



Universiteit  
Leiden  
The Netherlands

## Implementation and use of patient-reported outcome measures in routine nephrology care

Willik, E.M. van der

### Citation

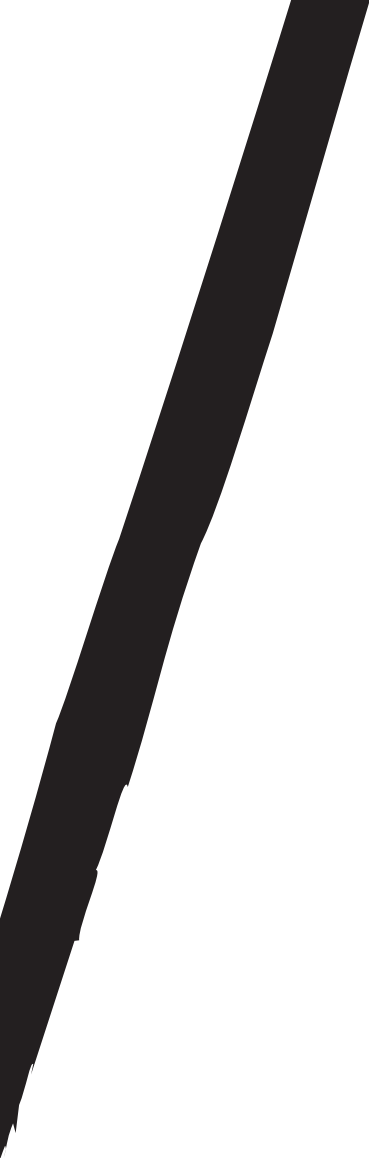
Willik, E. M. van der. (2023, June 8). *Implementation and use of patient-reported outcome measures in routine nephrology care*. Retrieved from <https://hdl.handle.net/1887/3619965>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3619965>

**Note:** To cite this publication please use the final published version (if applicable).



# **Patient-reported outcome measures (PROMs): making sense of individual PROM scores and changes in PROM scores over time**

Esmee M. van der Willik, Caroline B. Terwee, Willem Jan W. Bos,  
Marc H. Hemmelder, Kitty J. Jager, Carmine Zoccali, Friedo W.  
Dekker, Yvette Meuleman

*Nephrology (Carlton). 2021;26(5):391-399.*



## Abstract

Patient-reported outcome measures (PROMs) are increasingly being used in nephrology care. However, in contrast to well-known clinical measures such as blood pressure, healthcare professionals are less familiar with PROMs and the interpretation of PROM scores is therefore perceived as challenging. In this paper, we provide insight into the interpretation of PROM scores by introducing the different types and characteristics of PROMs, and the most relevant concepts for the interpretation of PROM scores. Concepts such as minimal detectable change, minimal important change and response shift are explained and illustrated with examples from nephrology care.

## Introduction

Over the last decades, a shift towards a more value-based and patient-centred healthcare has taken place, resulting in a stronger focus on patient-reported outcomes (PROs) such as health-related quality of life (HRQOL) and symptom burden.<sup>1,2</sup> PRO measures (PROMs) are nowadays introduced in nephrology care and may be used at individual level for personalized care and at aggregated level to evaluate healthcare quality. The use of PROMs at individual level as part of personalized care has been considered of great added value, as it may provide insight into patients' perceived health and their needs, and enhance patient-professional communication and shared decision making.<sup>3,4</sup> Ultimately, PROMs can be used to improve symptom management, HRQOL and other outcomes of healthcare.<sup>5,6</sup> To achieve such goals, knowledge about PROMs and the interpretation of PROM scores are needed. In contrast to well-known clinical outcomes such as blood pressure, healthcare professionals and researchers are not yet familiar with PROMs and the interpretation of PROM scores is therefore perceived as challenging. For example: What does a symptom burden score of 27 mean? Is a HRQOL-score of 36 normal for a certain patient or in a certain situation? Is a change in HROQL-score of 4 points clinically relevant? And why does the change in PROM score not always reflect the clinical change in health status?

In this paper, we provide insight into the interpretation of PROM scores by introducing the different types and characteristics of PROMs, and by presenting the most relevant concepts for the interpretation of PROM scores (i.e. minimal detectable change, minimal important change, and response shift), illustrated with examples from nephrology care.

## Patient-reported outcome measures

PROs are *outcomes* on aspects of patients’ perceived health, which includes a variety of concepts, for instance: HRQOL, functional status or symptom burden. PROs can be best measured by asking the patient himself and are reported by the patient himself (support may be offered when filling in PROMs, as long as responses reflect the patient’s perspective). PROMs are *questionnaires* that assess these aspects of perceived health. PROMs do not include experiences with, or perceptions and evaluations of healthcare provision; for this purpose, other measures are used, namely patient-reported experience measures (PREMs). Table 1 provides an overview of the terms used in this article.

