ANALYSIS OF LONGITUDINAL PATIENT REPORTED OUTCOMES DATA WITH CLASSICAL TEST THEORY AND RASCH-BASED METHODS: AN APPLICATION ON HEALTH-RELATED QUALITY OF LIFE IN BREAST CANCER PATIENTS

Myriam Blanchin[1], Jean-Benoit Hardouin, Angélique Bonnaud-Antignac, Véronique Sébille

EA 4275, Biostatistics, Pharmacoepidemiology and Subjective Measures in Health Sciences, University of Nantes, France

Abstract *Patient Reported Outcomes (PRO) can be analysed by several approaches, most commonly based on the Classical Test Theory (CTT) or on the Rasch model. The most adequate strategy to analyse longitudinal PRO data, taking into account the latent characteristic of what PRO are intended to measure as well as the specificity of longitudinal designs with repeated measurements remains to be identified. This study compares two methods of analysis of longitudinal PRO data based on CTT or on Rasch model approaches. Health-related quality of life data of 100 breast cancer patients allocated in three groups was evaluated at three time points. Both methods led to similar conclusions regarding tests of time and group effects and seem to be adequate for the analysis of longitudinal PRO data.*

Keywords: *Longitudinal data, Patient Reported Outcomes, Classical Test Theory, Rasch Model, Breast cancer.*

1. INTRODUCTION

Assessment of patient-reported outcomes (PRO) such as health-related quality of life, patient satisfaction or pain is more and more used in health sciences. Finding treatments that increase life span has been a major issue in the past. From now on, improving quality of life is a part of the medical care, in particular in cancer care (Duska and Dizon, 2014; Montazeri, 2008). Quality of life is used as an endpoint to contribute to improve treatments or to enhance care of long-term survivors. As for many patient-reported outcomes, quality of life is assessed through questionnaires. The development and assessment of PRO measures are based either on Classical Test

[1] Correspondence to: Myriam Blanchin, myriam.blanchin@univ-nantes.fr

Theory (CTT) either on Item Response Theory (IRT). The CTT is the most common approach and relies on a score computation by aggregating item responses. This computed score is used to estimate the 'true score', representing the outcome of interest, through a linear relation. In IRT, the outcome of interest is represented by a latent variable. In this type of models, the probability to answer to an item is a function of the latent variable and item parameters. A widespread model of IRT for dichotomous items is the Rasch model due to its simplicity and its psychometric properties (Fischer and Molenaar, 1995; Rasch, 1980).

In health sciences, several approaches - CTT, Rasch model family, parametric or non parametric IRT models - are now commonly used for the development, validation and reduction of questionnaires (Blackmon et al., 2015; Franchignoni et al., 2015; Friedrich et al., 2015; Wilburn et al., 2015). CTT remains the most used approach at the analysis stage. However, when the questionnaire used to collect PRO data was validated with a Rasch model, it seems more appropriate to analyse the data with a Rasch model, to base the analysis on the same grounds than the validation. Furthermore, the analysis should be enhanced by the psychometric properties of the family of Rasch models such as the specific objectivity or the parameters invariance (Embretson and Reise, 2000; Rasch, 1977).

While assessing a PRO in clinical and epidemiological studies, the interest is often in studying the impact of the disease and treatments on this outcome during treatment period but also in the long-term follow-up. In this setting, longitudinal studies, where outcomes are measured repeatedly over time on the same subject, are widely used. Longitudinal data allows measuring change over time. In this context, linear mixed models (Fitzmaurice et al., 2009; Verbeke and Molenberghs, 2000) have become very popular as they are adapted to the analysis of correlated data and are a good way to deal with missing data.

Regarding the analysis of health related quality of life data evaluated in a longitudinal context, we have to choose a method of analysis that is based either on CTT approach or on the Rasch model and that takes into account the feature of longitudinal data. Both approaches have been compared in simulation studies and have shown comparable performance in case of complete data or data subject to ignorable dropout (Blanchin et al., 2011a). However, a Rasch-based method should perform better in case of intermittent missing items according to de Bock et al. (2015). This

study compares two methods of analysis of longitudinal PRO data on a real dataset. The analysis was performed on health-related quality of life of breast cancer patients data evaluated at three time points with the cancer-specific questionnaire European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) (Aaronson et al., 1993).

