



Longitudinal analysis of health-related quality of life in cancer clinical trials: methods and interpretation of results

Caroline Bascoul-Mollevi^{1,2,3} · Antoine Barbieri⁴ · Céline Bourcier^{5,2} · Thierry Conroy^{6,7} · Bruno Chauffert⁸ · Mohamed Hebbar⁹ · William Jacot^{10,2} · Beata Juzyna¹¹ · Hélène De Forges¹² · Sophie Gourgou^{1,3} · Franck Bonnetain^{3,13,14} · Célia Touraine^{1,3} · Amélie Anota^{3,13,14}

Accepted: 6 August 2020 / Published online: 18 August 2020
© Springer Nature Switzerland AG 2020

Abstract

Purpose Health-related quality of life (HRQoL) is assessed by self-administered questionnaires throughout the care process. Classically, two longitudinal statistical approaches were mainly used to study HRQoL: linear mixed models (LMM) or time-to-event models for time to deterioration/time until definitive deterioration (TTD/TUDD). Recently, an alternative strategy based on generalized linear mixed models for categorical data has also been proposed: the longitudinal partial credit model (LPCM). The objective of this article is to evaluate these methods and to propose recommendations to standardize longitudinal analysis of HRQoL data in cancer clinical trials.

Methods The three methods are first described and compared through statistical, methodological, and practical arguments, then applied on real HRQoL data from clinical cancer trials or published prospective databases. In total, seven French studies from a collaborating group were selected with longitudinal collection of QLQ-C30. Longitudinal analyses were performed with the three approaches using SAS, Stata and R software.

Results We observed concordant results between LMM and LPCM. However, discordant results were observed when we considered the TTD/TUDD approach compared to the two previous methods. According to methodological and practical arguments discussed, the approaches seem to provide additional information and complementary interpretations. LMM and LPCM are the most powerful methods on simulated data, while the TTD/TUDD approach gives more clinically understandable results. Finally, for single-item scales, LPCM is more appropriate.

Conclusion These results pledge for the recommendation to use of both the LMM and TTD/TUDD longitudinal methods, except for single-item scales, establishing them as the consensual methods for publications reporting HRQoL.

Keywords Health-related quality of life · Cancer · Clinical trial · Longitudinal analysis · Generalized linear mixed model · Time to deterioration

Introduction

Health-related quality of life (HRQoL) is a multidimensional, subjective, and dynamic concept, incorporating at least three functional domains: physical, psychological, and social functioning, as well as symptoms due to disease and treatment [1].

In cancer clinical trials, HRQoL is often prospectively assessed using the European Organization for Research and Treatment of Cancer (EORTC) Core Quality of Life Questionnaire (QLQ-C30) [2]. The QLQ-C30 is a widely used cancer questionnaire composed of 30 ordinal items assessing 15 scales: global health status/QoL (GHS/QoL), five functional domains (physical, role, cognitive, emotional and social), three multi-item symptoms scales (fatigue, pain, nausea and vomiting), and six single-items symptom scales (diarrhea, constipation, insomnia, appetite loss, dyspnea, and perceived financial impact). For each domain, a raw score is first estimated as the average of all contributing items then standardized by linear transformation on a scale from 0 to 100 according to the scoring procedure recommended

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11136-020-02605-3>) contains supplementary material, which is available to authorized users.

✉ Caroline Bascoul-Mollevi
Caroline.Mollevi@icm.unicancer.fr

Extended author information available on the last page of the article

by the EORTC [3]. A high score on functional scales and GHS/QoL represents a high/healthy level of function and high global HRQoL, respectively, whereas a high score on symptom scales indicates a high level of symptomatology. The QLQ-C30 is now used in numerous studies, enabling comparison of results. It is also often associated with disease-specific modules.

In cancer clinical trials, HRQoL is generally collected at different assessment times predefined in the study protocol. To evaluate the impact of the treatment on the change in HRQoL over time, at least three assessments of HRQoL are recommended [4, 5]: at baseline (before the start of treatment), during treatment, and at the end of treatment.

