Selection of item response theory models for the longitudinal analysis of health-related quality of life in cancer clinical trials

Antoine Barbieri, Jean Peyhardi, Christian Lavergne, Thierry Conroy, Sophie Gourgou, Caroline Mollevi

▶ **To cite this version:**

Antoine Barbieri, Jean Peyhardi, Christian Lavergne, Thierry Conroy, Sophie Gourgou, et al.. Selection of item response theory models for the longitudinal analysis of health-related quality of life in cancer clinical trials. 2015. hal-01246368v1

HAL Id: hal-01246368

https://hal.science/hal-01246368v1

Preprint submitted on 18 Dec 2015 (v1), last revised 21 Nov 2016 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Selection of item response theory models for the longitudinal analysis of health-related quality of life in cancer clinical trials

Antoine Barbieri[*,1,2], Jean Peyhardi[1,3], Christian Lavergne[1,4], Thierry Conroy[5,6], Sophie Gourgou[2] and Caroline Mollevi[2]

Abstract

Statistical researches regarding health-related quality of life is a major challenge to better evaluate the impact of the treatments on their everyday life and to improve patients' care. In the literature, the mixed models based on the item response theory (IRT) are proposed to analyze directly HRQoL data from the questionnaires given to the patients. First, we use a recent classification of regression models for categorical data to discuss about a selection of IRT models for the longitudinal analysis of health-related quality of life in cancer clinical trials. Through methodological and practical arguments and an illustration on real data, the adjacent and cumulative models seem particularly suitable for this specific application. Then, a simulation study is carried out to compare the linear mixed model classically used to the most suitable proposed models. These simulations are a complement of other works concerning the comparison between classical test theory models and IRT models. This study is performed on the random part of mixed models and shows the IRT models are more precise. In opposite to the linear mixed model currently used, the IRT models are sensitive to the model used to generate the data. Between two IRT models used on the data, we recommend to consider that which does not take into account the random effect. If both considered the random effect, the choice of model depends to preference of user following the arguments of the first part of this work.

Keywords: Health-related quality of life; Item response theory; Mixed models; Ordinal categorical data; Longitudinal analysis.

1 Introduction

Endpoints refer to biological and clinical measures to assess the efficiency of new therapeutic strategies. The overall survival endpoint is the gold standard to show a clinical benefit of these strategies and treatments. Therapeutic treatments being more efficient and increasing the patients' lifetime, the overall survival endpoint may become insufficient to show a significant difference between two treatments. It is then necessary to consider a longer follow-up or a larger cohort of patients to have a sufficient number of events and a good statistical power[15], both representing considerable costs. Thus, to conclude to the

[*]Corresponding author: Antoine.Barbieri@umontpellier.fr
[1]Institut Montpelliérain Alexander Grothendieck (IMAG), Université de Montpellier
[2]Institut régional du Cancer Montpellier (ICM) - Val d'Aurelle, Unité de Biométrie
[3]CIRAD, AGAP and Inria, Virtual Plants, France
[4]Université Paul-Valéry Montpellier 3, Montpellier, France
[5]National Quality of Life in Oncology Platform, France
[6]Institut de Cancérologie de Lorraine, Nancy, France

benefit of a new treatment, other endpoints have emerged and the health-related quality of life (HRQoL) is currently one of the most important. In cancer clinical trials, the patient-reported outcomes are increasingly used to analyze a clinical benefit for medical decision-making[15]. The HRQoL endpoint may seem more pertinent to show the interest of a new therapy in some cases such as the palliative or geriatric situations. However, there are conceptual and methodological brakes underlying to the concept and the assessment of HRQoL. First of all, HRQoL is a multidimensional concept regarding the physical, psychological and social functions as well as symptoms associated with the disease and treatments. Another conceptual brake is the subjectivity of its measurement. Indeed, patients report their feelings about their HRQoL thanks to self-reported questionnaires. Both arguments preclude the use of HRQoL as sole primary endpoint in clinical trials.

In oncology, HRQoL is assessed using a general questionnaire for a set of different cancers, and an additional specific questionnaire associated with each type of cancer[1, 9]. Each questionnaire decomposes the HRQoL to measure several under-concepts (dimensions of HRQoL) which themselves comprise one or several items. The items are built on the Likert scales in which the response categories are ordinal categorical. Thus, considering several items for a given dimension, HRQoL data are composed of multiple ordinal responses. Also, the questionnaires are filled by the subjects themselves, and collected at different times defined in the trial protocol (usually at inclusion, during treatment and follow-up). These repeated measures are used to assess the evolution of the subject's HRQoL over time. In Europe, these questionnaires are developed and validated by the European organization for research and treatment of cancer (EORTC). The standard questionnaire currently used in oncology is the EORTC Quality of Life Questionnaire - Core 30 (EORTC QLQ-C30) [1], together with the scoring procedure proposed by the EORTC[14]. The score is then calculated for each dimension and for each subject, corresponding to the average of the item responses for a single dimension, and expressed on a scale ranging from 0 to 100. The interpretation is such that high functional scores reflect good functional capacities and a good HRQoL level, and conversely, high symptomatic scores represent strong symptoms and point out difficulties. The use of scoring procedures is common in practice because the statistical methods for quantitative variables are more powerful and easier to implement and interpret [19]. But, in a Likert scale, the gap which separates each adjacent category of response ("not at all", "a little", "quite a bit" and "very much") may not be the same, and the HRQoL score calculation does not take into account this characteristic.

The longitudinal statistical models classically used in oncology are performed on the summary score through using the linear mixed models (LMM) or time-to-event models[3]. In the LMM, the variable associated with the HRQoL score is considered as a Gaussian variable while it presents the characteristics of an ordinal variable, being non-continuous and bounded. These models allow taking into account the correlation introduced by repeated measurements on the same patient (collection of the HRQoL questionnaires over time) and different covariates such as time, treatment group, age... However, the use of the LMM for HRQoL analysis is scientifically questionable, due to obvious violations in their key Gaussian homoscedastic distributional assumption. Also, the LMM do not consider the bounded characteristic of the HRQoL score, ranged between 0 and 100. Moreover, many symptomatic dimensions are composed of only one item, the HRQoL score has exactly the same properties than ordinal categorical data, and using the LMM is not appropriated. Thus, if a ceiling or floor effect is observed, the categorical feature is even more marked when one of the two extreme categories is over-represented. The second approach for the longitudinal analysis of HRQoL is based on the time-to-event models: the time-to-deterioration (TTD) and the time-until-definitive-deterioration (TUDD) [20]. Survival approaches are often used and

thus well-known in the oncologic field, and are appreciated for their easiness to interpret result and their good understanding by clinicians. In these models, an event is classically defined by the (definitive or not) deterioration of the HRQoL score between baseline and a follow-up time, given a minimally clinically important difference [4]. The lack of homogeneity of the methods used for the HRQoL data analyses in different oncology clinical trials is also a real obstacle to the comparison of results. Indeed, the LMM and TUDD approaches show results which may sometimes seem contradictory, and with different interpretations, but they may also be complementary. An example can be taken comparing two similar cancer clinical trials investigating the effect of bevacizumab. In the first trial [10], HRQoL analysis through TUDD showed that the bevacizumab group had a later deterioration of HRQoL compared with patients in the standard group. Conversely, in the second trial [16], HRQoL analysis using the LMM showed a worse HRQoL overtime in the bevacizumab group.

