LONGITUDINAL ANALYSIS OF THE HEALTH-RELATED QUALITY OF LIFE IN ONCOLOGY INTEGRATING THE OCCURRENCE OF THE RESPONSE SHIFT EFFECT

Amélie Anota[1]

Methodology and Quality of Life in Oncology Unit (EA 3181) and French National Platform Quality of Life and Cancer, University Hospital of Besançon, France

Francesco Cottone

Italian Group for Adult Hematologic Diseases (GIMEMA) Data Center and Health Outcomes Research Unit, Italy

Antoine Barbieri

Biometrics Unit - CTD INCa, Institut régional du Cancer de Montpellier (ICM) Val d'Aurelle, and Institut montpelliérain Alexander Grothendieck (IMAG), University of Montpellier, France

Caroline Bascoul-Mollevi

Biometrics Unit - CTD INCa, Institut régional du Cancer de Montpellier (ICM) Val d'Aurelle, and INSERM U1194, Integrative cancer research for personalized medicine in GI oncology, Institut de Recherche en Cancérologie de Montpellier (IRCM), France

Fabio Efficace

Italian Group for Adult Hematologic Diseases (GIMEMA) Data Center and Health Outcomes Research Unit, Italy

Franck Bonnetain

Methodology and Quality of Life in Oncology Unit (EA 3181) and French National Platform Quality of Life and Cancer, University Hospital of Besançon, France

---

[1] Amélie Anota, email: aanota@chu-besancon.fr

Abstract. *The aim was to investigate how to take into account the occurrence of the response shift effect according to the linear mixed model for repeated measures (LMMRM), the time to health-related quality of life score deterioration (TTD) and the longitudinal partial credit model (LPCM). Three analyses were conducted per method on data from a study on early breast cancer patients: 1) using the then-test score as the reference, 2) on prospective measures and 3) proposing an alternative method using prospective measures. The LMMRM was the most impacted method by the occurrence of the RS effect. Alternative methods investigated seemed to reduce the bias for LMMRM, increase the bias for LPCM and not have any impact on TTD. Further analyses are still needed to confirm these first results on both real data and simulations.*

Keywords: *Health-related quality of life, oncology, longitudinal analysis, Response shift, then-test method.*

1. INTRODUCTION

Health-Related Quality of Life (HRQoL) is a key endpoint in oncology clinical trials aiming to ensure the clinical benefit for the patient when assessing new treatments or therapeutic strategies (Osoba, 2011). However, results of HRQoL are still underutilized in clinical practice to change the patient's standard of care, mainly due to the complexity of HRQoL concept and its analysis. In fact, HRQoL is a subjective and dynamic concept depending on the patient's adaptation to the disease and reflected by the occurrence of a response shift effect. This effect can be defined as "a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement (i.e. scale recalibration); (b) a change in the respondent's values (i.e. the importance of component domains constituting the target construct, [reprioritization]) or (c) a redefinition of the target construct (i.e. reconceptualization)" (Sprangers and Schwartz, 1999). To illustrate these three components of response shift, you can consider a woman who is diagnosed for breast cancer. We ask this woman to rate her fatigue level on a 0 to 100 scale where 100 reflects the highest level of fatigue imaginable. At the time of diagnosis, the patient reports a fatigue level of 80/100. After beginning her treatment consisting of a surgery followed by a chemotherapy, we ask again to this patient to assess her fatigue level. She rates her fatigue level to 50/100. This is an illustration of the recalibration effect. The reprioritization component can be illustrated by a change in the importance of some domains as regard to HRQoL. For example, this woman can attach importance to her physical activity doing a lot of sports before her cancer was diagnosis. After her cancer was declared and her treatment began, she gives less importance to her physical well-being. Conversely, she gets close to her family and friends and thus her social well-being becomes more importance as regard to her HRQoL. This is an example of the reprioritization. Finally, the reconceptualization is a most conceptual component of the response shift effect. It can be illustrated by a person who engages in spiritual practice after the cancer diagnosis (Park et al., 1990). Response shift effect can bias the longitudinal analysis if it is not adequately taken into account. Particularly, a differential impact of the occurrence of the response shift effect between two treatment arms can bias the conclusion of some randomized clinical trials by over or underestimating the treatment effect.

Several statistical and methodological approaches have been proposed to identify its occurrence (Guilleux et al., 2015; Korfage et al., 2007; Oort et al., 2005; Schwartz and Sprangers, 1999). The then-test method is often considered as a gold standard to assess the recalibration component of the response shift effect. It consists to introduce in the study design a retrospective measurement of HRQoL level (Schwartz and Sprangers, 1999). The test involves asking patients post treatment to provide their current levels (post-test) but also their pre-test levels in retrospect (then-test). This method is based on the assumption that patients rate their HRQoL post-test and pre-test levels with the same criteria, since the assessments occur at the same time point. The recalibration component of response shift should thus be taken into account when comparing post-test and then-test scores. Comparing the mean of the pre-test and then-test scores explores recalibration component of response shift.

Structural equations modeling (MacCallum and Austin, 2000) have also been proposed to take into account all the three components of response shift (i.e. recalibration, reprioritization and reconceptualization) but are mainly applied on the MOS SF-36 generic questionnaire (Oort, 2005; Oort et al., 2005). Finally, item response theory (IRT) models (De Ayala, 2013) have also been proposed to take into account both recalibration and reprioritization component of the response shift effect, using the then-test method (Anota et al., 2014b) or the structural equations modeling (Guilleux et al., 2015) as a gold standard to assess the response shift effect, but research is still ongoing on these models.

At this time, few studies have been done on the way to deal with the occurrence of this effect in longitudinal analysis. Some first studies were done on the time to deterioration approach: first depending on the then-test method (Hamidou et al., 2014) and then by choosing an alternative HRQoL score as the reference such as the best previous score (Anota et al., 2013). In this way, it is essential to explore several methods to take into account the response shift effect according to the main statistical methods used for the longitudinal analysis of HRQoL in oncology. Some convergent results could then ensure the validity of the methodology, i.e. to produce some robust results not bias by the response shift effect.

The objective of this paper was to explore several methods to take into account the occurrence of the recalibration component of the response shift effect according to three statistical methods for the longitudinal analysis: the linear mixed model for repeated measures (LMMRM) (Cnaan et al., 1997; Fairclough, 2010), the time to HRQoL score deterioration (TTD) (Bonnetain et al., 2010) and the longitudinal partial credit model (LPCM) based on the IRT approach (Bacci, 2008). All these methods were applied on data from a multicenter prospective cohort study on early breast cancer patients where the then-test method was used as a standard to identify recalibration (Dabakuyo et al., 2013).

2. METHODS

2.1 PATIENTS AND ELIGIBILITY CRITERIA

A prospective, multicenter, randomized cohort study was performed in four French centers. All women initially hospitalized between February 2006 and February 2008 for diagnosis or treatment of primary or suspected breast cancer were eligible for inclusion. Women with cancer other than breast cancer, already undergoing breast cancer treatment, or with a previous history of cancer were excluded. Written informed consent was obtained from every participant and the protocol was approved by Ethics committee. The complete design of this study was extensively described elsewhere (Dabakuyo et al., 2013).

2.2 HEALTH-RELATED QUALITY OF LIFE ASSESSMENT

HRQoL was evaluated using the European Organization of Research and Treatment of Cancer (EORTC) QLQ-C30 cancer specific questionnaire (Aaronson et al., 1993) and its breast cancer module QLQ-BR23 (Sprangers et al., 1996) at four time points: at diagnosis (T_1), at the end of the initial hospitalization (T_2), at three (T_3) and six months (T_4) after the first hospitalization. Moreover, the then-test method was used as a gold standard to assess the recalibration component of the response shift effect. Then, at each follow-up, patients also had to fill out retrospective questionnaires (for both QLQ-C30 and BR23 questionnaires):

- At T_2 and at T_3, patients had to reassess their baseline HRQoL level
- At T_4, patients had to reassess their HRQoL level at three months (T_3).

The 30-item QLQ-C30 questionnaire measures five functional scales (physical, role, emotional, cognitive and social functioning), global health status (GHS), financial difficulties and eight symptom scales (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea) (Aaronson et al., 1993). Each item is constructed on a 4-point Likert scale ("Not at all"/"Quite a

Bit"/"Somewhat"/"Very Much") coded 0 to 3 except the two last items assessing GHS which are on a 0-6 scale.

The 23-item QLQ-BR23 module contains 23 items assessing four functional scales (body image, sexual functioning, sexual enjoyment, future perspectives) and four symptom scales (systemic therapy side effects (STSE), breast symptoms, arm symptoms, upset by hair loss) specific to breast cancer (Sprangers et al., 1996). Each item is constructed on a 4-point Likert scale ("Not at all"/"Quite a Bit"/ "Somewhat"/"Very Much") coded 0 to 3.

Scores for each dimension were calculated if at least half of the items were answered according to the recommendations of the EORTC Scoring Manual (Fayers et al., 1999). These scores vary from 0 (worst) to 100 (best) for the functional dimensions and GHS, and from 0 (best) to 100 (worst) for the symptom dimensions.

2.3 STATISTICAL ANALYSIS

2.3.1 DESCRIPTIVE ANALYSIS AND STUDY POPULATION

Baseline variables were described using means and standard deviations (SD) for continuous variables and percentages for qualitative variables. The number of HRQoL questionnaires completed at each measurement time was reported.

A descriptive analysis of the percentage of item responses according to surgery (mastectomy vs. no mastectomy) was performed at each measurement time.

All included patients with a confirmed breast cancer and with at least one available baseline score and the retrospective measure performed post-surgery were included in the longitudinal analysis, according to the modified intent to treat principle (Fisher et al., 1990).

Missing data profile was already studied and extensively described elsewhere (Anota et al., 2014b).

Five HRQoL dimensions were selected for the analyses based on:

- The number of items
- The number of response categories per item
- The potential impact of the response shift effect on these HRQoL dimensions based on the literature.

In fact, a previous simulation study aiming to compare these three statistical approaches (LMMRM, LPCM and TTD) for the longitudinal HRQoL analysis showed that the number of items and response categories per item influence the statistical power of the test of a treatment by time interaction (Anota et al., 2014a). Moreover, several studies highlighted that the response shift effect may affect the fatigue assessment of patients (Visser et al., 2000), particularly in breast cancer patients (Andrykowski et al., 2009). Hinz et al. showed that the response shift effect has an important impact on the global health status dimension of the EORTC QLQ-C30 questionnaire (Hinz et al., 2012). A previous study also highlighted the occurrence of a recalibration effect on the future perspective dimension of the QLQ-BR23 in breast cancer patients (Dabakuyo et al., 2013).

According to these criteria, we chose to retain global health status (2 items with 7 response categories), fatigue (3 items with 4 response categories), pain (2 items with 4 response categories), body image (4 items with 4 response categories) and future perspective (1 item with 4 response categories) dimensions.

We also expected an impact of response shift effect on both pain and body image dimensions.