**Table 1. Overview of terms used in this article**

|                                         |                                                                                                                                                                                                       |
|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Patient reported outcome (PRO)          | Outcomes on aspects of patients’ perceived health, reported from the patient’s perspective. E.g. health-related quality of life (HRQOL), functional status or symptom burden.                         |
| Patient reported outcome measure (PROM) | Questionnaire to measure one or multiple PROs (i.e. uni- or multidimensional PROM). PROMs are often classified as either a generic PROM or a specific PROM (i.e. for a certain disease or condition). |
| PROM score                              | Score for a PRO as measured by a PROM (i.e. the result from a PROM), which can be a score for one item or multiple items.                                                                             |
| Interpretability                        | “The degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to an instrument’s quantitative scores or change in scores.” <sup>7</sup>               |
| Minimal detectable change (MDC)         | A parameter of reliability that is defined as the “smallest change in score that can be detected beyond measurement error.” <sup>8</sup>                                                              |
| Minimal important change (MIC)          | “The smallest change in score in the construct to be measured which patients perceive as important.” <sup>8</sup>                                                                                     |
| Response shift                          | “A change in the meaning of one’s self-evaluation, which can be a result of recalibration, reprioritization and/or reconceptualization of the PRO.” <sup>9</sup>                                      |

Various types of PROMs exist and knowledge about certain characteristics of the PROMs is required to properly interpret PROM scores. Therefore, we will briefly introduce different types and characteristics of PROMs and elaborate on how they relate to the interpretation of PROM scores.

## Generic and specific PROMs

PROMs can roughly be classified as either generic or specific for a certain disease, condition or treatment. Generic PROMs measure a wide variety of health aspects and usually include aspects of people's health that are widely relevant (e.g. functional status or HRQOL in its broadest sense). Generic PROMs can therefore be used in any population, hereby enabling comparisons across populations or treatments, and are very suitable for heterogeneous and multimorbid populations (e.g. the elderly patient with chronic kidney disease (CKD) who often suffers from multiple comorbid conditions such as diabetes mellitus and cardiovascular disease). A disadvantage of this broad applicability is that it often goes with less precise PRO estimates and that nuances or small differences in PROs between or within specific populations may remain undetected.

Specific PROMs are tailored to a certain disease, condition or treatment, and address issues that are relevant to a specific group of patients, for example symptom burden related to CKD or related to immunosuppressive treatment after kidney transplantation.<sup>10, 11</sup> By tailoring to particular conditions, specific PROMs are usually better able to detect smaller or more specific differences or changes in PROs (e.g. a change in intensity or type of itching). Hence, specific PROMs are particularly suitable for comparisons within a population, but not for comparisons across populations. A disadvantage of a specific PROM is that relevant outcomes may be missed due to the focus on a certain disease or condition, for instance in heterogeneous populations with multiple comorbid conditions.

Whether a generic or a specific PROM is suitable depends on various aspects, including which PRO you aim to measure (e.g. disease specific symptoms or general functional status), the setting and purpose of measuring the PRO (e.g. is comparison within or also across populations of interest?), the diversity and characteristics of the population of interest (e.g. heterogeneity of the population), and the availability and quality of instruments (i.e. are high-quality and validated generic and/or specific PROMs available?). In practice, a combination of generic and specific PROMs is often used; either combined into one PROM such as the 36-item Kidney Disease Quality of Life (KDQOL-36) measuring generic HRQOL and kidney disease specific burden<sup>12, 13</sup>, or as separate PROMs for instance a combination of the SF-12 to measure generic HRQOL<sup>14</sup> and the Dialysis Symptom Index (DSI)<sup>10</sup> to measure kidney disease specific symptoms. The latter combination is used since 2018 in Dutch dialysis care<sup>3</sup>, for which the selection of the DSI has been described in detail elsewhere.<sup>15</sup>

## Scoring systems of PROMs

A standard PROM scoring system or scale does not exist, not even when PROMs are measuring the same PRO. In contrast to other measures (e.g. temperature and distance) that can be measured on the same scale (e.g. Celsius and meters), PROMs use varying scales and scoring methods.

Table 2 presents an example of three PROMs that measure HRQOL (PROMIS Profile-29), symptoms (DSI) or both (KDQOL-36) to illustrate the variety in measurement characteristics across PROMs. The PROMs differ for many features, such as the domains being measured (also for the same PRO, i.e. HRQOL), the number of questions, response options, scales and scoring methods. As a result of the differences in features, PROMs often also differ in the interpretation of scores. For example: although the DSI and the KDQOL-36 both measure disease specific symptoms, PROM scores are not directly comparable due to different scoring systems (e.g. score range, method and direction; Table 2). A KDQOL-36 symptom burden score of 71 represents a reasonable health status similar to that of an average patient with CKD.<sup>13, 16</sup> However, a DSI symptom burden score of 71 represents an extremely high symptom burden that is twice as high as in an average dialysis patient.<sup>3</sup>

## 7

## Measurement properties of PROMs

Measurement properties such as validity and reliability provide essential information about the quality of the PROM in certain populations and settings. The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) taxonomy describes which aspects should be considered to judge the quality of the PROM.<sup>7</sup> Good measurement properties are a prerequisite for PROMs to be useful and reasonably interpretable. However, measurement properties such as validity and reliability itself provide insufficient insight into the meaning of scores, i.e. the interpretation of PROM-scores.

