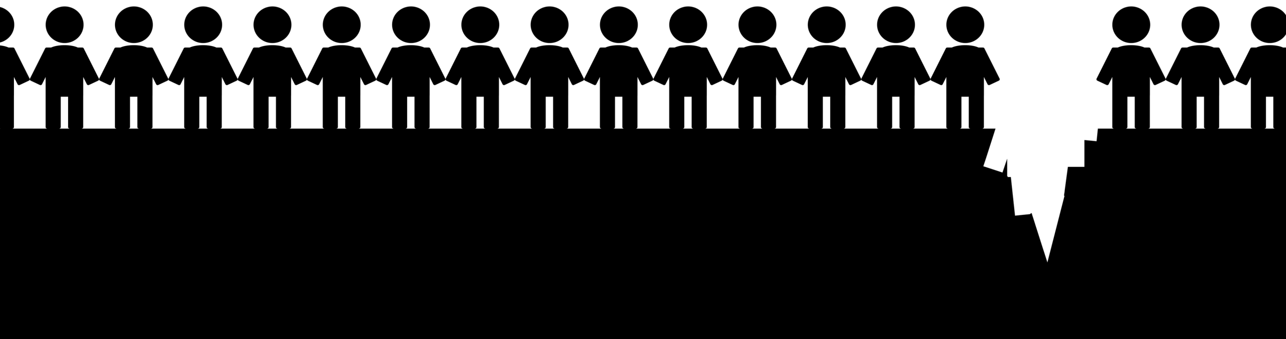
Mind the gap : explanations for the differences in utilities between respondent groups

Peeters, Y.

2

Health State Valuations of Patients and the General Public Analytically Compared

Peeters, Y. & Stiggelbout, A.M. (2010). Health State Valuations of Patients and the General Public Analytically Compared: A Meta-Analytical Comparison of Patient and Population Health State Utilities. Value in Health, 13, 306-309.



Abstract ***Objectives:*** *To obtain quality-adjusted life-years, different respondent groups, such as patients or the general public, may be asked to value health states. Until now, it remains unclear if the respondent group has an influence on the values obtained. We assessed this issue through metaanalysis.* ***Methods:*** *A literature search was performed for studies reporting valuations given by patients and nonpatients. Studies using indirect utility instruments were excluded.* ***Results:*** *From 30 eligible studies, 40 estimators were retrieved revealing a difference between respondent group (Cohen's* $d = 0.20, p < 0.01$*). When elicitation methods were analyzed separately, patients gave higher valuations than nonpatients using the time trade-off (TTO)* ($N = 25$, *unstandardized* $d = 0.05, p < 0.05$) *and the visual analog scale (VAS)* ($N = 22$, *unstandardized* $d = 0.04, p < 0.05$)*. When the standard gamble was used, no difference was seen* ($N = 24$, *unstandardized* $d = 0.01, p = 0.70$)*.* ***Conclusion:*** *In contrast with Dolders et al., our results show that patients give higher valuations than members of the general public. For future cost-utility analyses, researchers should be aware of the differential effects of respondent group for the elicitation methods TTO and VAS.*

2.1 Introduction

Valuations used in decision analyses and cost-utility analyses can be given by different groups, such as patients or the general public. Three studies have investigated the effect of response group by summing results of empirical studies,[16],[4446]. Two of these studies, a review, and a meta-analysis of prostate cancer utilities, found higher valuations given by patients. The third, a meta-analysis on varying patient groups, did not find any difference. The latter two included indirect utility instruments like the European Quality of Life Five Dimensions EQ-5D-tariff[11] or Health Utilities Index Mark (HUI)[47] and included multiple health state valuations from the same study sample.

In studies using indirect utility instruments, only patients are approached to participate, members of the public are not included as a separate sample. Such studies calculate health state utilities of members of the general public from patients' answers to a short questionnaire. These answers are put in a model captured from an earlier study[11] which generates the utility values of the general public. Therefore, including more than one study using indirect utility instruments leads to multiple health state valuations from the same subject sample, which is a violation of the assumption of independent data points. This may have led to a distortion of the standard error, an inflated sample size, and an overrepresentation of certain studies.[48] The aim of our study was to investigate through meta-analysis the influence of the respondent group on valuations avoiding this bias.

