Validity, reliability, and ceiling and floor effects of the PROMIS Preference score (PROPr) in patients with rheumatological and psychosomatic conditions

Christoph Paul Klapproth (christoph-paul.klapproth@charite.de)
Charité - University Medicine Berlin

Felix Fischer
Charité - University Medicine Berlin

Marie Merbach
Charité - University Medicine Berlin

Rose Matthias
Charité - University Medicine Berlin

Alexander Obbarius
Charité - University Medicine Berlin

Research Article

Keywords: Patient Reported Outcomes, Quality of Life, Pain, Health Status, Outcome Measures, QALY, PROPr, EQ-5D, ICER

Posted Date: May 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-478767/v1

Validity, reliability, and ceiling and floor effects of the PROMIS Preference score (PROPr) in patients with rheumatological and psychosomatic conditions

Authors: Klapproth CP[1], Fischer F[1], Merbach M[1], Rose M[1, 2], Obbarius A[1, 3]

[1] Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117, Berlin, Germany

[2] Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, USA

[3] Dornsife Center for Self-Report Science, University of Southern California, Los Angeles, USA;

Corresponding Author:

Dr. med. Christoph Paul Klapproth, M.D.

Health Outcomes Research

Department for Psychosomatic Medicine

Charité University Medicine Berlin

Charitéplatz 1

10117 Berlin

Germany

E-mail: christoph-paul.klapproth@charite.de

Phone: +49 160 555 47 92

Phone: +49 30 450 653 353

Abstract (248 words, limit 350)

Background: The PROMIS Preference score (PROPr) is a new generic preference-based health-related quality of life (HRQoL) score that can be used as a health state utility (HSU) score for quality-adjusted life years (QALYs) in cost-utility analyses (CUAs). It is the first HSU score based on item response theory (IRT) and has favorable psychometric properties. The PROPr combines the seven PROMIS domains: cognition, depression, fatigue, pain, physical function, sleep disturbance, and ability to participate in social roles and activities. It was developed based on preferences of the US general population. We aimed to validate the PROPr in a patient sample and to compare it to the EQ-5D.

Methods: We collected PROPr and EQ-5D-5L data from 141 patients treated in the rheumatology and psychosomatic departments. We evaluated the convergent validity, reliability, known-groups construct validity, and ceiling and floor effects of the PROPr and compared those characteristics to those of the EQ-5D.

Results: The mean PROPr (0.26, 95% CI: 0.23; 0.29) and the mean EQ-5D (0.44, 95% CI: 0.38; 0.51) differed significantly ($d = 0.18$, $p < 0.001$). The Pearson correlation coefficient between the two scores was $r = 0.72$. The PROPr and EQ-5D demonstrated high reliability and similar discrimination power across sex, age, and conditions. While the PROPr showed a floor effect, the EQ-5D showed a ceiling effect.

Conclusion: PROPr and EQ-5D measure HSU differently. The PROPr shows a broader definition of perfect health, with a wider range of measurements at the top end. The previous results of excellent validity and precision were confirmed.

Key Words: Patient Reported Outcomes, Quality of Life, Pain, Health Status, Outcome Measures, QALY, PROPr, EQ-5D, ICER

Significance and Innovations

- PROPr shows good validity, high reliability, and no ceiling but floor effects in a patient sample
- PROPr yields considerably lower health state utility values than the EQ-5D, invariant to age, sex, and conditions in a patient sample
- PROPr's confidence intervals are narrower, and its variance is smaller across age, sex, and conditions in a patient sample
- The use of the PROPr as a health state utility score would greatly affect QALYs in cost-utility analyses

1. Background

The burden of autoimmune diseases, such as rheumatological conditions, was shown to be growing for the past decades. Its costs for both national health plans and society, grew accordingly(1–3). Therefore, in light of budget constraints of the national health plans, additionally stressed by the global COVID-19 pandemic, researchers in rheumatology will need to prove the value of new and more sophisticated treatments to Health Technology Assessement (HTA) agencies, who are responsible for reimbursement decision making and thus patients' access to new treatments.

These decisions rely on the new treatment's value, or cost-effectiveness, measured by the incremental cost effectiveness ratio (ICER) in costs per quality-adjusted life years (QALYs) gained compared to the standard of care (SOC)(4–6):

$$\text{ICER} = \frac{\text{Costs (new treatment)} - \text{Costs (SOC)}}{\text{QALY (new treatment)} - \text{QALY (SOC)}} = \frac{\Delta \text{ Costs}}{\Delta \text{ QALY}}$$

A QALY is the number of remaining life years multiplied by a health state utility (HSU or u) value between 0 and 1, where 0 represents death and 1 represents perfect health(7). With the remaining number of life years assumed constant in both treatments, the ICER can be expressed as:

$$\text{ICER} = \frac{\Delta \text{ Costs}}{\text{u (new treatment)} - \text{u (SOC)}} = \frac{\Delta \text{ Costs}}{\Delta \text{ u}}$$

Many HTA agencies adopted an ICER threshold. For example, the National Institute for Health and Care Excellence (NICE) in England and Wales has a threshold of 30,000 pounds per QALY gained; new treatments costing more will not be reimbursed(8–10). Thus, there are two ways for a new treatment to be cost-effective: having a low price or having a high utility (or both). It is therefore crucial for patient access to new treatments that HSU assessments be valid, precise, reliable, and responsive to change.

The HSU is obtained from preference-based health-related quality of life (HRQoL) scores, such as the SF-6D or the EuroQol EQ-5D index value (EQ-5D)(11). The choice of score has an impact on the ICER(5). The EQ-5D has five dimensions: mobility, self-care, usual activities, pain/discomfort, and depression/anxiety(12). Each is measured at five levels, defining 5^5 or 3125 health states. These health states are assigned a value between 0 and 1 or even below

0 (“worse than dead”) by preference elicitation methods such as standard gamble (SG) or time trade-off (TTO)(11,13).

The EQ-5D is endorsed by many HTA agencies(4,8). It demonstrates good psychometric performance in terms of construct validity and responsiveness in some conditions (e.g., rheumatoid arthritis or depression), while it performs poorly in others (e.g., mental health conditions)(11). In some conditions (i.e., COPD and cardiovascular diseases), the results were inconsistent(11).