Interest in the HRQoL endpoint is growing rapidly in cancer clinical trials and the standardization of its analysis has become an important challenge to make robust hypotheses for the design of future trials. It is essential to find a suitable methodology to analyze HRQoL data, taking into account the data properties (repeated measurements of the ordinal multiple responses). In our study, we first focused on the different and most adapted models to analyze HRQoL from raw data, *i.e.* directly on the item responses. Studies on psychometric properties from questionnaires such as the one used for HRQoL have been ongoing for a long time [12, 26], known as the item response theory (IRT). The IRT models link the individual's item responses and the latent variable which represents the studied HRQoL concept. They are generalized linear mixed models (GLMM) for ordinal data with a particular parameterization of the linear predictor. The interest for this kind of model to analyze the data, including the longitudinal analyzes, is growing [34, 22, 17]. However, to our knowledge, there is no work that discusses of the choice of one of the different IRT models over the others for HRQoL longitudinal analysis. We propose in section 2 a selection of these models, through the clinical expectations and practical and methodological arguments. This section is illustrated though the use of a multicenter randomized phase III clinical trial in first-line metastatic pancreatic cancer patients. Section 3 then presents a simulation study to compare the models described in section 2 and the classical LMM. As some comparison simulations have already been performed between the IRT models and the LMM on their capacity to detect fixed effects, we focus in section 3 on the sensibility of these models to detect the random effects.

2 Mixed models for the analysis of HRQoL raw data

HRQoL raw data are repeated measurements of ordinal multiple responses. The GLMM for ordinal categorical data seem well suitable to analyze this kind of data. The use of random effects takes into account the inter-patient variability and the correlation between the repeated measurements for each single patient. Recently, the IRT models have been increasingly used to analyze data deriving from self-questionnaires made of polytomous responses [21, 3]. These models are GLMM for ordinal data with a specific parameterization of the linear predictor due to the taking into account of the multiple responses. For ordinal data, three families of regression models are described: the families of adjacent models [28, 2], cumulative models [33, 29], and sequential models [35, 13]. Each model can be defined by the choice of three components, the ratio of probabilities, the cumulative distribution function (CdF), and the parameterization of the linear predictor [31]. Concerning HRQoL longitudinal analysis, we focus in this section on the mixed regression models for ordinal data with parameterization from the IRT. Many

IRT models are proposed for the analysis of this kind of data, often with no explanation regarding the choice of one model over another. After a presentation of the IRT parameterization used concerning the linear predictor, we compare different polytomous IRT models on the basis of the link function (ratio of probabilities and the cumulative distribution function) using methodological and practical arguments.

2.1 The IRT parameterization of the linear predictor

The IRT probabilistic models emerged following the works of Georg Rasch [32] on dichotomous responses, and were then extended to polytomous ordinal responses [33, 28, 35]. These models link the individual's item responses to the latent variable which represents a concept not directly measurable. In our case, the concept is HRQoL, and more precisely one specific HRQoL dimension. However, general IRT models presented in the literature do not consider covariates and are not appropriate for longitudinal analyses, even if it now begins to be explored [27, 36, 3, 17]. Considering the three families of adjacent, cumulative and sequential models, there are three associated famous IRT models [8, 5], respectively the (generalized) partial credit model [28, 30], the graded response model [33], and the sequential model [35].

Let $\pi_{im}^{(j)} = \Pr\left(Y_{ij} = m|\eta_{im}^{(j)}\right)$ denote the probability to response the category $m \in \{0, \ldots, M_j\}$ for the subject i at the item j given the linear predictor $\eta_{im}^{(j)}$. The IRT is based on the assumption that a concept exists but it cannot be directly measured. It is assumed that the item or the set of several associated items measures one unidimensional latent concept θ, which represents one HRQoL dimension. From the IRT, the specific parameterization of the linear predictor $\eta_{im}^{(j)}$ is built into two parts: the individual part and the item part. The best-known way is to consider the following decomposition:

$$\eta_{im}^{(j)} = \alpha_j \left(\theta_i - \delta_{jm}\right), \tag{1}$$

where θ_i is associated with an unidimensional variable representing the latent value for the subject i, δ_{jm} and α_j being the item parameters. Generally called "difficulty parameter", δ_{jm} is the intercept associated with the item j for the category $m \in \{1, \ldots, M_j\}$. The parameter α_j is called the "discrimination parameter" of item j, and represents the sensitivity of each probability of response according to the value of the latent trait. In HRQoL study in oncology, the analysis is carried out for each HRQoL dimension. Given one HRQoL dimension, the discrimination parameters could be considered equals to one for each item associated with this HRQoL dimension. This corresponds to the following constraint parameterization:

$$\eta_{im}^{(j)} = \theta_i - \delta_{jm}. \tag{2}$$

These IRT models with no discrimination parameters take into account the multiple-item responses and the categorical data. Concerning the longitudinal analysis, several studies proposed to extend some IRT models using the linear decomposition of the latent variable θ with fixed and random effects [21, 36, 24]:

$$\theta_{iv} = X_{iv}\boldsymbol{\beta} + U_v\boldsymbol{\xi}_i, \tag{3}$$

with $\boldsymbol{\beta}$ associated with the fixed effects, $\boldsymbol{\xi_i}$ with the subject-specific random effects and the index v the current visit. In the equation (3), the latent variable θ is thus not considered as random effect but as a linear combination of mixed effects. For the whole parameterization presented in this study, the slope θ is common for all response categories. Given the common parameterization of the linear predictor,

each specific model is defined by its link function which relates the linear predictor and the probability of response. It is characterized by the two remaining components, the ratio of probabilities (r) and the CdF (F).

2.2 The probability ratio: structure of the models

The linear predictor is not directly related to the response probability but to a particular transformation ratio previously mentioned [31].