Using appropriate methods to analyze such longitudinal data is essential, but the analysis strategies used are still not homogenous [4, 6]. The choice of an appropriate methodology for longitudinal analyses of HRQoL would enable homogenization of results across different therapeutic situations and tumor sites, thereby ensuring greater comparability of the results between trials [7, 8].

Two main longitudinal statistical approaches are used to analyze HRQoL in cancer clinical trials, namely linear mixed models (LMM) [9] or time-to-event modeling [10]. Time-to-event modeling, time-to-HRQoL score(s) deterioration, seems more accessible and intuitive for clinicians, but LMM are more widely used in practice, even though they are more complex. Both approaches use the standardized score recommended by the EORTC. In recent years, the longitudinal partial credit model (LPCM), an alternative strategy based on generalized linear mixed models (GLMM) for categorical data from Item Response Theory (IRT) has been proposed [11, 12]. LPCM has also previously been compared with the two former in a large simulation study [13]. To complement these results, the objective of this article is to evaluate these methods and to propose recommendations to standardize longitudinal analysis of HRQoL data in cancer clinical trials. The methods are first described and compared via statistical, methodological, and practical arguments, then applied on real HRQoL data from selected clinical cancer trials or published prospective databases in a variety of therapeutic situations and tumor sites. The advantages and disadvantages of the methods are discussed, and recommendations are proposed.

Materials and methods

Study selection

In total, seven French studies from a collaborating group were selected according to the following criteria: published randomized phase 2/3 clinical trials or prospective cohort studies comparing two treatments or groups of patients in

adjuvant, advanced, or palliative setting in different cancer sites with longitudinal collection of HRQoL data with the QLQ-C30.

The scoring procedure recommended by the EORTC was used to calculate the standardized scores. To allow comparison between studies, all analyses were performed on the modified intent-to-treat (mITT) population, i.e., including all ITT patients with HRQoL data available at baseline [4].

Statistical analysis

Longitudinal analyses were performed using the three approaches described above. First, LMM that modeled the change in HRQoL score over time for each domain were used [12, 13]. This model combined fixed effects, i.e., group effect, time effect (time was considered as a continuous variable) and group-by-time interaction effect (difference in HRQoL change between groups); and random effects, i.e., random intercept and random slope. The random effects take into account the correlation between the different observations for a same patient and represent the individual deviation from the average intercept and average slope. Finally, the model was the following:

$$Y_i(t) = \beta_0 + \beta_1 t + \beta_2 \text{grp}_i + \beta_3 \{ \text{grp}_i \times t \} + u_{0i} + u_{1i} t + \varepsilon_i(t)$$

where $Y_i(t)$ denotes the HRQoL score for patient i at time t and should be normally distributed, $\varepsilon_i(t) \sim N(0, \sigma^2)$ represents the error term. The vector of the random effects u_{0i} and u_{1i} is assumed to be normally distributed with a mean of zero and an unconstrained covariance matrix.

Most of the time, the group effect was null, i.e., there was no difference between groups at baseline as is usual in randomized clinical trials, and no fixed group effect was kept in the model.

Second, the time-to-event approach was used, in which the deterioration of the HRQoL score is considered as an event. Due to substantial variability in the event definitions, a first set of recommendations were made regarding the definition of the time to deterioration [10]. Accordingly, for the adjuvant setting, we considered the time to first deterioration (TTD), defined as the time from randomization/inclusion to the study to the observation of the first clinically significant deterioration of the HRQoL as compared to the baseline score. Patients without significant deterioration were censored at the time of the last HRQoL assessment. Patients with only a baseline score (i.e., with no follow-up) were censored one day after baseline. For the advanced or metastatic settings, we considered the time until definitive deterioration (TUDD), defined as the time from randomization/inclusion to the study to the observation of the first clinically significant deterioration of the HRQoL score as compared to the baseline score, with no further clinically

significant improvement as compared to the baseline score. Patients without clinically significant deterioration and those with deterioration but which was followed by a significant improvement are censored at the time of the last HRQoL assessment. Note that the TTD/TUDD approaches assume that right-censoring is independent of time to deterioration. Thus, the right-censored patients must be comparable to the patients still at risk regarding their risk of HRQoL deterioration. Finally, the responder threshold to qualify an individual change in TDD/TUDD was fixed at ten points, as usually considered for EORTC HRQoL questionnaires [1, 14]. Sensitivity analyses were then performed considering the best previous (instead of baseline) score as the reference score [10], and death as an event (only added for TUDD compared to the baseline score).