However, the EQ-5D is criticized for its coarseness and limited range of measurement, indicated by ceiling effects(14). Surprisingly, at least to our knowledge, there is little debate about the consequences on the ICER: coarseness makes Δu subject to chance: it can be either underestimated, resulting in a high ICER and a denial of reimbursement, or overestimated, resulting in a low ICER and public welfare loss. A ceiling effect can be expressed as $u(SOC) \longrightarrow 1$, and consequently, $\Delta u \longrightarrow 0$ and ICER$\longrightarrow \infty$, resulting in denial of reimbursement and limiting patient access to new treatments. Furthermore, when condition-specific instruments are used for assessment, mapping is needed to predict HSU, resulting in further loss in precision, exacerbating the impact on the ICER(15–17).

These shortcomings motivated the development of a new score based on instruments from the Patient Reported Outcome Information System (PROMIS), called the PROMIS Preference (PROPr) score(14,18–22). The PROMIS is a common metric for a broad collection of item banks based on item response theory (IRT), which covers a wide range of different aspects of physical, mental, and social health (e.g., physical function, depression, pain).

These item banks give researchers the flexibility to collect domains of interest with instruments of their choice (e.g., predetermined short forms or computer adaptive tests) to tailor the measurement range to a specific population(23–27). PROMIS scores can be derived from legacy measures anchored on the PROMIS metric, such as the PHQ-9 for depression(28). PROMIS therefore provides a framework to assess HRQoL, which can in turn be used to construct health states for economic evaluations.

Seven PROMIS domains were included in the PROPr: cognition, depression, fatigue, pain interference, physical function, sleep disturbance, and ability to participate in social roles and activities. The PROPr was valuated with the preferences of the US population using online SG(14,18–22). It leverages the excellent psychometric properties of the PROMIS item banks: high validity and reliability, avoidance of floor and ceiling effects, high sensitivity to

change, high precision, and small sample size requirements(23–27). Therefore, the PROPr has the potential to overcome limitations such as coarseness and ceiling effects. The known-groups construct validity and convergent validity of the PROPr have already been shown in US general population samples(19,21).

In this study, we seek (1) to confirm the validity and reliability of the PROPr in a sample of patients with rheumatological and psychosomatic conditions, (2) to compare the measurement properties of the PROPr and the EQ-5D, which is known to perform well in this patient group(11), (3) to investigate the causes of the differences between the two scores by analyzing their ceiling and floor effects and (4) to discuss the consequences for the ICER that arise from the different conceptualizations of the two scores.

2 Patients and Methods

2.1 Samples

We conducted a cross-sectional study and invited patients undergoing inpatient treatment at the Department of Rheumatology and Immunology and the Department of Psychosomatic Medicine at Charité - Universitätsmedizin Berlin to participate. Patients participated between March 2018 and August 2018 (rheumatology department) and between October 2018 and June 2019 (psychosomatic department). Informed consent was obtained from all patients directly as none of them had legal guardians. A set of various HRQoL questionnaires was administered to the patients, including 14 PROPr items, the 5 EQ-5D-5L items, and sociodemographic questions. In rheumatology, items were administered in a paper-and-pencil form. Data were collected electronically in the psychosomatic department. There were no inclusion or exclusion criteria regarding condition, age, sex or kind of treatment applied. Reasons for exclusion were previous participation in this study, impaired vision, illiteracy, language barrier, and inability to use the tablet.

This study was approved by Charité's Ethics Committee (EA/133/17) and was conducted in accordance with the Declaration of Helsinki.

2.2 Measures

PROMIS Preference score (PROPr)

The PROPr is a new preference-based HRQoL score developed and copyrighted by the PROMIS Health Organization. It is based on the PROMIS framework and is therefore the first HSU based on IRT. It aggregates seven PROMIS domains: cognition, depression, fatigue, pain, physical function, sleep disturbance, and ability to participate in social roles and activities. We used the two recommended items from each of the 7 domains (i.e., 14 items in total, Table 1)(14,18–22). Both items for cognition and ability to participate in social roles and activities and one fatigue item had not yet passed the translation process into German at the time of our survey; therefore, we used the preliminary translations. In the interim, the final versions of three of these five items were released and were very similar to the preliminary versions (see appendix, Table A1).

Each of these items is measured on five levels (e.g., “never”, “rarely”, “sometimes”, “often”, “always” or “not at all”, “a little bit”, “somewhat”, “quite a bit”, and “very much”) and, except for physical function, refers to the past 7 days. The responses (1 to 5 on a Likert scale) were transformed to PROMIS theta scores (mean = 0 +/- SD = 1) or T-Scores (mean = 50 +/- SD = 10) (http://www.healthmeasures.net/score-and-interpret/calculate-scores). For cognition, physical function, and ability to participate in social roles and activities, higher T-Scores (thetas) indicate more desirable outcomes. For depression, fatigue, pain, and sleep disturbance, lower T-Scores (thetas) indicate more desirable outcomes.

Theta scores were fed into a multi-attribute utility (MAUT) function to obtain the PROPr, ranging from -0.022 to 1.00. Negative values indicate a health state “worse than dead”. The MAUT function was derived from the preferences of a representative sample of the US population by online SG and is available online as the R code used in this study(21,29).

[Insert Table 1 here]

EQ-5D-5L index value

The EQ-5D is a preference-based instrument to measure HRQoL that consists of five health dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The EQ-5D-5L differentiates five levels for each dimension: “No problems” (score: 1), “Slide problems” (2), “Moderate problems” (3), “Severe problems” (4), and “Extreme problems” (5). The frame of reference is “Today”. All items and response options yield 5^5 or 3125 different health states. The value assigned to each health state was determined through cTTO following the standardized EuroQol VT protocol(30–32). We use the US value set, as the PROPr was also valuated in US preferences, avoiding systematic differences due to different valuation populations. A health state of 11111 (i.e., each of the five items is answered with ‘1’) has an HSU of 1.00 (perfect health). The worst health state, 55555, corresponds to an HSU of -0.573. Negative HSUs are considered “worse than dead”.