Concerning the structure of the models, the GLMM for ordinal data are considered, and thus the ordinal assumption among categories excluded the family of reference models defined for nominal responses, which is not presented. Let's consider the simple situation with one item with $(M+1)$ response categories and π_m the probability to choose the particular category $m \in \{0, \ldots, M-1\}$. $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_{M-1})$ is the truncated vector of probabilities of responses. The three model families for ordinal data are distinguished by the choice of the ratio of probabilities $\boldsymbol{r}(\boldsymbol{\pi}) = (r_0(\boldsymbol{\pi}), \ldots, r_{M-1}(\boldsymbol{\pi}))$. Each model is summarized by M equations $\{r_m(\boldsymbol{\pi}) = F(\eta_m^\star)\}_{m=0,\ldots,M-1}$ with $\eta_m^\star = \delta_m - \theta$, highlighting the decomposition of the link function into the ratio of probabilities and the CdF. Indeed, we may distinguish different ratios of probabilities for these different families, respectively, for the adjacent models,

$$r_m(\boldsymbol{\pi}) = \frac{\pi_m}{\pi_m + \pi_{m+1}}, \quad m = 0, \ldots, M-1; \tag{4}$$

for the cumulative models,

$$r_m(\boldsymbol{\pi}) = \pi_0 + \ldots + \pi_m, \quad m = 0, \ldots, M-1; \tag{5}$$

and, for the sequential models,

$$r_m(\boldsymbol{\pi}) = \frac{\pi_m}{\pi_m + \ldots + \pi_M}, \quad m = 0, \ldots, M-1.$$

In the literature, the ratios of adjacent and cumulative families for the IRT models are usually presented given the reverse permutation [33, 28, 5]. This permutation is defined as the reversal of category order [29]. Assuming that the considered CdF is symmetric, these models are invariant under this permutation [31]. In practice, this can be seen as an advantage for result interpretation. A lower item-response category reflects a lower level of the symptomatic dimensions whereas it represents a higher level of capacity for the functional dimensions. The reverse permutation for the functional dimensions makes it easier and intuitive for clinicians to present their results. This allows the homogenization of the result interpretation as it is done in the scoring procedure proposed by the EORTC (for functional dimensions, the score scale is reversed compared with the item responses categories order) [14]. Even if the sequential model seems suitable for HRQoL data, we preferred considering the adjacent and cumulative models given this practical argument with the reverse permutation. The choice of the F component is discussed in a general way in the next subsection. Based on this previous argument, the adjacent and cumulative models are thus preferred for the HRQoL raw data analysis.

The cumulative models also have additional properties, including that they are invariant when successive categories are gathered [29]. Thus, if one category is not observed, it can be combined with its successive categories without changing the model.

Table 1: Summary of the characteristics for the three model families

	Models		
	Adjacent	Cumulative	Sequential
Reversibility	yes	yes	no
Interpretation using the latent variable	no	yes	yes
Always defined	yes	$yes(no^1)$	yes

[1]: for some non odd proportional models

A first comparison between the two families of models can be carried out depending on the linear predictor parameterization used. Indeed, we differentiated the odd proportional models from the others. For the odd proportional models, a common slope (θ) is considered for all categories, else the slope is dependent of the category (θ_m). An advantage of the adjacent models is that there is no constraint affecting the model estimation. Nonetheless, the cumulative models have to respect constraints, which can make difficult the model estimation, particularly in the case of non-odd proportional models [31]. Let the simple parameterization of the linear predictor $\eta_m = \theta_m - \delta_m$ for $m \in \{1, \ldots, M\}$ where θ_m and δ_m are the slope and the intercept associated with the category m, respectively. Considering the odd proportional models ($\theta = \theta_1 = \ldots = \theta_M$), the cumulative models refer to the principle of thresholds [29, 23] with the constraint they have to be strictly increasing such as $-\infty < \delta_1 < \ldots < \delta_M < +\infty$. Considering the non-odd proportional models, the constraint then becomes $-\infty < \eta_M < \ldots < \eta_1 < +\infty$ which is more difficult to verify. In practice, for the longitudinal analysis of HRQoL data in oncology, odd proportional models are considered which makes them more easily accessible and understood. An advantage of the cumulative models is their interpretation through a continuous latent variable [29]. Indeed, the continuous latent variable $\tilde{Y}$ underlying the model exists and allows a direct link with the response variable Y, given the thresholds presumed to be strictly increasing ($-\infty = \delta_0 < \delta_1 < \ldots < \delta_M < \delta_{M+1} = +\infty$):

$$\Pr\left(Y = m|\theta, \boldsymbol{\delta}\right) = \Pr\left(\delta_m < \tilde{Y} \leq \delta_{m+1}\right), \quad m = 0, \ldots, M.$$

where $\tilde{Y} = \theta + \varepsilon$ and ε is the error term distributed following the CdF. Here, the latent variable $\tilde{Y}$ represents HRQoL and its interpretation is then equivalent to the interpretation of the response variable using a LMM. The cumulative models also allow the comparison of the proportions of the response categories for one specific item over time, or between different groups during a given time.

Table 1 summarizes some properties of these three families of models which are important for the analysis of HRQoL in clinical trials in oncology. In the HRQoL longitudinal analysis, odd proportional models are preferred. Under this parameterization, there is no difficulty to respect the cumulative models constraints and to estimate them. The adjacent models seem more flexible statistically but the interpretation of the results using the cumulative model is more intuitive. The sequential models is not taken into account in the next sections because they are not invariant under the reverse permutation. The adjacent and cumulative models, which both seem the most suitable for the longitudinal analysis of HRQoL, are then considered in the next sections.

2.3 The cumulative distribution function

The aim of the following section is to present the choice of model essentially based on the selection of the CdF (F), for the adjacent and cumulative model families. This choice is especially used to best fit the data [31]. We considered four CdF from two different kinds: the most commonly used symmetric distributions, the Logistic and Gaussian distributions (respectively associated with the logit link and the probit link functions), and the two well-known asymmetric distributions, the Gumbel min and Gumbel max distributions. The two later distributions are respectively defined by $F(u) = exp(-exp(-u))$ for the Gumbel max distribution and by $F(u) = 1 - exp(-exp(u))$ for the Gumbel min distribution. In the literature, the cumulative model is presented according to the use of several of the previously mentioned CdF [33, 13, 27], while the adjacent models are most often presented with the Logistic CdF. There is also equivalence between the adjacent models with Logistic CdF and the multinomial logit model for nominal data [2], which may question taking into account the ordinal characteristics of these models [31].

The choice of F with the adjacent and cumulative models was carried out through real data. A multicenter randomized phase III clinical trial [11], in first-line metastatic pancreatic cancer patients, PRODIGE4/ACCORD11, was used to illustrate the CdF selection. Three hundred and forty-two patients were randomly assigned to FOLFIRINOX (experimental arm) versus Gemcitabine (control arm) regimens. The detailed inclusion and exclusion criteria, the study design and protocol, the treatment, the compliance to the questionnaires, and the HRQoL analyses have previously been published [11, 18]. The patients filled the EORTC QLQ-C30 questionnaire themselves at different follow-up times defined in the protocol: at baseline, day 15, day 30, and at months 2, 4, 6, 8, and 10. The different measuring times reflected the longitudinal aspect of the HRQoL and allowed the assessment of the change of HRQoL for each dimension. The six following HRQoL dimensions were taken into account to represent all possible configurations concerning the number of items for each dimension: the global health status (QL2) which comprises two items with seven response categories; the physical functioning (PF2) with five items with four response categories; the cognitive functioning (CF) and the pain symptom (PA) with two items with four response categories; the fatigue (FA) and the diarrhea symptoms (DI) with respectively three and one items, all with four response categories. These HRQoL dimensions were analyzed independently. An exhaustive analysis of all HRQoL dimensions was previous published [6].