## Interpretation of PROM scores

The interpretability of a PROM has been defined as “the degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to an instrument’s quantitative scores or change in scores”.<sup>7</sup> The interpretability can be considered a characteristic of the PROM, meaning that one PROM may be easier to interpret than another PROM. The interpretation of PROM scores can be challenging, for instance due to the complexity of the PRO (e.g. HRQOL, which includes various physical, mental and social domains) or the PROM (e.g. a

complex scoring method). Luckily, there are some intuitive methods that may facilitate the interpretation of PROM scores which will be discussed below.

First, a discussion on PROs between the patient and the professional may provide insight into the individual's view on certain aspects of health, e.g. what is important to the patient and what is his frame of reference. The PROM items and also the overall PROM scores may facilitate this conversation, for instance by serving as a checklist or as a reason to start the conversation about (difficult) subjects.<sup>3</sup>

Second, group-level data may facilitate the interpretation of individual PROM-scores by providing insight into what is 'normal' and what may be expected. Descriptive information such as the mean, standard deviation and range in the population of interest gives an indication of the variability of scores (i.e. should scores be expected across the whole scale or on a smaller range?) and of what is 'normal' (e.g. is the score of a patient low, average or high as compared to other patients?). Comparison to norm- or reference scores of a general population or a population with a certain condition or treatment can be highly informative. For example: comparing a 65-year old dialysis patient's HRQOL-score of 40 to the average Dutch dialysis population (mean score: 36 (SD 11))<sup>3</sup> and the general 60-69 year old Dutch population (mean score: 51 (SD 9))<sup>18</sup> gives an idea of how the patient addresses his outcome in comparison to the reference population. Furthermore, descriptive information about floor- or ceiling effects, meaning that many individuals score at the lower (i.e. floor) or upper (i.e. ceiling) end of the scale, may be informative because differences below or above these limits cannot be observed. This may be valuable information to take into account when interpreting individual patient scores.

Third, it is insightful to compare PROM scores to scores of other measures. Since most PROMs are relatively new to clinical care, most users (both patients and healthcare professionals) are not yet sufficiently familiar with PROM scores. By comparing PROM scores to well-known (clinical) measures such as kidney function or laboratory measures and to patient- or disease characteristics, one may become more experienced with the scores and get a feeling for which scores are common for certain patients, conditions and situations (i.e. the scores get 'clinical or commonly understood connotations').

Finally, the interpretability of PROM scores may automatically improve over time when patients and professionals become more experienced in using and discussing PROM scores. In addition to these more intuitive aspects of interpreting PROM scores, there are also methodological concepts, i.e. benchmarks, that are relevant to the interpretability of *changes* in PROM scores, which will be discussed below.

**Table 2. Illustration of variation in characteristics across different patient-reported outcome measures**

|                                      | PROMIS Profile-29                                                                                                                                                          | KDQOL-36                                                                                                                                                                                      | DSI                                                                         |
|--------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| <b>PRO</b>                           | HRQOL                                                                                                                                                                      | Disease burden and HRQOL                                                                                                                                                                      | Symptom burden                                                              |
| <b>Target population<sup>5</sup></b> | People with or without (chronic) illness                                                                                                                                   | Patients with kidney disease                                                                                                                                                                  | Haemodialysis patients                                                      |
| <b>Type</b>                          | Generic                                                                                                                                                                    | Disease specific and generic <sup>#</sup>                                                                                                                                                     | Disease specific                                                            |
| <b>Domains</b>                       | Depression<br>Anxiety<br>Physical function<br>Pain interference<br>Fatigue<br>Sleep disturbance<br>Ability to participate in social roles and activities<br>Pain intensity | Disease specific:<br>Symptoms/problems<br>Effects of kidney disease<br>Burden of kidney disease<br>Generic <sup>#</sup> :<br>SF-12 Physical Health Composite<br>SF-12 Mental Health Composite | Symptom burden                                                              |
| <b>Number of questions</b>           | 29, or tailored to the patient <sup>§</sup>                                                                                                                                | 36                                                                                                                                                                                            | 30                                                                          |
| <b>Recall period</b>                 | In general/1 week                                                                                                                                                          | In general/4 weeks                                                                                                                                                                            | 1 week                                                                      |
| <b>Rating scale</b>                  | 5-point Likert scale, 0-10 scale (for pain intensity only)                                                                                                                 | Various scales: Yes/no, 3-, 5- or 6-point scale                                                                                                                                               | Yes/no (presence of symptoms), 5-point Likert scale (severity)              |
| <b>Item score</b>                    | 1 to 5 points or vice versa, so that a higher score represents more of the domain being measured.                                                                          | Item-scores are transformed to a 0-100 possible range. E.g. the 5-point scale has 0/25/50/75/100 points.                                                                                      | 0 points if symptom is not present; 1 to 5 points for severity <sup>^</sup> |
| <b>Total score (range)</b>           | T-score (roughly 0-100)                                                                                                                                                    | 0-100                                                                                                                                                                                         | 0-150 <sup>^</sup>                                                          |
| <b>Scoring method</b>                | IRT-based scoring                                                                                                                                                          | Disease specific: average score<br>Generic <sup>#</sup> : norm-based scoring algorithm                                                                                                        | Sum score <sup>^</sup>                                                      |

|                                    |                                                                                                                                                                                            |                                                                                                                                                                                             |                                                  |
|------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------|
| <b>Meaning of score direction</b>  | Higher scores represent more of the domain being measured. E.g. a higher score on fatigue means a worse fatigue, and a higher score on physical function means a better physical function. | Higher scores represent a more favourable health state. E.g. a higher score on symptoms means a lower symptom burden, and a higher score on physical health means a better physical health. | Higher scores represent a higher symptom burden. |
| <b>Norm- or reference standard</b> | General US population: mean 50, SD 10                                                                                                                                                      | Disease specific: n/a. Generic#: General US population: mean 50, SD 10                                                                                                                      | N/a                                              |

<sup>§</sup> The target population is the population for which the PROM was originally developed and is not necessarily the only population for which the questionnaire is used and considered suitable.