2.3 Statistical analysis

First, we investigated the convergent validity between the PROPr and the EQ-5D by calculating Pearson correlations, as HSU scores are measured on an interval scale: $r > 0.7$ refers to high, $r > 0.5$ refers to intermediate and $r < 0.5$ refers to low convergent validity(33).

Second, to determine reliabilities, we assessed Cronbach’s α, where $\alpha > 0.70$ was defined as high reliability(34).

Third, for known-groups construct validity, we investigated how differences between the two scores are affected by age, sex, and condition. We modeled HSU using a linear regression model with interaction terms. The regression equations were defined as HSU = intercept + instrument + age/sex/condition + instrument*age/sex/condition. We hypothesized that there is a significant instrument effect, as indicated by earlier research. We modeled interaction terms between instrument and age, sex, and condition to investigate whether the effects of age/sex/conditions on HSU depend on the instrument used.

Fourth, we compared ceiling and floor effects on subscale and score levels. If more than 15% of the sample scored within the top (bottom) 20% of the respective measurement scale, this defines a significant ceiling (floor) effect. This is the equivalent of scoring the highest or lowest score on a 5-point Likert scale that was used at the item level. Moderate ceiling (floor) effects are defined by > 10%, minor by > 5%, and negligible by < 5% of the sample(35).

We used Microsoft Excel 2016, IBM SPSS Statistics version 25 and R 4.0.0 with packages eq5d version 0.7.0, ggplot2 version 3.3.0, psych version 1.9.12.31, tidyverse 1.3.0, lme4 1.1-23, lmerTest 3.1-3, and reshape2 version 1.4.4 for analyses.

3 Results

3.1 Sample

We used a combined patient sample of 141 patients receiving inpatient diagnostics and treatment in our rheumatology and psychosomatic departments. In the rheumatology (psychosomatic) department, of 236 (157) patients screened, 10 (14) patients were not eligible to participate. A total of 162 (90) patients agreed to participate and received questionnaires after informed consent was obtained. A total of 118 (58) sets were returned, while 44 (32) participants either declined to participate after receiving questionnaires or were discharged early and did not return a questionnaire. In rheumatology, of these 118 returned sets, 83 completed all PROPr and EQ-5D items needed for this study. Skipped items may be explained by a high response burden, as our items were part of a larger survey comprising various HRQoL questionnaires. Tablet administration in the psychosomatic department prevented participants from skipping items, as only complete questionnaires were submitted to the database. Sample characteristics can be obtained from Table 2. In both samples, two-thirds of participants were female, and almost two-thirds lived with a partner. More than 85% of both samples were of German nationality. Ethnicity was not assessed, but as most Germans are Caucasian, nationality can serve as a proxy for ethnicity. Approximately one-third had a master's, bachelor's or doctoral degree, with rheumatology patients having a higher level of education than psychosomatic patients. One out of five was working part- or full-time, while more than one-third was unable to work for health reasons. Patients from the rheumatology department were more likely to be unable to work for health reasons than patients from the psychosomatic department. In the rheumatology department, arthritis (rheumatoid arthritis, psoriasis arthritis, ankylosing spondylitis, and gout arthritis), systemic sclerosis (SScl), systemic lupus erythematosus (SLE), and vasculitis (small and large vessel vasculitis, polymyalgia rheumatica) were the most frequent conditions. In the psychosomatic department, all patients had persistent somatoform pain disorder (PSPD, ICD-10 code F45.4) as the primary condition.

[Insert Table 2 here]

3.2 Convergent validity, distribution, and reliability

Table 2 shows that patients from the psychosomatic department reported a lower HRQoL than patients from the rheumatology department in all EQ-5D and PROPr domain scores and had significantly lower HSU scores. The PROPr showed narrower confidence intervals in both samples. For the remaining analysis, we used the total sample. The mean PROPr (0.26, 95% CI: 0.23; 0.29) score was significantly lower ($d = 0.18$, $p < 0.001$) than the mean EQ-5D (0.44, 95% CI: 0.38; 0.51). The correlation was $r = 0.72$ ($p < 0.001$), indicating high convergent validity.
On the subscale level, we observed moderate to high correlations: $r = -0.58$ (95% CI: -0.68; -0.47) for EQ-5D-5L mobility vs PROMIS physical function, $r = -0.60$ (95% CI: -0.70; -0.49) for EQ-5D-5L self-care vs PROMIS physical function, $r = 0.65$ (95% CI: 0.54; 0.74) for EQ-5D-5L pain/discomfort vs PROMIS pain interference, $r = 0.70$ (95% CI: 0.61; 0.78) for EQ-5D-5L anxiety/depression vs PROMIS depression, and $r = -0.66$ (95% CI: -0.74; -0.55) for EQ-5D-5L usual activities vs PROMIS ability to participate in social roles and activities.
Scatterplots and histograms of the EQ-5D and the PROPr illustrate the positive correlation (Figure 1). While the distribution of the PROPr is positively skewed, the distribution of the EQ-5D scores is negatively skewed. The median and mean PROPr scores are closer (Mn = 0.22, M = 0.26) than the median and mean EQ-5D scores (Mn = 0.50, M=0.44). Following a Shapiro-Wilk test, the hypothesis of normal distribution was rejected for both scores ($p < 0.001$). The EQ-5D measurement range (-0.425; 1.00) is wider than the PROPr measurement range (-0.004; 0.955). For both scores, Cronbach's α was estimated at $\alpha = 0.83$, showing high reliability.

[Insert Figure 1 here]

Figure 1: Scatterplot and histograms of PROPr and EQ-5D

3.3 Known-groups construct validity

Figure 2 illustrates the discrimination of the EQ-5D and the PROPr across sex and age. For both sex groups, the EQ-5D shows higher values than the PROPr. The PROPr's narrower confidence intervals indicate higher precision. The two scores show similar trends in both groups.

[Insert Figure 2 here]

Figure 2: Health state utility measured in both EQ-5D and PROPr, dependent on sex and age. Lines are Loess smoothers. The light background represents confidence bands.

Figure 3 illustrates how the two scores differentiate between the five most frequent conditions. The PROPr consistently measures HSU lower than the EQ-5D and has smaller confidence intervals. The hierarchy is the same: the best HSU is assigned to patients with SLE and the worst to patients with PSPD. Patients with systemic sclerosis had higher HSU than those with vasculitis and arthritis.