Third, LPCM [11–13] that considered the item responses instead of the score over time were used. A LPCM can be seen as a GLMM with a multinomial logit link function. It models the probability that a individual i selects category k of item j (k varies from 1 to m_j with m_j the number of possible response categories for item j) at visit t given her/his latent trait $\theta_i^{(t)}$ presenting her/his level of HRQoL at time t (time was considered as a continuous variable), and the difficulty parameters $\delta_{j,1}, \dots, \delta_{j,m_j}$:

$$P(X_{i,j} = k \mid \theta_i^{(t)}, \delta_{j,1}, \dots, \delta_{j,m_j}) = \frac{\exp(k\theta_i^{(t)} - \sum_{p=1}^k \delta_{j,p})}{\sum_{h=1}^{m_j} \exp(h\theta_i^{(t)} - \sum_{p=1}^h \delta_{j,p})}$$

The latent variable, assumed to be normally distributed, was linearly decomposed similarly to the first approach (LMM) with fixed and random effects:

$$\theta_i^{(t)} = \beta_0 + \beta_1 t + \beta_2 \text{grp}_i + \beta_3 \{ \text{grp}_i \times t \} + u_{0i} + u_{1i} t$$

LPCM is based on three fundamental IRT assumptions [15], namely unidimensionality (the latent trait is a scalar), monotonicity (the item response functions are increasing), and local independence (the item responses are conditionally independent given the latent trait). Statistical longitudinal analyses were performed using SAS software, Stata commands [16, 17] and R package QoLR [18].

Results

Methodological and practical comparison

Table 1 summarizes the main features of each of the three longitudinal approaches; considering both methodological and practical arguments related to the response variable, modeling, results, as well as the interpretation and readability of the results.

Using the LMM, the outcome is the HRQoL score, which is considered as a continuous variable, while the number of possible values of the HRQoL score depends on the number of items contributing to the scale. For example, for single-item scales (six single-item symptoms for QLQ-C30) with four response categories, only four values exist for the corresponding HRQoL score. Figure 1 illustrates how the 30 items of the QLQ-C30 are distributed to calculate the 15 scale-specific HRQoL scores. Time-to-event modeling approaches raise the same concern: a change of one unit in single-item scales that have four response categories corresponds to a HRQoL score difference of 33 points. Thus, a particular attention should thus be paid to the distribution of the EORTC scores in order to use the appropriate individual threshold to quality the deterioration instead of systematically consider a difference of ten points per scale. Only the LPCM approach can avoid such pitfalls by considering the response to the items as outcome instead of the HRQoL score. A limit of the LMM approach is the Gaussian assumption: the score variable could have a non-symmetrical distribution and the LMM treats the score as a continuous instead of a categorical variable. In this regard, LPCM seems more appropriate [11], but also has three strong assumptions. Another advantage is that it makes it possible to directly use the response to the items, and not only the summary HRQoL score. Indeed, patients can obtain the same HRQoL score with different responses to the items. However, few adapted programs are available to manage GLMM with both random intercept and slope. A SAS program using PROC NLMIXED and a Stata program using the *glamm* procedure (<https://www.glamm.org/>) give similar results but the Stata *glamm* procedure is time-consuming.