A 5-point change/difference in HRQoL score was considered as the minimal clinically important difference (MCID) (Osoba et al., 1998). All p-values < 0.05 were considered as statistical significant. No adjustment on multiple tests was done.

2.3.2 DETECTION OF THE RESPONSE SHIFT EFFECT OCCURRENCE USING THE THEN-TEST METHOD

For each score, the mean difference (MD) between the then-test score obtained at T2 and the baseline score was calculated and described as mean (SD). The existence of a significant recalibration was tested with a Wilcoxon non-parametric test. The effect size was calculated in order to assess the magnitude of the recalibration component of the response shift effect and was defined as the mean change score between the then-test and the corresponding pre-test dividing by the SD of patients at the prospective measurement time.

2.3.3 LONGITUDINAL ANALYSIS

Three statistical methods for the longitudinal analysis were explored and we interested on the impact of response shift effect to detect a surgery effect (mastectomy versus no mastectomy) on each HRQoL score. We focused on this surgery effect because we assumed that a response shift effect could differentially impact patients according to their surgery (mastectomy or not). Thus, according to the methods explored, we analyzed the effect of response shift effect on the determination of the time effect, surgery effect and treatment by time interaction effect. For each statistical method, three analyses were done:

- The first analysis was made considering the retrospective measure of baseline HRQoL level realized post-surgery as the reference (gold standard to take into account the recalibration component of the response shift effect)
- The second analysis was made on prospective measures (crude analysis)
- And finally, one alternative method was proposed based only on the prospective measures to take into account the occurrence of the recalibration component of the response shift effect in conjunction with the statistical method for the longitudinal analysis. The objective is to explore alternative method to take into account the response shift effect without the then-test method

LMMRM: A linear mixed model for repeated measure (LMMRM) was applied on HRQoL scores integrating the four measurement times. Three fixed effects were introduced in the model: a surgery effect (mastectomy versus no mastectomy), a time effect and an interaction between time and surgery. Moreover, some random effects on patient (individual deviance from average intercept) and time (individual deviance from average time effect) were also added to the model. Several structures of the variance-covariance matrix were tested (unstructured, first order autoregressive, heterogeneous first order autoregressive, compound symmetry and heterogeneous compound symmetry). The choice of the best structure was made according to the Akaike Information Criteria (Akaike, 1998).

The model considered can be written as follows:

$$Y_n^{(t)} = \mathbf{c} + \lambda \times \mathbf{a}_n + \gamma \times t + \nu \times \mathbf{a}_n \times t + \mathbf{u}_{0,n} + \mathbf{u}_{1,n} \times t + \varepsilon_n^{(t)}$$

$$(u_{0,n}, u_{1,n}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right)$$

and $\varepsilon_n^{(t)} \sim N(0, \sigma^2)$ independent where:

- $Y_n^{(t)}$ is the score of the patient n at time t,
- $\mathbf{c}$ is a constant,
- $\mathbf{a}_n$ is the surgery status of patient n (equal to 0 if no mastectomy was realized or 1 if a mastectomy was done),
- λ is a fixed surgery effect,

- γ is a fixed time effect,
- ν is a fixed interaction effect between treatment arm and time,
- $\mathbf{u}_{0,n}$ is a random intercept on patient n,
- $\mathbf{u}_{1,n}$ are random slopes for time,
- Σ is the covariance matrix of random effects $(u_{0,n}, u_{1,n})$
- $\varepsilon_n^{(t)}$ is the residual of patient n at time t
- σ^2 is the residual variance.

Estimation of the parameters was done using a maximum likelihood method based on the Newton-Raphson algorithm.

For this statistical model, the alternative method proposed to take into account the occurrence of the response shift effect was to adjust on the prospective score obtained post-surgery.

Moreover, if the surgery effect was not significant in each of the three analyses, all analyses were repeated without the surgery effect.

We reported estimate, standard error (SE) and *p*-value of Wald test for each effect. For each analysis, a positive (respectively negative) estimated value of the regression coefficients means:

- for surgery effect: patients with a mastectomy presented a higher score (respectively lower score) on average than other patients whatever the measurement time;
- regarding time effect: GHS, functional or symptomatic level increase (or decrease) over time;
- regarding interaction effect:patients with a mastectomy presented an increase of (a decrease of) GHS, functional or symptomatic level over time as compared to other patients.

LPCM: A longitudinal partial credit model (LPCM) was applied on items of each HRQoL dimension integrating the four measurement times and integrating the same effects as in the LMMRM. This model was completely described in previous publication (Anota et al., 2014a).

The model considered can be written as follows:

$$P\left(X_{n,j,t}=k|\theta_n^{(t)},\delta_{j,i},...,\delta_{j,m_j}\right)=\frac{\exp(k\theta_n^{(t)}-\sum_{i=1}^{k}\delta_{j,i})}{\sum_{h=1}^{m_j}\exp(h\theta_n^{(t)}-\sum_{i=1}^{h}\delta_{j,i})}$$

$$\theta_n^{(t)}=\lambda\times\mathbf{a}_n+\gamma\times t+\nu\times\mathbf{a}_n\times t+\mathbf{u}_{0,n}+\mathbf{u}_{1,n}\times t$$

$$(u_{0,n},u_{1,n})\;N\left(\begin{pmatrix}0\\0\end{pmatrix},\Sigma\right)$$

where:

- $X_{n,j,t}$ is the answer of the patient n to the item j at time t,
- $\theta_n^{(t)}$ is the latent trait of the patient n at time t,
- $\delta_{j,i},...,\delta_{j,m_j}$ are item difficulty parameters for item j with m_j+1 possible responses for the item.

The alternative method explored to take into account the occurrence of the recalibration component of the response shift effect only using prospective measures was to consider a change $\pi_{j,i}$ in each item response category parameter at T_2 which remained constant for following measurement times; thus $\delta_{j,i}+\pi_{j,i}$ was the new item difficulty parameter for item j and response category i since T_2.

If the *p*-value associated with trend parameters $\pi_{j,i}$ was significant, then a significant recalibration was highlighted.

Moreover, if the surgery effect was not significant in each of the three analyses, all analyses were repeated without the surgery effect.

We reported estimate, SE and *p*-value of Wald test for each effect. For each analysis, a positive (respectively negative) estimate means that:

- Patients with a mastectomy chose higher response categories for items as compared to other patients regarding mastectomy effect;
- Patients chose higher response categories over time for time effect;
- Patients with a mastectomy chose higher response categories over time as compared to patients with no mastectomy for interaction effect.

High response categories reflect a high GHS level, a low functional level and a high symptomatic level, except for sexual functional and sexual enjoyment for which a high response category means a high sexual level. Thus, interpretation of estimates for LMMRM and IRT are opposite for functional scales (except for sexual scales): for example, regarding time effect, a positive estimate corresponds to an improvement of the functional level over time while it means a deterioration of the functional scale for IRT (higher response categories over time).

TTD: The time to HRQoL score deterioration (TTD) approach belongs to timeto-event models and requires a definition of the event, i.e. the deterioration. Events can be defined according to the chosen reference score, MCID, missing scores, including all-cause death or not. In that study, we defined TTD as the time from inclusion in the study to the first deterioration with a 5-point MCID as compared to the reference score (Hamidou et al., 2011). This definition is recommended for adjuvant setting (Anota et al., 2013).

Since a high score corresponds to a high level of functioning on a functional scale, but corresponds to a strong presence of symptoms for a symptomatic scale, “deterioration” was defined as a decrease on the functional scale or GHS, and as an increase on the symptomatic scale.

Intermittent missing data were ignored, considering that the patient’s HRQoL level remained unchanged since the last available HRQoL assessment.

The alternative method proposed to take into account the occurrence of the response shift effect was to consider the best previous score as the reference score.

The TTD curves were estimated according to the Kaplan-Meier method and described using median with 95% confidence interval (95%CI). The TTD curves were compared according to the surgery type (mastectomy versus no mastectomy) using the log-rank test. Univariate Cox regression model was performed to estimate Hazard Ratio (HR) with 95%CI (Sasieni, 2005).

For each statistical method, results of each analysis will be compared to those obtained for the reference analysis using the then-test method in terms of *p*-value and direction of the effect.

All analyses were done on R software (version 3.2.1) (Team, 2014) using QoLR package and SAS software (version 9.4, SAS Institute Inc, Cary, NC).

3. RESULTS

3.1. POPULATION

Between February 2006 and February 2008, 381 patients were included in the four participating centers. Mean age was 58.4 (SD = 11) years. Three hundred and thirty-seven (88.5%) patients had confirmed breast cancer and thus were retained in this present study. Figure 1 flowchart represents the number of HRQoL questionnaires (prospective and retrospective) completed at each measurement time.

The number and percentage of response category per item at each measurement time according to surgery is given in Table 1.

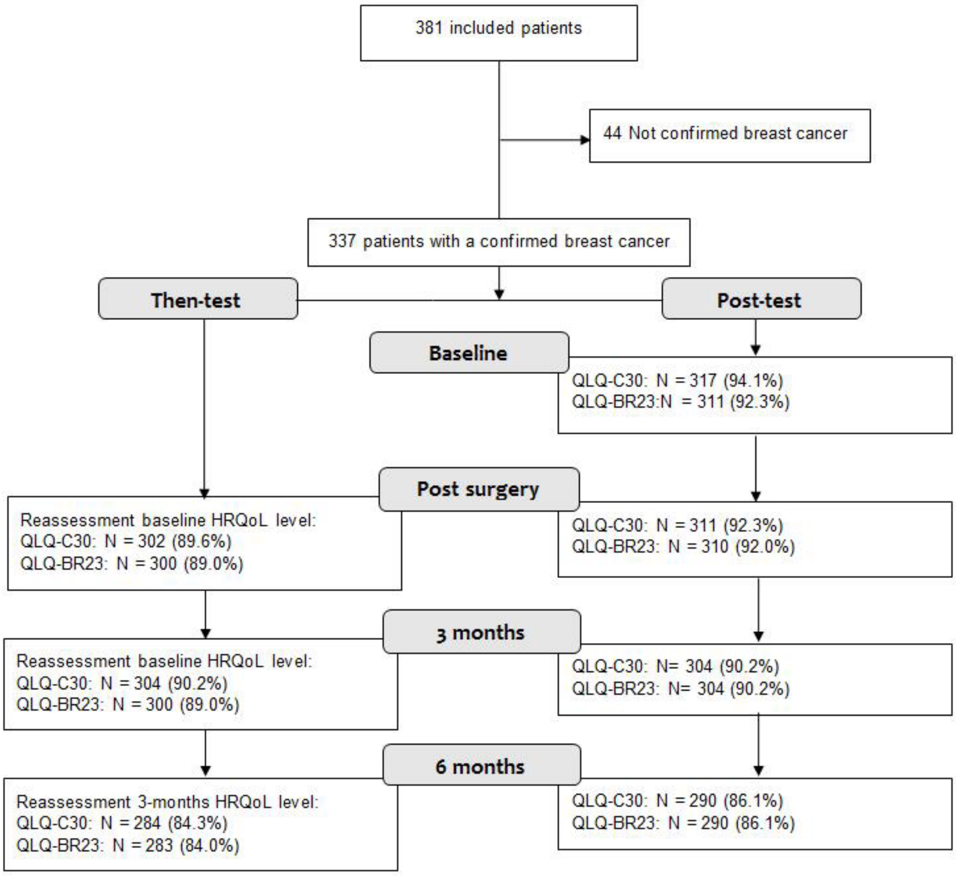

Figure 1: Flowchart of the number of questionnaires received at each measurement time