2. METHODS

2.1. QUALITY OF LIFE AND COPING OF WOMEN TREATED FOR A BREAST CANCER AND THEIR CAREGIVER

Health-Related Quality of Life of breast cancer patients was assessed in a longitudinal study taking place in the Nantes Institut de Cancérologie de l´Ouest. This study focused not only on the patients' quality of life but also on the one of their designed referent person, named as their caregiver. In fact, the disease and its treatments can have an impact on the quality of life of the patient and his/her family. Furthermore, the social support may also be associated with the patients' quality of life. As the occurrence of a cancer has huge implications in the life of the patient and his/her family, different strategies to cope with this event can be used either by the patient or by his/her caregiver. This study aimed at studying the influence of the coping strategies, the characteristics of patients and their caregiver and the quality of life of the caregiver on the quality of life of breast cancer patients. The longitudinal setting of the study intended to investigate the evolution of the quality of life of the patients and their caregivers as well as their coping strategies.

Patients diagnosed with primary breast cancer were recruited. Each patient was asked to choose a referent person, called a caregiver, to participate in the study. It was generally her husband or another member of the family. Both patients and caregivers were informed about the study and an informed written consent was requested for both. Measurements were made about 2 or 3 weeks after diagnosis (t1), at the end of the chemotherapy and/or radiotherapy treatments (t2) and six months after treatments (t3). Quality of life of 100 patients and their caregivers was assessed at t1, 82 patients at t2 and 79 patients at t3. At each time, the patients and caregivers quality of life was evaluated using the European Organization of Research and Treatment of Cancer Quality of Life Questionnaire C30 (EORTC QLQ-C30) (Aaronson et al., 1993) and the Duke Health Profile (Parkerson et al.,

1990). Furthermore, the Ways of Coping Checklist (Cousson et al., 1996) also assessed at each time three coping strategies: problem-centered (strategies directed at solving the impact of the stressful event), emotion-centered (efforts directed at affect regulation) and social support-centered (seeking social support).

A typology of patients (Bonnaud-Antignac et al., 2012) taking into account the socio-demographic characteristics and coping of the patients and caregivers as well as the quality of life of the caregivers has been realised with a Multiple Correspondence Analysis (MCA) followed by a Hierarchical Cluster Analysis (HCA). Four clusters composed respectively of 52, 28, 18 and 2 patients have been identified. In the following analysis, the group of 2 patients has not been retained. The first cluster was composed of patients older than the total sample and using less emotion-centered and problem-centered coping strategies. Their caregivers were more often aged less than 45 years, had a perceived health better than the mean of the total sample, and had a lesser use of the three coping strategies, i.e. problem-centered coping, emotion-centered and social support-centered. The second cluster was composed of patients with high levels of emotion-centered and problem-centered coping strategies. The caregiver was often present during the announcement of the cancer and generally presented a weak health status and more often incapacities and depression. Last, these caregivers used more often the three coping strategies than the complete sample. The third cluster was composed of patients whose ages ranged from 45 to 54 years and none of them were over 65. These patients used more frequently the three coping strategies than the complete sample. Their caregivers were more frequently between 45 and 54 years and a large majority were employed.

2.2. ANALYSIS

This analyis focuses on the evaluation of the quality of life of the patients through the EORTC QLQ-C30. The 30 items of the QLQ-C30 covers the functioning and symptoms of cancer patients divided in six functioning scales: physical functioning, role functioning, emotional functioning, cognitive functioning, social functioning and health-related quality of life and nine symptom scales: fatigue, pain, dyspnea and gastro-intestinal problems. Scales are scored according to the EORTC guidelines (Fayers et al., 2001). The scores range from 0 to 100. A high scale score indicates a high level of

functioning or symptomatology. In this analysis, the three scales with the greatest number of items were studied: physical functioning (PF, 5 items), emotional functioning (EF, 4 items) and fatigue (FAT, 3 items). Two methods to analyse data from quality of life of breast cancer patients based either on CTT either on the Rasch model were compared: the Score and Mixed Models (SMM)) and the Longitudinal Partial Credit Model (LPCM).

CTT-based method: Score and Mixed Models

The Score and Mixed Models method was based on the CTT approach. It consisted first in computing the score of each patient at each time for each of the three dimensions according to the EORTC guidelines. A linear mixed model was then used.

