Applied Health Economics and Health Policy
https://doi.org/10.1007/s40258-023-00837-1

ORIGINAL RESEARCH ARTICLE

Different Frameworks, Similar Results? Head-to-Head Comparison of the Generic Preference-Based Health-Outcome Measures CS-Base and EQ-5D-5L

Xin Zhang[1] · Karin M. Vermeulen[1] · Paul F. M. Krabbe[1]

Accepted: 12 September 2023

Abstract

Objective We compared two generic, preference-based health-outcome measures: the novel patient-centered Château-Santé Base (CS-Base), entailing a multi-attribute preference response framework, and the widely used EQ-5D-5L, regarding effects of different measurement frameworks and different descriptive systems.

Methods We conducted a cross-sectional study using a random sample of patients (3019 reached, 1988 included) in the USA with various health conditions. The CS-Base (12 attributes, each with four levels), EQ-5D-5L and the 5D-4L (an ad hoc, multi-attribute preference response-based measure that includes five attributes similar to the EQ-5D-5L, but with four levels) were used as health-outcome measures. We compared the proportions of problems reported on health attributes, statistical robustness and face validity of coefficients, attribute importance, differentiation between health states based on health-state values obtained with these measures, and user experience.

Results All the CS-Base and 5D-4L coefficients had logical orders and significant differences from the reference level ($p < 0.001$). Substantial differences were observed in the CS-Base and 5D-4L coefficients between all levels on all attributes, while subtle differences were seen in those of the EQ-5D-5L. Attribute importance of usual (daily) activities were lowest or second lowest in all the three health-outcome measures. Attributes with the highest importance in the CS-Base, 5D-4L, and EQ-5D-5L were respectively mobility, anxiety/depression, and pain/discomfort. Four attributes are similar between the CS-Base and EQ-5D-5L, eight are exclusive to CS-Base. Of the eight, vision and hearing had the highest importance. Health-state values showed a smoother distribution with minimal discontinuity in the CS-Base and EQ-5D-5L than in the 5D-4L. In user experience evaluation, both CS-Base and the 5D-4L showed mean scores above 50 (indicating positive evaluation) in terms of the description of health and ease of understanding.

Conclusions This study demonstrated that CS-Base and 5D-4L, which are grounded in the multi-attribute preference response framework, produced statistically robust coefficients, with better face validity than those for the EQ-5D-5L. CS-Base and the EQ-5D-5L outperformed the 5D-4L in differentiating between health states. Notwithstanding differences in content, measurement frameworks, and estimated coefficients, the computed health-state values were similar between CS-Base and EQ-5D-5L.

Key Points for Decision Makers

Château-Santé Base (CS-Base) coefficients showed better face validity than the EQ-5D-5L.

Notwithstanding differences in content, measurement frameworks, and estimated coefficients, the computed health-state values were similar between CS-Base and EQ-5D-5L.

✉ Paul F. M. Krabbe
p.f.m.krabbe@umcg.nl

[1] Department of Epidemiology, University Medical Center Groningen, P.O. Box 30 001, 9700 RB Groningen, The Netherlands

Published online: 12 October 2023

1 Introduction

The importance of measuring health outcomes in healthcare has been well acknowledged [1, 2]. Apart from standard health indicators (e.g., life-years, body temperature), subjective outcomes (e.g., well-being, quality of life, experienced/perceived health status) are considered crucial for understanding health outcomes [3]. Particular methods need to be applied to obtain metric measures for these subjective outcomes. Within the literature, the term "quality of life" is often used interchangeably with "health-related quality of life" or "health status" [4]. In this paper, we use the term "health status." Health status is the main outcome of this paper. We define it as the overall health condition or quality of life as perceived by individuals (often patients) themselves. However, health outcome is a broader concept encompassing all the objective and subjective health indicators, including health status.

Various health-outcome measures have been developed to measure health status and used different measurement frameworks as elaborated below. A measurement framework, some researchers refer to it as a conceptual model [5], broadly outlines the overall approach or system for (subjective phenomenon) measurement. It may include the theoretical basis for the measurement, the methods used to collect data, who reported the data, and the statistical model used to analyze the data. A measurement model, in contrast, is a specific statistical or mathematical representation used to operationalize the measurement framework in practice.

A first type of measurement framework is used to develop profile measures (questionnaires), using a classical test theory [6]. They cover one or more health domains. Multiple attributes are usually bundled together to measure a specific health domain. They measure the frequency or intensity of separate health domains, expressed as separate scores. Examples of profile measures include the SF-36 [7], NHP [8], and EORTC-QLQ-C30 [9].

A second type of measurement framework used to develop health outcome measures is the index-based framework (non-preference based), which is often used in clinical settings. When using these measures, a single score (index) is assigned to the overall health state measured by the specific instrument [10]. It should be distinguished from "health status," as a health state is a structured description of health status based on a set of attributes with a limited number of levels of severity, which forms a specific health outcome measure. A total score is obtained by rating separate attributes, whose scores are then summed. Examples of index measures include the Barthel Index [11] and the neonatal Apgar score [12].

A third type of measurement framework entails the use of preference methods [13]. Preference-based measures generate a value for the overall health state. An advantage of preference-based measures is the inclusion of the relative importance (weights) assigned to the attribute levels. These weights for attribute levels can be combined to produce a single index that expresses the (social) value of a health state [14]. Such values can be meaningful in many research situations, such as monitoring patients' health, assessing healthcare interventions, conducting a cost-effectiveness analysis, and comparing health status across different populations. A representative of a preference-based health-outcome measure is the widely applied EQ-5D-5L [15].

Preference methods originate from diverse scientific fields, including decision science [16], health economics [17], marketing [18], psychometrics [19], public health [20], and clinimetrics [21]. Well-known methods used in health economics include the standard gamble and time trade-off (TTO). The TTO is widely used for deriving health-state values. However, these conventional preference methods are associated with theoretical and empirical drawbacks relating to time preferences, are cognitively demanding, and are prone to interviewer effects and loss aversion [22].

Apart from the measurement framework used, another key element of a health-outcome measure is its content (health attributes). The choice of health attributes to be incorporated is an important consideration. Health-outcome measures should be constructed to adequately capture the overall prominent health concerns of the target population. It is increasingly recognized that the selection of attributes should be based on patients' perspectives [23]. However, many existing measures are not patient centered in their design, instead reflecting health professionals' views [24]. This could result in the omission of health attributes of high relevance to patients or an emphasis on irrelevant attributes. In this paper, we defined patients as individuals who have experience with one or more diseases or health complaints. This can be distinguished from a sample of the general population, which we define as individuals from a region or country, without selection based on specific characteristics (e.g., sociodemographic backgrounds or health conditions) [25]. The general population encompasses individuals with or without experience with any disease or health complaint.

The EQ-5D-5L is the most widely used, generic preference-based health-outcome measure. Its brevity (inclusion of only five attributes) is one of its advantages. However, questions arise as to whether the five attributes describe and measure health status sufficiently and whether its content reflects what patients deem important, as the attributes were not selected by patients but by researchers

[26]. In addition, although the descriptions of health states in EQ-5D-5L studies are reported by patients, subsequent computations of associated values are derived from separate valuation studies using preferences collected from samples of the general population.

The Château-Santé Base (CS-Base) is a new generic, preference-based health-outcome measure that is fully patient centered in its development and construction. The CS-Base health attributes are fully selected based on patient input [27]. It was designed and developed using a novel measurement framework entailing a new method for eliciting preferences in combination with a specific statistical routine [28]. Within this measurement framework, the description of health states and the preference tasks are both reported by patients. A previous study showed that using this novel measurement framework, the CS-Base produced promising results [29].

The purpose of this study was to compare CS-Base with the established EQ-5D-5L, to explore the effects of their different measurement frameworks and descriptive systems (content). Our main research questions were as follows. Do their weights (coefficients) of attribute levels have similar statistical robustness and face validity, estimated with different measurement frameworks? Do their health-state values have the similar ability to differentiate between health states, derived based on the combined effects of different descriptive systems and different measurement frameworks?

2 Methods

2.1 Sample

We conducted a population-based cross-sectional study. Respondents were US adult patients with one or more diseases or health complaints. They were registered with Dynata, a market research company based in Rotterdam, in the Netherlands. We asked Dynata to collect a sample that is nationally representative in terms of age, sex, and education. Dynata distributed an online survey for our study to patients through their system. The survey carried a link to our study tasks running in a mobile application (described below). Those who completed the survey received a small sum as financial compensation from Dynata. The amounts were decided according to the company's agreements with the groups of patients. Data were collected in January and February 2022. Patients' demographic data were provided by Dynata.