Table 1: Number and percentage of response category per item at each measurement time according to surgery

	Time 1			Time 1 retrospective			Time 2			Time 3			Time 4		
	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing
Pain															
Q9															
1	74 (61.2)	134 (64.5)	4 (50.0)	72 (59.5)	124 (59.6)	3 (37.5)	46 (38.1)	78 (37.5)	1 (12.5)	57 (47.1)	67 (32.2)	2 (25.0)	42 (34.7)	67 (32.2)	3 (37.5)
2	27 (22.3)	45 (21.6)	3 (37.5)	26 (21.5)	40 (19.2)	3 (37.5)	49 (40.5)	76 (36.5)	4 (50.0)	38 (31.4)	86 (41.4)	4 (50.0)	49 (40.5)	69 (33.2)	0
3	7 (5.8)	7 (3.4)	1 (12.5)	11 (9.1)	7 (3.4)	1 (12.5)	18 (14.9)	17 (8.2)	1 (12.5)	12 (9.9)	26 (12.5)	0	16 (13.2)	30 (14.4)	0
4	5 (4.1)	3 (1.4)	0	2 (1.7)	8 (3.9)	0	4 (3.3)	6 (2.9)	1 (12.5)	5 (4.1)	3 (1.4)	0	3 (2.5)	4 (14.4)	0
missing	8 (6.6)	19 (9.1)	0	10 (8.3)	29 (13.9)	1 (12.5)	4 (3.3)	31 (14.9)	1 (12.5)	9 (7.4)	26 (12.5)	2 (25.0)	11 (9.01)	38 (18.3)	5 (62.5)
Q19															
1	76 (62.8)	150 (72.1)	5 (62.5)	77 (63.6)	133 (63.9)	6 (75.0)	47 (38.8)	97 (46.6)	2 (25.0)	63 (52.1)	85 (40.9)	3 (37.5)	59 (48.8)	86 (41.4)	2 (25.0)
2	27 (22.3)	33 (15.9)	3 (37.5)	21 (17.4)	25 (12.0)	0	42 (34.7)	53 (25.5)	3 (37.5)	27 (22.3)	63 (30.3)	3 (37.5)	30 (24.8)	54 (26.0)	1 (12.5)
3	3 (2.5)	5 (2.4)	0	8 (6.6)	13 (6.3)	1 (12.5)	18 (14.9)	19 (9.1)	1 (12.5)	13 (10.7)	23 (30.3)	0	18 (14.9)	24 (11.5)	0
4	4 (3.3)	3 (1.4)	0	3 (2.5)	3 (1.4)	0	5 (4.1)	5 (2.4)	1 (12.5)	8 (6.6)	23 (11.1)	0	4 (3.3)	6 (2.9)	0
missing	11 (9.1)	17 (8.2)	0	12 (9.9)	34 (16.4)	1 (12.5)	9 (7.4)	34 (16.5)	1 (12.5)	10 (8.3)	11 (5.3)	2 (25.0)	10 (8.3)	38 (18.3)	5 (62.5)
Fatigue															
Q10															
1	51 (42.2)	93 (44.7)	5 (62.5)	56 (46.3)	91 (43.8)	4 (50.0)	25 (20.7)	49 (23.6)	1 (12.5)	19 (15.7)	27 (13.0)	1 (12.5)	22 (18.2)	39 (18.8)	1 (12.5)
2	45 (37.2)	72 (34.6)	3 (37.5)	36 (29.8)	64 (30.8)	1 (12.5)	56 (46.3)	88 (42.3)	3 (37.5)	60 (49.6)	91 (43.8)	3 (37.5)	56 (46.3)	84 (40.4)	2 (25.0)
3	8 (6.6)	15 (7.2)	0	12 (9.9)	17 (8.2)	2 (25.0)	22 (18.2)	31 (14.9)	2 (25.0)	19 (15.7)	42 (20.2)	1 (12.5)	22 (18.2)	39 (18.8)	0
4	4 (3.3)	7 (3.4)	0	4 (3.3)	6 (2.9)	0	11 (9.1)	9 (4.3)	1 (12.5)	12 (9.9)	20 (9.6)	0	9 (7.4)	10 (4.8)	0
missing	13 (10.7)	21 (10.1)	0	13 (10.7)	30 (14.4)	1 (12.5)	7 (5.8)	31 (14.9)	1 (12.5)	11 (9.1)	28 (13.5)	3 (37.5)	12 (9.9)	36 (17.3)	5 (62.5)
Q12															
1	63 (52.1)	103 (49.5)	5 (62.5)	63 (52.1)	103 (49.5)	4 (50.0)	35 (28.9)	70 (33.7)	2 (25.0)	35 (28.9)	46 (22.1)	1 (12.5)	38 (31.4)	66 (31.7)	2 (25.0)
2	38 (31.4)	69 (33.2)	3 (37.5)	29 (24.0)	60 (28.9)	3 (37.5)	58 (47.9)	84 (40.4)	5 (62.5)	52 (43.0)	82 (39.4)	5 (62.5)	51 (42.2)	71 (34.1)	1 (12.5)
3	7 (5.8)	11 (5.3)	0	14 (11.6)	10 (4.8)	0	17 (14.1)	15 (7.2)	0	20 (16.5)	39 (18.8)	0	11 (9.1)	25 (12.0)	0
4	3 (2.5)	6 (2.9)	0	2 (1.7)	6 (2.9)	0	5 (4.1)	8 (3.9)	0	5 (4.1)	15 (7.2)	0	9 (7.4)	10 (4.8)	0
missing	10 (8.3)	19 (9.1)	0	13 (10.7)	29 (13.9)	1 (12.5)	6 (5.0)	31 (14.9)	1 (12.5)	9 (7.4)	26 (12.5)	2 (25.0)	12 (9.9)	36 (17.3)	5 (62.5)
Q18															
1	40 (33.1)	73 (35.1)	0	42 (34.7)	80 (38.5)	3 (37.5)	19 (15.7)	50 (24.0)	1 (12.5)	17 (14.1)	25 (12.0)	1 (12.5)	19 (15.7)	36 (17.3)	1 (12.5)
2	51 (42.2)	95 (45.7)	7 (87.5)	49 (40.5)	74 (35.6)	3 (37.5)	64 (52.9)	95 (45.7)	1 (12.5)	57 (47.1)	85 (40.9)	4 (50.0)	60 (49.6)	90 (43.3)	2 (25.0)
3	16 (13.2)	16 (7.7)	1 (12.5)	17 (14.1)	14 (6.7)	1 (12.5)	24 (19.8)	29 (13.9)	4 (50.0)	25 (20.7)	53 (25.5)	1 (12.5)	25 (20.7)	33 (15.9)	0
4	5 (4.1)	7 (3.4)	0	3 (2.5)	8 (3.9)	0	9 (7.4)	5 (2.4)	2 (25.0)	14 (11.6)	15 (7.2)	0	6 (5.0)	10 (4.8)	0
missing	9 (7.4)	17 (8.2)	0	10 (8.3)	32 (15.4)	1 (12.5)	5 (4.1)	29 (13.9)	0	8 (6.6)	30 (14.4)	2 (25.0)	11 (9.1)	39 (18.8)	5 (62.5)

Table 1: continued

	Time 1			Time 1 retrospective			Time 2			Time 3			Time 4		
	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing
GHS															
Q29															
1	3 (2.5)	0	0	0	1 (0.5)	0	1 (0.8)	2 (1.0)	0	1 (0.8)	0	0	1 (0.8)	0	0
2	0	3 (1.4)	1 (12.5)	4 (3.3)	3 (1.4)	0	3 (2.5)	2 (1.0)	0	5 (4.1)	5 (2.4)	0	1 (0.8)	2 (1.0)	0
3	10 (8.3)	10 (4.8)	1 (12.5)	14 (11.6)	10 (4.8)	1 (12.5)	15 (12.4)	17 (8.2)	1 (12.5)	15 (12.4)	30 (14.4)	1 (12.5)	10 (8.3)	16 (7.7)	0
4	19 (15.7)	45 (21.6)	1 (12.5)	19 (15.7)	41 (19.7)	0	34 (28.1)	51 (24.5)	0	27 (22.3)	47 (22.6)	0	25 (20.7)	37 (17.8)	0
5	32 (26.5)	47 (22.6)	3 (37.5)	30 (24.8)	45 (21.6)	3 (37.5)	30 (24.8)	51 (24.5)	4 (50.0)	35 (28.9)	63 (30.3)	3 (37.5)	41 (33.9)	59 (28.4)	1 (12.5)
6	36 (29.8)	57 (27.4)	2 (25.0)	32 (26.5)	56 (21.6)	3 (37.5)	27 (22.3)	44 (21.2)	2 (25.0)	21 (17.4)	27 (13.0)	2 (25.0)	25 (20.7)	42 (20.2)	2 (25.0)
7	12 (9.9)	28 (13.5)	0	11 (9.1)	22 (10.6)	0	5 (4.1)	11 (5.3)	0	8 (6.6)	8 (3.9)	0	8 (6.6)	16 (7.7)	0
missing	9 (7.4)	18 (8.7)	0	11 (9.1)	30 (14.4)	1 (12.5)	6 (5.0)	30 (14.4)	1 (12.5)	9 (7.4)	28 (13.5)	2 (25.0)	10 (8.3)	36 (17.3)	5 (62.5)
Q30															
1	2 (1.7)	2 (1.0)	0	0	0	0	0	1 (0.5)	0	1 (0.8)	0	0	0	0	0
2	2 (1.7)	4 (1.9)	1 (12.5)	3 (2.5)	6 (2.9)	0	7 (5.8)	7 (3.4)	0	6 (5.0)	9 (4.3)	0	3 (2.5)	4 (1.9)	0
3	11 (9.1)	12 (5.8)	2 (25.0)	13 (10.7)	8 (3.9)	1 (12.5)	11 (9.1)	19 (9.1)	1 (12.5)	10 (8.3)	28 (13.5)	1 (12.5)	8 (6.6)	15 (7.2)	0
4	13 (10.7)	35 (16.8)	1 (12.5)	20 (10.7)	43 (20.7)	1 (12.5)	29 (24.0)	49 (23.6)	0	30 (24.8)	45 (21.6)	0	25 (20.7)	35 (16.8)	0
5	32 (26.5)	53 (25.5)	2 (25.0)	31 (25.6)	42 (20.2)	2 (25.0)	32 (26.5)	47 (22.6)	4 (50.0)	36 (29.8)	59 (28.4)	3 (37.5)	39 (32.3)	55 (26.4)	2 (25.0)
6	39 (32.2)	51 (24.5)	2 (25.0)	28 (23.1)	57 (27.4)	2 (25.0)	27 (22.3)	45 (21.6)	2 (25.0)	22 (18.2)	30 (14.4)	2 (25.0)	27 (22.3)	44 (21.2)	1 (12.5)
7	10 (8.3)	32 (15.4)	0	14 (11.6)	20 (9.6)	1 (12.5)	7 (5.8)	11 (5.3)	0	8 (6.6)	9 (4.3)	0	9 (7.4)	19 (9.1)	0
missing	9 (7.4)	19 (9.1)	0	12 (9.9)	32 (15.4)	1 (12.5)	8 (6.6)	29 (13.9)	1 (12.5)	8 (6.6)	28 (13.5)	2 (25.0)	10 (8.3)	36 (17.3)	5 (62.5)
BI[1]															
Q9															
1	91 (75.2)	140 (67.3)	3 (37.5)	82 (67.8)	137 (65.9)	4 (50.0)	63 (52.1)	113 (54.3)	4 (50.0)	49 (40.5)	85 (40.9)	2 (25.0)	47 (38.9)	94 (45.2)	0
2	14 (11.6)	31 (14.9)	2 (25.0)	20 (16.5)	27 (13.0)	1 (12.5)	30 (24.8)	40 (19.2)	3 (37.5)	34 (28.1)	51 (24.5)	2 (25.0)	33 (27.3)	45 (21.6)	3 (37.5)
3	5 (4.1)	5 (2.4)	1 (12.5)	2 (1.7)	6 (2.9)	1 (12.5)	12 (9.9)	12 (5.8)	0	10 (8.3)	25 (12.0)	1 (12.5)	15 (12.4)	23 (11.1)	0
4	0	1 (0.5)	0	1 (0.8)	0	0	5 (4.1)	5 (2.4)	0	16 (13.2)	16 (7.7)	0	16 (13.2)	8 (3.9)	0
missing	11 (9.1)	31 (14.9)	2 (25.0)	16 (13.2)	38 (18.3)	2 (25.0)	11 (9.1)	38 (18.3)	1 (12.5)	12 (9.9)	31 (14.9)	3 (37.5)	10 (8.3)	38 (18.3)	5 (62.5)