As mentioned previously, the adjacent models have always been presented using the Logistic CdF [28, 30, 13, 2, 21, 3]. But, the different response probabilities can be presented from the adjacent ratio and according to a general CdF (F). As commonly seen in the IRT models, and because of the reverse permutation property, the adjacent model is used in descending order. Let's $F_m = F(\eta_m)$ whatever $m \in \{1, \ldots, M\}$, the response probabilities for the adjacent model are deduced as:

$$\begin{cases} \pi_0 = & \dfrac{1}{1 + \sum_{m=1}^{M} \prod_{k=1}^{m} \left(\frac{F_k}{1-F_k}\right)} \\ \pi_m = & \dfrac{\prod_{k=1}^{m} \left(\frac{F_k}{1-F_k}\right)}{1 + \sum_{m=1}^{M} \prod_{k=1}^{m} \left(\frac{F_k}{1-F_k}\right)}, \quad m = 1, \ldots, M \end{cases} \tag{6}$$

For the cumulative model, the probabilities are defined from the equation (5) and given the CdF F as:

$$\left\{\begin{array}{rcl} \pi_0 & = & 1 - F(\eta_1) \\ \pi_m & = & F(\eta_m) - F(\eta_{m+1}), \quad m = 1, M-1 \\ \pi_M & = & F(\eta_M) \end{array}\right. \tag{7}$$

Regarding the HRQoL longitudinal analysis with the individual i, the visit v, the item j and its number of response categories M_j, the following linear predictor is considered in the analyses:

$$\left\{\begin{array}{l} \eta_{ivm}^{(j)} = \theta_{iv} - \delta_{jm}, \quad m = 1, \ldots, M_j \\ \theta_{iv} = (t_v - t_0)\beta_1 + x_i (t_v - t_0)\beta_2 + \xi_{i0} + (t_v - t_0)\xi_{i1} \end{array}\right. \tag{8}$$

where:

- t_v is the date of the visit v, and t_0 is the date of baseline
- $x_i = 1$ if the patient i belongs the experimental group, $x_i = 0$ if the patient i belongs the control group
- β_1 is the slope (HRQoL evolution) associated with the control group
- $\beta_1 + \beta_2$ is the slope (HRQoL evolution) associated with the experimental group
- ξ_{i0} and ξ_{i1} are respectively the subject-specific random effects associated with the intercept and the slope such as $(\xi_{i0}, \xi_{i1})' \sim \mathcal{N}(\mathbf{0}, \Sigma)$, Σ being the unstructured covariance matrix

The fixed effect associated with the treatment arm is not considered (group effect at the baseline (t_0)) because there should be no difference at baseline in a randomized clinical trial.

Given the use of the same linear predictor parameterization (equation (8)), the most suitable model choice was based on the likelihood value. Analyses were performed using the SAS software (version 9.3) [25, 8]. Table 2 shows the different values of -2ℓ (ℓ representing the log-likelihood) according to the model and CdF used. Of course, the choice of F depends on the data analyzed. Concerning the fatigue and pain symptoms, the cumulative model with a Logistic CdF was chosen, while the Gumbel max CdF was selected for the physical functioning (adjacent model) and the global health status (cumulative model). The cumulative models with a Gaussian CdF and Gumbel min CdF were respectively preferred for the analysis of the cognitive functioning and the diarrhea symptom. Moreover, in most cases, the cumulative model was preferred to the adjacent model.

Of note, whatever the choice of CdF, the HRQoL evolution was similar. The different CdF allowed testing different slopes of F in function of theta. The results also showed that there were some difficulties to estimate the model when the Gumbel distributions were used, and particularly for the Gumbel min CdF. However, the reverse permutation property for the advantage of the intuitive interpretation in practice was due to the symmetric CdF. Thus, the Logistic or the Gaussian CdF was preferred.

The constraints on the item parameter in the cumulative models allow an interpretation through the latent variable (e.g. comparing the proportions of the response categories for one specific item over time or between different groups during a fixed time). Figure 1 illustrates the evolution concerning the probability of response either over time (Figure 1a) or between group (Figure 1b). This example is based on the first item of the pain symptom of the clinical trial previously presented. The probability (π_m) for

Table 2: The estimated -2ℓ associated with adjacent and cumulative models with different cumulative distribution functions F. Six HRQoL dimensions are used: Global health status (QL2), physical functioning (PF2), Cognitive functioning (CF), Fatigue (FA), Pain symptom (PA) and Diarrhea symptom (DI). The reported results are. For the empty boxes, the model did not converge.

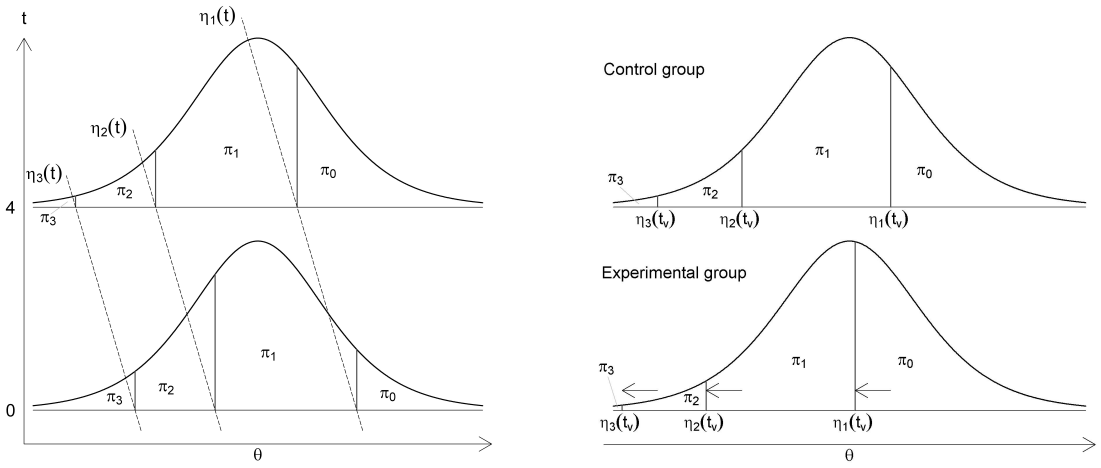
Family of models	CdF F	HRQoL dimensions					
		QL2	PF2	CF	FA	PA	DI
Adjacent Models	Logistic	6633.2	9319.2	3826.5	6930.0	5008.8	2238.2
	Gaussian	6651.4	9315.7	3823.5	6948.4	5012.0	2237.8
	Gumbel min	-	9400.5	3834.5	-	5014.6	2239.8
	Gumbel Max	6643.2	9299.4	3819.7	6927.0	5015.6	2239.2
Cumulative Models	Logistic	6558.1	9323.5	3818.0	6913.4	4998.0	2220.5
	Gaussian	6628.0	9331.1	3812.5	6947.0	5009.4	2227.8
	Gumbel min	6642.2	9500.6	3835.4	6990.0	5004.9	2216.5
	Gumbel Max	6553.9	9309.8	3821.0	6924.9	5031.9	2246.8