2.2 Health-Outcome Measures

Three health-outcome measures were used in this study: CS-Base, the EQ-5D-5L, and the 5D-4L. We initially intended to compare CS-Base and the EQ-5D-5L. However, these measures are based on different measurement frameworks. The CS-Base is grounded on the multi-attribute preference response (MAPR) framework, in which a novel preference-based method (described below) is applied, and the values are derived from patients [28]. The EQ-5D-5L uses a measurement framework that involves the TTO valuation technique based on responses from the general population. This discrepancy could have led to incomparability of the results of the two health-outcome measures. Therefore, we created another experimental health-outcome measure, the "5D-4L", which, like CS-Base, is based on the MAPR framework; its descriptive system is similar to the EQ-5D-5L but with four levels. The application for modifying the EQ-5D-5L was approved by the EuroQol Group (agreement number: 161826). By comparing the CS-Base with the EQ-5D-5L, we were able to explore the combined effect of using different measurement frameworks and different descriptive systems (content). By adding the 5D-4L, a direct comparison with the CS-Base for the effect of using different descriptive systems but the same measurement framework could be made. Additionally, a direct comparison for the effect of using different measurement frameworks between 5D-4L and EQ-5D-5L could be made, as their descriptive systems are similar. All the three health-outcome measures entailed two tasks: a descriptive task to describe the health state and a preference task. For the descriptive task, the three health-outcome measures used similar assessment methods, all of which involved patients describing their health today based on the descriptive systems of the measures. For the preference task, the same assessment method was used for CS-Base and the 5D-4L and a different one for the EQ-5D-5L (described below).

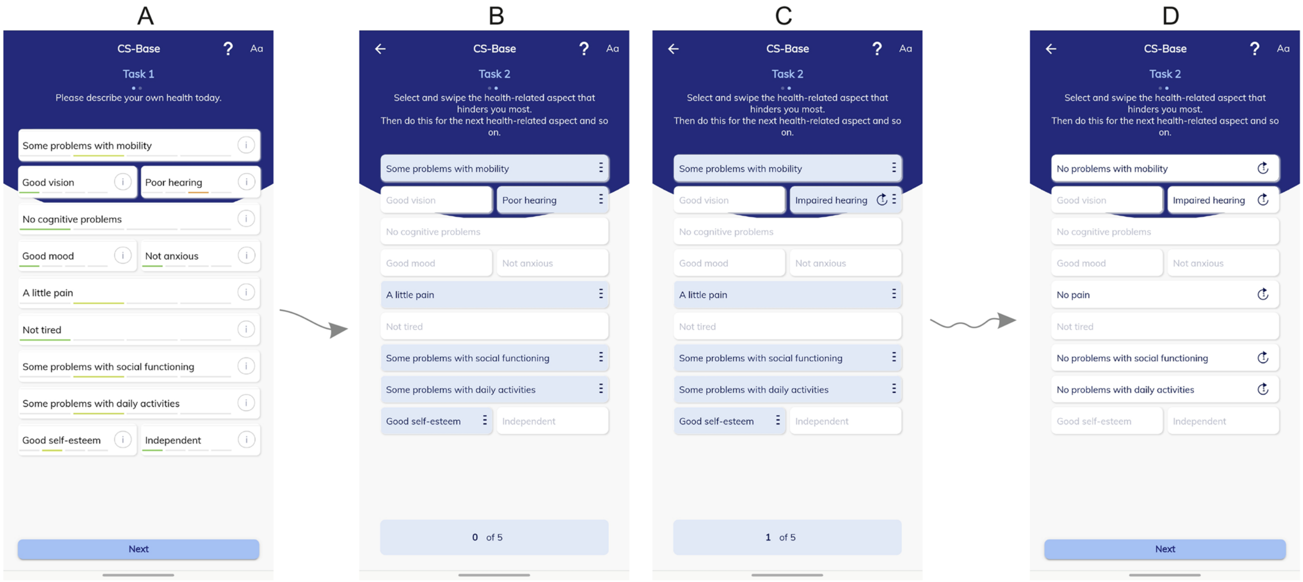
2.2.1 CS-Base

The CS-Base is a generic health-outcome measure comprising 12 health attributes: mobility, vision, hearing, cognition, mood, anxiety, pain, fatigue, social functioning, daily activities, self-esteem, and independence. Each attribute has four levels, with level 1 as optimal, indicating no problems and level 4 as the most suboptimal level, indicating the most severe problems. For Task 1, patients described their health today, by indicating the level of their problems for each health attribute. Thus, an overall description of their health states, expressed as 12 digits, was generated. A state of full health was expressed as "111111111111," and the worst health state was expressed as "444444444444". The total number of possible health states that can be generated by CS-Base is 16,777,216 (4^{12}, 12 attributes each with four levels). If the full health state (no suboptimal levels) is described in Task 1, the patients will not proceed to Task 2, their health assessment using this measure will be terminated and concluded at this point.

For Task 2, the "Drop-Down" (DD) method [29] within the MAPR framework was used for CS-Base and the 5D-4L. Patients were presented with their own health state (assessed in Task 1) and asked to select a suboptimal level (2, 3, or 4) for the attribute that hindered them the most. This was done by clicking or swiping this attribute and dropping down one level lower (e.g., from level 3 to 2, indicating worse to better health status, Fig. 1). Patients then selected the suboptimal level for another attribute that hindered them the second most. We set the maximum number of DD selections at five (patients could make between one and five selections). Each DD produced a health state that could be ranked as better than the initial health state from Task 1 (levels >1 had to be selected for at least two attributes; otherwise, it would not be possible to make a trade-off between attributes and come up with a preference-based choice. If an attribute was assessed at level 3 or higher, patients could drop down more than once for this attribute). The initial health state (Task 1) was ranked as the worst state. Trade-offs were made between the levels of multiple attributes (i.e., patients asked themselves whether a certain level of a certain attribute was worse than any level of another attribute). The DD method thus produced health states that were ranked as ordinal data for the analysis (Appendix 1 in the Electronic Supplementary Material [ESM]).

2.2.2 EQ-5D-5L

The EQ-5D-5L was developed by the EuroQol Research Foundation (www.euroqol.org) as a relatively simple generic health-outcome measure that could be used in clinical studies to provide utilities for health states used in economic evaluations. The descriptive system comprises five attributes: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each attribute has five levels (1 = no problems, 2 = slight problems, 3 = moderate problems, 4 = severe problems, and 5 = extreme problems/unable to). An overall description of a health state using the EQ-5D-5L was expressed as five digits (e.g., "12223"). The total number of possible health states that could be generated by the EQ-5D-5L is 3125 (5^5, five attributes, each with five levels). In this study, we did not include a preference task for the EQ-5D-5L, we used the EQ-5D-5L US tariff to compute health-state utilities of patients in this study [30]. In the US



CS-Base ? Aa
Task 1
Please describe your own health today.
Some problems with mobility
Good vision
Poor hearing
No cognitive problems
Good mood
Not anxious
A little pain
Not tired
Some problems with social functioning
Some problems with daily activities
Good self-esteem
Independent
Next

CS-Base ? Aa
Task 2
Select and swipe the health-related aspect that hinders you most.
Then do this for the next health-related aspect and so on.
Some problems with mobility
Good vision
Poor hearing
No cognitive problems
Good mood
Not anxious
A little pain
Not tired
Some problems with social functioning
Some problems with daily activities
Good self-esteem
Independent
0 of 5

CS-Base ? Aa
Task 2
Select and swipe the health-related aspect that hinders you most.
Then do this for the next health-related aspect and so on.
Some problems with mobility
Good vision
Impaired hearing
No cognitive problems
Good mood
Not anxious
A little pain
Not tired
Some problems with social functioning
Some problems with daily activities
Good self-esteem
Independent
1 of 5

CS-Base ? Aa
Task 2
Select and swipe the health-related aspect that hinders you most.
Then do this for the next health-related aspect and so on.
No problems with mobility
Good vision
Impaired hearing
No cognitive problems
Good mood
Not anxious
No pain
Not tired
No problems with social functioning
No problems with daily activities
Good self-esteem
Independent
Next

Fig. 1 Screenshots from the HealthSnApp depicting the use of the Drop-Down (DD) method. Patients first assessed and described their current health status under Task 1. They were subsequently directed to Task 2 (DD method). In Task 2, their health states, assessed in Task 1 (**A**), were presented (**B**). They then made multiple selections (one to five times) of attributes that hindered them most and dropped down to one level lower by swiping it. For example, the patient dropped down to one level lower for the "hearing" attribute for the first of the selections (**C**).

tariff, the composite TTO (cTTO) technique was used as the preference method.