<sup>§</sup> PROMIS questionnaires can be applied as Computerized Adaptive Test (CAT) per domain, whereby the computer selects items based on the patient's responses to previous questions. The number of questions usually depends on a predetermined threshold for the precision of the measurements and may therefore vary across patients and measurements.

<sup>#</sup> The generic part of the KDQOL-36 is the 12-item short form (SF-12) health survey.

<sup>^</sup> In the original development paper of the DSI<sup>10</sup>, a 0-4 scale was used for severity and no guidance for an overall score was provided. Therefore, the symptom burden score is often calculated according to the method presented in this table, which was previously described by Abdel-Kader et al. (2009).<sup>17</sup>

Abbreviations: DSI, Dialysis Symptom Index; KDQOL-36, 36-item Kidney Disease Quality of Life; PROMIS, Patient-Reported Outcomes Measurement Information System; IRT, Item Response Theory; n/a, not available.

## Minimal detectable change (MDC)

Suppose that a patient with advanced CKD fills in the Short Form-36 (SF-36) twice with a 6 months interval between the two measurements. The HRQOL results show a decrease of 5 points at the physical component score (hereafter called 'physical HRQOL') and a decrease of 2 points at the mental component score (hereafter called 'mental HRQOL'). Can we then speak of a real deterioration in HRQOL? In other words, do we observe an actual change or is it possibly just random variation? To answer this question we need to know whether the observed change is larger than the minimal detectable change (MDC), also known as the smallest detectable change or the minimal real change. The MDC is a parameter of reliability and is defined as the "smallest change in score that can be detected beyond

measurement error”.<sup>8</sup> Thus, the MDC reflects the threshold at which a change in score can be considered statistically significant.

The MDC should be estimated in persons who have *not* changed over time (e.g. clinically stable patients) using a test-retest design, because this demonstrates the random variation (i.e. measurement error) in score *within* persons (see Box 1 for the method to calculate the MDC). In patients with conservatively managed stage 5 CKD, Erez et al. (2016) found an MDC of 4.2 and 7.0 for the SF-36 physical and mental HRQOL, respectively.<sup>19</sup> Using these thresholds in our example, the observed change of 5 points for physical HRQOL is larger than the MDC and can therefore be considered a statistically significant change. The observed change of 2 points in mental HRQOL is smaller than the MDC and can therefore not be distinguished with 95% confidence from no change – i.e. the change in mental HRQOL may be due to random variation and thus cannot be considered a true change.

Taken together, the MDC helps with the interpretation of PROM scores over time by distinguishing real changes from what is probably random variation. Although some literature is available<sup>19,20</sup>, more research on MDC is needed to facilitate interpretation of changes in PROM scores for different PROMs and in different patients and settings within nephrology care.<sup>21</sup>

## 7

### Minimal important change (MIC)

If the observed change in our example of 5 points on physical HRQOL is likely a true change, can we then assume that this change is relevant to patients? And, if a decrease of 2 points does not demonstrate a real change in mental HRQOL, can we then also assume that this change is not meaningful for patients? To answer this question we need to know whether the observed change is larger than the minimal important change (MIC) or minimal clinically important change, in the literature also referred to as the minimal (clinically) important difference. MIC has been defined as “the smallest change in score in the construct to be measured which patients perceive as important”.<sup>8</sup>

There are several methods for estimating the MIC, some of which are briefly discussed in Box 2. The MIC is not a fixed characteristic of a PROM and can vary across populations and settings. For example: characteristics of the population (e.g. mild or severe conditions), the direction of change (i.e. improvement or deterioration) and the study design and analysis used to estimate MIC (e.g. different anchors or definitions of importance) can influence the MIC.<sup>8</sup> Some literature is available that can provide a cautious indication of the MIC of some PROMs (e.g. SF-36) that might be used in nephrology care.<sup>19,22</sup> However, in order to interpret changes in PROM scores clearly, more information is needed about the MIC in patients with

**Box 1. Measuring minimal detectable change (MDC)**

The MDC is a statistical parameter based on the measurement error (Standard Error of Measurement; SEM). The MDC can be determined in individuals who have not changed using a test-retest design, and can be calculated using the following formula:  $1.96 * SD_{\text{change}}$ , which equals  $1.96 * \sqrt{2} * SEM$ .<sup>8</sup>

**Box 2. Measuring minimal important change (MIC)**

The MIC can be assessed using an anchor-based approach, for which several methods exist. In the literature also distribution-based approaches have been described<sup>23</sup>; however, these methods do not involve the *importance* of change and are therefore considered less suitable. In this box, we briefly touch upon the most common (anchor-based) methods to define MIC.