Let $S_i^{(t)}$ be the standardised score of patient i $(i = 1, ..., N)$ at time t $(t = 1, 2, 3)$ computed on a functional scale:

$$S_i^{(t)} = \left(1 - \frac{\frac{1}{J}\sum_j y_{ij}^{(t)} - 1}{range}\right) * 100 \tag{1}$$

where $y_{ij}^{(t)}$ is the response of patient i to item j $(j = 1, ..., J)$ at time t and *range* is the difference between the maximum raw score (raw score$=\frac{1}{J}\sum_j y_{ij}$) and the minimum raw score. For PF, EF and FAT, the raw score ranged between 1 and 4 so that the *range* was 3. Scores $S_i^{(t)}$ ranged from 0 to 100. For functional scales (EF end PF), 0 indicated a poor quality of life and 100 indicated a high level of functioning or an excellent quality of life.

Similarly, $S_i^{(t)}$ is the standardised score of patient i $(i = 1, ..., N)$ at time t $(t = 1, 2, 3)$ computed on a symptom scale:

$$S_i^{(t)} = \frac{\frac{1}{J}\sum_j y_{ij}^{(t)} - 1}{range} * 100 \tag{2}$$

For symptom scales (FAT), 100 indicated a high level of symptomatology or problems.

A linear mixed model explaining the individual scores was then used to estimate the time and group effects. Mixed models are widely used for the analysis of data from longitudinal studies. They allow dealing with repeated measurements by specifying a structure for the correlation between

measurements from a same patient. Mixed models can also handle incomplete data. A mixed model is composed of both fixed effects and random effects, that is mixed effects. The mean response is modeled as a combination of fixed effects characterizing the population variables and random effects characterizing the subject-specific effects that are unique to a particular individual.

Let

- n_i be the number of observations on patient i, $i = 1...N$ and $M = \sum_{i=1}^{N} n_i$ be the total number of observations
- p and k be the number of fixed and random parameters respectively
- $\boldsymbol{Y}_i$ be the $(n_i \times 1)$ vector containing the responses for the patient i
- $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ be the $(n_i \times p)$ design matrix characterizing the fixed part of the model and the $(n_i \times k)$ design matrix characterizing the random variation in the model due to among-unit sources
- $\boldsymbol{\beta}$ be the $(p \times 1)$ vector of fixed effects parameters
- $\boldsymbol{b}_i$ be the $(k \times 1)$ vector of random effects parameters. $b_i \sim N_k(0, \boldsymbol{D})$
- $\boldsymbol{e}_i$ be the $(n_i \times 1)$ vector of error terms, characterizing variation due to within-unit and measurement error sources. $e_i \sim N_{n_i}(0, \boldsymbol{R}_i)$
- $\boldsymbol{\Sigma}_i$ be the $(n_i \times n_i)$ covariance matrix

$$\begin{aligned} &\boldsymbol{S}_i = (S_i^{(1)} S_i^{(2)} S_i^{(3)})' = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \\ &var(\boldsymbol{S}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i = \boldsymbol{\Sigma}_i \\ &\boldsymbol{S}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i) \end{aligned} \tag{3}$$

The parameters to be estimated in the model are the parameters $\boldsymbol{\beta}$ that characterise the mean and the covariance parameters $\boldsymbol{\omega}$ that characterise the variation of the model, the parameters that makes up $\boldsymbol{R}_i$ and $\boldsymbol{D}$. They were estimated using the REstricted Maximum Likelihood (REML) method in order to reduce the bias on covariance parameters in comparison to the use of maximum likelihood estimation (Fitzmaurice et al., 2004).

The model includes time effects, group effects and first-order interactions in the fixed part and an intercept and a slope in the random part. The structure of the covariance matrix could be unstructured (UN), first-order

autoregressive (AR(1)), first-order heterogeneous autoregressive (ARH(1)), compound symmetry (CS), heterogeneous compound symmetry (CSH). The Akaike's Information Criterion (AIC) for each covariance matrix structure were compared to choose the adequate structure. The fixed part of the model with the smallest AIC was then reduced if the test of the time×group interaction did not reject the nullity of the parameter at the level of 5%. An estimate of $\boldsymbol{\mu}$ could be given by $\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 & \hat{\mu}_2 & \hat{\mu}_3 \end{pmatrix}' = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Rasch-based method: Longitudinal Partial Credit Model

Different polytomous multidimensional latent-trait models for longitudinal data have been proposed (Fischer and Parzer, 1991; Meiser, 2007). The Longitudinal Partial Credit Model (LPCM) was based on longitudinal development of the Partial Credit Model (Masters, 1982).