2.2 Methods

2.2.1 Search and retrieval of studies

Studies reporting valuations given by patients and by members of the general public, professionals, or proxies (which we from now on refer to as "nonpatients") were retrieved through the computerized databases PsychInfo and PubMed. Studies published between 1970 and October 2008 were searched using preferences, utility, patient, public and, respectively, time trade-off (TTO), standard gamble (SG), or visual analog scale (VAS) as key words. With the so-called snowball method, the bibliographic information of De Wit et al.,[16] Dolders et al.,[44] Bremner et al.,[46] and other retrieved studies were searched for additional studies. With the database Web of Science, we retrieved studies for the citations of the already retrieved studies.

Abstracts were examined regarding the inclusion criteria. Studies were included if they reported valuations of both patients and nonpatients, used a standard utility method (TTO, SG, or VAS), included participants 18 years, and were written in English. Studies that used indirect instruments (classification systems), that investigated mental health states, or in which nonpatients answered what they thought the patient would have answered, were excluded.

2.2.2 Data extraction

A detailed coding system was used to extract data. From each study, the mean valuations and SDs for each evaluated health state were coded for every group. If these data were not reported, authors were contacted. We excluded studies when the authors did not respond after three attempts or could not reveal the mean valuations. If only the SDs were missing, we estimated these by the weighted sum of the SDs reported in the included studies. We further coded: elicitation method, nature of the nonpatient respondent group, and various types of information about the health state description used. With the elicitation method it was coded if the TTO, VAS, or SG was used. Non-patient respondent groups were coded as professionals/proxies or members of the general public. Information about the health state description included three aspects. First, it was recoded if the patients valued a description or if they valued their own experienced health state. Second, it was denoted what kind of health state description was used; a standard EQ-5D health state description, a standard HUI health state description, or a specifically developed health state description. Thirdly, it was coded if the health state description provided an illness label. Information of the retrieved studies was independently rated by two judges (A.M.S. and Y.P.) with satisfactory agreement for most variables (Cohen's κ between 1 and 0.77). Agreement on the variable "own health state or hypothetical health state" was low, (Cohen's $\kappa = 0.61$) in three of 30 ratings the judges disagreed. All dissimilar ratings were compared and discussed until agreement was found.

2.2.3 Statistical analyses

Before all meta-analyses, the standard mean differences and sample sizes were checked for outliers. One outlier for the sample size of nonpatients was found. Specifically, Smith et al.[49] included 567 nonpatients. Studies with larger sample size are given more weight as these are assumed to be more precise. In such weighted estimation, studies with extremely large sample size can define the entire meta-analysis if these are given according weights.[50] Therefore, we recoded this study sample into the highest nonextreme sample size of nonpatients ($N = 246$). Next, we compared the results obtained with the original sample size to those obtained with the recoded sample size. Because the results remained almost unchanged, we present the data including the original sample size.

One overall meta-analysis and three subanalyses by elicitation method were performed. Before any of the analyses, data within each of the retrieved studies were combined. If more than one health state was valued in one single study, a meta-analysis on the level of this primary study was performed. The differences between patients and nonpatients were estimated for each health state and were then combined into one estimator through metaanalysis. This estimated mean difference was then used as estimator for this study in the overall meta-analysis. In studies that included more than one respondent group in either the patient or the nonpatient group, estimations of both subsamples were included. The sample size of the other group was divided by two, and used twice to compare each of the subsamples. In studies using more than one elicitation method, a meta-analysis on the level of the primary study was performed.

Using the software Comprehensive Meta-Analysis (version 2.2.046),[51] the standard mean difference, Cohen's d, and 95% confidence interval were estimated.We used Cohen's d to control for the difference in the numerical scales of TTO, SG, and VAS. For each analysis by elicitation method, the unstandardized difference was estimated, instead of Cohen's d.

The homogeneity of the sample was checked with the Q-statistic.[52] If the sample of reports appeared to be heterogeneous, random effect models were used and moderator variables were analyzed to investigate if these could explain this heterogeneity. The significance of the six moderating effects was checked using the Q-statistic. A significant contrast means that the moderator variable explains some of the heterogeneity between the groups, but it does not necessarily imply that one of the subsamples is homogeneous. For each subsample, we again investigated the Q-statistic and Cohen's d. The Duval and Tweedie's trim and fill procedure[53] gave no indication for publication bias in the overall meta-analysis, nor in the subanalyses.