a patient to response the category m corresponds to the area under the curve delimited by the vertical lines. For the control group, Figure 1a shows that the probability for a patient to choose the categories 2 or 3 decreased over time while the probability to choose the category 0 increased. At baseline, the response proportion for the categories 0, 1, 2 and 3 were respectively $\pi_0 = 0.10$, $\pi_1 = 0.62$, $\pi_2 = 0.22$ and $\pi_3 = 0.06$ while after four months they were $\pi_0 = 0.29$, $\pi_1 = 0.61$, $\pi_2 = 0.08$ and $\pi_3 = 0.02$, respectively. The evolution of the proportions showed a decrease of the level of pain between the baseline and the 4-month visit, and, finally, a decrease of the latent trait over time. Likewise, Figure 1b shows the different response proportions between the two groups at four months. In the control group, the proportions were $\pi_0 = 0.29$, $\pi_1 = 0.61$, $\pi_2 = 0.08$ and $\pi_3 = 0.02$ for the categories 0, 1, 2 and 3, respectively, and $\pi_0 = 0.47$, $\pi_1 = 0.48$, $\pi_2 = 0.04$ and $\pi_3 = 0.01$, respectively, in the experimental group. The probability to response category 3 was the lowest whatever the group, but was even less likely for patients in experimental group than in control group. Contrarily, the probability to response category 0 was more likely in experimental group than in control group. The lag observed between the different borders corresponded to the difference between the two linear predictors associated with each group for a fixed time. One of the interests of this illustration concerns the clinical interpretation. The IRT models thus offer a complete analysis: the general analysis of a HRQoL dimension and the specific analysis of a specific item [12].

3 Study of simulations

In the previous section, we focused on the use of the mixed models for ordinal data analysis and their relevance in the HRQoL analysis in oncology was discussed. Some comparisons studies exist between these different approaches [7, 3], mainly on the fixed part of the mixed models. Anota et Al.[3] had shown an equivalent capacity to detect a fixed effect for the LMM and for one of the IRT models. Indeed, even if the LMM take into account the HRQoL score, which is a summary variable, this approach is at least equivalent to the IRT models in terms of power.

In our study, the adjacent and cumulative models with the same parameterization of the linear pre-



(a) *Evolution over time* ***(b)*** *Difference between groups*

Figure 1: *Interpretation of the cumulative model (with Logistic CdF) through its underlying latent variable for the first item concerning the pain symptom. The associated estimates are* $\beta_1 = -0.33$*,* $\beta_2 = -0.19$*,* $\delta_1 = -2.1$*,* $\delta_2 = 1$ *and* $\delta_3 = 2.75$*.* $\eta_m(t)$*, is the fixed part of the linear predictor given the category* $m \in \{0, 1, 2, 3\}$ *where* $t = t_v - t_0$*. (a): the different proportions of different responses (*π_m*) between two visits (*t_0 *and* t_4*) in the control group. (b): the different proportions of different responses (*π_m*) between the two groups for one given visit (*t_4*).*

dictor and the Logistic CdF were used (as usually in the IRT models). As presented by Peyhardi et al. [31] concerning the model for categorical data, the two IRT models can be defined by three components $(r; F; Z)$ such as $(adjacent; logistic; proportional)$ and $(cumulative; logistic; proportional)$ for the adjacent and the cumulative models, respectively.

The aim of the following section is to reinforce our study adding a comparison, using a simulation study, between the LMM and the IRT models on the random part of the mixed models. The datasets were simulated from an IRT model (adjacent and cumulative models). Regarding the parameterization, two subject-specific random effects ξ_{i0} and ξ_{i1} were considered, respectively associated with the intercept and the slope (equation (8)). Of course, the usefulness of the random effect introduction in the model is strongly associated with the observed data. As mentioned earlier, HRQoL is a subjective endpoint, and the individual random effect ξ_{i0} is thus entirely justified. Indeed, it is easy to imagine that each patient has a different level of HRQoL at baseline. The random slope is more questionable, indeed, the assumption that the specific HRQoL evolution of one single patient diverges from the average evolution for the whole population, is less obvious than the previous one. In this section, the capacity of the mixed models to detect the slope random effect was thus studied. No group effect was considered in this simulation study.

3.1 Design

The aim of the following simulations is to study the capacity of each model to detect the random effect ξ_{i1} associated with time (random slope). The two subject-specific random effects are considered independent

Table 3: Values of difficulty parameters used to simulate the data and expected responses at t_0 under each studied scenarios.

	Difficulty parameters	
Models	$\delta_1^{ne} = (-1.6, 1, 1.45)$	$\delta_1^{fa} = (-2.1, 1, 2.75)$
(r, F, Z)	$\delta_2^{ne} = (-0.8, 1.15, 1.9)$	$\delta_2^{fa} = (-1.25, 1.4, 3.3)$
$(adjacent, logistic, proportional)$	balanced responses	focus on center categories (1 and 2)
$(cumulative, logistic, proportional)$	focus on extreme categories (0,1 and 3)	balanced responses

where $\xi_{i0} \sim \mathcal{N}(0, \sigma_0^2)$ and $\xi_{i1} \sim \mathcal{N}(0, \sigma_1^2)$. The following model choice study is performed on the basis of the Bayesian information criteria (BIC) where two models were considered: $\mathcal{M}_1$ with the two random effects (ξ_0, ξ_1) and $\mathcal{M}_0$ excluding the random slope (ξ_1). For the IRT models, the linear decomposition of the latent trait θ_{iv} only took into account the time as a fixed effect. The two considered models are:

$$\begin{aligned} \mathcal{M}_1 &: \theta_{iv} = (t_v - t_0)\,\beta_1 + \xi_{i0} + (t_v - t_0)\,\xi_{i1} \\ \mathcal{M}_0 &: \theta_{iv} = (t_v - t_0)\,\beta_1 + \xi_{i0} \end{aligned} \tag{9}$$

The most frequent HRQoL dimension with two items ($j = 1, 2$) comprising four response categories ($m \in \{0, \ldots, M\}$ with $M = 3$), was used to design the simulation study. A sample size of three hundred subjects ($i = 1, \ldots, n$ with $n = 300$) and eight follow-up time ($v = 0, \ldots, 7$), as for the trial presented in the previous section, were considered. The datasets were simulated from a multinomial distribution. The different response probabilities $\left\{\pi_{ivm}^{(j)} = \Pr\left(Y_{iv}^{(j)} = m|\theta_{iv}, \delta_j\right)\right\}$ concerning the subject i for item j were determined by equation (6) for the adjacent model and by equation (7) for the cumulative model, given:

- the item parameters $(\delta_{j1}, \delta_{j2}, \delta_{j3})_{j=1,2}$;
- the latent trait (θ_{iv}) deduced in accordance with equation (9);
- the Logistic CdF,

$$F\left(\eta_{ivm}^{(j)}\right) = \frac{\exp\left(\eta_{ivm}^{(j)}\right)}{1 + \exp\left(\eta_{ivm}^{(j)}\right)},$$

  where $\eta_{ivm}^{(j)} = \theta_{iv} - \delta_{jm}$.