The EQ-5D range is wider, with a difference of 0.62 from 0.84 in patients with SLE to 0.22 in patients with PSPD. The PROPr's range between SLE (0.34) and PSPD (0.16) is only 0.18.

With the PROPr, for all conditions, the majority of the respective subsample scores a lower HSU than the population average of 0.5(21,22). With the EQ-5D, a considerable share of patients with SLE and vasculitis do not show a lower HSU than the population average of 0.9. As no recent HRQoL data measured by the EQ-5D-5L (US value set) are available for the US general population, we used studies from comparable countries and the 3L value set as a comparator, assuming values would be similar(36–40). Linear regression did not show significant interactions between instrument and age and sex ($p<0.05$). Hence, the difference between the EQ-5D and PROPr is consistent across age and sex. For conditions, due to our small sample size, the linear regression showed interaction terms that were not statistically significant but were nonetheless large.

[Insert Figure 3 here]

Figure 3: Health state utility measured in both EQ-5D and PROPr, dependent on condition

The dashed line at HSU = 0.9 refers to the estimated population average of the EQ-5D(36–40); the dashed-dotted line at HSU = 0.5 refers to the estimated population average of the PROPr(21,22); SLE = systemic lupus erythematosus, PSPD = persistent somatoform pain disorder.

3.4 Ceiling and floor effects

Table 3 shows ceiling and floor effects on the subscale and score levels. At the average subscale level, neither score shows significant floor effects, but the EQ-5D shows a significant ceiling effect. At the score level, the EQ-5D shows a significant ceiling effect (30.50%), while the PROPr shows a significant floor effect (41.84%).
PROPr physical function shows a mild ceiling effect (8.51%), while the two corresponding EQ-5D dimensions, mobility and self-care, show high ceiling effects (28.37% and 58.16%, respectively). EQ-5D pain/discomfort and PROPr pain interference showed similar ceiling effects (10.64% and 8.51%). The comparison of ceiling effects of the depression subscales (33.34% vs 15.6%) and usual activities/ability to participate in social roles and activities (20.58% vs 14.18%) also favored PROPr subscales.

[Insert Table 3 here]

4 Discussion

We investigated the measurement properties of the PROPr in a sample of patients with rheumatological and psychosomatic conditions and compared them to those of the EQ-5D.
Our first result confirms that the PROPr is a valid and reliable instrument for HSU in this patient group. Its psychometric properties of convergent validity, reliability, and known-groups construct validity are comparable with those in a general population sample(21,22).

Our second finding confirms that the PROPr yields a lower HSU than the EQ-5D. The mean difference was 0.18, which is smaller than that in the US general population (approximately 0.30). These differences were statistically invariant to age, sex, and condition. For age and sex, the interaction terms were small, though they were larger for conditions. Therefore, there could be an instrument effect in conditions that we were not able to detect due to the small sample size. The original PROPr validation used the EQ-5D-3L crosswalk value set, while we used the 5L value set(21,22). Therefore, the mean difference between PROPr and EQ-5D is lower with the 5L value sets, while the correlation remains comparable. Many 5L value sets yield lower HSUs than the corresponding 3L crosswalk value sets(38). Thus, the population average comparator might be lower than 0.9. The EQ-5D shows more considerable differences in only 3 of 5 conditions. The PROPr does so in all conditions, indicating higher sensitivity, as earlier results suggested(22).

Our third finding offers an explanation for the mean differences of the two scores: the PROPr shows no ceiling effect but a significant floor effect. Interestingly, while the PROPr moves from an approximate normal distribution with no floor effects in a general population sample to a positively skewed distribution with a floor effect in a patient sample, the EQ-5D is negatively skewed, showing ceiling effects in both the general population and patient samples(22,29). The floor effect of the PROPr was minor at the subscale level and cannot be resolved by choosing more sample-specific items. We attribute the floor effect to the theoretical requirement of PROPr's MAUT algorithm that HSU needs to have a lower boundary of 0(19,21). Ceiling effects are a well-known issue for the EQ-5D that remain unresolved in the 5L version(37). Our results suggest that the EQ-5D's ceiling effect does not result from its valuation technique but from its dimension items. For example, the PROPr physical function domain outperforms the corresponding EQ-5D dimensions mobility and self-care in terms of ceiling effects. Additionally, SG, the PROPr valuation method, yields 5-15% higher HSU than TTO, the EQ-5D valuation method, attributable to risk aversion(41). Thus, the difference could be larger if valuation was performed with the same method and without limited ranges of measurement at either end.

Finally, we discuss the consequences on the ICER if the PROPr is used as the HSU score, namely, SOC health states would have a lower HSU. The scale could measure improvements due to its wider range of measurement at the healthy end of the scale, possibly pushing the ICER under the threshold, making more treatments cost-effective. In our sample, the improvement of physical function could be measured with the PROPr but not with the EQ-5D. Physical function is considered one of the key issues for rheumatological patients; therefore, an HSU score should

capture its improvement(42,43). However, the PROPr may not capture the full extent of severity, as the floor effect suggests, though the ICER is not needed to measure deterioration but improvement. Both unjustified denial of reimbursement and welfare loss could be avoided. Additionally, the PROPr includes 3 additional domains, making HTA more comprehensive and more sensitive to other symptoms.

Future research should investigate which score more accurately reflects the patients' judgment and the clinical judgment of physicians. Also, the valuation of PROPr in German and other European preferences is necessary for its use in European HTA.

Strengths and limitations

This study investigates the convergent validity, known-groups validity, ceiling and floor effects, and reliability of the PROPr in a patient sample, contributing to the knowledge about its measurement properties. A limitation is the small sample size, hardly reflecting the broad range of chronically ill patients. Therefore, our results need to be confirmed in other conditions and larger samples. Additionally, our data do not provide information about disease severity. Furthermore, this is a self-selected sample with a relatively high proportion of drop-outs and incomplete assessments, which might have led to selection bias. Future studies should aim to reduce the drop-out rate, such as by offering an incentive to respondents and lowering the response burden.