2.3 Results

The search yielded 36 studies of which 30 could be included in the analyses. Two studies were excluded due to differences in elicitation method used for patients and non-patients[54,55] and two studies were excluded since the reported data was already included in another study.[56,57] In another two studies the same group of non-patients was used.[58,59] We decided to divide the sample size of this group of non-patients by two and keep the estimations of both studies in the analyses.

Of the remaining 31 studies, five studies reported other data than mean valuations. The authors of these studies were contacted. From three of these studies the authors sent the mean valuations and standard deviations by mail.[60–63] Of one study additional not-reported data was sent.[63] No mean valuations and standard deviations could be retrieved from the other two studies.[64,65] In Appendix A data of the included studies is shown.[16,49,58–63,66–85] In 23 studies, participants rated more than one health state, and in 13 studies, more than one elicitation method was used. In these studies, meta-analyses on the level of the primary study were performed.

2.3.1 Overall meta-analysis

From the included 30 studies, 40 mean differences in health state valuations between patients and nonpatients, from now on referred to as “estimators,” were extracted. The total set of estimators was heterogeneous [$Q(39) = 398.25, p < 0.01$]. Using the random effects model, the overall combined effect size for the total set was significant (Cohen’s $d = 0.20, SD = 0.06, p < 0.01$). Patients gave higher valuations compared to nonpatients. Figure 2.1presents the standardized mean differences for each study. Two moderators showed a significant contrast (Table 2.1).

Patients’ and nonpatients’ valuations were more distinct when no label was provided than when it was. Furthermore, valuations were more similar between groups when they both valued a health state description than when patients valued their own health. In terms of heterogeneity, the Q-statistic reveals that all subsamples remain heterogeneous, except for the subsample of studies without illness label. We want to emphasize that this sample consisted of only three studies. Because this subsample was homogeneous, the fixed effect model was used to test the group difference. For each subsample, the group difference is reported as Cohen’s d.

2.3.2 Meta-analysis of studies by estimation method

The set of 25 TTO estimators was heterogeneous [$Q(24) = 263.85, p < 0.01$]. The overall combined effect size revealed a difference between the response groups unstandardized $d = 0.05, SD = 0.02, p < 0.05$). Moderator analyses showed a significant contrast between studies with own health and studies with a health state description [$Q(1) = 5.93, p < 0.01$]. When patients valued their own health ($N = 3$), their valuations were different from those of nonpatients (unstandardized $d = 0.24, p < 0.01$). When both groups valued a health state description ($N = 22$), the