Techniques for dealing with multiple comparisons are available with all three approaches, even though, in practice, type I error adjustment is rarely taken into account in the analysis, except when HRQoL is the primary endpoint [6]. Concerning the management of missing data, likelihood-based methods such as LMM or GLMM provided unbiased estimates under MCAR or MAR assumptions [19] contrary to the time-to-event approach. Non-informative missing data reduce only the statistical power in all three strategies [4]. Time-to-event analysis, as well as LMM and LPCM, provides biased estimations in case of informative dropout. Only joint modeling of HRQoL measurement and the missing data process can produce unbiased estimation [19]. The compliance over time should always be described and compared between treatment arms. Moreover, the reason for missing HRQoL forms is an important issue and should be recorded in clinical cancer trials, to make it possible to characterize the mechanism of missing data at least.

Concerning the interpretation and readability of the results, time-to-event analyses are more easily interpretable for clinicians because of their ubiquitous use in oncology.

Table 1 Methodological comparison of the three longitudinal methods

	LMM	LPCM	TTD/TUDD
Response variable			
Type of response variable	Quantitative (HRQoL score)	Ordinal categorical data (Item responses)	Censored (time-to-HRQoL score deterioration)
Single-item scale	Possible values of the score: 4	Number of categories: 4 {Not at all (0), A little (1), Quite a bit (3), Very much (4)}	Minimal deterioration: 33.33
2-item scale	7	4 (by item)	16.67
4 response categories (RF, CF, SF, NV, PA)	13	7 (by item) {1, 2, 3, 4, 5, 6, 7}	8.33
7 response categories (QL2)	10	4 (by item)	11.11
3-item scale (FA)	13	4 (by item)	8.33
4-item scale (EF)	16	4 (by item)	6.67
5-item scale (PF2)			
Modeling			
Hypotheses	Normality of the distribution Homoscedasticity Normality of random effects HRQoL score change over time	Unidimensionality Monotonicity Local independence Latent trait level change over time	Independent right-censoring Proportional hazards for the Cox model Time to occurrence of the HRQoL deterioration
Longitudinal approach	SAS (PROC mixed) Stata (mixed) R (lme4)	SAS (PROC nlmixed) Stata (glamm)	R (package QoLR) Stata (program qlqc30 and qlqc30_TTD)
Software and implementation			
Results			
Estimated parameters	Fixed effects coefficients (time and group-by-time interaction)	Fixed effects coefficients (time and group-by-time interaction) and difficulty parameters	Median time, TTD/TUDD rate at fixed measurement time and HR with 95% CI
Associated test	Wald test	Wald test	Log-rank test
Management of Multiple comparisons	Possibility of type I error adjustment	Possibility of type I error adjustment	Possibility of type I error adjustment
Management of Missing data	Unbiased results under MCAR or MAR assumptions	Unbiased results under MCAR or MAR assumptions	Unbiased results under MCAR assumptions
Interpretation and readability of results			
Level of interpretation	Dimension	Dimension and item	Dimension
Interpretation	An increase in the coefficient on functional scales or GHS/QoL represents an increase in the function level and overall HRQoL (improvement of HRQoL) An increase in the coefficient on symptomatic scales represents an increase in the symptomatology level (deterioration of HRQoL)	An increase in the coefficient on functional scales represents a deterioration in HRQoL (A reversal of the order of the item categories related to functional scales would allow a similar interpretation as with the LMM) An increase in the coefficient of the GHS/QoL represents an improvement in HRQoL An increase in the coefficient on symptomatic scales represents a deterioration of HRQoL	An increase in TTD / TUDD median time represents an improvement in HRQoL (increase on functional scales and the GHS/QoL or decrease on symptomatic scales)

Table 1 (continued)

	LMM	LPCM	TTD/TUDD
Representation/graphical outputs	Change in mean HRQoL score over time according to treatment arm	Change in latent trait level over time according to treatment arm (Representation of Item Response Functions by item)	Kaplan–Meier curve according to treatment arm
Clinical relevance	Change in HRQoL score, interpretation in accordance with MCID	Change in the probability of attributing a given response category MCID is not taken into account	MCID is included in the definition of the event
Clinical interpretation	Average variation of X points in the HRQoL score between two measurement times	Variation of the latent trait over time, interpretation of the item characteristic curves	Direct interpretation of the results