5 Conclusion

Our analysis confirms that PROPr and EQ-5D measure HSU differently. The PROPr shows some favorable properties, such as the avoidance of ceiling effects, high validity, and high precision, but shows a floor effect. The application of PROPr would greatly affect the ICER for new treatments of rheumatological conditions.

List of Abbreviations

3L	3 levels
5L	5 levels
CAT	Computerized adaptive testing
CI	Confidence interval
COPD	Chronic obstructive pulmonary disease
cTTO	composite time trade-off
CUA	Cost-utility analysis
EQ-5D	EuroQoL 5 Dimensions 5 Level index value
HRQoL	Health-Related Quality of life
HSU	Health state utility
HTA	Health Technology Assessments
ICD-10	International Classification of Disease, 10^{th} revision
ICER	Incremental Cost-Effectiveness Ratio
IRT	Item response theory
NICE	National Institute of Health and Clinical Excellence
M	Mean
MAUT	Multiattribute Utility Theory
Mn	Median
PHQ-9	Patient-Health Questionnaire-9
PROMIS	Patient Reported Outcome Measurement Information System
PROMIS-29 v2.0	PROMIS Profile 29 Version 2.0
PROPr	PROMIS Preference Score

PSPD	Persistent somatoform pain disorder
QALY	Quality-adjusted life years
SD	Standard Deviation
SG	Standard gamble
SLE	Systemic lupus erythematosus
SOC	Standard of care
SS	Sjoegren's syndrome
SSCl	Systemic sclerosis
TTO	Time Trade-off
u	Health state utility
US	United States of America
VAS	Visual Analogue Scale

Declarations

Ethical Approval and Consent to participate:

Ethical approval: All procedures performed in this study were approved by Charité's Ethics Committee (EA/133/17), in accordance with the ethical standards of the institutional and/or national research committee and the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study. Participation was voluntary.

Consent for publication: not applicable.

Availability of data and material: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request

Competing interests: Authors declare that they have no competing interests.

Funding: none

Authors' contributions: C.P.K. conducted the statistical analysis, wrote the main manuscript text, and prepared all figures and tables. C.P.K., A.O. and M.M. conducted the sampling. All authors reviewed the manuscript.

Acknowledgements: Not applicable.

References

1. Uhlig T, Moe RH, Kvien TK. The Burden of Disease in Rheumatoid Arthritis. Pharmacoeconomics. 2014;32(9):841–51.

2. Cross M, Smith E, Hoy D, Carmona L, Wolfe F, Vos T, et al. The global burden of rheumatoid arthritis: Estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis. 2014;73(7):1316–22.

3. Association AARD, Groups NC of AP. The Cost Burden of Autoimmune Disease: The Latest Front in the War on Healthcare Spending. Am Autoimmune Relat Dieseases Assoc [Internet]. 2011;14. Available from: www.aarda.org/pdf/cbad.pdf

4. EUnetHTA Joint Action 2, Work Package 7, Subgroup 3, Heintz E, Gerber-Grote A, Ghabri S, Hamers FF, Rupel VP, et al. Is There a European View on Health Economic Evaluations? Results from a Synopsis of Methodological Guidelines Used in the EUnetHTA Partner Countries. Pharmacoeconomics. 2016;34(1):59–76.

5. Kvamme MK, Lie E, Uhlig T, Moger TA, Kvien TK, Kristiansen IS. Cost-effectiveness of TNF inhibitors vs synthetic disease-modifying antirheumatic drugs in patients with rheumatoid arthritis: A Markov model study based on two longitudinal observational studies. Rheumatol (United Kingdom). 2015;54(7):1226–35.

6. Bang H, Zhao H. Median-based incremental cost-effectiveness ratio (ICER). J Stat Theory Pract. 2012;6(3):428–42.

7. Weinstein MC, Torrance G, McGuire A. QALYs: The basics. Value Heal [Internet]. 2009;12(SUPPL. 1):S5–9. Available from: http://dx.doi.org/10.1111/j.1524-4733.2009.00515.x

8. NICE. Guide to the Methods of Technology Appraisal [Internet]. NICE Guidelines. 2013. Available from: nice.org.uk/process/pmg9

9. Van Lier A, Van Hoek AJ, Opstelten W, Boot HJ, De Melker HE. Assessing the potential effects and cost-effectiveness of programmatic herpes zoster vaccination of elderly in the Netherlands. BMC Health Serv Res. 2010;10.

10. Thokala P, Ochalek J, Leech AA, Tong T. Cost-Effectiveness Thresholds: the Past, the Present and the Future. Pharmacoeconomics [Internet]. 2018;36(5):509–22. Available from: https://doi.org/10.1007/s40273-

017-0606-1

11. Brazier J, Ara R, Rowen D, Chevrou-Severac H. A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models. Pharmacoeconomics. 2017;35(s1):21–31.

12. Olsen JA, Lamu AN, Cairns J. In search of a common currency: A comparison of seven EQ-5D-5L value sets. Health Econ. 2018;27(January 2017):39–49.

13. Weernink MGM, Janus SIM, van Til JA, Raisch DW, van Manen JG, IJzerman MJ. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. Pharmaceut Med. 2014;28(4):175–85.

14. Hanmer J, Feeny D, Fischhoff B, Hays RD, Hess R, Pilkonis PA, et al. The PROMIS of QALYs. Health Qual Life Outcomes [Internet]. 2015;15–7. Available from: http://dx.doi.org/10.1186/s12955-015-0321-6

15. Mukuria C, Rowen D, Harnan S, Rawdin A, Wong R, Ara R, et al. An Updated Systematic Review of Studies Mapping (or Cross-Walking) Measures of Health - Related Quality of Life to Generic Preference - Based Measures to Generate Utility Values. Appl Health Econ Health Policy [Internet]. 2019;17(3):295–313. Available from: https://doi.org/10.1007/s40258-019-00467-6

16. Revicki DA, Kawata AK, Harnam N, Chen W-H, Hays RD, Cella D. Predicting EuroQol (EQ-5D) scores from the patient-reported outcomes measurement information system (PROMIS) global items and domain item banks in a United States sample. Qual Life Res. 2009;18(6):783–91.

