitivity analyses without any form of synthesis. Thus far, no attempts have been made to reconcile differences in estimated values between studies, or to incorporate consolidated data from sources other than the trial of the specific treatment being evaluated.

In health technology evaluations, using utility values taken from individual sources contrasts with the conventional approach to evaluation of efficacy and safety data. For efficacy and safety data, the conventional approach is to include all relevant data through appropriate use of meta-analysis (Dias, Welton, Sutton, & Ades, 2011). Meta-analysis is a broad term encompassing various methodologies – in utility data meta-regression has most often been applied, with examples in human immunodeficiency virus (HIV), stroke and renal disease. Recent discussion in the literature has considered whether this approach is appropriate, given the differences between valuation measures (for example, between the EQ-5D and SF-36), and acknowledged the need for further research in the area (Liem, Bosch, Arends, Heijenbrok-Kal, & Hunink, 2007; Peasgood & Brazier, 2015; Tengs & Lin, 2002, 2003).

The objective of this study was to use registry data to provide an internally consistent set of utility estimates i.e. a set of data across the entire pathway that has been drawn from the same source data and patients, and then synthesize all available data (including registry data) to provide utilities that can be used in health economic modelling. This was achieved by conducting a systematic review, augmented with analysis of primary data from the EMMOS registry and APEX clinical study, followed by meta-regression.

2 Methods

2.1 Definition of classes of therapy

Due to the number of different interventions received by patients in the literature, as well as varying definitions of therapy lines (for example, whether re-treatment is classed as a new line of new therapy), patient classification was simplified. Trials were recategorized based on the number of treatment classes a patient had previously received, from the categories of PI, IMID, chemotherapy and novel agents (those licensed within the past 5 years, even if technically members of other classes).

As treatment dosing varies between treatments (for example, bortezomib is given for a fixed period, while lenalidomide is dosed continuously), for simplicity utilities were not considered separately for whether a patient was on or off treatment.

2.2 Registry and trial data analysis

Individual level data were made available by Janssen from the EMMOS registry and the APEX clinical study (Mohty et al., 2015; Richardson et al., 2005). The EMMOS registry contains data from 2,521 patients in 22 countries in Europe and Africa, across all classes of MM treatment. The APEX clinical study enrolled 669 patients with relapsed MM who were randomised to either bortezomib or placebo. This constitutes a large dataset of previously published data which can be used as an input to the meta-regression as data is available throughout the treatment pathway.

The EQ-5D-3L results from each dataset were valued using the UK tariff (Kind, Dolan, Gudex, & Williams, 1998). The utility values were used as the dependent variable in a regression model with explanatory variables of classes of MM treatment previously received and rate of SCT(Kind et al., 1998). Generalised estimating equation regression was used to account for each patient having multiple correlated observations, whilst also producing estimates applicable at the population level (Hanley, Negassa, Forrester, & others, 2003). The specification of these models is described in the below equations. The variables are labeled similarly in both models. .

APEX trial:

$$U_{it}^{APEX} = \beta_2 C1_{it} + \beta_3 C2_{it} + \beta_6 SCT_{it} + \varepsilon_{it}$$

EMMOS Registry:

$$U_{it}^{EMMOS} = \gamma_1 NEW_{it} + \gamma_2 C1_{it} + \gamma_3 C2_{it} + \gamma_4 C3_{it} + \gamma_5 C4_{it} + \gamma_6 SCT_{it} + \epsilon_{it}$$

Where U_{it} represents the utility observation for individual i at time t, β and γ the coefficients of the regressions,NEW and CX dummy variables to represent the patient being newly diagnosed or having received X prior classes, SCT a dummy variable of whether a patient had received SCT at the time the observation was taken, and ε_i & ϵ the error term. An unstructured correlation matrix was used.

The APEX trial only enrolled patients with 1 prior treatment who were treated until progression on bortezomib, and therefore a less expansive regression was specified. This analysis of the APEX data was performed using a variety of patient characteristics (such as age, gender, and country), none of which improved model fit or proved predictive of patient utility. This finding is consistent with the literature and clinical practice where disease characteristics appear most important predictors of quality of life. Including SCT and progressive disease as predictors produced the lowest mean absolute error, and root mean squared error to 2 decimal places.

The results of the analysis of the EMMOS dataset were similar, with patient characteristics not predictive of health related quality of life and limitations in data preventing analysis by individual treatment as many treatments were given in

combination, on differing regimens. The model with the lowest mean absolute error and root mean squared error was again the use of the number of classes of therapy a patient had received, and whether a patient had received stem cell transplant. A test for interaction between the line of therapy and stem cell transplant was non-significant indicating that the effect of SCT on utility did not vary by line.

The results of the APEX and EMMOS analyses are then included in **Table** 1, where they act as inputs to the meta-regression**Table** 1.