With an anchor-based approach the MIC is determined by comparing the changes in the PROM score to another measure that defines a clinical relevant change (i.e. the anchor). For PROMs usually the patient's general rating of change serves as an anchor, in which the minimal relevant change is explicitly defined by the patient.<sup>8, 23</sup>

A relatively easy method to determine the MIC is the mean change method. With this method the MIC is defined as the mean change in PROM score in patients who consider themselves to be minimally importantly changed, according to the anchor (e.g. in patients who rate their health as 'slightly improved').<sup>8, 23</sup> Another method to determine the MIC is by use of receiver operating characteristic (ROC) analysis. The method is similar to the method known from diagnostic test research, whereby the PROM score is considered the diagnostic test and the anchor serves as a gold standard. The optimal ROC cutoff point gives the smallest chance of misclassifying importantly improved and not-improved patients and is therefore considered the MIC.<sup>8, 24</sup>

Furthermore, predictive modeling can be used. The outcome in this analysis is being either improved or not improved, which is defined based on the anchor. The change in PROM score is used as the predictor variable. The MIC is then determined using logistic regression analysis and is defined at the point where the change in PROM score is associated with a likelihood ratio of 1. An example of this method has been described in detail by Terluin et al. (2015).<sup>24</sup>

CKD in different stages and settings, and receiving different treatments.<sup>21</sup>

In patients with conservatively managed stage 5 CKD, Erez et al. (2016) report a MIC of 6.3 for the SF-36 score on physical HRQOL and 8.7 for the SF-36 score on mental HRQOL.<sup>19</sup> Comparing these thresholds to the observed changes in scores in our example of 5 and 2 for physical and mental HRQOL, respectively, shows that both observed scores are smaller than the MIC and are thus, on average, not considered important by patients. This example can be seen as a desirable situation: although statistically there is a decline in physical HRQOL, patients most likely do not perceive it as a relevant deterioration in their HRQOL.

However, the MIC gives an indication of what is *on average* considered important by an individual and should therefore be considered as a probability-threshold to interpret individual changes: if an individual change is larger than the MIC, the probability that this change is perceived important by the patient is greater than the probability that this change is perceived as not important.<sup>25</sup> The fact that the interpretation of the MIC involves probabilities, also indicates that this threshold may not apply to all individuals and that patients differ in which change they perceive as important. Therefore, it may be of added value to discuss the changes to gain insight into what is perceived important by the individual. On the other hand, the MIC may also facilitate the conversation, for example: it may be informative to the patient to explain which change in HRQOL may be expected (e.g. after kidney transplantation) and whether this change is, on average, considered important by patients.

Taken together, the results from our example can be considered positive with regard to both the MIC and the MDC: the MIC is larger than the MDC ( $6.3 > 4.2$  and  $8.7 > 7.0$  for physical and mental HRQOL, respectively<sup>19</sup>) and thus, both the physical and mental HRQOL scales of the SF-36 seem to be able to detect changes that are, on average, important to patients. If the MIC would be smaller than MDC, the PROM may not be able to distinguish with high certainty relevant changes from random variation. Consequently, important changes might be missed and it may thus be advisable to use a different PROM or to improve the initial PROM in such way that it has a smaller MDC (i.e. by reducing the measurement error), for purposes where a high certainty is important (e.g. evaluation of treatment strategies).

## Response shift

Another concept that is important for the interpretation of PROM scores is response shift, which refers to a change in the *meaning* of one's evaluation of the PRO (e.g. HRQOL) over time. This means that patients' answers to PROM questions change over time, not only because their health or HRQOL has changed, but

also because they might have changed their perception on what health or HRQOL means to them. For example: when Jason (male, 62y) started dialysis treatment, he experienced a deterioration in his health condition. Jason had to deal with vascular access problems and anemia, and it took several months to reach a hemoglobin level within the target range. Starting dialysis also had a major impact on his daily life: the sudden change in his schedule affected his ability to work and to participate in social activities. One might expect that such changes would impact Jason's HRQOL. However, contrary to what one might expect, after a couple of months Jason reported a HRQOL that was only slightly lower compared to his HRQOL at the start of dialysis. In this example a 'response shift' has occurred, that has been defined as "a change in the meaning of one's self-evaluation, which can be a result of *recalibration*, *reprioritization* and/or *reconceptualization* of the PRO".<sup>9</sup> Below these response shift inducing concepts are described and illustrated by means of Jason's example.

*Recalibration* refers to a change in an individual's frame of reference. In the example of Jason, his daily schedule and social life have changed considerably: since Jason started with dialysis treatment, he became more engaged in social comparison by talking to and sharing experiences with other patients treated with dialysis. Insights into the experiences of other patients, changed Jason's internal definition (i.e. his reference standard) of a poor HRQOL and consequently, Jason rates the HRQOL he had when he started dialysis higher now than he did before. Thus, new information and experiences can lead to a change in where a person positions himself on the scale, i.e. recalibration.