LMM linear mixed model, LPCM longitudinal partial credit model, TTD time to deterioration, TUDD time until definitive deterioration, PF physical functioning, RF role functioning, EF emotional functioning, CF cognitive functioning, SF social functioning, FA fatigue, NV nausea and vomiting, PA pain, DY dyspnea, SL insomnia, AP appetite loss, CO constipation, DI diarrhea, FI financial difficulties, GHS/QoL global health status, MCID minimal clinically important difference, MCAR missing completely at random, MAR missing at random, MNAR missing not at random, IRT item response function, ICC item characteristic curves

TTD allows a direct interpretation of the results in terms of clinically relevance with the integration of the responder threshold within the definition of deterioration. Note that the mean HRQoL change for LMM can be also interpreted in accordance with the group-level MID. Finally, the IRT-based model remains very difficult to interpret even for a statistician. The different graphical outputs available with the three approaches are illustrated in Fig. 2. For each of the three methods, a summary graph with all the scales, such as a forest plot showing the estimated effect and its 95% confidence interval, could be also prepared [20].

Selected randomized clinical trials and prospective cohorts

Randomized clinical trials or prospective cohorts from a French collaborating group were selected in a variety of therapeutic situation and tumor sites. The CO-HO-RT trial [21], APAD trial [22, 23], and Response Shift study [24] involved patients with adjuvant breast cancer; MIROX [25] involved metastatic colorectal cancer patients, PRODIGE5/ACCORD17 [26] involved advanced esophageal cancer patients, PRODIGE4/ACCORD11 [27] involved metastatic pancreatic cancer patients, and TEMAVIR [28] included patients with unresectable glioblastoma. Table 2 describes the clinical trials selected, including the trial acronym and ClinicalTrials.gov identifier, control, and experimental arms, primary endpoint, and details about HRQoL assessment.

Application on the selected databases

Table 3 summarizes the results obtained with the three different approaches. Specifically, we report for each method: the scales with a significant improvement/deterioration over time (LMM, LPCM) and the scales with a significant difference in the experimental group compared to the standard group (LMM, LPCM: group-by-time interaction effect, TTD/TUDD: hazard ratio). The number of significant scales as well as the interpretation is also given. Additionally, all the PRO results that should be reported in RCTs are given in Supplementary Table 1 (LMM and LPCM) and Supplementary Table 2 (TTD/TUDD) on an example (PRODIGE5/ACCORD17 trial).

For adjuvant situations, for all breast cancer clinical trials, the significant scales and their interpretation were the same between LMM and LPCM; except in one study, where LPCM found an additional significant scale. Similar results (except for one study) were also observed between the TTD approach based on a 10-point responder threshold compared to the baseline score or the best previous score. However, results were concordant between the LMM/LPCM and TTD approaches for the APAD trial only, although the number of significant scales was lower with the TTD approaches