For the cTTO technique, the respondents were asked to trade off duration of life against health status. This trade-off meant choosing a shorter life spent in full health state or living longer in a less optimal state of health. The number of years sacrificed for the sake of living in a full health state represents the value of the less optimal state. If the respondent indicated that a health state was "worse than being dead", an additional 10 years (lead time) was granted to elicit "worse than being dead" values. The most negative health-state utility elicited by cTTO was "−1.0". The smallest unit of time traded was 6 months. States of full health and being dead were used as anchors in the cTTO tasks.

2.2.3 5D-4L

The 5D-4L is an ad hoc, experimental health-outcome measure. It comprises five attributes similar to the EQ-5D-5L. However, the number of levels for each attribute was reduced to four (1 = no problems, 2 = slight problems, 3 = moderate problems, and 4 = severe problems) to make the measure comparable to CS-Base. Its preference task (DD) is based on the MAPR framework, same as the CS-Base. The total number of possible health states that could be generated by the 5D-4L is 1024 (4^5, five attributes each with four levels). We omitted the fifth level of the EQ-5D-5L for two reasons. First, this level is not frequently selected by respondents [31]. Second, empirical studies showed that the phrasing of the fourth level ("severely") and the fifth level ("extremely") often results in a preference inversion among respondents, with the fourth level considered to be worse than the fifth level [32]. By including the 5D-4L, and thus using the same measurement framework, more detailed and robust comparisons of the estimated coefficients, weights of attributes, and health-state values between the CS-Base and EQ-5D-5L could be achieved.

2.3 Mobile App and Data Collection

The patented mobile application HealthSnApp® (www.chateau-sante.info) was used to manage the three health-outcome measures. The HealthSnApp is a flexible tool with interactive software routines and the potential to perform on-the-fly analytics. It comprises two main routines for descriptive and preference tasks, respectively, and an optional module with various questions for additional evaluation. The responses to the two tasks using CS-Base and the 5D-4L were collected using the two main routines of the HealthSnApp. The EQ-5D-5L responses (descriptive task only) were collected using the app's evaluation module. This study comprised two arms to ensure the head-to-head comparison, each of which entailed the use of the three health-outcome measures but in a different order: CS-Base, the 5D-4L, and the EQ-5D-5L (Arm I) and the 5D-4L, CS-Base, and the EQ-5D-5L (Arm II). The online survey randomly directed patients to one of the two arms. After completing the tasks, user experience was evaluated through questions in the evaluation module of the app.

2.4 User Experience Evaluation

Three questions were used to evaluate patients' user experience of CS-Base and the 5D-4L (Appendix 2 in the ESM): *(1) This tool gives a good description of my health*, *(2) Descriptions of attributes in this tool are easy to understand. (3) Which of the two tools do you prefer?* The first two questions were rating questions (scored 0–100, where 0 = totally disagree and 100 = totally agree). The third question was a binary choice. The EQ-5D-5L was excluded in this evaluation because it did not follow a similar administration routine as the other two measures in the App.

2.5 Analysis

Our analysis includes four aspects of the health-outcome measures: proportions of problems reported on attributes, coefficients of the attribute levels, attribute importance, and health-state values. Coefficients and health state-values show the major results of health outcome measures to answer the research questions. The coefficients reflect the effect of measurement framework, the health-state values reflect the combined effect of measurement framework and descriptive system. The proportions of problems reported on attributes and attributes importance can be regarded as descriptives to compare the three measures.

We first compared coefficients of the three measures. For EQ-5D-5L, the coefficients from the US EQ-5D-5L study were used. For CS-Base and 5D-4L, the ordinal response data (ranked health states) obtained using the preference task (DD) were performed using a rank-ordered logit choice model (cmrologit, Stata 17.0), which yields parameter estimates presented as regression coefficients. These coefficients reflect the weights of attribute levels. The first level of each attribute (level 1: no problems or an optimal condition) was the reference level. Regression coefficients were estimated for the remaining three levels (2, 3, and 4) using dummy variables (12 × 3 for CS-Base, 5 × 3 for the 5D-4L). No constants were included. Because of insufficient responses to level 4 of the attribute hearing, it was not possible to estimate the coefficient for this level using the data from this study alone. Fortunately, we were able to supplement the analysis with data from another study [29] that employed the same outcome measure and a similar design. By combining the datasets of the two studies, we were able to estimate the coefficient for level 4 of hearing. Notably, out of the 36 coefficients reported in this paper, only one pertains to level 4 of hearing and is based on the combined dataset. (for a detailed sample description and coefficients for this combined dataset, see Appendix 3 in the ESM).

Attribute importance was calculated using a variation of the coefficient range method [33]. The range between the coefficients for individual levels was obtained and converted to a proportion.

$$I_{attribute(i)} = \frac{maxC_i - minC_i}{\sum_j (maxC_i - minC_i)/J}, \quad (1)$$

where C_i represents the coefficients of the individual levels of attribute *i,* and *j* represents the number of attributes in the health-outcome measure. The range of coefficients for each attribute *i*, was divided by the weighted sum of the ranges of all attributes (weight based on number of attributes, *J*). This calculation yielded the attribute importance ($I_{attribute}$).

The estimated coefficients were used to compute the values for distinct CS-Base and 5D-4L health states. For the EQ-5D-5L, the value set from the US study was used, in which the lowest value was −0.573 (state 55555) and 0.0 = "dead". To allow for consistent comparisons of the three measures, their original values were rescaled into 0.0 to 1.0, where 0.0 for all the three measures stands for the lowest value in their original scales (worst health state) and 1.0 stands for the highest value (full health). Noticeably, after the rescaling for the EQ-5D-5L, "0.0" did not refer to the state "dead" but to state "55555". Only values were rescaled in this study, coefficients were not. In the context of the US EQ-5D-5L value set, since "values" are anchored on the location of dead (0.0), they are often called "utilities". When comparing the measures in this study, we use the term "value". Spearman's correlation was used to test the relations of the health-state values for each combination of two out of the three health-outcome measures. We used Stata 17.0, and CorelDraw 22.0 software to compute and visualize our results.

Frequencies and proportions were used to describe patients' demographic characteristics, health conditions, problems reported on health attributes of the three measures, and the responses to the binary question in the evaluation of user experiences. Means were used to describe the scores of responses to rating questions. To test the difference of user experience between CS-Base the 5D-4L, a paired-sample t-test was used for rating questions, the McNemar's test was used to test proportions of patients expressing their preference for one out of the two measures (binary question). A two-sample t-test for the rating questions and the McNemar's test for the binary question were done separately for CS-Base and 5D-4L, to explore the order effects (different order between two study arms) on the user experience.

3 Results

3.1 Completion

The survey was sent to 3019 patients of whom 2170 completed the CS-Base and 5D-4L response tasks. For the estimation of CS-Base coefficients one patient was excluded because of a technical issue of the software, an additional 270 patients reported full health state (Task 1), thus no DD response was generated. Therefore, the coefficients were estimated based on responses of 1899 patients (with 11,987 ranked health states generated). For the estimation of 5D-4L coefficients, one patient was excluded because of a technical issue of the software, another 479 patients reported full health state (Task 1), thus no DD response generated. Therefore, the coefficients were estimated based on responses of 1690 patients (with 7759 ranked health states generated). The EQ-5D-5L coefficients were estimated based on responses of 1062 patients, according to the US EQ-5D-5L study. Regarding the comparison of the three measures, among the 2170 patients, 24 patients did not complete the sociodemographic questions, 157 patients did not respond to the EQ-5D-5L, and one patient had a missing record on one attribute in the 5D-4L. Therefore, 1988 patients were included in the comparison.

3.2 General Information of Patients

For the 1988 patients who completed all three health-outcome measures, the mean age was 46 years (range 18–94 years, Table 1). More than half of the patients were women (*n* = 1142; 57%), the majority were white American/Caucasian (1594, 80%). More than half of the patients (1154, 58%) were high school graduates and 606 (30%) patients received a higher education level than this. The most frequently reported disease or health complaints were pain (*n* = 1066, 54%), fatigue/sleep problems (*n* = 816, 41%), mental health problems (*n* = 653, 33%), respiratory disease (*n* = 437, 22%), and diabetes mellitus (*n* = 340, 17%).

