ARTICLE IN PRESS

VALUE IN HEALTH ∎ (2017) ∎∎∎–∎∎∎

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Selecting Health States for EQ-5D-3L Valuation Studies: Statistical Considerations Matter

Zhihao Yang, MSc[1,2,*], *Nan Luo, PhD*[3], *Gouke Bonsel, PhD*[1,4], *Jan Busschbach, PhD*[1], *Elly Stolk, PhD*[4]

[1]*Erasmus Medical Center, Erasmus University Rotterdam, Rotterdam, The Netherlands;* [2]*Guizhou Medical University, Guiyang, Guizhou, China;* [3]*National University of Singapore, Singapore;* [4]*EuroQol Office, Rotterdam, The Netherlands*

ABSTRACT

Background: For many countries, the three-level EuroQol five-dimensional questionnaire (EQ-5D-3L) value sets have been established to estimate health state utilities. To generate these value sets, researchers first collect values for a subset of preselected health states from a panel representing the general public, and then use a prediction algorithm to generate values for all 243 states. High prevalence of a health state in daily practice has historically been a key criterion in selecting a subset of health states as the observed set. More recently, other criteria have been suggested, especially approaches based on statistical criteria such as randomization and orthogonality. **Objectives:** To evaluate the validity and accuracy of both the earlier and newer criteria, in terms of prediction of values for all the health states and of the values of common health states in particular. **Methods:** We used a pre-existing data set that contained visual analogue scale values from 126 students, each of whom valued all 243 EQ-5D-3L states. Then, we generated a series of designs and subsequently modeled the data with respect to each design. Some of these designs were used in the past; for example, the Measurement and Valuation of Health approach was included. Others were newly generated. The performance of different designs was evaluated in terms of the lowest root mean squared error for all health states taken together, and separately for common and rare states. Classification as common or rare was based on the frequency of the states' occurrence in three patient and population data sets pooled together (N = 5269). **Results:** The orthogonal design with 54 health states produced the lowest root mean squared errors. Over-representation of common health states in a design did not improve the estimations for these states. The published designs performed the worst, whereas the random selection designs were good on average. Nevertheless, the performance of the random selection designs showed more variance compared with orthogonal designs, because some of the former designs did not display appropriate balance. **Conclusions:** The published designs gave rise to large estimation errors for the extrapolated EQ-5D-3L health states. The orthogonal design focusing on statistical efficiency showed its superiority. Overall, when weighing up design properties, increased statistical efficiency outweighs an increased error rate, if any, in rare health states.

Keywords: common health state, EQ-5D-3L, orthogonal design, value set.

Introduction

The three-level EuroQol five-dimensional questionnaire (EQ-5D-3L) is the most widely used preference-based health-related quality-of-life questionnaire [1]. The EQ-5D descriptive system consists of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with three ordinal severity levels each (no problems, some problems, and extreme problems), thus defining 243 distinct health states [2]. The key feature of a preference-based instrument is that health state "values" (some prefer "utilities" or "index values") are derived for all its health states, which indicate how good or bad each health state is. These numbers are assumed to have ratio properties and can be used to estimate quality-adjusted life-years (QALYs).

The QALY is a preferred health outcome measure in cost-effectiveness studies around the world [3–6]. The results of cost-effectiveness studies could be biased if values of health states cannot be well estimated. Nevertheless, arriving at a set of 243 values for all separate states is a challenge: the valuation methods used can be demanding to the respondent, leading to data collection methods in which every respondent usually values only a defined subset of all 243 states. From such a data subset, typically, parametric regression analysis enables prediction/extrapolation of the values for all the health states. Different design choices (selection of subsets) have been documented in other areas [7,8], but in the literature on health state valuation, it is an open question as to how to select health states for inclusion in the subset for direct valuation. Different desirable properties have been identified that cannot all be satisfied at the same time. In the absence of straightforward statistical rules, selection has thus far been consensus-based. Hence, researchers take a leap of faith when estimating a value set while the advantages and

* *Address correspondence to:* Zhihao Yang, Erasmus Medical Center, Erasmus University Rotterdam, Na 2019, 80 Wytemaweg, Rotterdam 3015 CN, The Netherlands.

E-mail: E-mail address: z.yang@erasmusmc.nl

http://dx.doi.org/10.1016/j.jval.2017.09.001

disadvantages of their design choice are unknown. In this article, we aim to find the best design (subset of health states) for EQ-5D valuation studies by comparing the performance of different desirable design properties in prediction accuracy.

A key EQ-5D valuation study in the context of design choices was the Measurement and Valuation of Health (MVH) study conducted in the United Kingdom. Its consensus-based design has been replicated frequently [9]. In the MVH study, the following design criteria were selected [9]:

1. The set of states should spread widely over the valuation space so as to include most combinations of levels across the five dimensions.
2. Prima facie implausible states were excluded to sustain the credibility of the task and to reduce errors in assigned values.
3. All plausible combinations of dimension levels were to be included to allow identification of (first-level) interactions.
4. All respondents should value two out of the five mildest states that were most prevalent, and all respondents had to value the anchor state 33333 so that for every individual the utility range was known (11111 was defined as 1.0).
5. Some of the selected health states had been used in previous studies, thus maintaining a link to these studies.