[1] Body image

Table 1: continued

	Time 1			Time 1 retrospective			Time 2			Time 3			Time 4		
	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing	no mastectomy	mastectomy	missing
Q10															
1	93 (76.9)	142 (68.3)	3 (37.5)	86 (71.1)	137 (65.9)	3 (37.5)	69 (57.0)	115 (55.3)	3 (37.5)	52 (43.0)	92 (44.2)	1 (12.5)	50 (41.3)	95 (45.7)	0
2	11 (9.1)	25 (12.0)	2 (25.0)	16 (13.2)	25 (12.0)	2 (25.0)	29 (24.0)	36 (17.3)	3 (37.5)	31 (25.6)	51 (24.5)	5 (62.5)	30 (24.8)	42 (20.2)	3 (37.5)
3	5 (4.1)	9 (4.3)	1 (12.5)	1 (0.8)	7 (3.4)	1 (12.5)	7 (5.8)	15 (7.2)	0	12 (9.9)	22 (10.6)	0	14 (11.6)	22 (10.6)	0
4	1 (0.8)	1 (0.5)	0	1 (0.8)	0	0	8 (6.6)	3 (1.4)	0	17 (14.1)	15 (7.2)	0	16 (13.2)	11 (5.3)	0
missing	11 (9.1)	31 (14.9)	2 (25.0)	17 (14.1)	39 (18.8)	2 (25.0)	8 (6.6)	39 (18.8)	2 (25.0)	9 (7.4)	28 (13.5)	2 (25.0)	11 (9.1)	38 (18.3)	5 (62.5)
Q11															
1	84 (69.4)	137 (65.9)	4 (50.0)	76 (62.8)	125 (60.1)	4 (50.0)	54 (44.6)	109 (52.4)	3 (37.5)	46 (38.0)	99 (47.6)	3 (37.5)	47 (38.8)	97 (46.6)	1 (12.5)
2	19 (15.7)	27 (13.0)	0	20 (16.5)	30 (14.4)	1 (12.5)	35 (28.9)	40 (19.2)	1 (12.5)	30 (24.8)	48 (23.1)	0	30 (24.8)	47 (22.6)	1 (12.5)
3	2 (1.7)	9 (13.0)	1 (12.5)	5 (4.1)	11 (5.3)	0	11 (9.1)	17 (8.2)	2 (25.0)	18 (14.9)	23 (11.1)	1 (12.5)	16 (13.2)	18 (8.7)	1 (12.5)
4	3 (2.5)	5 (2.4)	1 (12.5)	3 (2.5)	5 (2.4)	1 (12.5)	10 (8.3)	5 (2.4)	1 (12.5)	18 (14.9)	13 (6.3)	2 (25.0)	17 (14.1)	8 (3.9)	0
missing	13 (10.7)	30 (14.4)	2 (25.0)	17 (14.1)	37 (17.8)	2 (25.0)	11 (9.1)	37 (17.8)	1 (12.5)	9 (7.4)	25 (12.0)	2 (25.0)	11 (9.1)	38 (18.3)	5 (62.5)
Q12															
1	83 (68.6)	126 (60.6)	4 (50.0)	80 (66.1)	121 (58.2)	4 (50.0)	55 (45.5)	105 (50.5)	3 (37.5)	37 (30.6)	94 (45.2)	3 (37.5)	35 (28.9)	90 (43.3)	1 (12.5)
2	17 (14.1)	34 (16.4)	0	18 (14.9)	26 (12.5)	1 (12.5)	30 (24.8)	43 (20.7)	1 (12.5)	37 (30.6)	49 (23.6)	0	35 (28.9)	48 (23.1)	1 (12.5)
3	2 (1.7)	12 (5.6)	1 (12.5)	2 (1.7)	14 (6.7)	0	10 (8.3)	13 (6.3)	1 (12.5)	16 (13.2)	16 (7.7)	2 (25.0)	22 (18.2)	20 (9.6)	1 (12.5)
4	2 (1.7)	4 (1.9)	1 (12.5)	3 (2.5)	5 (2.4)	1 (12.5)	9 (7.4)	8 (3.9)	1 (12.5)	19 (15.7)	20 (9.6)	1 (12.5)	17 (14.1)	12 (5.8)	0
missing	17 (14.1)	32 (15.4)	2 (25.0)	18 (14.9)	42 (20.2)	2 (25.0)	17 (14.1)	39 (18.8)	2 (25.0)	12 (9.9)	29 (13.9)	2 (25.0)	12 (9.9)	38 (18.3)	5 (62.5)
FP[1]															
Q13															
1	10 (8.3)	19 (9.1)	1 (12.5)	28 (23.1)	26 (12.5)	2 (25.0)	10 (8.3)	29 (14.0)	0	23 (19.0)	36 (17.3)	1 (12.5)	21 (17.4)	43 (20.7)	0
2	51 (42.2)	73 (35.1)	1 (12.5)	38 (31.4)	81 (38.9)	1 (12.5)	57 (47.1)	82 (39.4)	2 (25.0)	45 (37.2)	75 (36.1)	1 (12.5)	46 (38.0)	80 (38.5)	1 (12.5)
3	22 (18.2)	58 (27.9)	0	28 (23.1)	35 (16.8)	2 (25.0)	29 (24.0)	41 (19.7)	2 (25.0)	22 (18.2)	45 (21.6)	3 (37.5)	30 (24.8)	31 (14.9)	0
4	24 (19.8)	32 (15.4)	4 (50.0)	15 (12.4)	29 (13.9)	2 (25.0)	19 (15.7)	24 (11.5)	3 (37.5)	23 (19.0)	26 (12.5)	1 (12.5)	15 (12.4)	21 (10.1)	2 (25.0)
missing	14 (11.6)	26 (12.5)	2 (25.0)	12 (9.9)	37 (17.8)	1 (12.5)	6 (5.0)	32 (15.4)	1 (12.5)	8 (6.6)	26 (12.5)	2 (25.0)	9 (7.4)	33 (15.9)	5 (62.5)

[1] Future perspectives

3.2 DETECTION OF THE RESPONSE SHIFT EFFECT

After surgery (Table 2), the recalibration effect was statistically significant for both fatigue (*p*-value=0.03) and future perspective dimensions (*p*-value < 0.01) but only clinically significant for future perspectives (MD = 7.4).

Table 2: Recalibration component of response shift effect assessed with the then-test method after surgery among patients with a confirmed breast cancer and with both baseline score and then-test score post-surgery

	Baseline HRQoL		Post-surgery HRQoL		Post-surgery then-test		Post-surgery then-test minus baseline score		
Scores	**N**	**mean (SD)**	**N**	**mean (SD)**	**N**	**mean (SD)**	**mean (SD)**	**P**	**effect size**
GHS	280	69.1 (19.6)	274	63.1 (20.0)	280	68.3 (20.0)	-0.8 (16.7)	0.55	-0.04
Fatigue	278	22.8 (23.2)	270	32.4 (24.1)	278	21.4 (23.6)	-1.5 (18.2)	0.03	-0.06
Pain	285	13.3 (21.6)	281	25.6 (25.1)	285	13.9 (23.0)	0.5 (19.0)	0.90	0.02
Body image	262	90.8 (16.8)	251	81.7 (24.7)	262	90.0 (17.5)	-0.8 (11.9)	0.67	-0.05
Future perspectives	261	48.0 (30.7)	251	53.7 (29.6)	261	55.4 (31.9)	7.4 (30.6)	<0.01	0.24

3.3 LONGITUDINAL ANALYSIS

3.3.1 LMMRM

All results obtained for the LMMRM analyses are summarized in Table 3.