3.3 Health Attributes

Four similar attributes were included in all three health-outcome measures: mobility, pain/discomfort (pain in CS-Base), anxiety/depression (anxiety in CS-Base), and usual activities (daily activity in the CS-Base). Based on the 1988 patients, the percentages of patients reporting problems on

Table 1 Number of patients for each demographic subgroup and health condition (*N* = 1988)

Characteristics	*N* (%)
Sex, *N* (%)	1988 (100)
Female	1142 (57)
Male	846 (43)
Age (years), mean (SD)	46 (17)
Age (years), *N* (%)	1988 (100)
18–27	293 (15)
28–37	458 (23)
38–47	369 (19)
48–57	293 (15)
58–67	294 (15)
68–77	235 (12)
≥78	46 (2)
Ethnicity, *N* (%)	1983 (100)
Asian/Asian-American	44 (2)
Black/African American	174 (9)
Hispanic or Latino American	112 (6)
Native American/Inuit/Alaskan	30 (2)
Native Hawaiian/Pacific Islander	17 (1)
White American/Caucasian	1594 (80)
Other	12 (1)
Education, *N* (%)	1988 (100)
More than high school	606 (30)
High school graduate	1154 (58)
Less than high school	228 (12)
Main health conditions, *N* (%)	1979 (100)
Pain	1066 (54)
Fatigue/sleep problems	816 (41)
Mental health problems	653 (33)
Respiratory diseases	437 (22)
Diabetes mellitus	340 (17)
Hearing or vision loss	337 (17)
Eczema	218 (11)
Gastrointestinal disease	208 (10)
Heart disease	170 (9)
Rheumatism	88 (4)
Cancer	84 (4)
Stroke	65 (3)
Epilepsy	58 (3)
Other diseases	171 (9)

SD standard deviation

these four attributes in CS-Base and in the 5D-4L were similar at 31%, 66%, 60%, and 42% and 31%, 65%, 60%, and 36% respectively (Appendix 4 in the ESM). However, they were lower than the percentages of patients reporting problems on these attributes using the EQ-5D-5L (45%, 77%, 70%, and 51%, respectively). Only a few patients (≤8%) reported problems at level 5 for the EQ-5D-5L attributes. In all three health-outcome measures, problems with pain were most frequently reported (≥65%). Self-care was exclusively an EQ-5D-5L attribute (not part of CS-Base), and problems were the least frequently reported on it in both 5D-4L (19%) and the EQ-5D-5L (28%). However, among patients who reported full health state using CS-Base but reported impaired health states using EQ-5D-5L, nearly half (44%) reported problems on self-care (Appendix 5 Table A5.2 in the ESM). Of the eight exclusive CS-Base attributes, problems were reported on five attributes by almost half of patients: fatigue (63%), self-esteem (55%), social function (43%), mood (47%), and hearing (47%). Fewer patients reported problems on the other three attributes: independence (29%), vision (24%), and cognition (20%).

3.4 Coefficients and Attributes Importance

All the CS-Base and 5D-4L coefficients were negative and followed a logical order (e.g., increasing negative: no problems > slight problems > moderate problems > severe problems, Table 2). Negative coefficients implied that a particular level was worse than the reference, which in our study was the first level of each health attribute. Moreover, the less preferable a level was considered, the higher its coefficient was in a negative direction. Differences between all levels on all attributes were observed for all the CS-Base (Fig. 2A) and 5D-4L (Fig. 2B) coefficients. For the EQ-5D-5L (Fig. 2C) coefficients, such differences were only noticeable between levels 3 and 4, while very subtle between levels 2 and 3, and levels 4 and 5. All the CS-Base coefficients had smaller confidence intervals than the 5D-4L coefficients. Only level 4 of cognition showed a large confidence interval.

The attribute importance of usual activities (daily activities in CS-Base) was the lowest (or second lowest) for all three outcome measures. The most important attributes differed between the three measures, they were mobility, anxiety/depression, and pain/discomfort in the CS-Base, 5D-4L, and EQ-5D-5L respectively. In both 5D-4L and the EQ-5D-5L, the two double-barreled attributes, namely pain/discomfort and anxiety/depression had the highest and second highest importance respectively, whereas self-care and usual activities had the lowest and second lowest importance. Of the eight exclusive CS-Base attributes, vision and hearing had the highest importance. The importance of other exclusive CS-Base attributes were very similar.

3.5 Health States and Values

The number of actual health states reported (Task 1) by patients in this study using CS-Base, the 5D-4L, and the EQ-5D-5L were 1472, 329, and 483, respectively. For the total sample, the mean value of health states reported using

Table 2 Coefficients of attribute levels and attribute importance in the CS-Base, 5D-4L, and EQ-5D-5L

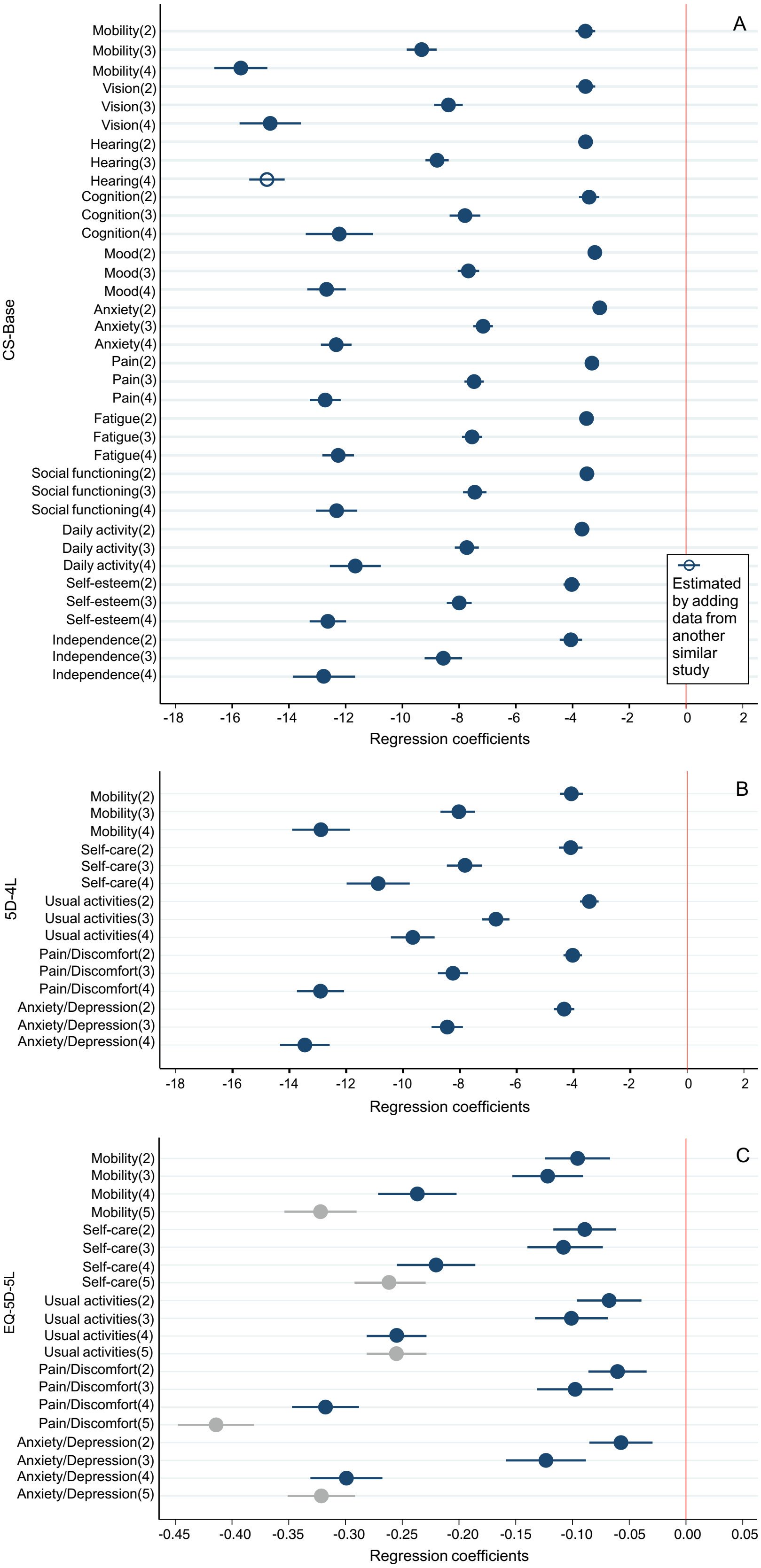
CS-Base (*N* = 1899)				5D-4L (*N* = 1690)				EQ-5D-5L (*N* = 1062)		
Attribute levels	Attribute importance	Coefficient	SE	Attribute levels	Attribute importance	Coefficient	SE	Attribute importance	Coefficient	SE
MO2	1.275	−3.543	0.177	MO2	1.107	−4.075	0.206	0.939	−0.096	0.015
MO3		−9.315	0.269	MO3		−8.078	0.309		−0.122	0.016
MO4		−15.689	0.477	MO4		−12.890	0.517		−0.237	0.018
				MO5		–	–		−0.322	0.016
PA2	0.987	−3.311	0.113	PD2	1.114	−4.027	0.166	1.471	−0.060	0.013
PA3		−7.467	0.173	PD3		−8.243	0.271		−0.098	0.017
PA4		−12.714	0.278	PD4		−12.905	0.423		−0.318	0.015
				PD5		–	–		−0.414	0.017
AN2	0.975	−3.037	0.117	AD2	1.147	−4.330	0.182	1.097	−0.057	0.014
AN3		−7.148	0.176	AD3		−8.445	0.281		−0.123	0.018
AN4		−12.327	0.276	AD4		−13.455	0.445		−0.299	0.016
				AD5		–	–		−0.321	0.015
DA2	0.839	−3.663	0.135	UA2	0.781	−3.442	0.167	0.777	−0.068	0.015
DA3		−7.724	0.216	UA3		−6.741	0.248		−0.101	0.016
DA4		−11.655	0.457	UA4		−9.657	0.392		−0.255	0.013
				UA5		–	–		−0.255	0.013
VI2	1.166	−3.537	0.174	SC2	0.851	−4.098	0.212	0.715	−0.089	0.014
VI3		−8.370	0.257	SC3		−7.840	0.314		−0.107	0.017
VI4		−14.653	0.551	SC4		−10.875	0.567		−0.220	0.018
				SC5		–	–		−0.261	0.016
HE2	1.178	−3.536	0.130							
HE3		−8.771	0.209							
HE4[a]		−14.761	0.319							
MD2	0.993	−3.209	0.128							
MD3		−7.666	0.193							
MD4		−12.667	0.347							
CO2	0.925	−3.410	0.182							
CO3		−7.787	0.276							
CO4		−12.219	0.603							
FA2	0.920	−3.502	0.114							
FA3		−7.537	0.181							
FA4		−12.258	0.284							
SF2	0.926	−3.489	0.132							
SF3		−7.443	0.210							
SF4		−12.313	0.371							
SE2	0.903	−4.023	0.149							
SE3		−7.991	0.224							
SE4		−12.622	0.328							
IN2	0.914	−4.054	0.200							
IN3		−8.550	0.337							
IN4		−12.759	0.560							