2.3 Literature review

To identify utilities in MM, a systematic review was conducted in MEDLINE, Embase, the Cochrane Library, MEDLINE In-Process and EconLit on 27 January 2016. All papers with a title or abstract indicating that the paper included preference-based utility values (from the EORTC, EORTC-8D, EQ-5D, SF-6D, SF-36, or HUI3) were included. Values derived from clinician opinion, vignette studies or custom scales were excluded.

2.4 Synthesis using meta-regression

To perform the synthesis of utility values, two distinct approaches were used: a frequentist meta-regression and a Bayesian statistical model with different specifications of each model giving a total of five model. Each model was then run twice: the first time using all available values (including utilities generated using other generic tools, and non-UK values), and the second time including only EQ-5D values meeting the NICE reference case (EQ-5D values, scored using the UK tariff) (NICE, 2008). A fifth model was then run using the Bayesian model with preferred data but with vague priors to see the impact this had on results.

2.5 Frequentist meta-regression

The treatment-associated utility was likely to be influenced by the proportion of patients in each study to have received an SCT, which would be expected to increase with the number of pre-treatments received – failing to account for this would likely generate biased predictions. Therefore, a meta-regression was specified with dummy variables for the number of previous treatment classes received, and the proportion of patients in each study to have received an SCT was included as a covariate. The reference

category was an unknown number of previous treatment lines, or multiple lines. In the instances of unreported SCT proportions (and no further information available), the mean SCT percentage for that number of previous treatment classes was assumed – this was based on clinical opinion, and assessment of the available evidence (presented in tabular format).

Information on the number of observations and the variance of the utilities estimated within each study were used as inputs to mixed-effects model using maximum-likelihood estimation – implemented using the *metafor* package within *R* (R Core Team, 2017; Viechtbauer & others, 2010). The results of the regression model were then nonparametrically bootstrapped to account for non-normality in distributions of coefficients. This step was performed using the *boot* package within R (Canty & Ripley, 2016; Davison & Hinkley, 1997). At each iteration, the nonparametric bootstrapping process randomly extracted a sub-sample of the full dataset and attempted to estimate the regression model described in the below equation. Failed regression attempts, that is sub-samples which did not have at least one observation for each previous treatment class, and consequently could not be estimated, were discarded, and the parameters from successfully estimated regression predictions for line-associated utilities were collected. Thus:

$$U_j = \beta_1 GENERAL + \beta_2 NEW_j + \beta_3 C1_j + \beta_4 C2_j + \beta_5 C3_j + \beta_6 C4_j + \beta_7 SCT\%_j + \lambda_j + \varepsilon_j$$

Where the model is moderated by the proportion of patients in each observation have had an SCT, U_j is reported utility in study j, and λ_j represents the between study heterogeneity.

The model was fitted using the Paule-Mandel estimator due to the small number of observations, and the fitted values were then graphically presented to demonstrate the uncertainty surrounding the health state utility estimates. From the estimated regression model, utility could be predicted using the coefficient of the appropriate number of treatment classes and the percentage of patients with SCT in the study. The resulting models are Model 1 including all methodologically sound utility data, and Model 2 which includes only EQ-5D utilities.

2.6 Bayesian statistical model

The Bayesian statistical model that was used to estimate utility using the number of treatment classes received and rate of SCT – as with the frequentist model. The main difference however being that the 'general disease' utilities were used as priors for one previous class of treatment (which otherwise would not be included in the analysis). This judgement was made based on the description of the patients in the paper rather than estimated as a separate health state in the model thus using the data to inform the health states. Thus:

$$U_j\beta_1 NEW_j + \beta_2 C1_j + \beta_3 C2_j + \beta_4 C3_j + \beta_5 C4_j + \beta_6 SCT\%_j + \varepsilon_j$$

Where $C1$ to $C4$ represent the number of prior lines a patient has received The Bayesian model was also specified without an intercept, as number of previous classes of treatment is mutually exclusive, with a proportion of patients also having experienced SCT. In this case comparing utility decrements as opposed to utility estimates, particularly for later in the pathway, would not have been intuitive.

Other than the prior for β_2 (which used the general disease utilities), all other priors were set to be informative with an upper bound of the 95% confidence interval of the data set to the mean utility of observations taken from patients with fewer classes of treatment, and a lower bound of 0.4 to represent the lowest plausible utility value. This resulted in priors of Normal(mean 0.6, standard deviation 0.12) for newly diagnosed patients, Normal(0.51, 0.06) for patients who had received two classes of treatment, Normal(0.52, 0.06) for patients who had received three classes, and of Normal(0.50, 0.05) for patients who had received four classes. Where multiple values were available to use as priors, these were combined through random effects inverse variance meta-analysis before use in the model. A random effects model was selected to allow for the effect to vary between studies. As with the frequentist analysis where the rate of SCT was not known for a study, this was assumed to be the mean of data from other studies for that stage of treatment for which the rate was known. To ensure the model successfully reached convergence to the underlying posterior distribution 300,000

simulations were used, with 50,000 as a warm up per chain (which were discarded), for a total of 500,000 simulations analysed.