*Reprioritization* refers to a change in personal values. In Jason's case, acceptance of not being able to work and positive experiences with peer support could have encouraged Jason to shift his focus towards other aspects in life and set new life goals. Prior to dialysis, Jason mainly focused on professional accomplishments but after starting dialysis treatment, family relationships and being able to help others became more important to Jason. This illustrates how experiences can change people's self-evaluation and the value of certain aspects in life, and thus in *the extent to which aspects contribute* to a PRO, i.e. reprioritization.

*Reconceptualization* is a redefinition of the concept of interest. In the example of Jason, this could mean that his personal meaning of HRQOL has changed. By accepting the new daily routine and by appreciating a different way of participation in society, Jason may have realized that other factors determine his HRQOL. For Jason, being able to offer support to less fortunate peers contributes to a good HRQOL and having a certain employment status does no longer determine his HRQOL and consequently, his definition of HRQOL has changed. Hence, new ex-

periences can induce a change in *which aspects* contribute to a PRO and thus in one's definition of the PRO, i.e. reconceptualization.

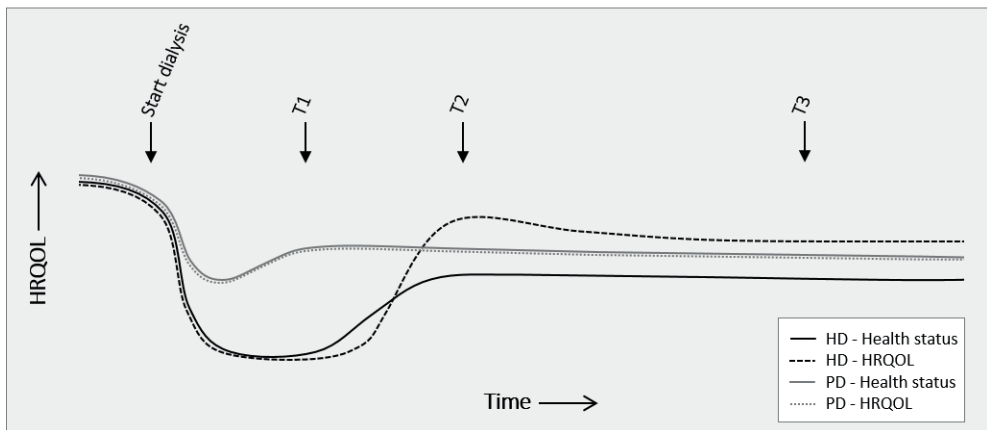
Changes in internal standards, personal values or conceptualization of PROs may result in a response shift and thus in an experienced HRQOL that differs from what would be expected based on one's change in clinical health status, that is, for instance, based on clinical parameters (i.e. a decline in health status does not automatically imply a decrease in HRQOL). Changes may be induced by certain health- or life-changing events (e.g. getting a diagnosis, the start of a treatment or the loss of a loved one) and can also occur more gradually over time, for instance in chronic diseases.<sup>9,26-28</sup> It is proposed that changes in health or in life may interact with the patient's characteristics (e.g. personality) and with mechanisms such as coping and social comparison, and consequently influence response shift.<sup>9</sup>

In the past decade, response shift has been investigated particularly in HRQOL research, but can occur in any PRO and when using any PROM as they all concern subjective self-evaluations. Nevertheless, PROs or PROMs that leave more room for personal interpretation are more sensitive to response shift compared to PROs or PROMs that are more unambiguously defined. For example: the question 'How is your sleep quality in general?' requires more consideration and evaluation from the patient than the question 'In the past week, did you sleep through the night without interruptions?', and the first question is therefore more prone to different interpretations over time.<sup>8</sup>

Response shift can complicate the interpretation of PROM scores over time. Therefore, it is important to know that this phenomenon exists, as it may explain unexpected findings (e.g. a stable HRQOL while clinical outcomes clearly show a deterioration in health). Response shift itself may also be a treatment goal, for instance in a treatment aimed at improved coping and self-management. Herein, response shift provides insight into the ability to adapt to a certain change in health. Furthermore, at the individual patient level, further investigation of and discussion about changes in internal standards, values and conceptualizations may help to interpret the patient's scores and guide decision-making.<sup>28</sup>

At a group level, it may also be informative to gain insight into response shift for instance by comparing treatment effects to inform decision making.<sup>26</sup> For example: let's compare HRQOL scores of patients treated with hemodialysis (HD) and peritoneal dialysis (PD) at several time-points during the first year of treatment (Figure 1). Theoretically, one may expect that HD impacts health status (e.g. based on clinical parameters) and HRQOL more severely compared to PD (e.g. due to the hospital visit 3 to 4 times a week). However, it is possible that PD patients will try to maintain their old way of life, while HD patients will try to adapt to their treat-

ment and to their new life. This may result in larger changes in internal standards, values and conceptualizations in HD patients compared to PD patients. As a result, HD patients may perceive a better HRQOL after some time (e.g. T2 in Figure 1), despite having a lower health status compared to PD patients. Such information is important for patients and professionals when drawing conclusions about treatment effects.



**Figure 1. Theoretical example of trajectories of health status and HRQOL in patients receiving HD and PD.** A response shift occurs in the HD patient between T1 and T2. Abbreviations: HRQOL, health-related quality of life; HD, haemodialysis, PD, peritoneal dialysis.