All *p*-values <0.001 (differences of each coefficient compared to the reference level [level 1])

AD anxiety/depression, *AN* anxiety, *CG* cognition, *DA* daily activity, *FA* fatigue, *HE* hearing, *IN* independence, *MD* mood, *MO* mobility, *PA* pain, *PD* pain/discomfort, *SC* self-care, *SE* self-esteem, *SF* social function, *UA* usual activities, *VI* vision

[a] The coefficient is based on the combined dataset. No coefficient could be estimated for level 4 of hearing, as only one patient reported at this level

Fig. 2 Distribution of coefficients (and their 95% confidence intervals) for the three health-outcome measures (A: CS-Base, B: 5D-4L, C: EQ-5D-5L). To facilitate a comparison with the 5D-4L, level 5 of the EQ-5D-5L coefficients is shown as grey dots in (**C**), thereby highlighting levels 2, 3, and 4. In (**A**), the coefficient of level 4 of hearing was estimated by adding data from another similar study. No coefficient could be estimated for this level based on the single dataset of this study, as only one patient reported at this level



CS-Base (0.81) was higher than that for the 5D-4L (0.78) and the EQ-5D-5L (0.79). Considering subgroups of patients with specific health complaints which are only represented by the CS-Base attributes (e.g., hearing, vision, cognition problems), the mean values decreased for all three measures in comparison to the total sample. The differences were more pronounced for the CS-Base than for the other two measures (Appendix 6 in the ESM). Numbers of patients reporting a full health state (value: 1.0) in the CS-Base, 5D-4L, and the EQ-5D-5L were 235 (12%), 426 (21%), and 199 (10%), respectively. Some patients reported full health state in the CS-Base, while they reported worse states in the 5D-4L or EQ-5D-5L. Values of some worse states were even lower than 0.5. By contrast, those reporting full health state using 5D-4L or EQ-5D-5L never reported such worse states with values lower than 0.5 using CS-Base (Appendix 5 Figure A5.1-3.3 in the ESM). The worst health state reported by six patients using the 5D-4L was “44444” (value: 0.0), while three patients reported their worst health states as “55555” (value: 0.0) using the EQ-5D-5L. The worst health state reported using CS-Base was “342444443344” (value: 0.17). The number of patients reporting health states with high values (≥0.8) using CS-Base (n = 1147, 58%), the EQ-5D-5L (n = 1129, 57%), and the 5D-4L (n = 1023, 51%) were almost equal. Fewer patients reported worse health states with values <0.4 using CS-Base (n = 29, 1%) compared with using the 5D-4L (n = 125, 6%) and EQ-5D-5L (n = 87, 4%). The values for health states reported by patients in this study showed a smoother distribution with minimal discontinuity in CS-Base and the EQ-5D-5L compared with the distribution in the 5D-4L (Fig. 3). The values for health states reported by patients in this study were compared for each pair of the three measures (Fig. 4). The correlation between the values was the highest for the CS-Base versus the 5D-4L (r = 0.741), lower for the 5D-4L versus EQ-5D-5L (r = 0.663) and the CS-Base versus EQ-5D-5L (r = 0.633).

3.6 User Experience Evaluation

No statistically significant differences were found between CS-Base and the 5D-4L for the two rating questions (good description, easy understanding, Appendix 7 Table A7.1 in the ESM). For the first question, the mean scores of the CS-Base and 5D-4L were 62.6 and 62.4 (p = 0.681). For the second question, the mean scores were 56.6 and 54.9 (p = 0.080). Responses to the binary question on the preferred tool indicated that the 5D-4L was preferred (57%, p < 0.001). Regarding a possible order effect (Appendix 7 Table A7.2 in the ESM), we found that for the rating of “good description”, both CS-Base (62.9 in Arm I: CS-Base–5D-4L, 62.4 in Arm II: 5D-4L–CS-Base, p = 0.672) and 5D-4L (62.3 in Arm I, 62.5 in Arm II, p = 0.873) had similar mean scores between two arms. For “easy-understanding”, CS-Base was rated higher in Arm I (65.4) than in Arm II (47.0, p < 0.001), reversely, 5D-4L was rated lower in Arm I (47.2) than in Arm II (63.0, p < 0.001). For the binary question regarding preference for one of the two instruments, CS-Base (51%) and 5D-4L (49%) were almost equally preferred in Arm I (p = 0.554); however, 5D-4L (65%) was more preferred than CS-Base (35%) in Arm II (p < 0.001).