The model was run with all utility data (Model 3), and then restricted to only UK EQ-5D utility data (Model 4). A final analysis was then conducted to assess the sensitivity of the Bayesian model to the priors used (Model 5). In this analysis, vague priors were used for all values of Normal(0.5,0.25), which practically bounds utilities between 0 and 1, and a prior for SCT used of Normal(0.06, 0.06) which practically bounds the impact of SCT to between -0.06 and 0.18 and indicates a likely positive impact with a reasonable degree of uncertainty.

The model was implemented in *R* for data processing and post-processing, and S*tan* to perform the Monte Carlo analysis. *Stan* allows fast computation of complex simulations using principles derived from physics. In addition to its speed, it presents a user-friendly interface, and can be called from within *R* using the package *rstan (Stan Development Team, 2016)*.

3 Results

3.1 Literature review

Figure 1 shows a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram for the systematic review, with 26 papers matching the inclusion criteria and 13 reporting methodologically appropriate utility values (10 of which were based on the EQ-5D in UK patients). When data were extracted from the published papers, 27 health state utilities were obtained (**Table** 1).

The results of the literature search show that utility appears to be poor on diagnosis, but increases as patients begin treatment, increasing again as patients move to a second class of treatment, before dipping slightly at three classes of previous treatment, and falling as patients have received all classes of treatment, including novel treatments. As would be expected, the proportion of patients who have received an SCT increases as patients become more heavily pre-treated (**Table** 1).

3.2 Analysis of registry and trial data

The EMMOS registry contains 9,080 completed EQ-5Ds from 2,445 patients. Data was very complete, with very low rates of missing data for variables used in our analysis (<3%) – records with missing data were therefore omitted from analysis. Analysis by the number of treatments received gave estimates for newly diagnosed patients of 0.459, increasing to approximately 0.6 while patients were receiving one to three classes of treatment, before decreasing to approximately 0.403 in patients who had received all classes of therapy (**Table** 1).

Similar results were seen in the APEX study (which only included patients with one and two previous classes of treatment). In the APEX study, data were available for 669 patients, who completed 1,568 EQ-5Ds pre-progression, and 944 post-progression. Analysis of the results of the completed EQ-5Ds showed that patients had a utility of 0.65 after one prior treatment, and on progression (assumed to be two prior classes as bortezomib had then been trialled) this decreased to 0.61 (**Table** 1). Being a regulatory study the data was highly complete (<2% missing data).

The analysis of patient data from both the EMMOS and APEX trials confirmed the role of SCT as an important predictor of patient health related quality of life - failure to include the rate of SCT in the regression led to counterintuitive results with utility appearing to increase throughout the disease pathway. This was as the increase in utility from SCT (which more patients have received in later lines) outweighed the increasing disutility associated with more previous classes of treatment. In the regressions the coefficient for SCT was 0.129 (standard deviation: 0.418) in the EMMOS study, and 0.056 (standard deviation: 0.010) in the APEX study. A test for interaction was performed in the EMMOS study to understand whether the effect varied by number of previous treatments, but the difference was not significant ($p>0.10$); supporting the assumption that the effect of SCT is independent of prior treatments.

3.3 EQ-5D vs all utilities analyses

The results of the literature search identified 13 papers with methodologically appropriate utilities, 10 of which used the EQ-5D in the UK population (as did the

EMMOS and APEX trials). Results estimated with all observations, and a sample limited to UK EQ-5D utilities are provided – the effect of including non-EQ-5D studies was to reduce the drop in utility as patients move through the disease pathway due to additional (higher) utilities coming from the additional 3 studies. However, with so few observations, it is not possible to conclude whether this is a true difference or due to a small sample.

3.4 Frequentist meta-regression

The results of the frequentist approach are presented in **Figure** 2. Models 1 and 2 both suggest that utility in newly diagnosed patients is low (0.529) and increases once patients are on treatment (0.659). Subsequent therapies are associated with sequentially lower levels of utility when adjusting for rate of SCT, decreasing to approximately 0.6 after patients have received three classes of treatment. Model 2, using EQ-5D values only (which we would expect to be more comparable), provides evidence to suggest that there is then a larger fall to 0.494 once patients have received all classes of treatment (**Table** 2). The limited number of studies in some areas, and the approach of omitting a study in each sample (through bootstrapping), lead to bimodal distributions (**Figure** 2); this is due to limited numbers of observations at later lines of therapy.

The non-zero SCT estimate in both Model 1 and Model 2 (mean 0.066, 95% interval: 0.056–0.17) suggests that trials with a higher proportion of SCT within their respective study samples have systematically higher utility values, even after adjusting for number of prior classes of therapy received. Consequently, the results of the bootstrapped meta-regressions indicate that SCT is associated with an improved level of utility – inkeeping with the results of the APEX and EMMOS studies.