Figure 2.1 The 40 mean differences from the 30 included studies

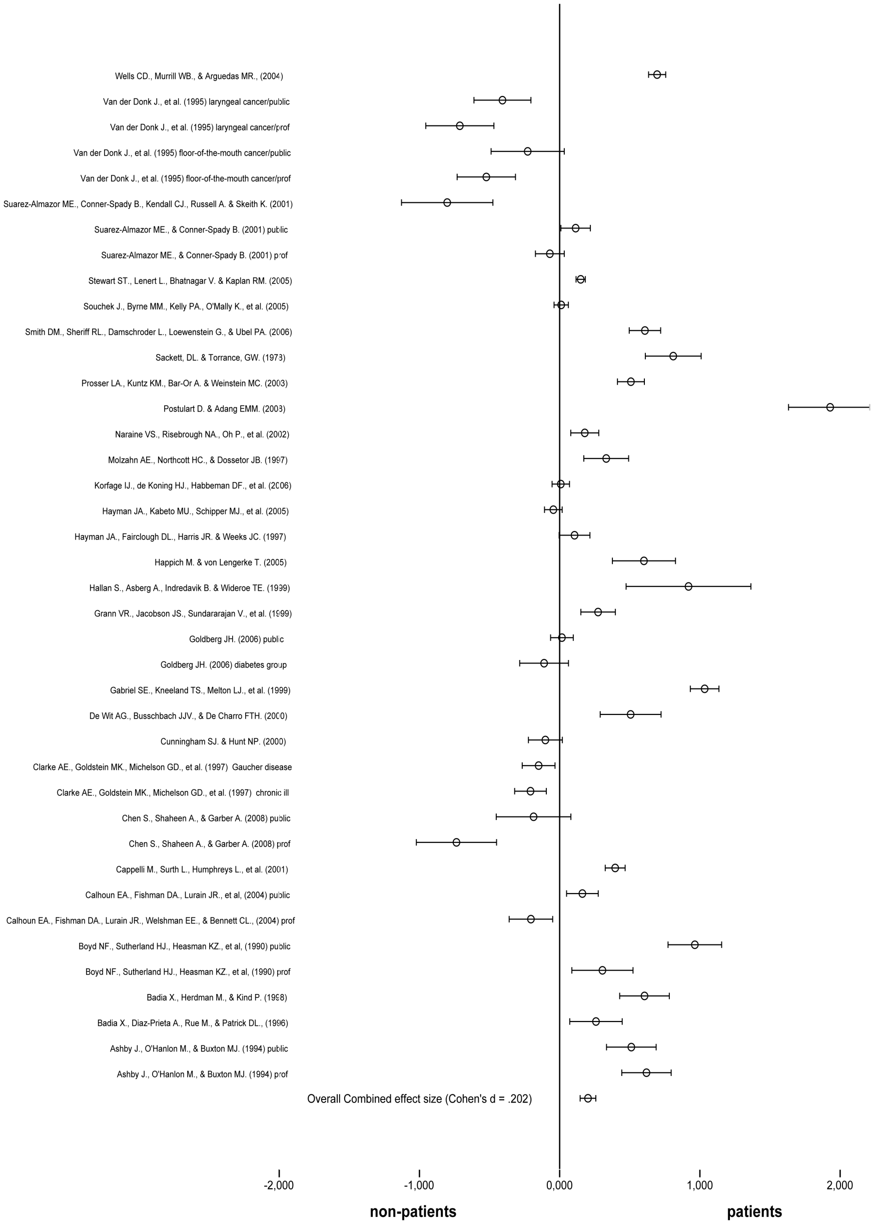


Table 2.1 Moderator variables; contrast for each of the moderators.

	N^d	Cohen's d	95% CI^a	Q^b	$Contrast^c$
Type of control group					2.54
Members of the general public	28	0.27‡	0.15 – 0.38	259.88**	
Professionals/proxies	12	0.00	−0.03 – 0.24	131.21**	
Own/hypothetical					4.63*
Scenario	34	0.14†	0.03 – 0.24	263.55**	
Own (patients valuing their own health)	6	0.67‡	0.20 – 1.15	54.71**	
Type of scenario					0.28
Description	28	0.12†	0.00 – 0.24	240.70**	
EQ-5D	6	0.18	0.00 – 0.36	17.56**	
Illness Label					4.81*
With Label	37	0.18‡	0.07 – 0.30	390.29**	
Without Label	3	0.46‡	0.24 – 0.68	1.87	
Non-patients actual/hypothetical					0.53
Actual (imagining health patient)	2	-0.20	−1.31 – 0.91	9.75**	
Hypothetical	38	0.22‡	0.10 – 0.33	388.22**	
Administration method					0.12
Computerized interview	7	0.27	−0.01 – 0.54	112.78**	
Interview without computer	30	0.21‡	0.07 – 0.35	271.62**	

$^a CI$ = Confidence Interval, $^b Q$ = heterogeneity statistic, ^{c}Contrast between sets of studies, in Q, ^{d} the total N does sometimes not add up to 41 due to missing data,*Contrast for the moderator variable is significant $p < 0.05$, † Effect size of the subsample is significant $p < 0.05$, ‡ Effect size of the subsample is significant $p < 0.01$, **Subsample is heterogeneous $p < 0.01$