17. Klapproth CP, van Bebber J, Sidey-Gibbons CJ, Valderas JM, Leplege A, Rose M, et al. Predicting EQ-5D-5L crosswalk from the PROMIS-29 profile for the United Kingdom, France, and Germany. Health Qual Life Outcomes [Internet]. 2020;18(1):1–13. Available from: https://doi.org/10.1186/s12955-020-01629-0

18. Hanmer J, Cella D, Feeny D, Fischhoff B, Hays RD, Hess R, et al. Selection of key health domains from PROMIS® for a generic preference-based scoring system. Qual Life Res. 2017;26(12):1–9.

19. Dewitt B, Feeny D, Fischhoff B, Cella D, Hays RD, Hess R, et al. Estimation of a Preference-Based Summary Score for the Patient-Reported Outcomes Measurement Information System: The PROMIS®-Preference (PROPr) Scoring System. Med Decis Mak. 2018;38(6):683–98.

20. Hanmer J, Cella D, Feeny D, Fischhoff B, Hays RD, Hess R, et al. Evaluation of options for presenting health-states from PROMIS ® item banks for valuation exercises. Qual Life Res [Internet]. 2018;27(7):1835–43. Available from: http://dx.doi.org/10.1007/s11136-018-1852-1

21. Hanmer J, Dewitt B. The Development of a Preference-based Scoring System for PROMIS® (PROPr): A Technical Report Version 1.4. 2017.

22. Hanmer J, Dewitt B, Yu L, Tsevat J, Roberts M, Revicki D, et al. Cross-sectional validation of the PROMIS- Preference scoring system. PLoS One. 2018;13(7):1–13.

23. Embretson SE, Reise SP. Item Response Theory For Psychologists. Psychology Press; 2013.

24. PROMIS Cooperative Group. PROMIS ® Instrument Maturity Model [Internet]. 2012. p. 1–4. Available from: http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers_2_0_MaturityModelOnly_508.pdf

25. Rupp AA, Zumbo BD. Understanding parameter invariance in unidimensional IRT models. Educ Psychol Meas. 2006;66(1):63–84.

26. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and promis: Assessment of physical function. J Rheumatol. 2014;41(1):153–8.

27. Hays RD, Revicki DA, Feeny D, Fayers P, Spritzer KL, Cella D. Using Linear Equating to Map PROMIS Global Health Items and the PROMIS-29 V2.0 Profile Measure to the Health Utilities Index Mark 3. Pharmacoeconomics. 34(10):1015–22.

28. Choi SW, Schalet B, Cook KF, Cella D. Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. Psychol Assess. 2014;26(2):513–527.

29. Hanmer J, Dewitt B. PROPr MAUT R code [Internet]. 2017 [cited 2020 Jun 4]. Available from: https://github.com/janelhanmer/PROPr/blob/master/Generic MAUT code 2017_09_02.R

30. Pickard AS, Law EH, Jiang R, Pullenayegum E, Shaw JW, Xie F, et al. United States Valuation of EQ-5D-5L Health States Using an International Protocol. Value Heal [Internet]. 2019;22(8):931–41. Available

from: https://doi.org/10.1016/j.jval.2019.02.009

31. Ludwig K, Schulenburg JG Von Der, Greiner W, Ludwig K. German Value Set for the EQ-5D-5L. Pharmacoeconomics [Internet]. 2018;36(6):663–74. Available from: https://doi.org/10.1007/s40273-018-0615-8

32. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. Pharmacoeconomics. 2016;34(10):993–1004.

33. Hott A, Liavaag S, Juel NG, Brox JI, Ekeberg OM. The reliability, validity, interpretability, and responsiveness of the Norwegian version of the Anterior Knee Pain Scale in patellofemoral pain. Disabil Rehabil [Internet]. 2019;0(0):1–10. Available from: https://doi.org/10.1080/09638288.2019.1671499

34. Bilbao A, García-Pérez L, Arenaza JC, García I, Ariza-Cardiel G, Trujillo-Martín E, et al. Psychometric properties of the EQ-5D-5L in patients with hip or knee osteoarthritis: reliability, validity and responsiveness. Qual Life Res [Internet]. 2018;27(11):2897–908. Available from: http://dx.doi.org/10.1007/s11136-018-1929-x

35. Gulledge CM, Smith DG, Ziedas A, Muh SJ, Moutzouros V, Makhni EC. Floor and Ceiling Effects, Time to Completion, and Question Burden of PROMIS CAT Domains Among Shoulder and Knee Patients Undergoing Nonoperative and Operative Treatment. JBJS Open Access. 2019;4(4):e0015.

36. Grochtdreis T, Dams J, König HH, Konnopka A. Health-related quality of life measured with the EQ-5D-5L: estimation of normative index values based on a representative German population sample and value set. Eur J Heal Econ [Internet]. 2019;20(6):933–44. Available from: https://doi.org/10.1007/s10198-019-01054-1

37. Martí-Pastor M, Pont A, Ávila M, Garin O, Vilagut G, Forero CG, et al. Head-to-head comparison between the EQ-5D-5L and the EQ-5D-3L in general population health surveys. Popul Health Metr. 2018;16(1):1–11.

38. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L Better Than EQ-5D-3L? A Head-to-Head Comparison of Descriptive Systems and Value Sets from Seven Countries. Pharmacoeconomics [Internet]. 2018;36(6):675–97. Available from: https://doi.org/10.1007/s40273-018-0623-8

39. Ko Y, Coons SJ. Self-reported chronic conditions and EQ-5D index scores in the US adult population. Curr Med Res Opin. 2006;22(10):2065–71.

40. Brennan DS TD. Comparing UK, USA and Australian values for EQ- 5D as a health utility measure of oral health. Community Dent Heal. 2015;32(3):180–4.

41. Busschbach JJ, Hessing DJ, De Charro FT. An Empirical Comparison of four Measurements of Quality of Life: Standard Gamble , Time Trade-off , the EuroQoL-Visual Analog Scale and the Rosser & Kind Matrix. In: The EuroQol Group after 25 years. 1992.

42. Liegl G, Gandek B, Fischer HF, Bjorner JB, Jr JEW, Rose M, et al. Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. 2017;1–12.

43. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: Responsiveness, floor and ceiling effects, and internet administration. J Rheumatol. 2011;38(8):1759–64.