3.5 Bayesian statistical model

Meta-analysing the ‘general disease’ and SCT utilities led to priors of Normal(0.689,0.427) for one previous treatment class and Normal(0.562, 0.039) for SCT based on the 4 and 3 studies respectively that gave relevant values. The resulting model coefficients, presented as Model 3 (using all utility estimates from generic

preference-based measures) and Model 4 (using only UK EQ-5D data) in **Table** 2 were similar to the frequentist analysis. These showed a large increase in utility for patients going from newly diagnosed to on treatment (0.530 to 0.661), before falling with each treatment class to reach 0.577 after three treatment classes, and then showing a precipitous drop to 0.471 (albeit with substantial uncertainty) once patients have received all treatment classes (**Table** 2). In the model, SCT was associated with increased utility, with a mean increase of 0.056 (95% credible interval 0.037 to 0.075), and none of the 500,000 simulations indicated that SCT would have a negative impact (**Figure** 3). There were no indications of problems with model convergence.

The results of the Bayesian model were similar in both Model 3 and Model 4. In Model 5, vague priors were used for all values using only the UK EQ-5D utilities (as in Model 4). The effect of this in the earlier disease stages was small changes at the second and third decimal place for the point estimates and credible intervals. However, where data were scarcer at later disease stages, the lack of informative priors lead to an increase in uncertainty resulting large credible intervals. For example, in patients who had received all classes of treatment, the 95% credible interval was 0.020–0.919, reflecting the uncertainty in the underlying data and that the model was unable to narrow the range of the prior.

4 Discussion

The results of the literature review, the analysis of registry data, and the meta-regressions all indicate that the utility of patients is low at diagnosis, and increases when patients are on treatment (likely due to symptom control). Subsequently, utility falls slightly as patients progress through the treatment classes, before falling further when patients have exhausted all existing treatment classes. Interestingly, the most uncertainty around utility values is for the one previous treatment class, and the three and four previous treatment classes – the causes of this uncertainty which we believe to be different. Based on the literature, it seems patients receiving their first treatment class are a highly heterogeneous group. Whilst there are a greater number of studies on

this group, and subsequently more observations in this study, these patients receive a wide variety of treatments. This is likely due to diversity in respective patient populations (as evidenced by the SCT rate ranging from 18.3% to 68.9%), with reported utility showing substantial variability (Acaster, Gaugris, Velikova, Yong, & Lloyd, 2013; Mohty et al., 2015). Conversely, patients receiving their second treatment class appear to exhibit less variability in reported health related quality of life. By the third and fourth treatment classes received (likely after having the disease for several years, having had re-treatment with some classes) there are relatively few values and small sample sizes, leading to uncertainty in health state utility estimates.

4.1 Role of SCT

Apparent in the data is the role of SCT, which is clearly linked to improved utility independent of the number of previous treatments. Taking the mean utilities from the systematic review, patients who failed their first treatment class and moved to a second treatment class were found to have higher utility. However, after taking into account the rise in SCT rate, the results were in line with what would have been expected: that utility decreases through the treatment pathway. The magnitude of the difference is also noteworthy – it was approximately 0.06 in both frequentist and Bayesian synthesis, approximately the level of a minimally important difference for the EQ-5D at the individual patient level (Pickard, Neary, & Cella, 2007).

The exact mechanism by which SCT increases utility is unknown. Nevertheless, we suggest two possible explanations. Firstly, only patients healthy enough to tolerate the intensive chemotherapy are eligible for SCT. Therefore, the higher utility among SCT patients may be the result of selection bias, where the fittest patients have undergone SCT. Secondly, it may be that SCT leads to a more benign disease form even when it fails to control the disease indefinitely (with patients going on to receive further treatments), and thus improve health related quality of life despite in patients subsequently receiving further treatment.

4.2 Choice of data source for economic modelling

Each dataset identified in our literature review includes values on only two levels of treatment which would be insufficient to populate a model, except those of end-stage myeloma, and is associated with substantial uncertainty around estimates. Only the EMMOS dataset is able to estimate utilities throughout the disease course (from newly diagnosed patients to those heavily pre-treated) from a single source, albeit still with uncertainty around point estimates. In the instance where use of data from differing sources is objected to by payers or decision makers, we suggest that the EMMOS dataset provides the most complete set of utility data in MM to date.