valuations of the two groups were similar (unstandardized $d = 0.02, p < 0.05$). The set of 24 SG estimators was heterogeneous [$Q(23) = 116.36, p < 0.01$]. There was no significant difference between response groups (unstandardized $d = 0.01, SD = 0.01, p < 0.05$); therefore, search for moderator factors was not performed. The set of 22 VAS estimators was heterogeneous [$Q(21) = 189.47, p < 0.01$]. A difference was seen between respondent groups (unstandardized $d = 0.04, SD = 0.02, p < 0.01$). Patients valued health states higher compared with nonpatients. A significant contrast was found between professionals/proxies and members of the general public [$Q(1) = 9.53, p < 0.01$]. Professionals/proxies ($N = 6$) did not value health states different from patients (unstandardized $d = -0.04, p < 0.05$), whereas members of the general public ($N = 16$) gave lower valuations compared with patients (unstandardized $d = 0.07, p < 0.01$).

2.4 Discussion

In this meta-analysis using 40 estimators from 30 studies, we found a small to moderate difference in valuations between patients and nonpatients. This finding contrasts with the findings of Dolders et al.[44] The exclusion of studies that used indirect instruments is unlikely to have caused this, as Dolders et al. did find a difference in valuations between respondent groups in studies using indirect instruments. A smaller number of included studies is not an explanation either, because we included 29 studies compared with only 11 by Dolders et al. From these 11 studies, seven studies were selected for the current meta-analyses; of the remaining four studies included in Dolders et al., three were based on indirect health state valuations, (the EQ-5D) and one study valued health states worse than death and reported that the majority of patients were unable to complete or understand the measurement tasks. Newly published studies ($N = 10$) included in our study may partly explain the difference. Finally, the difference might be explained by the inclusion of multiple effect sizes by Dolders et al.[44] which might have led to errors.

The results of the current study showed that states providing an illness label were rated more similar by patients and nonpatients than states not providing an illness label. Possibly, healthy subjects, like patients, will not use the whole utility continuum for labelled health states.[86] Another contrast was shown between studies in which patients valued their own health and studies in which patients valued a health state description. Valuations were more similar between groups when they both valued a description. This might be explained by a so-called loss aversion, patients giving higher valuations when they "own" a health state.[15] Initially, in three studies, the judges disagreed on this moderator variable, but after reading through the studies again, agreement was easily found. The disagreement was in two studies due to poor reporting and in one study due to a poor definition.

Only in the meta-analysis including studies for the VAS was an effect for the type of nonpatient group found. Valuations of professionals/proxies were more similar to those of patients than valuations of the general public, probably because of their experience with patients. In future meta-analyses, it may be worthwhile to start off by stratifying by both disease label and type of health state valued by patients (own health vs. scenarios), as these had moderating effects.

Despite the use of several moderator factors, all samples remained heterogeneous, except for three studies without illness label. Different explanations may be given for this heterogeneity. First, a great diversity was seen between the type and severity of the health states. As shown by Insinga and Fryback,[23] the difference between valuations given by different respondent groups may depend on the severity of the health state. Second, patients as well as members of the general public differ in the extent of their experience with different health states, which creates heterogeneous groups.[17] Unfortunately, we were not able to control for the differences in experience and the choice of the particular health states.

In this study, multiple significance tests were carried out, which might have led to multiplicity. Using Bonferroni correction, the main results of the elicitation subsamples remained the same. Correcting the moderator variables in the overall metaanalysis and in the meta-analysis of studies by elicitation method, nonsignificant contrasts for all samples were found. However, it has been argued that tests performed to investigate heterogeneity should not be adjusted for multiple testing.[87] Given our results, future studies should take the impact of respondent group into account. Which respondent group should assign health state valuations depends on the research question of the study. For cost-utility analysis, the implications of our findings can be best illustrated using the unstandardized differences. Mean unstandardized difference in studies using the TTO or the VAS was 0.05 and 0.04 with a 95% confidence interval of 0.01-0.08 for the TTO and 0.01-0.07 for the VAS. The influence of such a difference on a cost-utility ratio depends on other characteristics included in the analysis, for example the period for which the effect of treatment lasts. In studies using the SG, no effect of respondent group was seen, probably due to ceiling effects caused by risk aversion.[15] Given the small sample sizes and different findings between the meta-analyses, we feel that we cannot claim implications for the findings of the moderator analyses. These results should be corroborated in future research.

We would like to thank those authors who provided additional information for their cooperation.