Appendix

[Insert Table A1 here]

Figures

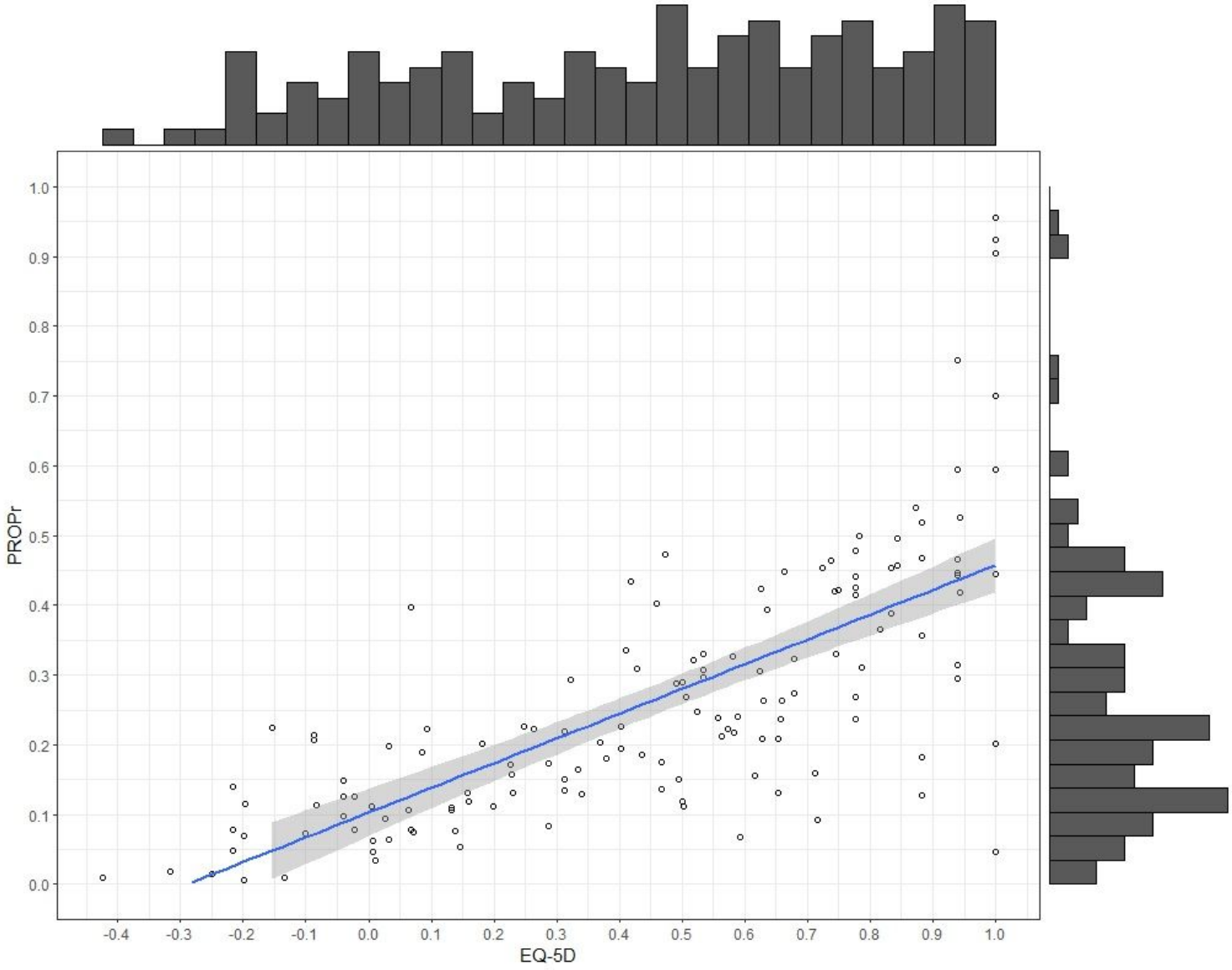

Figure 1

Scatterplot and histograms of PROPr and EQ-5D

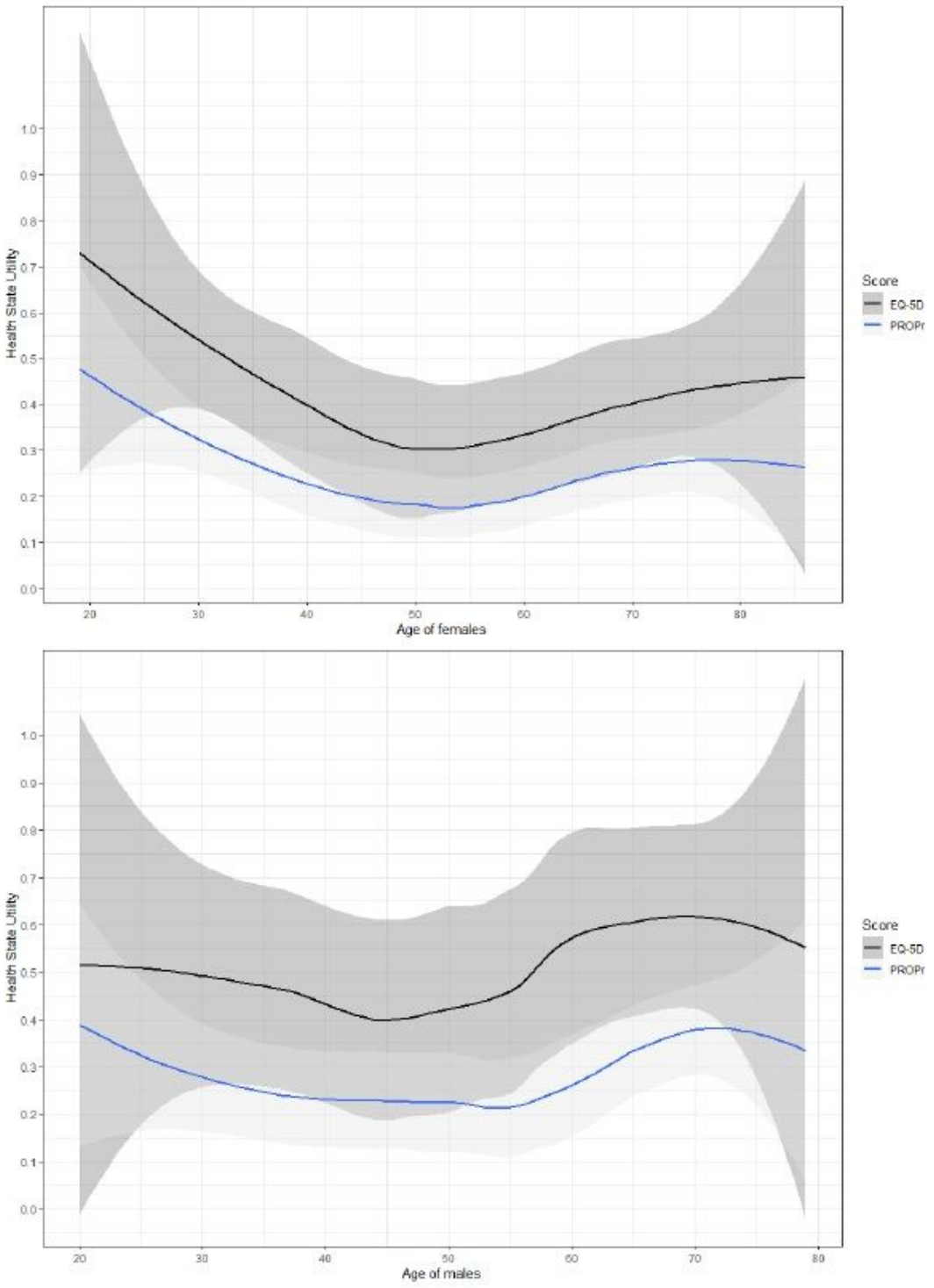

Figure 2

Health state utility measured in both EQ-5D and PROPr, dependent on sex and age. Lines are Loess smoothers. The light background represents confidence bands.

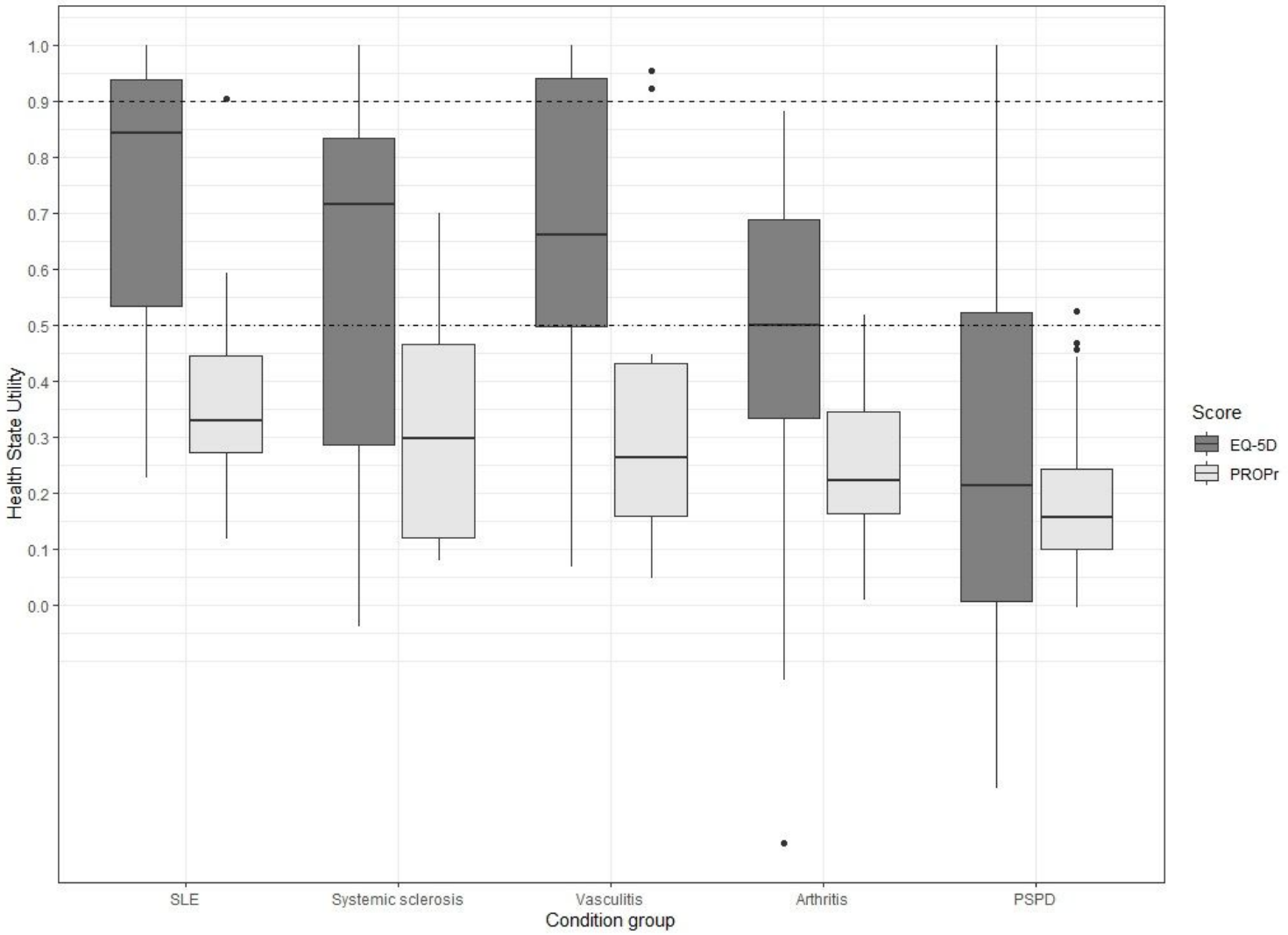

Figure 3

Health state utility measured in both EQ-5D and PROPr, dependent on condition The dashed line at HSU = 0.9 refers to the estimated population average of the EQ-5D(36–40); the dashed-dotted line at HSU = 0.5 refers to the estimated population average of the PROPr(21,22); SLE = systemic lupus erythematosus, PSPD = persistent somatoform pain disorder.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PROPrEQ5DBMCrheumatologyAppendixTable1A.docx