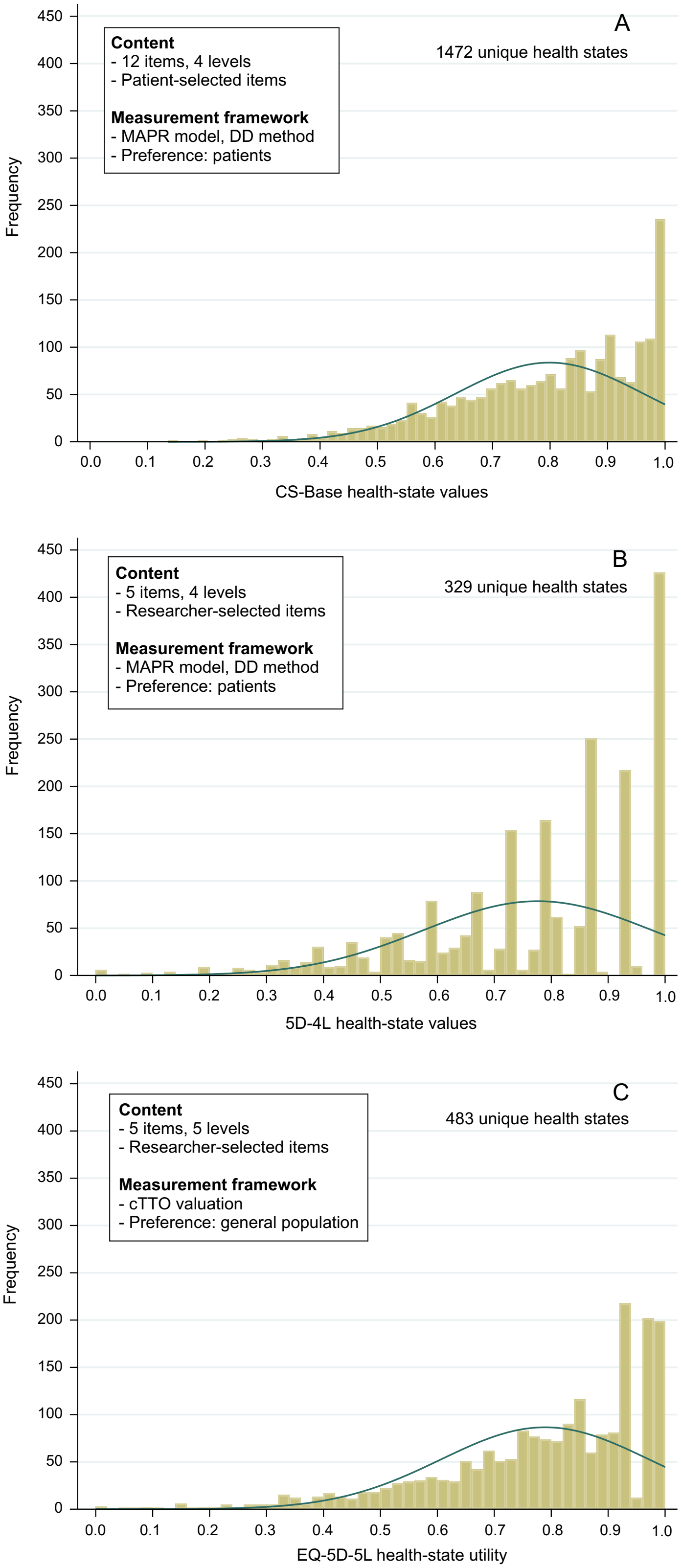
4 Discussion

Our study entailed a head-to-head comparison of the CS-Base and EQ-5D-5L, to explore the effects of their different content (descriptive systems) and measurement frameworks. This was achieved through comparing the statistical robustness of and face validity of their estimated coefficients, and their ability to differentiate between health states.

Health problems were reported by nearly half of the patients on five of the eight exclusive CS-Base attributes that are not part of the EQ-5D-5L. Noticeably, as we stated in the results section, the other four CS-Base attributes are similar to those of the EQ-5D-5L, but their descriptions are not exactly the same. All the eight attributes resulted in similar or higher importance compared to the four similar attributes of CS-Base and EQ-5D-5L. These observations indicate that important health attributes may not be appropriately captured by the EQ-5D-5L. Other studies have consistently revealed similar findings. For instance, a critical overview of reviews revealed no evidence to support EQ-5D validity and responsiveness in mental health and in diseases of the nervous system [34]. EQ-5D bolt-on (health domains added to an established health-outcome measure) studies indicated that hearing was highly important among various candidate bolt-on attributes (e.g., sleep, relationships, and tiredness) [35, 36]. In the well-known Health Utilities Index Mark 3 System, hearing, vision, and cognition were included as part of the eight attributes in total [37]. Therefore, extending the EQ-5D’s descriptive system using “bolt-ons” could lead to broader coverage of health status. Our findings revealed that vision and hearing could be prominent candidates for an expanded EQ-5D descriptive system. Apart from these two sense-related attributes, the remaining six exclusive CS-Base attributes were almost equally important at a moderate level. This finding is reasonable, as these attributes were deliberately selected by patients according to their importance in a previous study [27].

Considering the distribution of health-state values, CS-Base and the EQ-5D-5L appeared to be more sensitive than the 5D-4L in differentiating health states. Many patients

Fig. 3 Distribution of values for health states reported by patients in this study derived by the three health-outcome measures



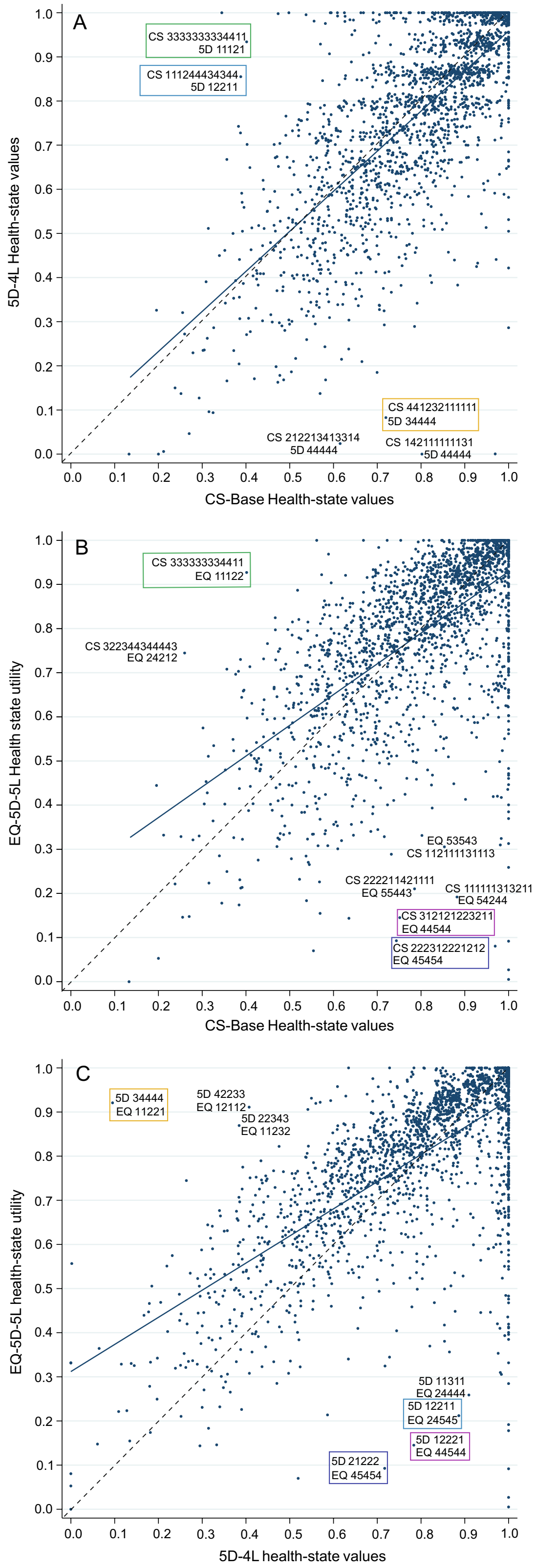


Fig. 4 Comparison of values for health states reported by patients in this study for each pair of the three measures. **A** CS-Base vs 5D-4L; **B** CS-Base vs EQ-5D-5L, **C** 5D-4L vs EQ-5D-5L. The *blue line* in each of the three sub-figures represents the best-fit line of the estimated correlation. The *dashed grey line* stands for the perfect (expected) best-fit line of a correlation coefficient of $r = 1.0$. Outliers are placed in different *colored rectangles*, with rectangles of the *same color* indicating that the assessments belong to the same patient

reported impaired health states using CS-Base and the EQ-5D-5L, while reporting a full health state with the 5D-4L. This finding may indicate that these patients did not actually experience full health, but that it is likely that the 5D-4L had a ceiling effect. The outperformance of CS-Base over 5D-4L can be attributed to the larger set of attributes. The outperformance of EQ-5D-5L over 5D-4L is more difficult to explain, as the extra fifth level in the EQ-5D-5L was the only factor distinguishing it from the 5D-4L. If patients indeed have complaints on a health attribute, it is unlikely that they would report problems using the five-level descriptive system but not the four-level descriptive system. We posit that the different formats of the two instruments presented to patients may be an explanation for this finding. In the 5D-4L, all attributes were presented to patients simultaneously on one screen, whereas in the format of the EQ-5D-5L in this study, attributes were presented one by one, each on a separate screen. These two formats resemble the two administration modes of EQ-5D-5L: the paper version with all attributes presented on one page, and the mobile version where each of the five attributes appears on a separate screen. A study comparing these two administration modes showed that more respondents reported problems on all the five attributes using the mobile version (attribute by attribute) than using the paper version [38]. This aligns with our finding that more patients reported problems on all the attributes using EQ-5D-5L compared with using 5D-4L.