Let $Y_{ij}^{(t)}$ be the response of individual i on item j at time t, the probability of positive response h $(h = 1, ..., m)$ on item j at time t for patient i belonging to the group g $(g = 1, 2, 3)$ is

$$P(Y_{i[g]j}^{(t)} = h^{(t)} | \theta_i^{(t)}, \delta_{jp}, g) = \frac{exp(h^{(t)}(\theta_{i[g]}^{(t)} + \beta_{gp[g]}) - \sum_{p=1}^{h} \delta_{jp})}{\sum_{l=0}^{m} exp(l(\theta_{i[g]}^{(t)} + \beta_{gp[g]}) - \sum_{p=1}^{l} \delta_{jp})} \qquad (4)$$

$$\boldsymbol{\theta_{i[g]}} = (\theta_{i[g]}^{(1)}, ..., \theta_{i[g]}^{(3)})' \sim N_3 \left(\boldsymbol{\mu}_{i[g]}, \boldsymbol{\Sigma}_i\right)$$

where $\theta_{i[g]}^{(t)}$ is the person parameter representing the ability of the individual i at time t. The distribution of the person parameter depends on the group g of the individual i. The model is a mixed-effects logistic regression with a covariate, as the latent traits are considered as random variables and the model includes group effects $\beta_{gp[g]}$. Item parameters δ_{jp}, distribution parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of $\boldsymbol{\theta}$ and the group effects $\beta_{gp[g]}$ were estimated using marginal maximum likelihood estimation (MML). The item parameters were assumed to be constant with time. The constraint of identifiability was the nullity of the latent trait mean at time 1 for group 3. For comparison with the Score and Mixed Models method, item responses of EF and PF were reversed and all item responses ranged from 0 to 3.

As for the SMM method, the structure of the covariance matrix could be UN, AR(1), ARH(1), CS or CSH. We used AIC to evaluate models that showed an acceptable fit to the data. The model retained was the model with the smallest AIC.

Comparison of the methods

The time effect between times t and t' ($t, t' = 1, 2, 3$, $t \neq t'$) was estimated by $\hat{d}_{t,t'} = \hat{\mu}_{t'} - \hat{\mu}_t$. The test of a global time effect used an approximate F-test where the null hypothesis is the equality of the means at each time. Similarly, the group effect between two groups g and g' ($g, g' = 1, 2, 3$, $g \neq g'$) was estimated by $\hat{d}_{gp_{g,g'}} = \hat{\mu}_{g'} - \hat{\mu}_g$. The test of group effect used a similar F-test.

3. RESULTS

Models with the smallest AIC were selected for each dimension and each method. Every models retained contained no random effects and were of the form:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i \tag{5}$$

3.1. PHYSICAL FUNCTIONING DIMENSION

The Physical Functioning (PF) dimension of the QLQ-C30 consisted of 5 items (a high score indicates a high level of functioning and a better quality of life). Estimations of parameters and their standard errors for both methods Longitudinal Partial Credit model (LPCM) and Score and Mixed models (SMM) are shown in Table 1. The value of the global F-tests of time and group effect and their p-values are also shown. Results were obtained with a CSH covariance matrix structure for both methods.

Both methods gave comparable results. The test of time effect concluded to a global time effect at 5 % level. The time effect between time 1 and time 2 and between time 1 and time 3 is negative whereas the time effect between time 2 and time 3 is not significant. The physical functioning decreased between the diagnosis (t_1) and the end of treatments (t_2). It then stays stable between the end of treatments (t_2) and the sixth months after the end of treatment (t_3). Last, a global decrease of the physical functionning was observed on the overall period of the study showing a deterioration of the quality of life on this dimension. There was no significant global group effect. The estimated group effects between two groups were low and not significantly different from 0. Each of the three groups has approximately the same levels on the physical functioning dimension.

Table 1: Parameters estimations of the dimension "Physical Functioning" of the QLQ-C30 for the methods Longitudinal Partial Credit model (LPCM) and Score and Mixed models (SMM). Estimations of the variance at time t ($\hat{\sigma}_t^2$), correlation coefficient ($\hat{\rho}$), time effect between t and t' ($\hat{d}_{tt'}$), group effect between group g and g' ($\hat{d}_{gp_{gg'}}$) for $t = 1, 2, 3$ et $g = 1, 2, 3$. Test statistic and p-value of the test of global time/group effect.