While the EMMOS registry provides an extraordinary volume of data (over 9,000 completed EQ-5Ds), the advantage of meta-regression is the synthesis of all available data to provide a coherent set of health state utilities, which are as robust and as generalisable as possible. Consequently, we recommend that the meta-regression values should be preferred to values from individual studies in future economic evaluations, or at a minimum incorporated into sensitivity analyses. Although there may be concern regarding the synthesis of values from different sources, by using only papers with methodologically appropriate values we believe this concern should be ameliorated. Further restricting sources to only papers that meet the NICE reference case of EQ-5D values using the UK tariff (Model 2 and Model 4) strengthens this approach.(NICE, 2008, 2013)

As new values are made available (with the completion of ongoing trials), this analysis can also be updated. To this end, we have made the results of our data extraction and source code available as online appendices to this paper. The code has been written to automatically accommodate the addition of more values, provided they are added to the data extraction table in the same format. We suggest that such openness is required for transparency and the development of best practice. This updating is particularly important as there are few values in the later stages of disease (and thus high uncertainty). Whilst not the objective of this paper, a model combining the individual and aggregate level data may also be possible to construct.

4.3 Frequentist vs Bayesian analysis

In our analyses the frequentist and Bayesian models gave similar results for the synthesis of values. Investigating further, the similar results are due to relatively weak priors being used in the Bayesian analysis, thus letting the data drive the results of the analysis. Arbitrarily removing studies / adding hypothetical studies and experimenting with different priors (data not shown), differences are seen between the approaches where data is conflicting, or where there is a large variation in results between studies – in these cases, the information encoded in the priors may be used to reconcile the estimates.

Despite the similarity in this instance, our preference is the Bayesian model, particularly Model 4 (EQ-5D data only) where the inputs are more homogenous (with not much data lost as a cost). There are two reasons for the choice of preferred model. Firstly, the Bayesian models sample from the distributions of the studies and consequently have face validity in that smooth distributions are simulated and presented (**Figure** 3). This contrasts with the nonparametric bootstrapping used in the frequentist analysis, which resulted in the presentation of multimodal distributions (**Figure** 2). The second advantage of the Bayesian analysis is that it can use priors to incorporate all data and prior beliefs. In the model we have constructed, this allows us to use 'general disease' utilities identified in the systematic review as priors for the one and two previous treatment class groups – the likely disease stage of patients in the studies even if the exact percentage breakdowns are not given. Equally, where priors are not available, these can (and have) be left vague. The effect of the priors can be seen in the difference between Model 4 and Model 5. Model 5 is based on the same data but with uninformative priors, leading to an increase in uncertainty beyond that which is plausible based on our prior knowledge of the structure of utility data. Model 5 therefore demonstrates that the priors in our analysis have acted as intended by constraining values to reasonable bounds, yet letting data determine the conclusions of the analysis.

The typical disadvantages of Bayesian analysis include difficulty of implementation, and increased computational burden associated with estimation. Whilst these critiques can be true, *Stan* allows for easy processing. Our model consists of approximately 30 lines

of code (available in the online Appendix), compared approximately 100 of lines included in the frequentist approach code due to the requirement for non-parametric bootstrapping. Similarly, the runtime (on a standard laptop) for the Bayesian analysis is under a minute, compared to approximately 30 minutes for the frequentist analysis (again due to bootstrapping). This difference is driven by the requirement for bootstrapping in the frequentist approach, versus the highly efficient *Stan* code – indeed it is likely the relatively simple model had converged before the 500,000 simulations used, and thus the analysis could have been performed faster to the same degree of accuracy. Although the appropriate solution to any particular analysis is likely to depend on the nature of the data and form/availability of prior information, based on our inputs and results, a Bayesian approach should be considered as an option. We believe that it is the first time this approach has been taken, with the proof of concept demonstrated alongside the equivalent frequentist analysis, showing better performance on all metrics – speed, flexibility, face validity and interpretability.

4.4 Other considerations

Whilst many of the areas discussed apply across many areas of economic evaluation (for example the techniques highlighted could be used with systematic reviews of efficacy values), there are some areas which are specific to utility values.

The first of these is that utilities are bounded by 1 (and potentially by zero). Whilst not an issue in our example (no studies had a reasonable chance of sampling over 1, should this be an issue, other distributions could be considered – notably a beta distribution (which in inherently capped at 1). Utility data from individuals is also notoriously multimodal, with EQ-5D data showing many patients with a utility of 1, with then, whilst such data is possible to model, it should not prove to be an issue for meta-regression, as only the mean values are used.

A further issue to consider is the number of studies available, and number of explanatory variables used (in our case, health states). Whilst no studies exist in utilities per se, a simulation study of linear regression in general found a minimum of 2 subjects per variable (which would be studies in the case of utility meta-regression) to be desirable (Austin & Steyerberg, 2015).

4.5 Limitations

The main limitation of the work presented is that it relies on the underlying data. As the treatment of MM has evolved when new treatments have become available, the definition of 'lines of treatment' and what constitutes relapse/progression has become somewhat complex and varied. Although the definition of lines has now been standardised, this will only apply for papers published in the future and, as a result our analysis, could only consider the treatment classes patients had received (Rajkumar, Richardson, & San Miguel, 2015). Similarly, due to the limited number of studies identified, it was not possible to estimate the differences in utility of each treatment available – either between classes of treatment or within classes of treatment – these may be a driver of economic models in certain circumstances. The limited volume of data available is also apparent in the multimodal distributions from the frequentist bootstrapped regressions resulted in jagged distributions - particularly in later classes where few studies have been reported.