Table 3: Results of the linear mixed model for repeated measure taking into account the recalibration component of the response shift effect

			then-test post-surgery as the reference measure			prospective measures			adjusting on the score obtained post-surgery		
Dimension	**N**	**Effect**	**Estimate**	**SE**	**P**	**Estimate**	**SE**	**P**	**Estimate**	**SE**	**P**
GHS	273	time	-1.06	0.58	0.07	-1.28	0.55	0.02	-1.30	0.55	0.02
		interaction	0.44	0.78	0.58	0.25	0.76	0.75	0.42	0.72	0.56
Fatigue	271	time	5.24	0.71	<0.01	4.72	0.71	<0.01	5.21	0.72	<0.01
		interaction	-0.63	1.03	0.54	-0.54	1.03	0.60	-1.61	0.98	0.10
Pain	278	surgery	4.28	2.74	0.12	5.46	2.55	0.03	1.70	1.77	0.34
		time	4.55	0.69	<0.01	4.93	0.65	<0.01	4.73	0.66	<0.01
		interaction	-1.87	1.11	0.09	-2.43	1.04	0.02	-2.18	1.05	0.04
Body image	257	time	-4.15	0.80	<0.01	-4.24	0.82	<0.01	-5.13	0.84	<0.01
		interaction	-4.42	1.23	<0.01	-4.80	1.26	<0.01	-2.73	1.15	0.02
FP	256	time	2.59	0.93	<0.01	4.75	0.84	<0.01	4.53	0.84	<0.01
		interaction	-2.07	1.29	0.11	-1.85	1.23	0.14	-0.69	1.11	0.54

SE: standard Error GHS: Global Health Status; A heterogeneous first order autoregressive structure for the covariance matrix was retained for each analysis and each score except for nausea and vomiting for which an unstructured structure was retained based on the AIC criteria. Results which are not consistent with those obtained using the then-test method are highlighted in grey in the Table

Using the then-test obtained post-surgery as the reference measure (reference analysis):

- a significant time effect was highlighted for fatigue (estimate = 5.24), pain (estimate = 4.55), body image (estimate = -4.15) and future perspective (estimate = 2.59) dimensions (*p*-values < 0.01) and
- a significant interaction between time and surgery was only highlighted for body image dimension (estimate = -4.43; *p*-value < 0.01). Results obtained using the prospective measures (crude analysis) were consistent to these results using the then-test method except that:
- a significant surgery effect (mastectomy vs. no mastectomy) was highlighted for pain dimension (estimate = 5.46, *p*-value = 0.03) while this effect was not significant in the reference analysis (estimate = 4.28, *p*-value = 0.11);
- a significant time effect was highlighted for GHS (estimate = -1.28, *p*-value = 0.02) while this trend was not significant in the reference analysis (estimate = -1.06, *p*-value = 0.07) and
- a significant interaction effect between pain and time was highlighted (estimate = -2.43, *p*-value = 0.02) while this effect was not significant in the reference analysis (estimate = -1.87, *p*-value = 0.09).

Regarding the results obtained adjusting on the score obtained post-surgery (alternative method to take into account the RS effect), the same divergence was observed as compared to the results obtained with the then-test method except that the surgery effect of pain was not statistically significant (estimate = 1.70, *p*-value = 0.34).

3.3.2 LPCM

Results of the LPCM analyses are presented in Table 4 for fixed effects and in Table 5 for item difficulty parameters and trends.

Table 4: Results of the linear partial credit model taking into account the recalibration component of the response shift effect

			then-test post-surgery as the reference measure			prospective measures			adjusting on the score obtained post-surgery		
Dimension	**N**	**Effect**	**Estimate**	**SE**	**P**	**Estimate**	**SE**	**P**	**Estimate**	**SE**	**P**
GHS	273	time	-0.16	0.06	0.09	-0.15	0.07	0.04	0.18	0.07	0.01
		interaction	0.04	0.08	0.63	0.06	0.09	0.52	0.04	0.09	0.66
Fatigue	271	time	0.49	0.07	<0.01	0.49	0.07	<0.01	0.11	0.09	0.20
		interaction	-0.03	0.09	0.73	-0.04	0.10	0.73	-0.03	0.11	0.75
Pain	278	mastectomy	0.48	0.29	0.10	0.57	0.26	0.03	0.64	0.28	0.03
		time	0.48	0.07	<0.01	0.46	0.07	<0.01	0.09	0.08	0.31
		interaction	-0.21	0.10	0.04	-0.25	0.09	<0.01	-0.29	0.11	<0.01
Body image	257	time	0.42	0.09	<0.01	0.34	0.09	<0.01	-0.02	0.11	0.83
		interaction	0.40	0.12	<0.01	0.46	0.12	<0.01	0.50	0.13	<0.01
Future perspectives	256	time	-0.25	0.09	<0.01	-0.43	0.08	<0.01	-0.37	0.11	<0.01
		interaction	0.19	0.12	0.11	0.16	0.11	0.13	0.16	0.11	0.13

SE: standard Error GHS: Global Health Status; Results which are not consistent with those obtained using the then-test method are highlighted in grey in the Table

Table 5: Item difficulty and trend parameters obtained for each longitudinal partial credit model

		Item difficulty parameter δ estimates (then-test method)						Item difficulty parameter δ estimates (crude analysis)						Item difficulty δ and trend π parameters estimates (alternative method)					
		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
GHS																			
Item 29	δ	-5.61	-4.72	-2.72	-0.96	0.71	3.19	-4.70	-4.54	-2.57	-0.93	0.60	2.81	-6.10	-5.27	-3.48	-1.29	0.15	2.91
	π													2.15	1.32	1.60*	1.02*	1.33*	0.86*
Item 30	δ	-7.42	-3.89	-2.76	-0.98	0.63	3.10	-5.65	-3.55	-2.50	-1.02	0.53	2.67	-5.06	-4.10	-2.86	-1.55	0.23	2.72
	π													-0.82	1.10*	0.93*	1.23*	1.11*	0.91*
Fatigue																			
Item 10	δ	-0.78	2.75	4.01				-0.77	2.75	4.01				-0.09	3.66	4.25			
	π													-1.77*	-1.73*	-0.91			
Item 12	δ	0.11	3.23	4.34				0.11	3.23	4.34				0.59	3.72	4.32			
	π													-1.39*	-1.23*	-0.60			
Item 18	δ	-1.09	2.63	4.43				-1.09	2.63	4.43				-0.76	3.12	4.38			
	π													-1.28*	-1.23*	-0.56			
Pain																			
Item 9	δ	0.74	3.20	4.83				0.71	3.10	4.66				1.46	3.63	3.93			
	π													-1.77*	-1.21*	0.42			
Item 19	δ	1.40	2.97	4.58				1.33	2.96	4.26				1.91	4.18	3.68			
	π													-1.50*	-1.98*	0.15			
Body image																			
item 9	δ	1.97	3.91	4.75				1.82	3.67	4.50				2.66	4.88	6.44			
	π													-1.65	-1.86	-2.59			
item 10	δ	2.18	4.03	4.56				2.03	3.70	4.31				2.93	4.32	5.85			
	π													-1.72	-1.18	-2.19			
item 11	δ	1.82	3.64	4.49				1.74	3.46	4.17				2.37	4.35	4.02			
	π													-1.39	-1.50	-0.33			
item 12	δ	1.67	3.65	4.12				1.55	3.46	3.87				2.31	4.04	4.75			
	π													-1.57	-1.17	-1.48			
Future perspectives																			
Item 13	δ	-2.53	0.91	2.36				-2.83	0.44	1.69				-2.89	0.40	1.65			
	π													0.19	0.16	0.17			

*: significant trend parameter at the statistical level of 5%

$\delta_{(j,i)}$ is the item difficulty parameter of item i and modality j at time T_1. Change $\pi(j,i)$ in each item response category parameter at T2 remained constant for following measurement times; thus $\delta_{(j,i)} + \pi(j,i)$ was the new item difficulty parameter for item j and response category i since T_2

Table 5: Item difficulty and trend parameters obtained for each longitudinal partial credit model

		Item difficulty parameter δ estimates (then-test method)						Item difficulty parameter δ estimates (crude analysis)						Item difficulty δ and trend π parameters estimates (alternative method)					
		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
GHS																			
Item 29	δ	-5.61	-4.72	-2.72	-0.96	0.71	3.19	-4.70	-4.54	-2.57	-0.93	0.60	2.81	-6.10	-5.27	-3.48	-1.29	0.15	2.91
	π													2.15	1.32	1.60*	1.02*	1.33*	0.86*
Item 30	δ	-7.42	-3.89	-2.76	-0.98	0.63	3.10	-5.65	-3.55	-2.50	-1.02	0.53	2.67	-5.06	-4.10	-2.86	-1.55	0.23	2.72
	π													-0.82	1.10*	0.93*	1.23*	1.11*	0.91*
Fatigue																			
Item 10	δ	-0.78	2.75	4.01				-0.77	2.75	4.01				-0.09	3.66	4.25			
	π													-1.77*	-1.73*	-0.91			
Item 12	δ	0.11	3.23	4.34				0.11	3.23	4.34				0.59	3.72	4.32			
	π													-1.39*	-1.23*	-0.60			
Item 18	δ	-1.09	2.63	4.43				-1.09	2.63	4.43				-0.76	3.12	4.38			
	π													-1.28*	-1.23*	-0.56			
Pain																			
Item 9	δ	0.74	3.20	4.83				0.71	3.10	4.66				1.46	3.63	3.93			
	π													-1.77*	-1.21*	0.42			
Item 19	δ	1.40	2.97	4.58				1.33	2.96	4.26				1.91	4.18	3.68			
	π													-1.50*	-1.98*	0.15			
Body image																			
item 9	δ	1.97	3.91	4.75				1.82	3.67	4.50				2.66	4.88	6.44			
	π													-1.65	-1.86	-2.59			
item 10	δ	2.18	4.03	4.56				2.03	3.70	4.31				2.93	4.32	5.85			
	π													-1.72	-1.18	-2.19			
item 11	δ	1.82	3.64	4.49				1.74	3.46	4.17				2.37	4.35	4.02			
	π													-1.39	-1.50	-0.33			
item 12	δ	1.67	3.65	4.12				1.55	3.46	3.87				2.31	4.04	4.75			
	π													-1.57	-1.17	-1.48			
Future perspectives																			
Item 13	δ	-2.53	0.91	2.36				-2.83	0.44	1.69				-2.89	0.40	1.65			
	π													0.19	0.16	0.17			

*: significant trend parameter at the statistical level of 5%

$\delta_{(j,i)}$ is the item difficulty parameter of item i and modality j at time T_1. Change $\pi(j,i)$ in each item response category parameter at T2 remained constant for following measurement times; thus $\delta_{(j,i)} + \pi(j,i)$ was the new item difficulty parameter for item j and response category i since T_2

Using the then-test obtained post-surgery as the reference measure (reference analysis):

- a significant time effect was highlighted for fatigue (estimate = 0.49), pain (estimate = 0.48), body image (estimate = 0.42) and future perspectives (estimate = -0.25) dimensions (*p*-value <0.01);
- a significant interaction between time and surgery was highlighted for both pain (estimate = -0.21; *p*-value = 0.04) and body image dimensions (estimate = 0.40, *p*-value <0.01). Results obtained using the prospective measures (crude analysis) were consistent to these results using the then-test method except that:
- a significant surgery effect was highlighted for pain dimension (estimate = 0.57; *p*-value = 0.03) while this effect was not significant in the reference analysis (estimate = 0.48; *p*-value = 0.10) and
- a significant time effect was highlighted for GHS (estimate = -0.15; *p*value = 0.04) while this effect was not significant in the reference analysis (estimate = -0.16; *p*-value = 0.09).