Method	longitudinal PCM			Score and Mixed Models		
Structure	CSH			CSH		
Parameter	Est.	S.e..	P-value	Est.	S.e.	P-value
$\hat{\rho}$	0.857	0.052	<0.0001	0.681	0.049	<0.0001
$\hat{\sigma}_1^2$	4.457	1.250	0.0006	244.850	35.731	<0.0001
$\hat{\sigma}_2^2$	3.472	0.929	0.0003	416.590	66.375	<0.0001
$\hat{\sigma}_3^2$	2.549	0.731	0.0007	273.200	44.425	<0.0001
$\hat{d}_{12}$	-1.242*	0.281	<0.0001	-7.853*	1.674	<0.0001
$\hat{d}_{23}$	0.222	0.227	0.3308	3.239	1.739	0.0644
$\hat{d}_{13}$	-1.020*	0.285	0.0005	-4.614*	1.455	0.0018
$\hat{d}_{gp_{12}}$	-0.380	0.452	0.4030	-0.401	3.508	0.9093
$\hat{d}_{gp_{23}}$	0.702	0.590	0.2371	1.897	4.497	0.6741
$\hat{d}_{gp_{13}}$	0.322	0.523	0.5394	1.496	4.053	0.7128
Test	Statistic	P-value		Statistic	P-value	
Time effect	10.100	0.0001		12.080	< 0.0001	
Group effect	0.740	0.480		0.100	0.909	

* The Student test is significant at 5%

Est.: Estimation, S.e.: standard error

3.2. EMOTIONAL FUNCTIONING DIMENSION

The dimension Emotional Functioning consists of 4 items (a high score indicates a high level of functioning and a better quality of life). Table 2 shows the estimations of the parameters and their standard errors for both methods LPCM and SMM. The value of the global F-tests of time and group effect and their p-values are also shown. Results were obtained with a CSH covariance matrix structure for LPCM method and CS for SMM method.

The methods LPCM and SMM presented comparable results. Both methods rejected the hypothesis of the nullity of the time effect at level 5%. The time effect estimations between any times were positive. The time effect between $t1$ and $t2$ was high and significantly different from 0 whereas the

Table 2: Parameters estimations of the dimension "Emotional Functioning" of the QLQ-C30 for the methods Longitudinal Partial Credit model (LPCM) and Score and Mixed models (SMM). Estimations of the variance at time t ($\hat{\sigma}_t^2$), correlation coefficient ($\hat{\rho}$), time effect between t and t' ($\hat{d}_{tt'}$), group effect between group g and g' ($\hat{d}_{gp_{gg'}}$) for $t = 1, 2, 3$ et $g = 1, 2, 3$. Test statistic and p-value of the test of global time/group effect.

Method	longitudinal PCM			Score and Mixed Models		
Structure	CSH			CS		
Parameter	Est.	S.e.	P-value	Est.	S.e.	P-value
$\hat{\rho}$	0.524	0.078	<0.0001	0.440	——	——
$\hat{\sigma}_1^2$	3.703	0.859	<0.0001	611.420	——	——
$\hat{\sigma}_2^2$	7.100	1.712	<0.0001	611.420	——	——
$\hat{\sigma}_3^2$	6.193	1.548	0.0001	611.420	——	——
$\hat{d}_{12}$	0.906*	0.316	0.0051	5.929*	2.844	0.0388
$\hat{d}_{23}$	0.325	0.361	0.3691	3.435	3.013	0.2561
$\hat{d}_{13}$	1.231*	0.320	0.0002	9.363*	2.911	0.0016
$\hat{d}_{gp_{12}}$	-1.419*	0.473	0.0034	-12.756*	4.796	0.0092
$\hat{d}_{gp_{23}}$	0.386	0.596	0.5186	1.999	6.096	0.7436
$\hat{d}_{gp_{13}}$	-1.033	0.545	0.0611	-10.756	5.480	0.0526
Test	Statistic	P-value		Statistic	P-value	
Time effect	8.710	0.0003		5.430	0.005	
Group effect	4.980	0.009		4.280	0.017	

* The Student test is significant at 5%
Est.: Estimation, S.e.: standard error
———: not available

time effect between $t2$ and $t3$ was not significantly different from 0. Quality of life on emotional functioning increased on the full period, sharply in the first period between the diagnosis and the end of treatments and slowly (but non-significantly) in the second period after the end of treatments. Hence, the level of emotional functioning increased on the overall period. Both methods rejected the hypothesis of the nullity of the group effect. The group effect estimation between group 1 and group 2 was significantly different from 0. The level of emotional functioning was higher in group 1 than in group 2. The group effect estimation between group 1 and group 3 was also high but not significantly different from 0.

3.3. FATIGUE DIMENSION

The dimension Fatigue (FAT) of the QLQ-C30 consists of 3 items measuring the level of symptoms of fatigue (a high score indicates a high level of symptomatology and a worse quality of life). Table 3 shows the estimations of the parameters and their standard errors for both methods LPCM and SMM. The value of the global F-tests of time and group effect and their p-values are also shown. Results were obtained with an AR(1) covariance matrix structure for LPCM method and a CSH structure for SMM method.