By analysing data from the EMMOS and APEX studies we are able to ensure that the results of the synthesis are consistent with the individual level data, which is not always the case (Lambert, Sutton, Abrams, & Jones, 2002). With further access to individual level data however more comprehensive analysis may be possible, including estimation of utility differences between treatments, or a more complex model that incorporate both aggregate and individual patient level data.

Conclusion

The work conducted in this paper highlights the advantages of synthesis of utility data in being able to produce a consistent set of values for use in economic modelling through a disease pathway. In the area of MM, we demonstrate the importance of factoring in the rate of SCT as an explanatory variable for differences in estimated utility as patients progress through different treatment classes.

The main areas of uncertainty highlighted in the analysis are the exact mechanism by which SCT increases utility, as well as the need for further data in the later stages of disease. Further research is also needed on the methodology for meta-analysis of utility

values, where we believe Bayesian models can add to the tools presently available to analysts.

References

Acaster, S., Gaugris, S., Velikova, G., Yong, K., & Lloyd, A. (2013). Impact of the treatment-free interval on health-related quality of life in patients with multiple myeloma: a UK cross-sectional survey. *Supportive Care in Cancer*, *21*(2), 599–607.

Ashaye, A., Altincatal, A., Bender, R., Zhang, J., & Panjabi, S. (2015). Estimating Eortc-8d Health State Utility Values From Eortc Qlq-C30 Scores In Relapsed Multiple Myeloma. *Value in Health*, *18*(7), A468.

Ashaye, A., Zhang, J., Bender, R., Altincatal, A., & Panjabi, S. (2015). Mapping Utility Scores from European organization for Treatment of Cancer Core-30 Questionnaire Scores (Eortc Qlq-C30) In Relapsed Multiple Myeloma. *Value in Health*, *18*(3), A208.

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Canty, A., & Ripley, B. (2016). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-18.

Crott, R., Versteegh, M., & Uyl-de-Groot, C. (2013). An assessment of the external validity of mapping QLQ-C30 to EQ-5D preferences. *Quality of Life Research*, *22*(5), 1045–1054.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Vol. 1). Cambridge university press.

Delea, T., El Ouagari, K., Rotter, J., Wang, A., Kaura, S., & Morgan, G. (2012). Cost-effectiveness of zoledronic acid compared with clodronate in multiple myeloma. *Current Oncology*, *19*(6), e392–e403. https://doi.org/10.3747/co.19.1004

Delforge, M., Minuk, L., Eisenmann, J.-C., Arnulf, B., Canepa, L., Fragasso, A., … others. (2015). Health-related quality of life in patients with newly diagnosed multiple myeloma in the FIRST trial: lenalidomide plus low-dose dexamethasone versus melphalan, prednisone, thalidomide. *Haematologica*, S1--S145.

Dias, S., Welton, N. J., Sutton, A. J., & Ades, A. E. (2011). NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. *University of Sheffield, Decision Support Unit*, 1–24.

Hanley, J. A., Negassa, A., Forrester, J. E., & others. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, *157*(4), 364–375.

Kharroubi, S. A., Edlin, R., Meads, D., Browne, C., Brown, J., & McCabe, C. (2015). Use of Bayesian Markov Chain Monte Carlo Methods to Estimate EQ-5D Utility Scores from EORTC QLQ Data in Myeloma for Use in Cost-Effectiveness Analysis. *Medical Decision Making*, *35*(3), 351–360.

Kind, P., Dolan, P., Gudex, C., & Williams, A. (1998). Variations in population health status: results from a United Kingdom national questionnaire survey. *BMJ*, *316*(7133), 736–741.

Lambert, P. C., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, *55*(1), 86–94.

Liem, Y. S., Bosch, J. L., Arends, L. R., Heijenbrok-Kal, M. H., & Hunink, M. M. (2007). Quality of life assessed with the Medical Outcomes Study Short Form 36-Item Health Survey of patients on renal replacement therapy: a systematic review and meta-analysis. *Value in Health*, *10*(5), 390–397.

Mohty, M., Terpos, E., Mateos, M., Palumbo, A., Lejniece, S., Beksac, M., … others. (2015). Frontline therapy for multiple myeloma (MM) in real-world clinical practice: Results from the third interim analysis of the multinational, non-interventional, observational EMMOS study. *Clinical Lymphoma, Myeloma and Leukemia*, *15*, e127–e128.

Naik, H., Howell, D., Qiu, X., Brown, C., Vennettilli, A., Irwin, M., … others. (2014). *Canadian cancer site-specific health utility values: Creating the basis for measuring value and costs of therapy.* American Society of Clinical Oncology.