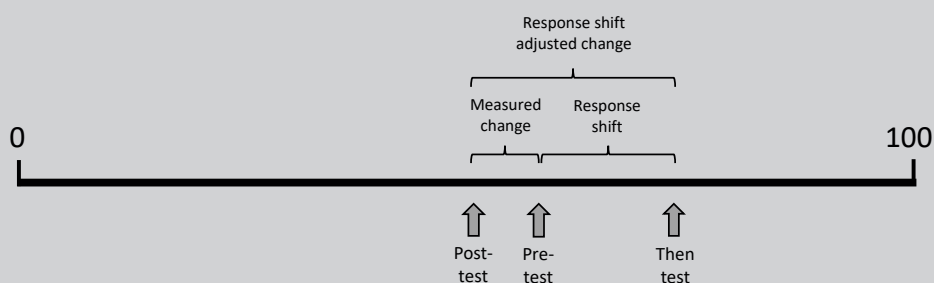
Furthermore, information about PRO-trajectories over time is also important when evaluating a patient's treatment, for example: the time-point at which the PRO was assessed could be informative to the interpretation of the PROM score.<sup>26</sup> Based on the trajectory comparison between HD and PD in Figure 1, different conclusions can be drawn, depending on the moment PROs are measured (start of dialysis, T1 or T2/T3). This example shows that a response shift may also occur later in the trajectory (e.g. between T1 and T2 in HD), and not directly after the life-changing event (e.g. start of dialysis).

Insight into the size and direction of the response shift can be informative, not only to explain unexpectedly small (or large) changes in PROM scores, but also to gain insight into the psychological change that may have occurred and the patient's ability to adapt. Several methods exist to determine response shift<sup>29</sup>, some of which are briefly discussed in Box 3.

### Box 3. Measuring response shift

Several methods exist to assess whether, how and to what extent response shift occurred. Barclay-Goddard et al. (2009) provided an overview of the methodologies to address response shift.<sup>29</sup> In this box, we briefly highlight some of the main approaches.

The most commonly used method is the *then-test*. In this method, the patient is asked to complete a PROM about his health status at two time-points, for instance at baseline (pre-test) and after 6 months (post-test). Additionally, the patient is asked at the post-test time-point to also complete the PROM for his health status at baseline (then-test). Since both the post-test and the then-test are completed at the same time-point, it may be assumed that the patient applied the same standards, values and concepts. Therefore, response shift can be assessed by comparing the pre-test and the then-test, and the difference between the post-test and then-test gives the response shift adjusted change (Figure 2).<sup>8, 29</sup>



**Figure 2. Then-test**

The then-test has also been applied in combination with *qualitative methods* (e.g. using an interview) to explore response shift.<sup>30</sup> An advantage of combining these methods is that both numerical value of the response shift (using the then-test) and in-depth insight into the patient's thoughts and considerations regarding his standards, values and concepts are assessed. Qualitative methods can also be applied independently to investigate mechanisms of reconceptualization, reprioritization and recalibration that induce response shifts, as was done by Elliott et al. (2014) in dialysis patients.<sup>28</sup>

Another method to gain insight into changes in the patient's standards, values and concepts is by the use of a *questionnaire* that enables patients to define their own meaning of the construct (e.g. HRQOL), such as the Schedule for the Evaluation of Individual Quality of Life (SEIQOL).<sup>30, 31</sup> Changes over time in the

patient's reference standard, or in which and to what extent domains contribute to the patient's HRQOL may indicate a response shift.

Furthermore, response shift can be investigated using a *statistical approach*, such as confirmatory factor analysis. With this method, the three response shift inducing concepts can be identified by comparing the factor structure of the PROM pre- and post-measurement, namely: recalibration (apparent from a mean change in the variables), reprioritization (by means of a change in importance – i.e. factor loadings – of domains) and reconceptualization (by means of a change in the number of identified domains).<sup>8, 29</sup>

## Conclusion

PROMs are instruments to assess aspects of the patient's perceived health, such as HRQOL or symptom burden. Different types of PROMs exist and knowledge about the characteristics of the PROM is necessary to interpret PROM scores and change scores. Information about the average and distribution of PROM scores in a reference population or in comparison to more familiar outcomes (e.g. laboratory measures) are indispensable to interpret and get used to PROM scores. Furthermore, the MDC and MIC are important to inform us about statistically and clinically relevant changes, respectively. Besides, one must be aware that response shift may occur, which may explain unexpectedly small (or large) changes in PROM scores. Finally, communication is important to interpret individual PROM scores; the best manner to interpret individual PROM scores and changes in PROM scores is through a discussion between the patient and the healthcare professional, in which the measures discussed in this paper (i.e. MDC, MIC and response shift) may have a facilitating role. Ideally, such measures are integrated into a dynamic report with individual PROM scores over time, enabling both patients and professionals to easily oversee which outcomes require attention and possibly intervention, and to evaluate treatment strategies at individual level. This will potentially increase the usability of PROMs in nephrology care for both patients and healthcare professionals.