The values of the parameters used were deduced from the pain symptom data of the clinical trial presented in the previous section. We considered two kinds of difficulty parameters: near $\delta^{ne} = (\delta_1^{ne}, \delta_2^{ne})$ and far $\delta^{fa} = (\delta_1^{fa}, \delta_2^{fa})$. These parameter values were chosen in order to illustrate several scenarios described in Table 3. The different scenarios are due with the different associations between the model used to simulate the data, $(adjacent, logistic, proportional)$ or $(cumulative, logistic, proportional)$, and the different values of the difficulty parameters considered. Table 3 shows the simulated responses expected at baseline ($t = 0$). The responses simulated across time depended of the considered coefficient β_1. Each scenario was simulated $N = 500$ times.

Concerning the LMM, the scoring procedure proposed by the EORTC was considered [14], and the score associated with a symptomatic dimension was first calculated using the simulated data. Let the

two simulated ordinal outcomes $Y_{iv}^{(1)}$ and $Y_{iv}^{(2)}$ concerning the individual i at the visit v, the related score was:

$$S_{iv} = \left(\frac{\sum_{j=1}^{J=2} Y_{iv}^{(j)}}{2} \right) \frac{100}{M}$$

Similarly to the parameterization in equation (9), we took into account the related choice model with:

$$\begin{aligned}
\mathcal{M}_1 : S_{iv} &= \beta_0^l + (t_v - t_0)\,\beta_1^l + \xi_{i0}^l + (t_v - t_0)\,\xi_{i1}^l + \varepsilon_{iv} \\
\mathcal{M}_0 : S_{iv} &= \beta_0^l + (t_v - t_0)\,\beta_1^l + \xi_{i0}^l + \varepsilon_{iv}
\end{aligned}$$

where β_0 is the fixed parameter associated with the intercept, the ξ are the random effects normally distributed with the mean equals to zero and $\varepsilon_{iv} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ the error term.

3.2 Results

Table 4 shows the capacity of the three models (adjacent model, cumulative model and LMM) to detect the random slope given different scenarios (Table 1). When we simulated the data under $\mathcal{M}_1$ (with a random slope) according to the random effect variances estimated from real data, each model detected the random slope (ξ_{i1}) in 100% of cases whatever the different situations. On the contrary, under $\mathcal{M}_0$, the simulated model $\mathcal{M}_0$ was correctly chosen in most cases. For all simulations under $\mathcal{M}_0$, the cumulative model seemed less suitable than the others because it seemed to detect the random slope although it was not included in the simulation step. Moreover, the IRT model, which was not use to simulate the data, wrongly detected this random effect given a negative value of β_1 and the difficulty parameter coefficients δ^{ne}. This could be explained by the fact that the difficulty parameters were not uniformly separated around zero and also because they were too close. Indeed, given $\beta_1 < 0$, the probabilities to observe the upper categories were very small over time and under-represented in comparison with the lower categories (as illustrated in Figure 1a). In the specific case where $\beta_1 = -0.3$, the IRT model which did not simulate the data could not explain the different outcomes only with the fixed effect and the random intercept, and it compensated the lack of information with the random slope. We then could expect symmetric results from β_1 (positive values) considering the opposite sign of the difficulty parameters because of the reverse permutation property of the IRT models.

On the contrary, the LMM was stable and thus allowed making the good choice of model whatever the β_1 values and the IRT model used to simulate the data. Concerning the IRT models and in the case where there was only one model out of the two detecting the random effect ξ_{i1}, the most suitable model seemed the one not detecting this random effect.

The capacity of the different models to detect the random slope when its variance value changes, is presented in Table 5. All models were sensitive to the signal-to-noise ratio. Indeed, the more β_1 increased, the less the random effect provided information. This was well characterized as the capacity to detect the random effect for greater variances when the signal was strong. In this case, the signal provided the essential information explaining the different responses. In the model comparison, the LMM was less sensitive than the IRT models. Indeed, the LMM detected the random slope for a greater variance of this one whatever the β_1 value. This result was expected because the LMM is based on the HRQoL score which is a summary variable with less information than the raw data. Thus, the IRT models are more sensitive in all cases. Comparing the two IRT models, the cumulative model had a

Table 4: Frequency (on $N = 500$ datasets) of the $\mathcal{M}_0$ selection according to the BIC, given $t_v = (0, 1, 2, 4, 6, 8, 10, 12)$ and $\sigma_0^2 = 1.5$. The *(adjacent,logistic,proportional)* model and the *(cumulative,logistic,proportional)* model denoted respectively by AM and CM.

		Scenarios											
		AM using δ^{ne}			CM using δ^{fa}			CM using δ^{ne}			AM using δ^{fa}		
σ_1^2	β_1	LMM	AM	CM	LMM	AM	CM	LMM	AM	CM	LMM	AM	CM
0.2	-0.3	0	0	0	0	0	0	0	0	0	0	0	0
0.2	0.3	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.5	97.67	99.29	56.49	100	94.63	92.98	100	61.33	95.71	100	99.66	89.54
0	-0.3	99.00	100	33.04	100	88.63	93.30	100	36.33	94.91	100	100	83.33
0	-0.2	100	99.62	49.28	100	94.56	93.81	100	71.67	95.77	100	99.64	79.02
0	-0.1	98.67	95.65	94.78	100	98.65	89.62	100	98.98	90.41	100	100	88.10
0	0.0	95.60	100	94.55	99.00	99.66	91.75	99.00	99.66	89.71	97.00	99.66	94.42
0	0.1	83.00	100	94.78	93.33	100	92.63	97.00	100	90.91	87.33	100	94.69
0	0.3	98.33	99.64	90.61	100	99.64	89.05	100	100	93.67	100	99.65	93.78
0	0.5	100	100	94.29	100	99.32	94.71	100	100	97.61	100	100	97.19

Table 5: Frequency (on $N = 500$ datasets) of the $\mathcal{M}_1$ selection according to the BIC, given $t_v = (0, 1, 2, 4, 6, 8, 10, 12)$ and $\sigma_0^2 = 1.5$. The *(adjacent,logistic,proportional)* model and the *(cumulative,logistic,proportional)* model denoted respectively by AM and CM.

		Scenarios											
		AM using δ^{ne}			CM using δ^{fa}			CM using δ^{ne}			AM using δ^{fa}		
β_1	σ_1^2	LMM	AM	CM	LMM	AM	CM	LMM	AM	CM	LMM	AM	CM
	0.01	0	2.33	24.92	0	5.03	6.94	0	2.69	3.70	0.33	6.44	24.75
	0.02	0	21.40	54.67	0	37.58	44.11	0	17.73	18.12	0	50.00	77.00
	0.03	0	61.00	90.97	0	75.67	80.00	0	41.33	45.58	0	86.33	98.33
1	0.05	0	97.67	99.66	0	100	100	0.33	89.00	90.00	0	99.33	100
	0.2	39.33	100	100	40.67	100	100	10.67	100	100	57.67	100	100
	0.5	100	100	100	100	100	100	100	100	100	100	100	100
	0.002	16.67	6.33	21.40	0	2.03	3.97	0	3.06	3.94	11.00	11.04	15.25
	0.005	72.33	86.33	92.67	30.67	55.33	59.00	0	32.33	46.00	85.67	87.33	91.67
0.3	0.008	97.67	100	100	86.00	97.3	98.00	4.00	76.33	88.33	99.33	99.67	100
	0.01	100	100	100	96.33	99.67	99.33	17.33	94.00	97.00	100	100	100
	0.02	100	100	100	100	100	100	96.67	100	100	100	100	100
	0.002	0.67	4.36	61.43	0	54.00	5.09	0	93.33	1.82	0	2.07	18.57
-0.3	0.005	5.67	62.33	79.00	0	95.67	40.40	0	99.67	33.22	0	56.00	48.33
	0.008	23.67	96.33	97.33	0	100	86.67	0	100	82.67	1.67	96.33	86.33

better capacity than the adjacent model to detect the presence of the random slope. On the contrary, in the specific case where $\beta_1 = -0.3$, the IRT model used to simulate the data was less efficient than the other IRT model which detected a random slope to remedy the lack of information. This was coherent with our previous results shown in Table 4. Finally, the more β_1 was close to zero, the more the models detected the random slope for a low variance.