Table 3: Parameters estimations of the dimension "Fatigue" of the QLQ-C30 for the methods Longitudinal Partial Credit model (LPCM) and Score and Mixed models (SMM). Estimations of the variance at time t ($\hat{\sigma}_t^2$), correlation coefficient ($\hat{\rho}$), time effect between t and t' ($\hat{d}_{tt'}$), group effect between group g and g' ($\hat{d}_{gp_{gg'}}$) for $t = 1, 2, 3$ et $g = 1, 2, 3$. Test statistic and p-value of the test of global time/group effect.

Method	**longitudinal PCM**			**Score and Mixed Models**		
Structure	**AR(1)**			**CSH**		
Parameter	**Est.**	**S.e.**	**P-value**	**Est.**	**S.e.**	**P-value**
$\hat{\rho}$	0.610	0.070	<0.0001	0.520	0.065	<0.0001
$\hat{\sigma}_1^2$	8.874	1.750	<0.0001	591.030	87.758	<0.0001
$\hat{\sigma}_2^2$	8.874	1.750	<0.0001	770.160	121.800	<0.0001
$\hat{\sigma}_3^2$	8.874	1.750	<0.0001	529.510	86.383	<0.0001
$\hat{d}_{12}$	1.572*	0.368	<0.0001	12.029*	2.838	<0.0001
$\hat{d}_{23}$	-0.838*	0.374	0.0274	-6.751*	2.894	0.0210
$\hat{d}_{13}$	0.733	0.427	0.0891	5.278*	2.594	0.0437
$\hat{d}_{gp_{12}}$	1.251	0.639	0.0530	9.662	4.959	0.0543
$\hat{d}_{gp_{23}}$	-1.075	0.808	0.1866	-9.161	6.306	0.1496
$\hat{d}_{gp_{13}}$	0.176	0.726	0.8091	0.502	5.671	0.9297
Test	Statistic	P-value		Statistic	P-value	
Time effect	9.500	0.0002		9.010	0.0002	
Group effect	1.990	0.143		2.040	0.136	

* The Student test is significant at 5%

Est.: Estimation, S.e.: standard error

For both methods, the test of time effect has concluded to the presence of a global time effect. The time effect between time 1 and time 2 was positive and significantly different from 0. The level of symptoms on the

fatigue dimension increased between the diagnosis (t_1) and the end of treatments (t_2). It then decreased a little in the post-treatments period. On the overall time of the study, the level of symptoms on the fatigue dimension has increased significantly according to the SMM method whereas the level of symptoms on the fatigue dimension remained stable according to the LPCM method. The test of group effect has concluded to the absence of a global group effect for both methods. No estimation of the group effect between two groups was significantly different from 0. The level of fatigue was approximately the same in the 3 groups.

4. DISCUSSION

Quality of life of breast cancer patients assessed with the EORTC-QLQC30 was analysed with two different methods, Longitudinal Partial Credit Model (LPCM) based on the Rasch model and Score and Mixed Models (SMM) based on CTT. The point was to compare these two approaches for the analysis of longitudinal PRO data. The global time and group effects on the three scales: Fatigue, Physical Functioning and Emotional Functioning were almost similar whatever the method used. As the methods are based on two different theories, the results of each single analysis cannot be compared. The results of SMM are expressed in terms of score and the results of LPCM in terms of latent variable. So, the comparison of the magnitude of the time or group effect would have made no sense. However, it was possible to compare the sign and the significance of the effects as well as the conclusions of the global tests of effects.

For the three scales studied, both methods have shown similar results. The same group or time effects were significant and the global tests led to the same conclusions. Three different forms of time effect were observed. On the Fatigue dimension, a large increase first and a small decrease then has lead to a rise on the full period (significant increase only regarding SMM). Quality of life of patients has decreased during the study on this dimension because the level of fatigue has globally increased. The quality of life on Physical Functioning dimension went down on the full period, decreasing in the first period and lightly rising in the second period. On the Emotional Functioning dimension, quality of life of the patients increased during the study, sharply first and slowly then. The evolution of the time effect is in general not constant with time in clinical and epidemiological studies and each time effect between consecutive times should be considered. In fact,

when the evolution of a PRO is studied during the treatment period and the follow-up, we can expect that the impact of the treatment will not be the same on the overall period of the study and this has to be taken into account in the modeling process.