The number of patients reporting a full health state using CS-Base was slightly higher than the number using the EQ-5D-5L. The exclusive EQ-5D-5L attribute, self-care, might account for this difference. Among the patients who reported a full health state using CS-Base but reported impaired health states using EQ-5D-5L, more patients reported problems on self-care compared with the total sample. However, fewer patients reported worse health states using the CS-Base than EQ-5D-5L, possibly because patients using CS-Base had more ways of reporting their health status given the broader range of attributes included. In general, it is difficult to obtain a clear understanding of the (counterintuitive) differences between the two outcome measures.

Apart from the content of the health-outcome measures used in this study, the measurement frameworks on which they are founded could have had an important effect on the results they produced. All the estimated coefficients of CS-Base and the 5D-4L showed robust and statistically significant results. Compared with the 5D-4L, CS-Base produced more precise coefficients at smaller confidence intervals. This result could be related to more DD responses collected in the CS-Base than in the 5D-4L. Because the 5D-4L only has five attributes, whereas CS-Base has 12 attributes, patients could select fewer health attributes, resulting in fewer DD responses. Similarly, the larger confidence interval at level 4 of cognition in CS-Base seems to have resulted from fewer collected responses at that level.

A comparison of the coefficients between the three health-outcome measures indicated that CS-Base and 5D-4L coefficients showed better face validity than EQ-5D-5L coefficients. Differences in coefficients in CS-Base and the 5D-4L were observed between all levels on all attributes. However, subtle differences (or even overlap) between levels in the EQ-5D-5L were observed. These findings could imply that the measurement framework used made a difference. The coefficients of the CS-Base levels were comparable between all the 12 attributes. This may indicate that using the numerical levels of the attributes instead of the estimated coefficients could produce comparable values. Nonetheless, this study had a minor impediment concerning the CS-Base coefficients. One of the 36 coefficients (level 4 of hearing) was estimated based on a combined dataset, by combining the data from this study and another study that employed a similar design, owing to insufficient responses to level 4 of hearing in the data from this study alone.

In the US EQ-5D-5L study, preference-based responses were collected using the cTTO valuation technique. As previously mentioned, the TTO technique is associated with both theoretical and empirical drawbacks that may produce less accurate results. In 2014, the EuroQol Group launched a standard protocol known as EuroQol Valuation Technology (EQ-VT) for valuation studies of the EQ-5D-5L [39]. This protocol entails the use of the cTTO valuation technique, supplemented by the discrete choice experiment. As part of this protocol, a hybrid statistical model was constructed to collectively process the responses from the cTTO and the discrete choice experiment. This protocol has been applied in studies conducted in some countries like the UK and Germany [40, 41]. Despite the potential advantages of the hybrid model, researchers in some countries (e.g., the USA and the Netherlands) have so far chosen to use only cTTO-based data to develop their EQ-5D-5L value set [30, 42]. Whereas the EuroQol Group advocates the use of value sets based on the EQ-VT (hybrid) protocol, such sets are currently not recommended by the National Institute for Health and Care Excellence [43]. This is because concerns were raised about the quality and reliability of the data collected in the UK valuation study for the EQ-5D-5L, and the methods used to model these data [44].

The measurement framework of CS-Base and the 5D-4L is grounded in the MAPR framework, and preferences are elicited using the DD method. The key advantage of the DD method is that patients are not asked to assess hypothetical health states, which can be difficult for them to imagine. They are only asked to assess their own health state and to indicate which attributes hindered them the most. The health states presented to patients through the drop-downs are states that they could probably imagine or may even have experienced. Thus, the DD method could be simpler to perform than the (c)TTO and even simpler than the discrete choice experiment method. However, compared with (c)TTO, it is less straightforward in generating utilities (where "dead" = 0.0), owing to the absence of "dead" or time duration in the tasks. For the MAPR framework, a separate study is needed to generate data that can be used to rescale the DD-based values (where the location of "dead" is unknown) to utilities [45]. A major benefit of the MAPR framework covered in the HealthSnApp is the integration of descriptions of patients' health status and the preference elicitation. In addition, the application of the HealthSnApp generates a set of exclusive values for each specific study.

User experience evaluation indicated that both CS-Base and the 5D-4L received a positive perception in terms of the description of health and the ease of understanding among the patients. The 5D-4L was somewhat preferred over CS-Base, possibly because of the brevity of the descriptive system (only five attributes). An ideal generic health-outcome instrument should be concise and easy to use, while providing an adequate description of the patient's health status. Therefore, in developing generic health-outcome measures, a balance between the brevity and enough health-related attributes to cover the generic health concept should be considered. In terms of order effect on user experience, a peculiar finding is that the first measure (CS-Base in Arm I and 5D-4L in Arm II) was perceived as easier to understand (based on the rating question "descriptions of attributes in this tool are easy to understand"). One explanation for this observation is that patients may have been more focused on the attribute descriptions when encountering the first measure. Confronted with the second measure they were likely to expect something similar, but different attributes were presented.

To summarize the effects of measurement frameworks and content (descriptive systems) based on each paired comparison of the three measures, for the CS-Base versus 5D-4L (using the same measurement framework), both produced statistically robust coefficients and showed good face validity. However, probably because of the more attributes included in the descriptive system, the CS-Base demonstrated better differentiation between health states. When comparing the 5D-4L to EQ-5D-5L (based on different measurement frameworks), the coefficients of the 5D-4L showed better face validity than those of the EQ-5D-5L. The EQ-5D-5L outperformed 5D-4L in differentiating health states with a similar descriptive system (similar attributes but different levels), but with different administration modes (one-screen and separate-screens display for attributes). As for the comparison of CS-Base versus EQ-5D-5L, notwithstanding differences in content, applied measurement frameworks, and estimated coefficients, the computed health-state values were rather similar between the CS-Base and EQ-5D-5L. This result may be attributed to our relatively large sample comprising patients with various health conditions. However, for patients with specific health complaints, the situation might be different. For example, lower health status was observed in both measures among patients with limited hearing, vision, and cognition. This decrease appeared to be larger for the CS-Base, which may be explained by the fact that the EQ-5D-5L has no exclusive attributes for these health domains.

A limitation of this study may be that the descriptive classification system of 5D-4L is not exactly the same as EQ-5D-5L. Although the five attributes and the description of 1-4 levels are the same, level 5 of the EQ-5D-5L was excluded from the 5D-4L. An additional explanation of the attributes was given in the description task (Task 1) of 5D-4L, similar to the CS-Base. Such an explanation was not part of the EQ-5D-5L. In this study, patients were presented with 5D-4L before EQ-5D-5L. They might be influenced by the explanation of the attributes when responding to the EQ-5D-5L. These differences between the 5D-4L and EQ-5D-5L imply that some elements of EQ-5D-5L could not be thoroughly compared in this study. Another concern that might remain in this study is that our patients were recruited by a market research company. Some of these patients may be less motivated in responding to the study tasks, which could affect the robustness of our results.

5 Conclusions

This study demonstrated that CS-Base and 5D-4L, which are grounded on the MAPR framework, produced statistically robust coefficients, with better face validity than those for the EQ-5D-5L. Overall, CS-Base and EQ-5D-5L performed similarly, outperforming the 5D-4L in differentiating between health states. Notwithstanding differences in content, measurement frameworks, and estimated coefficients, overall the CS-Base and EQ-5D-5L values were similar. For patients with specific health complaints, differences in values were found between the measures.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40258-023-00837-1.

Acknowledgments An earlier version of this paper was presented at the EuroQol plenary meeting 2022 (Chicago), 20–23 September, 2022. We thank the audience for the critical reflection and helpful contribution to the current version of the article.

Declarations

Funding Financial support for this study was provided by the PhD grant sponsored by the Chinese Scholarship Council and the University Medical Center Groningen. The following author is employed by the sponsors: Xin Zhang. The funding agreement ensured the independence of the authors in designing the study, interpreting the data, and writing and publishing the manuscript.

Conflicts of Interest Xin Zhang, Karin M. Vermeulen, and Paul F.M. Krabbe have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. As an extension of the CS-Base ePROM, MAPR framework, and the DD method presented in this paper, additional PROMs and software are developed by Paul F.M. Krabbe as part of academic/commercial activities. Commercial activities are part of the spin-off company Château Santé that is affiliated with the University of Groningen.