From the simulation results, we can propose a last argument in order to choose the best model. In case where only one of the two models detects the random slope, the model that does not detect the random effect should be preferred. This model is the most suitable to fit the data.

4 Discussion

We have explored the suitable mixed models for the longitudinal analysis of the HRQoL in oncology. This data coming from questionnaires through Likert scales, we focused on regression models for ordinal data. These models can be specified with three components, the linear predictor parameterization, the ratio of probabilities and the CdF [31]. In oncology, the analysis being performed on multiple-item measurements associated with one HRQoL dimension [14], the specific IRT parameterization of the linear predictor is thus used. The item parameters allow to distinguish the outcomes from different items which measure an unique unidimensional latent variable. This latent variable was decomposed linearly to take into account the different covariates in the fixed part of the model and to incorporate subject-specific random effects. The analysis with IRT models is the richer because they are based on raw data while LMM is based on summary variable commonly used [17]. The analysis can be made on one specific item through the item parameters or on the studied HRQoL dimension [12]. Indeed, these models take into consideration all available information from the data, it is why the use of this kind of model is more and more studied [17].

Then, concerning the choice of the model family, the cumulative and adjacent models are preferred. From the ratios of probabilities which characterize them and the symmetric CdF, the practical property of the reversibility seems important to interpret the results with clinicians in oncology. The cumulative models also assume an underlying continuous latent variable that is associated with a linear mixed regression model [29, 23]. This allows a better interpretation of the results such as the easy analysis of the evolution of the response proportions of the different categories over time or between groups, given one item. The adjacent models show the advantage not to have any constraint for the model estimation. Finally, the choice of the CdF essentially depends on the observed data and no specific rule exists yet. We here showed an example on clinical trial data where the model choice was performed thank to the likelihood. For the longitudinal analysis of HRQoL in oncology, we thus avoided the use of commonly symmetric CdF: the Logistic and the Gaussian distributions. We can precise that whatever CdF was used, there was a similar influence from the different factors.

The simulation study showed that the IRT model capacity to detect the random effect was better than the LMM currently used. This result seems natural because the LMM is based on the study of a summary variable with less information. Thus, the variability from data is also reduced. Of course, the usefulness of the random effect introduction in the model is strongly associated with the observed data. Moreover, the more the difficulty parameters were distinct and the influence of covariates was stronger, the less the random effect provided information. All these results confirmed that the IRT models allow a more detailed analysis to interpret the results from a specific dimension or item. Whatever the IRT model used to generate the data, the LMM remained competitive through these simulations. However, the IRT model, that did not generate data, seemed more sensitive to the random slope than the other IRT model used to simulate the dataset. Indeed, in some cases, it tended to detect the random slope while it did not exist. In case where one of the two models detects the random slope, we recommend the use of the model not detecting the effect as it is the most suitable to fit the data. On the opposite, if both models

are identically sensitive to its presence, we suggest using the IRT model following the previous practical and methodological arguments.

An aspect that remains to be discussed is the multidimensional aspect of HRQoL. Nowadays in oncology, the different dimensions are analyzed independently of one another, and this causes the problem of multiple tests. It is possible to imagine an unidimentional latent trait for two or several HRQoL dimensions. Indeed, the discrimination parameters could be considered as equal between items from the same dimension. We can thus think that this parameter can be associated with the dimension and no to with item. Finally, this allows to consider a set of HRQoL dimensions to assess a common concept like such as the symptomatic status concerning all symptomatic dimensions. Another approach to consider the all HRQoL dimensions would be the use of structural equation modeling. This would allow to summarize the information through some factors to explain the evolution of the global HRQoL and the potential structural link between the latent variables.

Acknowledgment

This study was supported by a grant from the French Public Health Research Institute (`www.iresp.net`) under the 2012 call for projects as part of the 2009-2013 Cancer Plan. We thank Dr. Hélène de Forges for her editorial assistance and UNICANCER for the data from PRODIGE4/ACCORD11 clinical trial which is used in this paper.

References

[1] N K Aaronson, S Ahmedzai, B Bergman, M Bullinger, A Cull, N J Duez, A Filiberti, H Flechtner, S B Fleishman, and J C de Haes. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute, 85(5):365–376, March 1993.

[2] Alan Agresti. Analysis of Ordinal Categorical Data. John Wiley & Sons, April 2010.

[3] Amélie Anota, Antoine Barbieri, Marion Savina, Alhousseiny Pam, Sophie Gourgou-Bourgade, Franck Bonnetain, and Caroline Bascoul-Mollevi. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. Health and Quality of Life Outcomes, 12, December 2014.

[4] Amélie Anota, Zeinab Hamidou, Sophie Paget-Bailly, Benoist Chibaudel, Caroline Bascoul-Mollevi, Pascal Auquier, Virginie Westeel, Frederic Fiteni, Christophe Borg, and Franck Bonnetain. 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? Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, November 2013.

[5] Silvia Bacci, Francesco Bartolucci, and Michela Gnaldi. A class of Multidimensional Latent Class IRT models for ordinal polytomous item responses. arXiv:1201.4667 [stat], January 2012. arXiv: 1201.4667.

[6] Antoine Barbieri, Amélie Anota, Thierry Conroy, Sophie Gourgou-Bourgade, Béata Juzyna, Franck Bonnetain, Christian Lavergne, and Caroline Bascoul-Mollevi. Applying longitudinal model from item response theory to assess the health-related quality of life in prodige 4 / accord 11 randomized trial. Accepted in Medical Decision Making, 2015.

[7] Myriam Blanchin, Jean-Benoit Hardouin, Tanguy Le Neel, Gildas Kubis, Claire Blanchard, Eric Mirallié, and Véronique Sébille. Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. Statistics in Medicine, 30(8):825–838, April 2011.

[8] Paul de Boeck and Mark Wilson. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer, New York, August 2004.

[9] D. F. Cella, D. S. Tulsky, G. Gray, B. Sarafian, E. Linn, A. Bonomi, M. Silberman, S. B. Yellen, P. Winicour, and J. Brannon. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 11(3):570–579, March 1993.