Longitudinal data are more difficult to analyse than cross-sectional data as the correlation between measures of the same patient makes the model more complex. The number of estimated parameters increases with the number of time points and groups. The choice of the model for the analysis can be limited as the convergence may not be achieved. In this example, it was not possible to achieve the convergence for the Rasch-based method when group-time interactions were added in the models. For the CTT-based method, interactions can be added for only some structures of covariance matrix. However, the interaction estimates did not turn out to be significantly different from 0 and we would probably have obtained the same results with the Rasch-based method. The choice of the best model was based on the AIC and led to retain the simplest models, without random effects.

Both methods led to the same conclusions and seem to be adequate for the analysis of longitudinal PRO data. However, the Rasch-based method might be preferred as the family of Rasch models might be a better way to deal with intermittent missing items (de Bock et al., 2015). The loss of information will be less in a Rasch-based method by using all available information whereas a CTT-based method needs to have an aggregated measure that may not exist if some items are missing for the patient. Furthermore, the specific objectivity property in the Rasch model family may ensure that the latent variable may be estimated consistently even for patients with missing items (Fischer, 1995). The imputation for missing data in CTT can help recovering a part of the lost information but can introduce bias in the analysis if the method of imputation is wrongly chosen (Fayers et al., 1998; Hamel et al., 2012). Besides, imputing for missing data is only possible if the proportion of missing items is not too large, generally half of the items in the guidelines for scoring quality of life scales.

Simulation studies regarding type I, power and estimations of time effect in the context of longitudinal PRO data have compared a CTT-based method and a Rasch-based method in different frameworks of missing data. For a complete data case, both approaches obtained similar results (unbiased estimates and good power) (Blanchin et al., 2011b). In the same

framework and for complete dropout case, both methods engendered poor power and biased estimates in case of informative missing data (Blanchin et al., 2011a). Moreover, for longitudinal data, in the presence of possibly informative intermittent missing data, a Rasch-based method appeared to be more powerful than CTT for identifying and quantifying a time effect in a single group of patients (de Bock et al., 2013). Furthermore, simulation studies regarding type I, power and estimations of group effect in the context of longitudinal PRO data evaluated in two groups of patients have also compared a CTT-based method (including a personal mean score imputation) and a Rasch-based method in the framework of intermittent missing items (de Bock et al., 2015). When data are subject to intermittent missing items, this study has shown that the Rasch-based method performs better than the CTT-based one. In fact, the group effect estimations were less often biased and the power of the tests of group effect was greater for the Rasch-based method. Furthermore, the better performance of the Rasch-based method compared to the CTT-based one increased with the proportion of intermittent missing items.

In our study, 14.3% of the patients dropped out from the study at time 2 and 10.7% of the remaining patients dropped out from the study at time 3. At each time of measurement, apart from the patients that did not answered to any of the items of the physical functioning, emotional functioning and fatigue dimensions, we observed a very small number of intermittent missing items (3.06% on 2 items of fatigue at time 1, 1.25% on 2 items of physical functioning and 1 item of fatigue at time 2 and 1.33% on 1 item of fatigue and 1 item of emotional functioning at time 3). This low rate of intermittent missing items may explain that both methods led to the same conclusions regarding time and group effects on the three dimensions of the QLQ-C30 studied.

REFERENCES

Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J., Filiberti, A., Flechtner, H., Fleishman, S.B., Haes, J.C.J.M.d., Kaasa, S., Klee, M., Osoba, D., Razavi, D., Rofe, P.B., 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.*

Blackmon, J.E., Liptak, C. and Recklitis, C.J. (2015). Development and preliminary validation of a short form of the Beck Depression Inventory for Youth (BDI-Y) in a sample of adolescent cancer survivors. In *Journal of Caner Survivorship: Research and Pratice*, 9 (1): 107-114.

Blanchin, M., Hardouin, J.B., Le Néel, T., Kubis, G. and Sébille, V. (2011a). Analysis of longitudinal Patient-Reported Outcomes with informative and non-informative dropout: Comparison of CTT and Rash-based methods. In *International Journal of Applied Mathematics & Statistics*, 24 (SI-11A): 107-124.

Blanchin, M., Hardouin, J.B., Neel, T.L., Kubis, G., Blanchard, C., Mirallié, E. and Sébille, V. (2011b). Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. In *Statistics in Medicine*, 30 (8): 825-838.