Ethics Approval This study was exempt from review by the Medical Ethics Review Committee at the University Medical Center of Groningen. It issues waivers stating that the pertinent Dutch Legislation (the Medical Research Involving Human Subjects Act) does not apply to survey research where participants fill out a single questionnaire, which does not pose additional risks. A third-party vendor, Dynata, collected data of this study from Dynata's panel of patients in the USA. Dynata complies with the ethical guidelines and standards set forth in the Market Research Society Code of Conduct.

Consent to Participate All participants provided informed consent through completion and return of the survey of the study. The purpose of the study was explained prior to data collection.

Consent for Publication All participants consented to have their data included in publications when registering with the research market company (Dynata).

Availability of Data and Material The datasets generated during the current study are not publicly available because of the storage on the dedicated Château-Santé server but are available from the corresponding author on reasonable request.

Code Availability The code is available upon reasonable request from the corresponding author.

Authors' Contributions XZ contributed to the data analysis and interpretation, figures preparation, and writing and reviewing the article. KMV helped with the data interpretation, and contributed to the manuscript writing and language phrasing and improvement. PFMK contributed to the study concept and design, as well as to the interpretation of the data, writing, and the critical reviewing of the article in terms of important intellectual content. He also approved the version to be published.

References

1. Testa MA, Simonson DC. Assessment of quality-of-life outcomes. N Engl J Med. 1996;334(13):835–40. https://doi.org/10.1056/NEJM199603283341306.
2. Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science. 1998;282(5387):245–6. https://doi.org/10.1126/science.282.5387.245.
3. Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med. 2016;374(6):504–6. https://doi.org/10.1056/NEJMp1511701.
4. Karimi M, Brazier J. Health, health-related quality of life, and quality of life: what is the difference? Pharmacoeconomics. 2016;34(7):645–9. https://doi.org/10.1007/s40273-016-0389-9.
5. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics. 2000;17:13–35. https://doi.org/10.2165/00019053-200017010-00002.
6. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med. 1993;118(8):622–9. https://doi.org/10.7326/0003-4819-118-8-199304150-00009.
7. Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE Jr. The SF-36 health survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability. Med Care. 1999;37(5 Suppl.):MS10–22. https://doi.org/10.1097/00005650-199905001-00002.
8. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract. 1985;35(273):185–8.
9. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365–376a. https://doi.org/10.1093/jnci/85.5.365.
10. Balinsky W, Berger R. A review of the research on general health status indexes. Med Care. 1975;13(4):283–93. https://doi.org/10.1097/00005650-197504000-00001.
11. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA. 1964;185(12):914–9. https://doi.org/10.1001/jama.1963.03060120024016.
12. Apgar VA. Proposal for a new method of evaluation of the newborn infant. Curr Res Anesth Analg. 1953;32(4):260–7. https://doi.org/10.1213/00000539-195301000-00041.
13. Bergner M. Measurement of health status. Med Care. 1985;23(5):696–704.
14. Thurstone LL. The method of paired comparisons for social values. J Abnorm Psychol. 1927;21:384–400.
15. Brooks R. The EuroQol Group after 25 years. Heidelberg: Springer; 2013.
16. Torrance GW. A generalized cost-effectiveness model for the evaluation of health programs [doctoral dissertation]. Buffalo (NY): State University of New York at Buffalo; 1971. Available from: http://hdl.handle.net/11375/5559. [Accessed 20 Sep 2023].
17. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ. 2006;15:393–402. https://doi.org/10.1002/hec.1069.
18. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics. 2008;26:661–77. https://doi.org/10.2165/00019053-200826080-00004.
19. Groothuis-Oudshoorn CGM, Heuvel ER, Krabbe PFM. A preference-based item response theory model to measure health: concept and mathematics of the multi-attribute preference response model. BMC Med Res Methodol. 2018;18(1):62. https://doi.org/10.1186/s12874-018-0516-8.
20. Murray CJ, Lopez AD. Quantifying disability: data, methods and results. Bull World Health Organ. 1994;72(3):481–94.
21. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J Clin Epidemiol. 1992;45(11):1201–18. https://doi.org/10.1016/0895-4356(92)90161-f.
22. Lugnér AK, Krabbe PFM. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. Expert Rev Pharmacoecon Outcomes Res. 2020;20(4):331–42. https://doi.org/10.1080/14737167.2020.1779062.
23. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity: establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 1-eliciting concepts for a new PRO instrument. Value Health. 2011;14(8):967–77. https://doi.org/10.1016/j.jval.2011.06.014.
24. Wiering B, de Boer D, Delnoij D. Patient involvement in the development of patient reported outcome measures: a scoping review. Health Expect. 2017;20(1):11–23. https://doi.org/10.1111/hex.1244.2.
25. Wikipedia. General population. 2020. Available from: https://en.wikipedia.org/wiki/General_population. [Accessed 2 Jun 2023].
26. EuroQol Group. EuroQol-a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199–208. https://doi.org/10.1016/0168-8510(90)90421-9.
27. Krabbe PFM, van Asselt ADI, Selivanova A, Jabrayilov R, Vermeulen KM. Patient-centered item selection for a new preference-based generic health status instrument: CS-Base. Value Health. 2019;22(4):467–73. https://doi.org/10.1016/j.jval.2018.12.006.
28. Krabbe PF. A generalized measurement model to quantify health: the multi-attribute preference response model. PLoS ONE. 2013;8(11): e79494. https://doi.org/10.1371/journal.pone.0079494.
29. Zhang X, Krabbe PFM. From simple to even simpler, but not too simple: a head-to-head comparison of the Better-Worse and Drop-Down methods for measuring patient health status. Revision submitted.
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 Health. 2019;22(8):931–41. https://doi.org/10.1016/j.jval.2019.02.009.
31. 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. 2018;36(6):675–97. https://doi.org/10.1007/s40273-018-0623-8.
32. Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? Qual Life Res. 2015;24(7):1759–65. https://doi.org/10.1007/s11136-014-0897-z.

33. Selivanova A, Buskens E, Krabbe PFM. Head-to-head comparison of EQ-5D-3L and EQ-5D-5L health values. Pharmacoeconomics. 2018;36(6):715–25. https://doi.org/10.1007/s40273-018-0647-0.
34. Finch AP, Brazier J, Mukuria CW. Validity and responsiveness of preference-based measures: a critical overview of reviews. Value Health. 2015;18:A744. https://doi.org/10.1016/j.jval.2015.09.2868.
35. Finch AP, Brazier J, Mukuria C. Selecting bolt-on dimensions for the EQ-5D: testing the impact of hearing, sleep, cognition, energy, and relationships on preferences using pairwise choices. Med Decis Making. 2021;41(1):89–99. https://doi.org/10.1177/0272989X20969686.
36. Yang Y, Rowen D, Brazier J, Tsuchiya A, Young T, Longworth L. An exploratory study to test the impact on three “bolt-on” items to the EQ-5D. Value Health. 2015;18(1):52–60. https://doi.org/10.1016/j.jval.2014.09.004.
37. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. Med Care. 2002;40(2):113–28. https://doi.org/10.1097/00005650-200202000-00006.
38. Mulhern B, O’Gorman H, Rotherham N, Brazier J. Comparing the measurement equivalence of EQ-5D-5L across different modes of administration. Health Qual Life Outcomes. 2015;26(13):191. https://doi.org/10.1186/s12955-015-0382-6.
39. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445–53. https://doi.org/10.1016/j.jval.2014.04.002.
40. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27(1):7–22. https://doi.org/10.1002/hec.3564.
41. Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. Pharmacoeconomics. 2018;36(6):663–74. https://doi.org/10.1007/s40273-018-0615-8.
42. Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19(4):343–52. https://doi.org/10.1016/j.jval.2016.01.003.
43. NICE. Position statement on use of the EQ-5D-5L value set for England (updated October 2019). Oct 2019. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l. [Accessed 9 Jun 2023].
44. EuroQol. QC report for England. June 03, 2019. Available from: https://eq-5dpublications.euroqol.org/download?id=0_80133&fileId=81961. [Accessed 18 Jul 2023].
45. Krabbe PFM, Jabrayilov R, Detzel P, Dainelli L, Vermeulen KM, van Asselt ADI. A two-step procedure to generate utilities for the Infant health-related Quality of life Instrument (IQI). PLoS ONE. 2020;15(4): e0230852. https://doi.org/10.1371/journal.pone.0230852.