Regarding the results obtained by re-estimating item difficulty parameter at time T2 for each response category (alternative method to take into account the RS effect), the same divergences were observed as compared to the results obtained with the then-test method. Moreover, the time effects for fatigue (estimate= 0.15; *p*-value = 0.20), pain (estimate = 0.09; *p*-value = 0.31) and body image (estimate = -0.02; *p*-value = 0.83) dimensions were not detected with a corresponding *p*-value >0.05.

3.3.3 TTD

All results obtained for the TTD analyses according to each reference score are summarized in Table 6.

Using the then-test obtained post-surgery as the reference measure (reference analysis), patients with a mastectomy tended to present a shorter TTD deterioration for 3/5 selected dimensions with a HR (mastectomy vs. no mastectomy) >1. Patients who undergone a mastectomy presented a significantly shorter TTD of body image (HR = 1.56 (95%CI 1.16; 2.11)) and future perspectives dimensions (HR = 1.52 (95%CI 1.00; 2.31)).

Considering the baseline score as the reference score (crude analysis), patients with a mastectomy tended to present a shorter deterioration for 2/5 selected dimensions with a HR > 1. Only the body image effect highlighted in the reference analysis remained significant using the baseline score as the reference with a HR = 1.84 (95%CI 1.36; 2.47) and *p*-value < 0.01. Considering the best previous score as the reference score (alternative method to take into account the response shift effect), patients with a mastectomy tended to present a shorter deterioration for all selected dimensions with a HR > 1. Using this reference score, and as for the crude analysis, only the body image effect highlighted in the reference analysis remained significant using the baseline score as the reference with a HR = 1.93 (95%CI 1.44; 2.59) and *p*-value < 0.01.

Table 6: Results of the time to deterioration method for each health-related quality of life score and according to three reference scores

		Then-test-surgery				Baseline score				Best previous score			
	N	events	median (95%CI)	HR (95%CI)	P	events	median (95%CI)	HR (95%CI)	P	events	median (95%CI)	HR (95%CI)	P
GHS													
No mastectomy	167	111	3.06 (2.92; 3.22)	1	0.15	120	2.92 (0.23; 3.03)	1	0.92	136	2.92(0.23; 3.02)	1	0.77
Mastectomy	106	58	3.15 (2.90; NA)	0.80(0.58; 1.10)		72	2.89 (0.20; 3.06)	0.99 (0.74; 1.32)		83	2.83 (0.20; 2.99)	1.04 (0.79; 1.37)	
Fatigue													
No mastectomy	167	125	2.89 (0.23; 3.02)	1	0.59	125	2.83 (0.26; 3.02)	1	0.90	136	2.79 (0.26; 2.99)	1	0.84
Mastectomy	104	80	0.39 (0.16; 2.99)	1.08 (0.81; 1.42)		74	0.36 (0.13; 2.99)	0.98 (0.74; 1.31)		85	0.36 (0.13; 2.96)	1.03 (0.78; 1.35)	
Pain													
No mastectomy	171	121	2.99 (2.66; 3.09)	1	0.77	127	2.89 (0.23; 3.02)	1	0.49	135	2.86 (0.23; 3.02)	1	0.82
Mastectomy	107	72	2.89 (0.23; 3.15)	0.96 (0.72; 1.28)		73	1.58 (0.16; 3.06)	0.91 (0.68; 1.21)		87	1.58 (0.16; 3.02)	1.03 (0.79; 1.35)	
Body image													
No mastectomy	156	93	3.06 (3.02; 5.95)	1	<0.01	95	3.19 (2.99; 5.88)	1	<0.01	97	3.02 (2.99; 3.42)	1	<0.01
Mastectomy	101	80	2.99 (2.69; 3.02)	1.56 (1.16; 2.11)		82	1.58 (0.16; 2.96)	1.84 (1.36; 2.47)		87	1.58 (0.16; 2.96)	1.93 (1.44; 2.59)	
Future perspectives													
No mastectomy	157	47	7.39 (NA; NA)	1	0.04	40	6.64 (6.64; NA)	1	0.36	68	6.08 (5.95; 7.39)	1	0.59
Mastectomy	99	42	6.28 (5.95; NA)	1.52 (1.00; 2.31)		31	7.00 (7.00; NA)	1.24 (0.78; 1.98)		51	6.05 (5.98; 6.47)	1.10 (0.76; 1.59)	

Results which are not consistent with those obtained using the then-test method are highlighted in grey in the Table

TTD curves are given in Figure 2 to Figure 6.

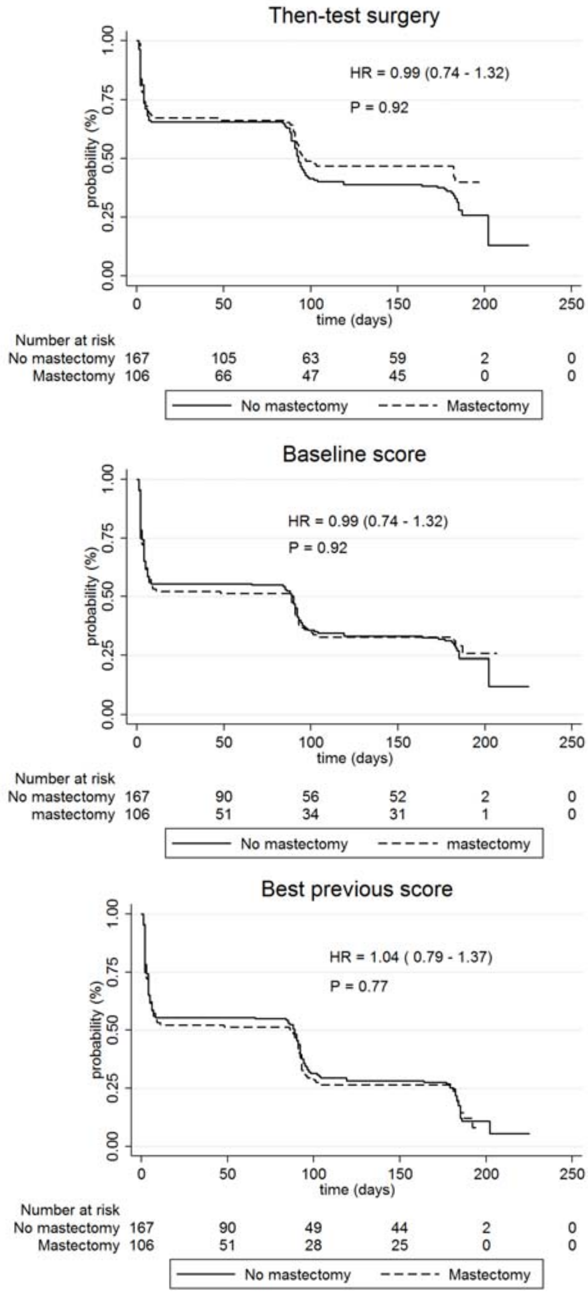

Figure 2: Kaplan-Meier survival curve of the time to global health status deterioration according to each reference score and surgery

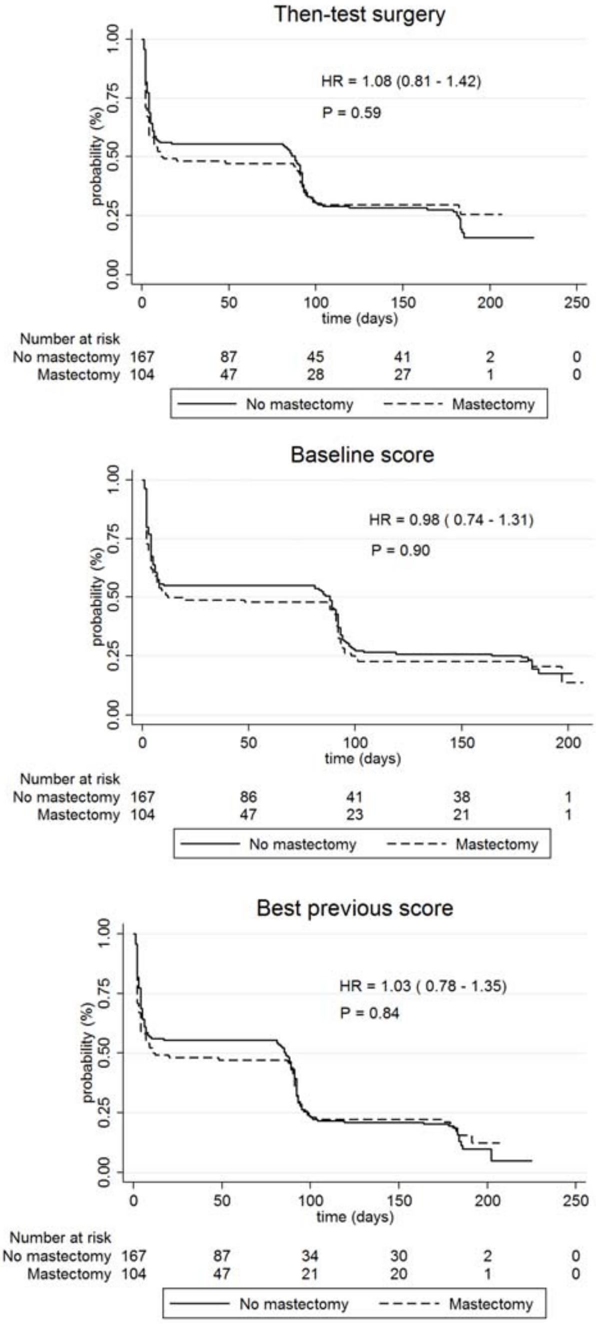

Figure 3: Kaplan-Meier survival curve of the time to fatigue deterioration according to each reference score and surgery