[10] Olivier L. Chinot, Wolfgang Wick, Warren Mason, Roger Henriksson, Frank Saran, Ryo Nishikawa, Antoine F. Carpentier, Khe Hoang-Xuan, Petr Kavan, Dana Cernea, Alba A. Brandes, Magalie Hilton, Lauren Abrey, and Timothy Cloughesy. Bevacizumab plus Radiotherapy–Temozolomide for Newly Diagnosed Glioblastoma. New England Journal of Medicine, 370(8):709–722, February 2014.

[11] Thierry Conroy, Françoise Desseigne, Marc Ychou, Olivier Bouché, Rosine Guimbaud, Yves Bécouarn, Antoine Adenis, Jean-Luc Raoul, Sophie Gourgou-Bourgade, Christelle de la Fouchardière, Jaafar Bennouna, Jean-Baptiste Bachet, Faiza Khemissa-Akouz, Denis Péré-Vergé, Catherine Delbaldo, Eric Assenat, Bruno Chauffert, Pierre Michel, Christine Montoto-Grillot, and Michel Ducreux. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. The New England journal of medicine, 364(19):1817–1825, May 2011.

[12] Maria Orlando Edelen and Bryce B. Reeve. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16 Suppl 1:5–18, 2007.

[13] Ludwig Fahrmeir and Gerhard Tutz. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, April 2001.

[14] P. M. Fayers, N. K. Aaronson, K. Bjordal, M. Groenvold, D. Curran, and A. on behalf of the EORTC Quality of Life Group. Bottomley. EORTC QLQ-C30 Scoring Manual (3rd edition), volume Brussels: EORTC 2001. 2001.

[15] F. Fiteni, V. Westeel, X. Pivot, C. Borg, D. Vernerey, and F. Bonnetain. Endpoints in cancer clinical trials. Journal of Visceral Surgery, 151(1):17–22, February 2014.

[16] Mark R. Gilbert, James J. Dignam, Terri S. Armstrong, Jeffrey S. Wefel, Deborah T. Blumenthal, Michael A. Vogelbaum, Howard Colman, Arnab Chakravarti, Stephanie Pugh, Minhee Won, Robert Jeraj, Paul D. Brown, Kurt A. Jaeckle, David Schiff, Volker W. Stieber, David G. Brachman, Maria

Werner-Wasik, Ivo W. Tremont-Lukats, Erik P. Sulman, Kenneth D. Aldape, Walter J. Curran, and Minesh P. Mehta. A Randomized Trial of Bevacizumab for Newly Diagnosed Glioblastoma. New England Journal of Medicine, 370:699–708, February 2014.

[17] Rosalie Gorter, Jean-Paul Fox, and Jos WR Twisk. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Medical Research Methodology, 15(1):55, July 2015.

[18] Sophie Gourgou-Bourgade, Caroline Bascoul-Mollevi, Françoise Desseigne, Marc Ychou, Olivier Bouché, Rosine Guimbaud, Yves Bécouarn, Antoine Adenis, Jean-Luc Raoul, Valérie Boige, Jocelyne Bérille, and Thierry Conroy. Impact of FOLFIRINOX Compared With Gemcitabine on Quality of Life in Patients With Metastatic Pancreatic Cancer: Results From the PRODIGE 4/ACCORD 11 Randomized Trial. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, December 2012.

[19] Leonardo Grilli and Carla Rampichini. Multilevel Models for Ordinal Data. In Ron S. Kenett and Silvia Salini, editors, Modern Analysis of Customer Surveys, pages 391–411. John Wiley & Sons, Ltd, 2011.

[20] Zeinab Hamidou, Tienhan S. Dabakuyo, Mariette Mercier, Jean Fraisse, Sylvain Causeret, Hervé Tixier, Marie-Martine Padeano, Catherine Loustalot, Jean Cuisenier, Jean-Marc Sauzedde, Marc Smail, Jean-Philibert Combier, Patrick Chevillote, Christian Rosburger, Patrick Arveux, and Franck Bonnetain. Time to Deterioration in Quality of Life Score as a Modality of Longitudinal Analysis in Patients with Breast Cancer. The Oncologist, 16(10):1458–1468, October 2011.

[21] Jean-Benoit Hardouin, Etienne Audureau, Alain Leplège, and Joël Coste. Spatio-temporal Rasch analysis of quality of life outcomes in the French general population. Measurement invariance and group comparisons. BMC Medical Research Methodology, 12(1):182, November 2012.

[22] Jean-Benoit Hardouin, Myriam Blanchin, Mohand-Larbi Feddag, Tanguy Le Néel, Bastien Perrot, and Véronique Sébille. Power and sample size determination for group comparison of patient-reported outcomes using polytomous Rasch models. Statistics in Medicine, pages n/a–n/a, March 2015.

[23] D. Hedeker and R. D. Gibbons. A random-effects ordinal regression model for multilevel analysis. Biometrics, 50(4):933–944, December 1994.

[24] Catherine Huber, Nikolaos Limnios, Mounir Mesbah, and Mikhail Nikulin. Mathematical Methods in Survival Analysis, Reliability and Quality of Life. John Wiley & Sons, March 2013.

[25] S. A. S. Institute. SAS/STAT 9.3 User's Guide: Mixed Modeling (Book Excerpt). SAS Institute, July 2011.

[26] Peyman Jafari, Zahra Bagheri, Seyyed MT Ayatollahi, and Zahra Soltani. Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQLTM 4.0 Generic Core Scales in school children. Health and Quality of Life Outcomes, 10(1):27, March 2012.

[27] Li C Liu and Donald Hedeker. A mixed-effects regression model for longitudinal multivariate ordinal data. Biometrics, 62(1):261–268, March 2006.

[28] Geoff Masters. A rasch model for partial credit scoring. Psychometrika, 42(2):149–174, 1982.

[29] P. MCCULLAGH. Regression models for ordinal data (with discussion). Journal of the Royal Statistical Society, Series B, 42:109–142, 1980.

[30] Eiji Muraki. A Generalized Partial Credit Model: Application of an EM Algorithm. Applied Psychological Measurement, 16(2):159–176, January 1992.

[31] J. Peyhardi, C. Trottier, and Y. Guédon. A new specification of generalized linear models for categorical responses. Biometrika, page asv042, November 2015.

[32] Georg Rasch. Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks Paedagogiske Institut, 1960.

[33] Fumi Samejima. Estimation of Latent Ability Using a Response Pattern of Graded Scores1. ETS Research Bulletin Series, 1968(1):i–169, 1968.

[34] Andrew C. Titman, Gillian A. Lancaster, and Allan F. Colver. Item response theory and structural equation modelling for ordinal data: Describing the relationship between KIDSCREEN and Life-H. Statistical Methods in Medical Research, October 2013.

[35] Gerhard Tutz. Sequential item response models with an ordered response. British Journal of Mathematical and Statistical Psychology, 43(1):39–55, 1990.

[36] Josine Verhagen and Jean-Paul Fox. Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. Statistics in Medicine, 32(17):2988–3005, July 2013.