NICE. (2008). NICE guide to the methods of technology appraisal. Retrieved from https://www.nice.org.uk/process/pmg9/

NICE. (2009). Lenalidomide for the treatment of multiple myeloma in people who have received at least one prior therapy. Retrieved February 24, 2017, from https://www.nice.org.uk/guidance/ta171

NICE. (2013). Guide to the methods of technology appraisal. Retrieved June 8, 2017, from https://www.nice.org.uk/process/pmg9/chapter/the-reference-case

NICE. (2014). Bortezomib for induction therapy in multiple myeloma before high-dose chemotherapy and autologous stem cell transplantation. Retrieved February 24, 2017, from https://www.nice.org.uk/guidance/ta311

NICE. (2017). Pomalidomide for multiple myeloma previously treated with lenalidomide and bortezomib. Retrieved March 15, 2017, from https://www.nice.org.uk/guidance/ta427

Palumbo, A., & Cerrato, C. (2013). Diagnosis and therapy of multiple myeloma. *The Korean Journal of Internal Medicine*, *28*(3), 263–273. https://doi.org/10.3904/kjim.2013.28.3.263

Peasgood, T., & Brazier, J. (2015). Is meta-analysis for utility values appropriate given the potential impact different elicitation methods have on values? *PharmacoEconomics*, *33*(11), 1101–1105.

Pickard, A. S., Neary, M. P., & Cella, D. (2007). Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health and Quality of Life Outcomes*, *5*(1), 70. https://doi.org/10.1186/1477-7525-5-70

Proskorovsky, I., Lewis, P., Williams, C. D., Jordan, K., Kyriakou, C., Ishak, J., & Davies, F. E. (2014). Mapping EORTC QLQ-C30 and QLQ-MY20 to EQ-5D in patients with multiple myeloma. *Health and Quality of Life Outcomes*, *12*(1), 35.

Quinn, C., Hirji, I., Shingler, S. L., & Davis, C. (2015). Mapping Health State Utility Values From Eortc Data Collected From A Clinical Trial Population With Relapsed/Refractory Multiple Myeloma. *Value in Health*, *18*(7), A468–A468.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rajkumar, S. V., Richardson, P., & San Miguel, J. F. (2015). Guidelines for determination of the number of prior lines of therapy in multiple myeloma. *Blood*, *126*(7), 921–922.

Richardson, P., Sonneveld, P., Schuster, M., Irwin, D., Stadtmauer, E., Facon, T., … others. (2005). Bortezomib Continues Demonstrates Superior Efficacy Compared with High-Dose Dexamethasone in Relapsed Multiple Myeloma: Updated Results of the APEX Trail. *Blood*, *106*(11), 2547–2547.

Stan Development Team. (2016). *RStan: the R interface to Stan*. Retrieved from http://mc-stan.org/

Tengs, T. O., & Lin, T. H. (2002). A meta-analysis of utility estimates for HIV/AIDS. *Medical Decision Making*, *22*(6), 475–481.

Tengs, T. O., & Lin, T. H. (2003). A meta-analysis of quality-of-life estimates for stroke. *Pharmacoeconomics*, *21*(3), 191–200.

Uyl-de Groot, C., Buijt, I., Gloudemans, I., Ossenkoppele, G., Berg, H., & Huijgens, P. (2005). Health related quality of life in patients with multiple myeloma undergoing a double transplantation. *European Journal of Haematology*, *74*(2), 136–143.

Viechtbauer, W., & others. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, *36*(3), 1–48.

Figures & Tables

Figure 1: PRISMA diagram of included papers

Key: PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Figure 2: Nonparametric bootstrapped meta-regression of treatment line and utility in MM patients, accounting for moderation via SCT (Model 2)

Key: MM, multiple myeloma; SCT, stem cell transplant.

Figure 3: Density plot of Bayesian statistical model (Model 4)

Table 1: Utility values identified in the systematic review and included after methodological review