## References

1. Porter ME. What is value in health care? *The New England journal of medicine* 2010;363(26):2477-81
2. Black N. Patient reported outcome measures could help transform healthcare. *BMJ (Clinical research ed)* 2013;346:f167-f
3. van der Willik EM, Hemmelder MH, Bart HAJ, et al. Routinely measuring symptom burden and health-related quality of life in dialysis patients: first results from the Dutch registry of patient-reported outcome measures. *Clinical Kidney Journal* 2021;14(6):1535-44
4. Noonan VK, Lyddiatt A, Ware P, et al. Montreal Accord on Patient-Reported Outcomes (PROs) use series - Paper 3: patient-reported outcomes can facilitate shared decision-making and guide self-management. *J Clin Epidemiol* 2017;89:125-35
5. Basch E, Deal AM, Kris MG, et al. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *J Clin Oncol* 2016;34(6):557-65
6. Kotronoulas G, Kearney N, Maguire R, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. *J Clin Oncol* 2014;32(14):1480-501
7. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology* 2010;63(7):737-45
8. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge University Press: 2011;
9. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med* 1999;48(11):1507-15
10. Weisbord SD, Fried LF, Arnold RM, et al. Development of a symptom assessment instrument for chronic hemodialysis patients: the Dialysis Symptom Index. *J Pain Symptom Manage* 2004;27(3):226-40
11. Dobbels F, Moons P, Abraham I, Larsen CP, Dupont L, De Geest S. Measuring symptom experience of side-effects of immunosuppressive drugs: the Modified Transplant Symptom Occurrence and Distress Scale. *Transpl Int* 2008;21(8):764-73
12. Hays RD, Kallich JD, Mapes DL, Coons SJ, Carter WB. Development of the kidney disease quality of life (KDQOL) instrument. *Qual Life Res* 1994;3(5):329-38
13. Kidney Disease Quality of Life Instrument (KDQOL) Santa Monica, CA, RAND Health Care. [https://www.rand.org/health-care/surveys\\_tools/kdqol.html](https://www.rand.org/health-care/surveys_tools/kdqol.html). Accessed July 17, 2020
14. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care* 1996;34(3):220-33
15. van der Willik EM, Meuleman Y, Prantl K, et al. Patient-reported outcome measures: selection of a valid questionnaire for routine symptom assessment in patients with advanced chronic kidney disease - a four-phase mixed methods study. *BMC nephrology* 2019;20(1):344-

16. Hays RD, Kallich JD, Mapes DL, et al.: Kidney Disease Quality of Life Short Form (KDQOL-SF), Version 1.3: A Manual for Use and Scoring, Santa Monica, CA, RAND, 1997 Contract No.: P-7994
17. Abdel-Kader K, Unruh ML, Weisbord SD. Symptom burden, depression, and quality of life in chronic and end-stage kidney disease. *Clin J Am Soc Nephrol* 2009;4(6):1057-64
18. Mols F, Pelle AJ, Kupper N. Normative data of the SF-12 health survey with validation using postmyocardial infarction patients in the Dutch population. *Qual Life Res* 2009;18(4):403-14
19. Erez G, Selman L, Murtagh FE. Measuring health-related quality of life in patients with conservatively managed stage 5 chronic kidney disease: limitations of the Medical Outcomes Study Short Form 36: SF-36. *Qual Life Res* 2016;25(11):2799-809
20. Ware J, Ma K, Keller SD. SF-36 Physical and Mental Health Summary Scales: a User's Manual. 1993;8:23-8
21. Aiyegbusi OL, Kyte D, Cockwell P, et al. Measurement properties of patient-reported outcome measures (PROMs) used in adult patients with chronic kidney disease: A systematic review. *PloS one* 2017;12(6):e0179733-e
22. Finkelstein FO, van Nooten F, Wiklund I, Trundell D, Cella D. Measurement properties of the Short Form-36 (SF-36) and the Functional Assessment of Cancer Therapy - Anemia (FACT-An) in patients with anemia associated with chronic kidney disease. *Health Qual Life Outcomes* 2018;16(1):111
23. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56(5):395-407
24. Terluin B, Eekhout I, Terwee CB, de Vet HCW. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. *Journal of Clinical Epidemiology* 2015;68(12):1388-96
25. de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol* 2010;63(1):37-45
26. Hartog ID, Willems DL, van den Hout WB, et al. Influence of response shift and disposition on patient-reported outcomes may lead to suboptimal medical decisions: a medical ethics perspective. *BMC Med Ethics* 2019;20(1):61
27. Barclay-Goddard R, King J, Dubouloz CJ, Schwartz CE. Building on transformative learning and response shift theory to investigate health-related quality of life changes over time in individuals with chronic health conditions and disability. *Arch Phys Med Rehabil* 2012;93(2):214-20
28. Elliott BA, Gessert CE, Larson PM, Russ TE. Shifting responses in quality of life: people living with dialysis. *Qual Life Res* 2014;23(5):1497-504
29. Barclay-Goddard R, Epstein JD, Mayo NE. Response shift: a brief overview and proposed research priorities. *Qual Life Res* 2009;18(3):335-46
30. Westerman MJ, Hak T, Sprangers MA, Groen HJ, van der Wal G, The AM. Listen to their answers! Response behaviour in the measurement of physical and role functioning. *Qual Life Res* 2008;17(4):549-58
31. O'Boyle C. The Schedule for the Evaluation of Individual Quality of Life (SEIQoL). *International Journal of Mental Health* 1994;23:3-23