Using these criteria the MVH study selected 42 health states, where the health data of the respondents themselves were used to determine common health states. The valuation of these 42 states was used to predict the values of all 243 states, including the 201 not included in the data subset [10]. In later research, the possibility that the MVH approach contained redundant values was investigated [2], but without returning to the desirability of the requirements mentioned earlier that were originally imposed on this approach.

Although the face validity of the aforementioned sampling criteria appears apparent, a disadvantage is the lack of attention to the statistical properties of the resulting design. From a statistical point of view, desired properties of a design are level balance and level pair balance (i.e., orthogonality). These two properties allow for statistical decomposition of all separate dimension effects dependent on the level of other dimensions. Nevertheless, both properties were ignored to adhere to the desirable properties emphasized in the MVH design. Whether this design reflects the best compromise with given resources is hitherto unknown.

Any comparison of design strategies requires quantitative criteria on what constitutes the "best" strategy. Bonsel et al. [11] systematically compared designs on the mean absolute error criterion, exploiting the possibilities of a "saturated data set" that contained observed values of all 243 EQ-5D-3L health states. The emphasis was on questions around the type (flat, stressed, mild, and severe) and number of health states that may be selected for valuation exercises (affecting estimation bias), and on whether the introduction of interaction terms in the model increased or reduced the risk of mis-specifying the value of out-of-sample health states. In their study, random selection of health states worked better than selection of states of specific types (as mentioned earlier) and the predictions became acceptable when the sampling ratio reached 10% (in EQ-5D-3L, that is 25 health states). In addition, the main-effects model appeared to be a crude but "safe" tool for estimation bias analysis [11].

The present study built on the work outlined in the last paragraph, with a specific focus on strategies for selecting health states. There were two competing design principles at issue that both have their merits. First, historically, health states for inclusion in a design tended to be hand-picked, on the basis of the properties of those health states (such as whether they were common) and the easily recognizable properties of the set as a whole (spanning the value range), but without an eye for the statistical properties of the set. Second, optimization of statistical properties was an alternative route to follow, with an orthogonal design looking like a promising alternative to the designs that had been used historically.

This study aimed to assess the comparative performance of the designs created while giving different weight to these principles. For this purpose, we compared designs on the basis of root mean squared errors (RMSEs), and also taking into account that from a user's perspective, misprediction of common health states could be considered as a mistake to be penalized more heavily than misprediction of rare health states (defined later). At first glance, over-representation of common health states (with assumed better face validity and data quality compared with rare health states) could lead to more accurate estimations for the common health states that shared the mild levels. Nevertheless, this could also lead to reduced statistical efficiency compared with balanced designs. Thus, it would be unknown whether the best prediction of a common state was achieved.

Although we were aware that using more design choices (the number of respondents, the number of health states per respondent, and the use of blocks [11]) may affect misprediction, this study focused on the aforementioned two design principles. Hence, "design" refers only to the deliberate selection of a subset of health states. Our research questions were addressed by testing various designs, using a pre-existing saturated data set with observations on all health states for reference purposes and by using different RMSE-based performance measures with and without focus on common health states.

Methods

Research Strategy

We used an existing data set with visual analogue scale (VAS) values from 126 students, each of whom valued all 243 EQ-5D-3L states. We generated a series of designs and subsequently modeled data subsets derived from each design.

Some of these designs were used previously, for example, the MVH subset; others were newly generated on the basis of our proposed design strategies. The performance of the different designs was evaluated in terms of the lowest RMSEs for all health states taken together and for common and rare states separately. Common health states were defined in terms of the frequency of their occurrence in the three reference data sets (see later).

Existing Data Set with VAS Values

In 1996, a students' panel (n = 126) provided EQ-VAS values for all 243 EQ-5D-3L health states. Such a data set is called "saturated," and if the resulting data set was regarded as suitable, no regression was needed to generate a value for each health state.

The EQ-VAS was displayed as a standard vertical 20-cm scale to record an individual's rating for a health state. By using the EQ-VAS, each health state could be valued on a scale from 0 (the worst score) to 100 (the best score). The students received 41 sheets of paper, each containing six EuroQol health states, except for the last sheet that contained one health state plus the states "unconscious" and "dead." The states 11111 and 33333 were valued twice at the outset before valuing all other states, and then with all the other health states. At the first time the values of 11111 and 33333 were used as anchors, whereas at the second time these values were used for analysis. The EQ-VAS was shown on a separate paper. The standard abbreviations for the health states (e.g., 11132) were printed above the health states to provide a shortcut for the stimuli. Next to the standard abbreviations, the students were able to fill in the values