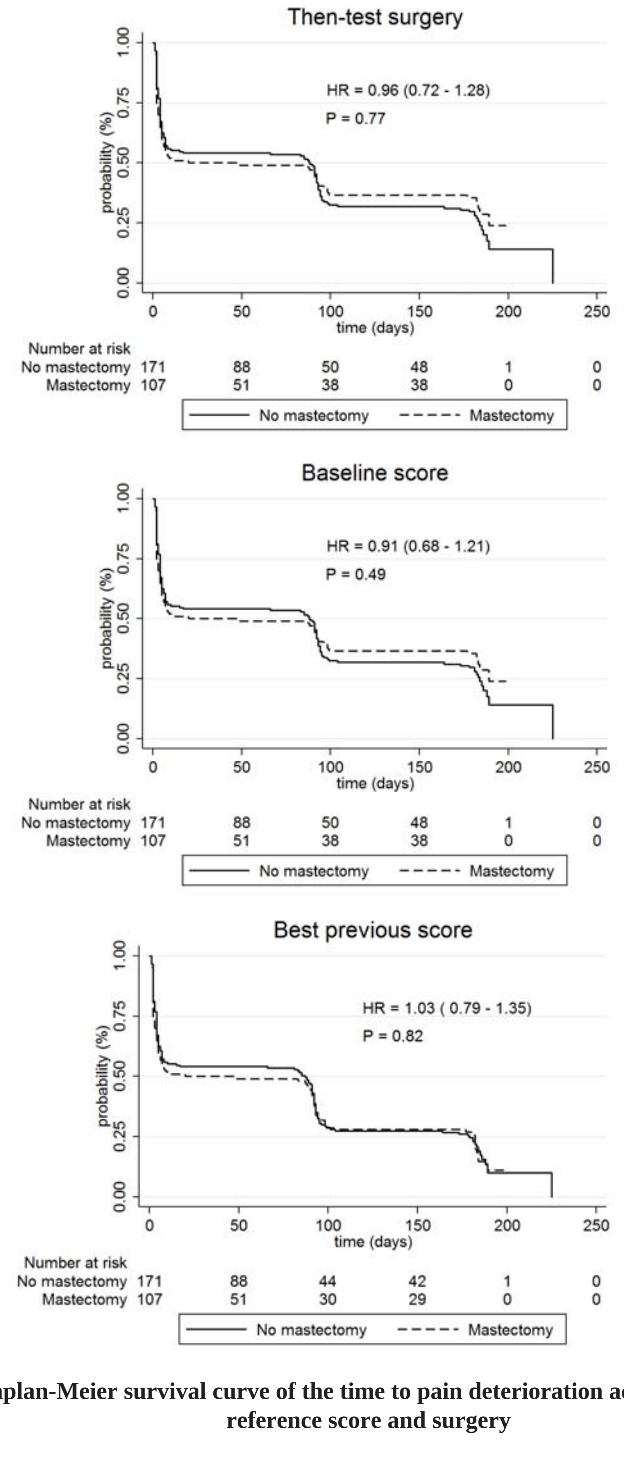

Figure 4: Kaplan-Meier survival curve of the time to pain deterioration according to each reference score and surgery

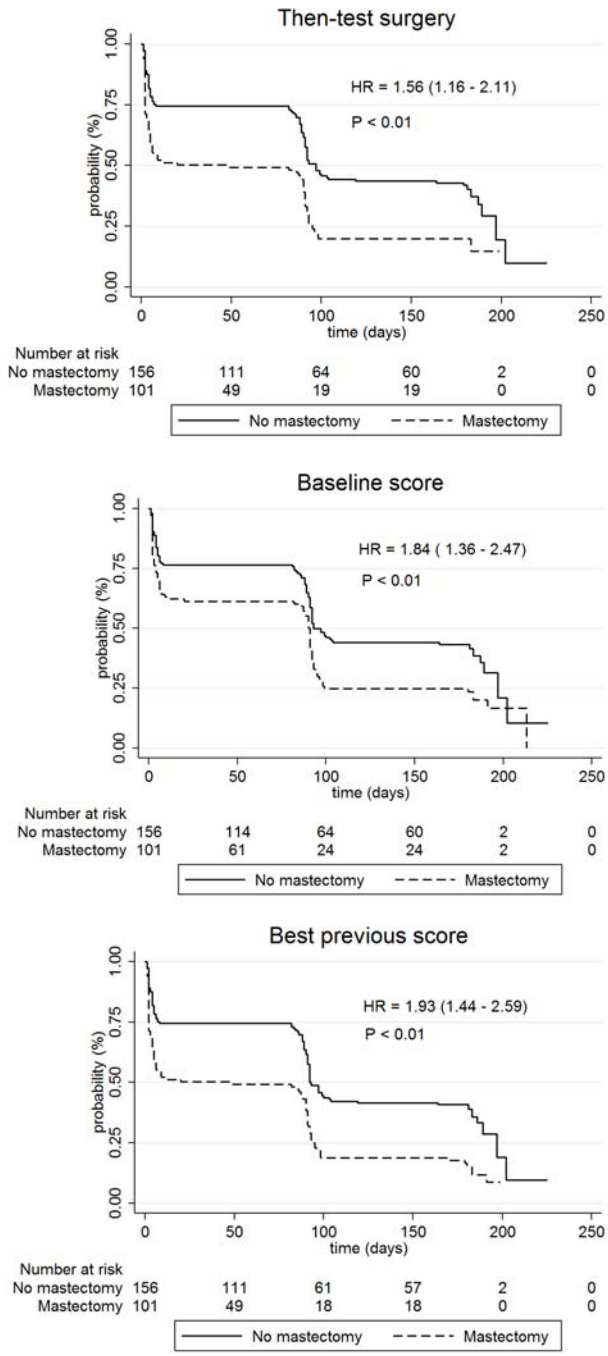

Figure 5: Kaplan-Meier survival curve of the time to body image deterioration according to each reference score and surgery

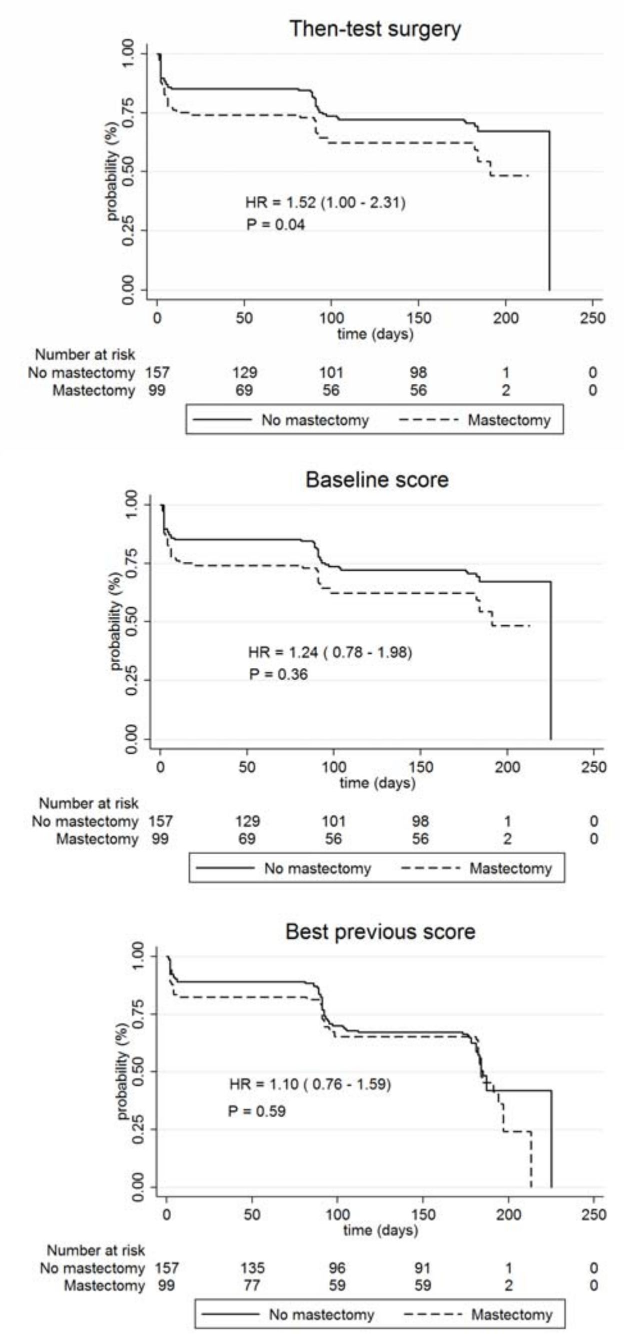

Figure 6: Kaplan-Meier survival curve of the time to future perspectives deterioration according to each reference score and surgery

4. DISCUSSION

This study was the first one exploring several methods to take into account the occurrence of the response shift effect in conjunction with three longitudinal analysis models using the then-test method as the gold standard.

It is important to apply several statistical methods for a longitudinal analysis since no recommendations have been proposed to analyze HRQoL data in oncology. Thus, some consistent results obtained between these methods allow to validate observed trends. In particular, regarding the reference analysis using the then-test method, both mixed models (LMMRM and IRT) showed convergent results for all HRQoL dimensions (same significant effects, same direction for estimate regarding HRQoL level) except that a significant interaction between time and surgery was highlighted for pain dimension in the IRT model (estimate = -0.21; *p*-value = 0.04) while the corresponding effect was not significant in the LMMRM (estimate = -1.87; *p*-value = 0.09) even if in the same direction (indicating a decrease of pain level over time for patients who undergone a surgery as compared to other patients). Results obtained by TTD cannot be directly compared to mixed models since they are not focused on same effects. In fact, TTD is a time-to-event model focused on the event "deterioration". It can highlight a longer or a shorter time to deterioration for patients who undergone a mastectomy as compared to other patients. In this way, TTD approach can complement results obtained using mixed models. Using this approach, a shorter TTD of body image (HR = 1.56 (%CI 1.16; 2.11)) and future perspectives (HR = 1.52 (95%CI 1.00; 2.31)) was highlighted for patients who had a mastectomy as compared to other patients.

For each longitudinal analysis method, the comparison between the crude analysis and the reference analysis allowed to highlight the impact of the occurrence of the recalibration component of the response shift effect on the longitudinal analysis. The LMMRM seems to be the statistical method the most impacted by the occurrence of a response shift effect with three divergent results as compared to the reference analysis. The impact of the response shift effect was also important on the TTD method. In fact, among both dimensions for which the presence of a mastectomy was significantly associated with a shorter time to deterioration using the then-test method, only the impact on body image remained significant in the crude analysis. Finally, the impact of the response shift effect seems to be more moderate on the LPCM with two divergent results as compared to the reference analysis among all parameters estimated: a mastectomy effect wrongly detected for pain as well as a significant time effect for GHS.

For each longitudinal analysis model, one statistical method was explored in order to take into account the occurrence of the recalibration component of the response shift effect based on prospective measures only. Regarding the LMMRM, the analysis performed by adjusting on the score obtained post-surgery seems to reduce the bias due to the occurrence of the response shift effect but not with an optimal efficacy, since some discrepancies with the reference analysis were still observed. Regarding the TTD approach, the use of the best previous score as the reference score seems not to reduce the bias due to the occurrence of the response shift by only detecting the mastectomy effect body image dimension (HR = 1.93 (95%CI 1.44; 2.59), *p*-value < 0.01), not on future perspectives (HR = 1.10 (95%CI 0.76; 1.59), *p*-value = 0.59). Finally, the method explored in conjunction with the IRT model appears to increase the bias due to the occurrence of the response shift effect as compared to the crude analysis since all discrepancies highlighted in the crude analysis as compared to the reference analysis were still present and two additional time effects were not correctly highlighted with a *p*-value > 0.05.

All these conclusions are based on the assumption that the then-test method is the gold standard to capture the recalibration component of the response shift effect. However, the then-test method is increasingly questioned (Schwartz and Rapkin, 2012; Schwartz and Sprangers, 2010; Visser et al., 2005), mainly because it can induce a recall bias (Sprangers et al., 1999). Moreover, we only considered in the present paper the then-test performed at the post-surgery time window. The occurrence of a recalibration was also highlighted at three and six months and published in another paper (Anota et al., 2014b). In IRT models, this longitudinal occurrence of a response shift effect could be taken into account by re-estimating the trend of item difficulty parameter at each follow-up. Another study already investigated the ability of the IRT models to characterize the occurrence of the recalibration component of the response shift effect by varying item difficulty parameters (Guilleux et al., 2015). However, the main studies regarding IRT were realized on simulation studies while real data are required to validate such methodology. Moreover, it is also necessary to compare the results to a gold standard such as the then-test method.