Bonnaud-Antignac, A., Hardouin, J.B., Léger, J., Dravet, F. and Sébille, V. (2012). Quality of life and coping of women treated for breast cancer and their caregiver. What are the interactions? In *Journal of Clinical Psychology in Medical Settings*, 19 (3): 320-328.

Cousson, F., Bruchon-Shweitzer, M., Quintard, B., Nuissier, J. and Rascle, N. (1996). Analyse multidimensionnelle d'une échelle de coping: validation française de la W.C.C. (ways of coping checklist). In *Psychologie Française*, 41 (2): 155-164.

de Bock, E., Hardouin, J.B., Blanchin, M., Le Neel, T., Kubis, G., Bonnaud-Antignac, A., Dantan, E. and Sébille, V. (2013). Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. In *Statistical Methods in Medical Research.* doi:10.1177/0962280213515570.

de Bock, E., Hardouin, J.B., Blanchin, M., Le Neel, T., Kubis, G. and Sébille, V. (2015). Assessment of score- and Rasch-based methods for group comparison of longitudinal patient-reported outcomes with intermittent missing data (informative and non-informative). In *Quality of Life Research*, 24 (1): 19-29.

Duska, L.R. and Dizon, D.S. (2014). Improving quality of life in female cancer survivors: current status and future questions. In *Future Oncology*, 10 (6): 1015-1026.

Embretson, S.E. and Reise, S.P. (2000). The Trait Level Measurement Scale: Meaning, interpretations, and measurement-scale properties. In *Item Response Theory for Psychologists*, Multivariate Applications Series, 125-157. Lawrence Erlbaum Associates Inc, Mahwah, New Jersey.

Fayers, P., Aaronson, N.K., Bjordal, K., Groenvold, M., Curran, D. and Bottomley, A. (2001). *EORTC QLQ-C30 Scoring Manual, Third edition.* European Organisation for Research and Treatment of Cancer, Brussels, on behalf of the EORTC Quality of Life Group edn.

Fayers, P.M., Curran, D. and Machin, D. (1998). In complete quality of life data in randomized trials: missing items. In *Statistics in Medicine*, 17 (5-7): 679-96.

Fischer, G. (1995). Derivations of the Rasch model. In G.H. Fischer and I.W. Molenaar, eds., *Rasch Models: Foundations, Recent Developments, nd Applications*, 15-38. Springer-Verlag, New-York.

Fischer, G.H. and Molenaar, I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, New York.

Fischer, G. and Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. In *Psychometrika*, 56:637-651.

Fitzmaurice, G.M., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Chapman and Hall/CRC.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). Estimation and statistical inference. In *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics. Wiley-IEEE, Hoboken.

Franchignoni, F., Monticone, M., Giordano, A. and Rocca, B. (2015). Rasch validation of the Prosthetic Mobility Questionnaire: A new outcome measure for assessing mobility in people with lower limb amputation. In J*ournal of Rehabilitation Medicine*, 47 (5): 460-465.

Friedrich, O., Sipötz, J., Benzer, W., Kunschitz, E. and Höfer, S. (2015). The dimensional structure of the MacNew Health Related Quality of Life questionnaire: A Mokken Scale Analysis. In *Journal of Psychosomatic Research*, 79 (1): 43-48.

Hamel, J.F., Hardouin, J.B., Le Neel, T., Kubis, G., Roquelaure, Y. and Sébille, V. (2012). Biases and power for groups comparison on subjective health measurements. In *PLoS ONE*, 7 (10): e44695.

Masters, G.N. (1982). A Rasch model for partial credit scoring. In *Psychometrika*, 47 (2): 149-174.

Meiser, T. (2007). Rasch models for longitudinal data. In M. von Davier and C.H. Carstensen, eds., *Multivariate and Mixture Distribution Rasch Models*, Statistics for Social and Behavioral Sciences, 191-199. Springer, New York.

Montazeri, A. (2008). Health-related quality of life in breast cancer patients: A bibliographic review of the literature from 1974 to 2007. In *Journal of Experimental & Clinical Cancer Research*, 27: 32.

Parkerson, G.R., Broadhead, W.E. and Tse, C.K. (1990). The Duke Health Profile. a 17-item measure of health and dysfunction. In *Medical Care*, 28 (11): 1056-1072.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In *Danish Yearbook of Philosophy*, 14: 58-94.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, Berlin.

Wilburn, J., McKenna, S.P., Twiss, J., Kemp, K. and Campbell, S. (2015). Assessing quality of life in Crohn's disease: Development and validation of the Crohn's Life Impact Questionnaire (CLIQ). In *Quality of Life Research*, 24 (9): 2279-2288.