Table 2: Meta-analysis model parameters and 95% intervals

588 Table 1:

Author	Year	Line	EQ-5D	UK value set	Utility	SD	SCT Percent
Delea et al., 2012	2011	Newly diagnosed	Yes	Yes	0.485	0.375	Not reported
Delea et al., 2012	2011	First line	Yes	Yes	0.55	0.3	Not reported
Delea et al., 2012	2011	First line	Yes	Yes	0.55	0.3	Not reported
(Delea et al., 2012	2011	First line	Yes	Yes	0.66	0.26	Not reported
Delea et al., 2012	2011	First line	Yes	Yes	0.67	0.27	Not reported
Crott, Versteegh, & Uyl-de-Groot, 2013	2013	General disease	Yes	Yes	0.69	0.26	Not reported
Uyl-de Groot et al., 2005	2005	Newly diagnosed	Not reported	Not reported	0.6	0.33	0
Uyl-de Groot et al., 2005	2005	SCT	Not reported	Not reported	0.17	0.13	100
Uyl-de Groot et al., 2005	2005	First line	Not reported	Not reported	0.79	0.18	46.2
Kharroubi et al., 2015	2015	General disease	Yes	Yes 1	0.52	Not reported	Not reported
Acaster et al., 2013	2013	First line	Yes	Yes	0.63	0.26	8.3
Acaster et al., 2013	2013	First line	Yes	Yes	0.72	0.26	69.7
Acaster et al., 2013	2013	Second line	Yes	Yes	0.67	0.25	5.1
Acaster et al., 2013	2013	Third line	Yes	Yes	0.63	0.29	15.6
Quinn, Hirji, Shingler, & Davis, 2015	2015	Second line	Yes	Yes	0.603	0.03	Not reported
Quinn et al., 2015	2015	Second line	Yes	Yes	0.649	0.016	Not reported
Proskorovsky et al., 2014	2014	General disease	Yes	Yes	0.7	0.3	11.7
Naik et al., 2014	2014	General disease	Yes	No	0.71	0.14	Not reported
Delforge et al., 2015	2015	Newly diagnosed	Yes	Yes	0.53	0.01	Not reported
Delforge et al., 2015	2015	Second line	Yes	Yes	0.59	0.015	Not reported
Ashaye, Zhang, Bender, Altincatal, & Panjabi, 2015	2015	Second line	Yes	Yes	0.59	0.27	Not reported
Ashaye, Zhang, et al., 2015	2015	Second line	Yes	Yes	0.71	0.2	Not reported
Ashaye, Altincatal, Bender, Zhang, & Panjabi, 2015	2015	Second line	No	Yes	0.785	0.129	Not reported

Palumbo & Cerrato, 2013	2013	Third line	Yes	Yes	0.61	0.31	Not reported
Palumbo & Cerrato, 2013	2013	Fourth line	Yes	Yes	0.57	0.3	Not reported
Palumbo & Cerrato, 2013	2013	Third line	No	Yes	0.57	0.3	Not reported
Palumbo & Cerrato, 2013	2013	Fourth line	No	Yes	0.69	0.14	Not reported
Richardson et al., 2005	-	Second line	Yes	Yes	0.654	0.29	68.3
Richardson et al., 2005	-	Third line	Yes	Yes	0.619	0.312	90.4
Richardson et al., 2005	-	SCT	Yes	Yes	0.056	0.01	100
EMMOS (Mohty et al., 2015)	-	Newly diagnosed	Yes	Yes	0.459	0.396	0
EMMOS (Mohty et al., 2015)	-	First line	Yes	Yes	0.606	0.308	15.7
EMMOS (Mohty et al., 2015)	-	Second line	Yes	Yes	0.619	0.298	31
EMMOS (Mohty et al., 2015)	-	Third line	Yes	Yes	0.561	0.325	38.3
EMMOS (Mohty et al., 2015)	-	Fourth line	Yes	Yes	0.403	0.355	55.6
EMMOS (Mohty et al., 2015)	-	SCT	Yes	Yes	0.129	0.418	100
Mean general disease utility					0.655		11.7
Mean newly diagnosed utility					0.491		0
Mean first-line utility					0.627		24.6
Mean second-line utility					0.636		48
Mean third-line utility					0.610		67.6
Mean fourth-line utility					0.486		55.6
Mean SCT utility					0.093		100
Key: SCT, stem cell transplant; SD, standard deviation.							

589

590 Table 2:

Number of treatment classes received	Model 1: Meta-regression (all values)	Model 2: Meta-regression (EQ-5D only)	Model 3: Bayesian model (all values)	Model 4: Bayesian model (EQ-5D only) [preferred approach]	Model 5: Bayesian model (EQ-5D only) with weak priors
Newly diagnosed	0.529 (0.459–0.600)	0.529 (0.459–0.600)	0.530 (0.510–0.550)	0.530 (0.510–0.550)	0.530 (0.510–0.550)
One	0.659 (0.597–0.736)	0.659 (0.591–0.734)	0.646 (0.496–0.796)	0.620 (0.456–0.786)	0.626 (0.424–0.829)
Two	0.626 (0.591–0.707)	0.620 (0.590–0.650)	0.591 (0.569–0.613)	0.590 (0.568–0.612)	0.613 (0.523–0.704)
Three	0.599 (0.568–0.625)	0.606 (0.561–0.630)	0.568 (0.299–0.837)	0.578 (0.275–0.880)	0.603 (0.286–0.920)
Four (all)	0.599 (0.403–0.690)	0.494 (0.403–0.570)	0.607 (0.373–0.842)	0.469 (0.021–0.918)	0.497 (0.034–0.958)
Stem cell transplant	0.066 (0.056–0.170)	0.066 (0.056–0.170)	0.057 (0.037–0.076)	0.056 (0.037–0.076)	0.007 (-0.178–0.191)
Key: Values in parentheses are 95% confidence intervals for Models 1 and 2, and 95% credible intervals for Models 3-5					

591