The choice of these three statistical methods for the longitudinal analysis was based on the most frequently used to analyze longitudinal HRQoL data in oncology (for both LMMRM and TTD approaches) or the a priori most appropriate method (for IRT models). However, one limitation of the LMMRM is that it required the normally distribution of the score, which is not respected for the HRQoL scores of the EORTC questionnaire. Studies using such model to analyze longitudinal HRQoL data also generally do not check this hypothesis. The IRT model is more adapted to this kind of data which raised from questionnaire. In fact, they are based on items themselves and not on the observed score and they have some interesting properties such as the independence of the sample considered. However, some strong properties should be respected in order to apply such Rasch-family model (De Ayala, 2013), namely: the unidimensionality of the scale, the monotonicity of the latent trait and the local independence of the items. Studies using longitudinal IRT models generally do not check these assumptions which require to be respected at each measurement time. Another disadvantage of the IRT model is that a long time is required for the model to converge. This time is also increased when trend parameters for items difficulty were added to the model. One advantage of the TTD approach is that it allows to integrate the MCID into the definition of the deterioration, and then results are clinically meaningful. Conversely, it seems difficult at this time to integrate the MCID in the interpretation of the results of the IRT models. Regarding the LMMRM, it could be possible to interpret the results as regards to the MCID at the group level. Then, some research is still needed in order to produce some clinically meaningful results with some results easy to understand for the clinician.

Regarding the ability of these statistical methods to take into account the occurrence of the response shift, further research is also still needed. Methods investigated in this paper have to be applied on other data from clinical trials as well as on a simulation study in order to determine their ability to adequately take into account the occurrence of the response shift effect. Moreover, it would be interesting to propose a multidimensional analysis taking into account all the dimensions. This multidimensional analysis could thus explore the reprioritization component of the response shift effect.

To conclude, adjusting on the score obtained post-surgery seems to be a promising approach to use in conjunction to the LMMRM in order to reduce the bias due to the occurrence of the response shift effect. Regarding IRT models, both methods investigated in this paper did not provide satisfying results. Moreover, these models required a long time to converge and research is still needed for these models. Finally, the best previous score seemed to be a good alternative to take into account the occurrence of the response shift using the TTD approach but did not seem to reduce the bias due to the occurrence of the response shift effect.

REFERENCES

Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J., Filiberti, A., Flechtner, H., Fleishman, S.B., de Haes, J.C., Kaasa, S., Klee, M., Osoba, D., Razavi, D., Rofe, P., Schraub, S., Sneeuw, K., Sullivan, M. and Takeda, F. (1993). The European organization for research and treatment of cancer qlq-c30: A quality-of-life instrument for use in international clinical trials in oncology. In *Journal of the National Cancer Institute*, 85 (5): 365–376.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Parzen, E., Tanabe, K. and Kitagawa, G., eds., *Selected Papers of Hirotugu Akaike*, 199–213. Springer, New York.

Andrykowski, M.A., Donovan, K.A. and Jacobsen, P.B. (2009). Magnitude and correlates of response shift in fatigue ratings in women undergoing adjuvant therapy for breast cancer. In *Journal of Pain and Symptom Management*, 37 (3): 341–351.

Anota, A., Barbieri, A., Savina, M., Pam, A., Gourgou-Bourgade, S., Bonnetain, F. and Bascoul-Mollevi, C. (2014a). Comparison of three longitudinal analysis models for the health-related quality of life in oncology: A simulation study. In *Health and Quality of Life Outcomes*, 12 (1): 192.

Anota, A., Bascoul-Mollevi, C., Conroy, T., Guillemin, F., Velten, M., Jolly, D., Mercier, M., Causeret, S., Cuisenier, J., Graesslin, O., Hamidou, Z. and Bonnetain, F. (2014b). Item response theory and factor analysis as a mean to characterize occurrence of response shift in a longitudinal quality of life study in breast cancer patients. In *Health and Quality of Life Outcomes*, 12 (1): 32.

Anota, A., Hamidou, Z., Paget-Bailly, S., Chibaudel, B., Bascoul-Mollevi, C., Auquier, P., Westeel, V., Fiteni, F., Borg, C. and Bonnetain, F. (2013). Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: Do we need RECIST for quality of life to achieve standardization? In *Quality of Life Research*, 24 (1): 5–18.

Bacci, S. (2008). Analysis of longitudinal hrqol using latent regression in the context of Rasch modeling. In C. Huber, N. Limnios, M. Mesbah and N. M., eds., *Mathematical Methods in Survival Analysis, Reliability and Quality of Life*, 275–290. Wiley Online Library, Great Britain.

Bonnetain, F., Dahan, L., Maillard, E., Ychou, M., Mitry, E., Hammel, P., Legoux, J.L., Rougier, P., Bedenne, L. and Seitz, J.F. (2010). Time until definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. In *European Journal of Cancer*, 46 (15): 2753–2762.

Cnaan, A., Laird, N.M. and Slasor, P. (1997). Tutorial in biostatistics: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. In *Statistics in Medicine*, 16: 2349–2380.

Dabakuyo, T., Guillemin, F., Conroy, T., Velten, M., Jolly, D., Mercier, M., Causeret, S., Cuisenier, J., Graesslin, O., Gauthier, M. and Bonnetain, F. (2013). Response shift effects on measuring post-operative quality of life among breast cancer patients: A multicenter cohort study. In *Quality of Life Research*, 22 (1): 1–11.

De Ayala, R.J. (2013). *Theory and Practice of Item Response Theory*. Guilford Publications, New York.

Fairclough, D.L. (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials*. CRC press, New York.

Fayers, P.M., Aaronson, N.K., Bjordal, K., Curran, D., and Grønvold, M. (1999). Eortc qlq-c30 scoring manual. *Tech. rep.*, EORTC.

Fisher, L., Dixon, D., Herson, J., Frankowski, R., Hearron, M., and Peace, K.E. (1990). Intention to treat in clinical trials. In K. Peace and D. M., eds., *Statistical Issues in Drug Research and Development*, 115–162. CRC Press, NewYork.

Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C.E., Hardouin, J.B., and Sébille, V. (2015). Response shift algorithm in item response theory (rosali) for response shift detection with missing data in longitudinal patient-reported outcome studies. In *Quality of Life Research*, 24 (3): 553–564.

Hamidou, Z., Dabakuyo, T.S., Mercier, M., Fraisse, J., Causeret, S., Tixier, H., Padeano, M.M., Loustalot, C., Cuisenier, J., Sauzedde, J.M., Smail, M., Combier, J., Chevillote, P., Rosburger, C., Arveux, P. and Bonnetain, F. (2011).

Time to deterioration in quality of life score as a modality of longitudinal analysis in patients with breast cancer. In *The Oncologist*, 16 (10): 1458–1468.

Hamidou, Z., Dabakuyo-Yonli, T.S., Guillemin, F., Conroy, T., Velten, M., Jolly, D., Causeret, S., Graesslin, O., Gauthier, M., Mercier, M. and Bonnetain, F. (2014). Impact of response shift on time to deterioration in quality of life scores in breast cancer patients. In *Plos One*, 9 (5): e96848.

Hinz, A., Einenkel, J., Briest, S., Stolzenburg, J.U., Papsdorf, K. and Singer, S. (2012). Is it useful to calculate sum scores of the quality of life questionnaire eortc qlq-c30? In *European Journal of Cancer Care*, 21 (5): 677–683.

Korfage, I.J., de Koning, H.J. and Essink-Bot, M.L. (2007). Response shift due to diagnosis and primary treatment of localized prostate cancer: A then-test and a vignette study. In *Quality of Life Research*, 16 (10): 1627–1634.

MacCallum, R.C. and Austin, J.T. (2000). Applications of structural equation modeling in psychological research. In *Annual Review of Psychology*, 51 (1): 201–226.

Oort, F.J. (2005). Using structural equation modeling to detect response shifts and true change. In *Quality of Life Research*, 14 (3): 587–598.

Oort, F.J., Visser, M.R. and Sprangers, M.A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. In *Quality of Life Research*, 14 (3): 599–609.

Osoba, D. (2011). Health-related quality of life and cancer clinical trials. In *Therapeutic Advances in Medical Oncology*, 3 (2): 57–71.

Osoba, D., Rodrigues, G., Myles, J., Zee, B. and Pater, J. (1998). Interpreting the significance of changes in health-related quality-of-life scores. In *Journal of Clinical Oncology*, 16 (1): 139–144.

Park, C., Cohen, L.H. and Herb, L. (1990). Intrinsic religiousness and religious coping as life stress moderators for catholics versus protestants. In *Journal of Personality and Social Psychology*, 59 (3): 562.

Sasieni, P. (2005). Cox regression model. In P. Armitage and T. Colton, eds., *Encyclopedia of Biostatistics*. Wiley Online Library.

Schwartz, C.E. and Rapkin, B.A. (2012). Understanding appraisal processes underlying the thentest: A mixed methods investigation. In *Quality of Life Research*, 21 (3): 381–388.

Schwartz, C.E. and Sprangers, M.A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. In *Social Science & Medicine*, 48 (11): 1531–1548.

Schwartz, C.E. and Sprangers, M.A. (2010). Guidelines for improving the stringency of response shift research using the thentest. In *Quality of Life Research*, 19 (4): 455–464.

Sprangers, M., Groenvold, M., Arraras, J.I., Franklin, J., te Velde, A., Muller, M., Franzini, L., Williams, A., De Haes, H., Hopwood, P., Cull, A. and Aaronson, N. (1996). The European organization for research and treatment of cancer breast cancer-specific quality-of-life questionnaire module: First results from a three-country field study. In *Journal of Clinical Oncology*, 14 (10): 2756–2768.

Sprangers, M.A. and Schwartz, C.E. (1999). Integrating response shift into healthrelated quality of life research: A theoretical model. In *Social Science & Medicine*, 48 (11): 1507–1515.

Sprangers, M.A., Van Dam, F.S., Broersen, J., Lodder, L., Wever, L., Visser, M.R., Oosterveld, P. and Smets, E.M. (1999). Revealing response shift in longitudinal research on fatigue: The use of the thentest approach. In *Acta Oncologica*, 38 (6): 709–718.

Team, R.C. (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www. R-project.org/. ISBN 3-900051-07-0.

Visser, M.R., Oort, F.J. and Sprangers, M.A. (2005). Methods to detect response shift in quality of life data: A convergent validity study. In *Quality of Life Research*, 14 (3): 629–639.

Visser, M.R., Smets, E.M., Sprangers, M.A., and de Haes, H.J. (2000). How response shift may affect the measurement of change in fatigue. In *Journal of Pain and Symptom Management*, 20 (1): 12–18.