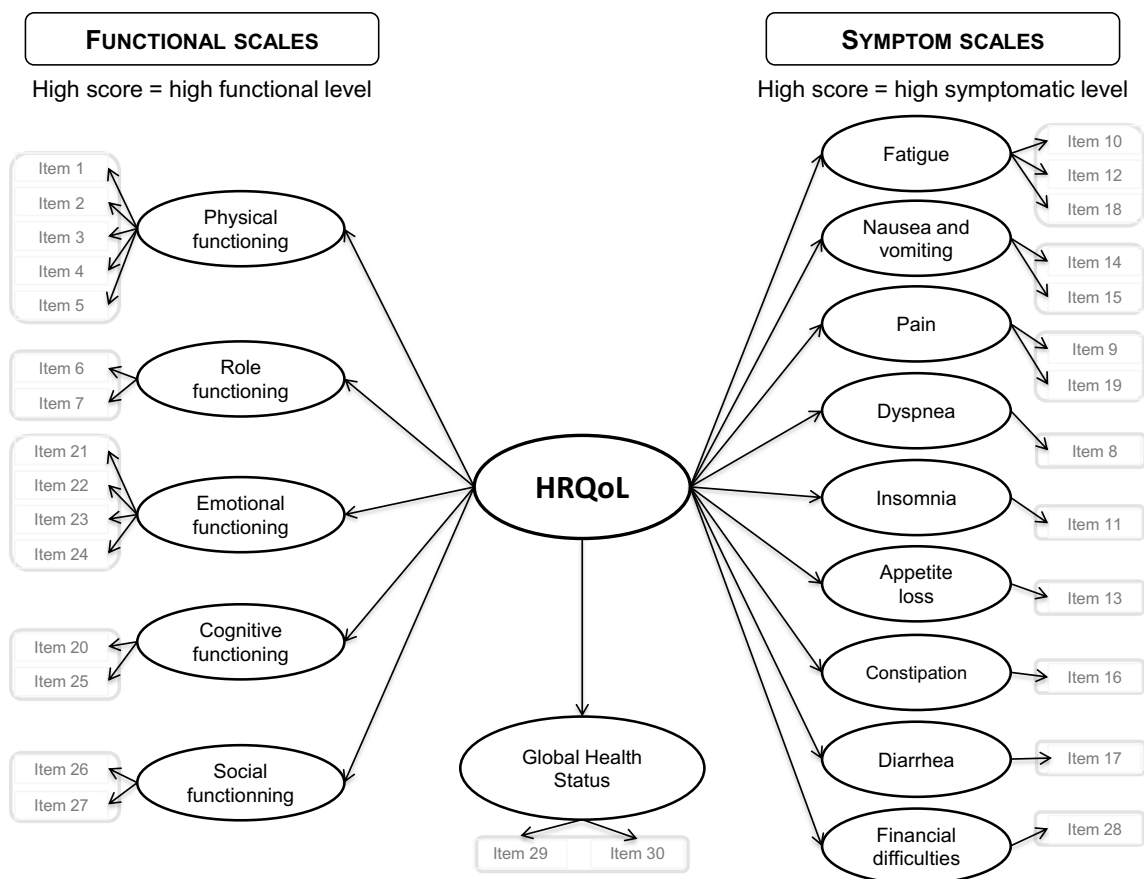


Fig. 1 Distribution of the 30 items in the HRQoL score calculation for EORTC QLQ-C30

(5 for LMM/LPCM vs 3 or 4 for TTD). In the other two studies, the TTD approach found only one significant scale, which was not among the scales found to be significant by the LMM and LPCM methods in three of the four cases.

For advanced disease, for the PRODIGE5/ACCORD17 trial in esophageal cancer and TEMAVIR trial in glioblastoma, we observed a similar number of significant scales for LMM and LPCM, always with the same interpretation. Moreover, only one or two scales were significant considering the TUDD, and the scales identified with LMM/LPCM and TUDD were always different.

For metastatic disease, whatever the tumor site (colorectal and pancreas cancer), the results were similar between LMM and LPCM. Indeed, the following scales: pain for LMM and fatigue for LPCM in PRODIGE4/ACCORD11 and MIROX trials, respectively, were non-significant but, always, at the limit of significance. Finally, the number and type of significant scales were different between LMM/LPCM and TUDD. Moreover, among the three event definitions considered for the TUDD approach, the results were also discordant, in particular when death was added as an event.

Discussion

This article compares the two most common methods for longitudinal analysis of HRQoL in cancer clinical trials, the LMM and TTD/TUDD approaches, and an alternative strategy based on IRT, namely the LPCM, through statistical, methodological, and practical arguments.

From a statistical point of view, the LPCM approach is more suited than LMM and TTD/TUDD to the construction of EORTC questionnaires. Indeed, the HRQoL scores for dimensions based on few items are considered as continuous variables whereas in fact, they present the characteristics of ordinal variables [11, 12]. However, a previous simulation study comparing these three approaches found that the LMM was the most powerful method in all the scenarios considered, ahead of the LPCM [13]. This study also found that the statistical power of the TTD/TUDD approach was low, especially for single-item scales (even with a large sample size), but the case where death or drop-out was integrated into the event definition was not considered. Finally, the LMM is a well-established approach, more intuitive and easy to perform, contrary to LPCM, which is difficult to understand and to interpret, even for a statistician, and not implemented

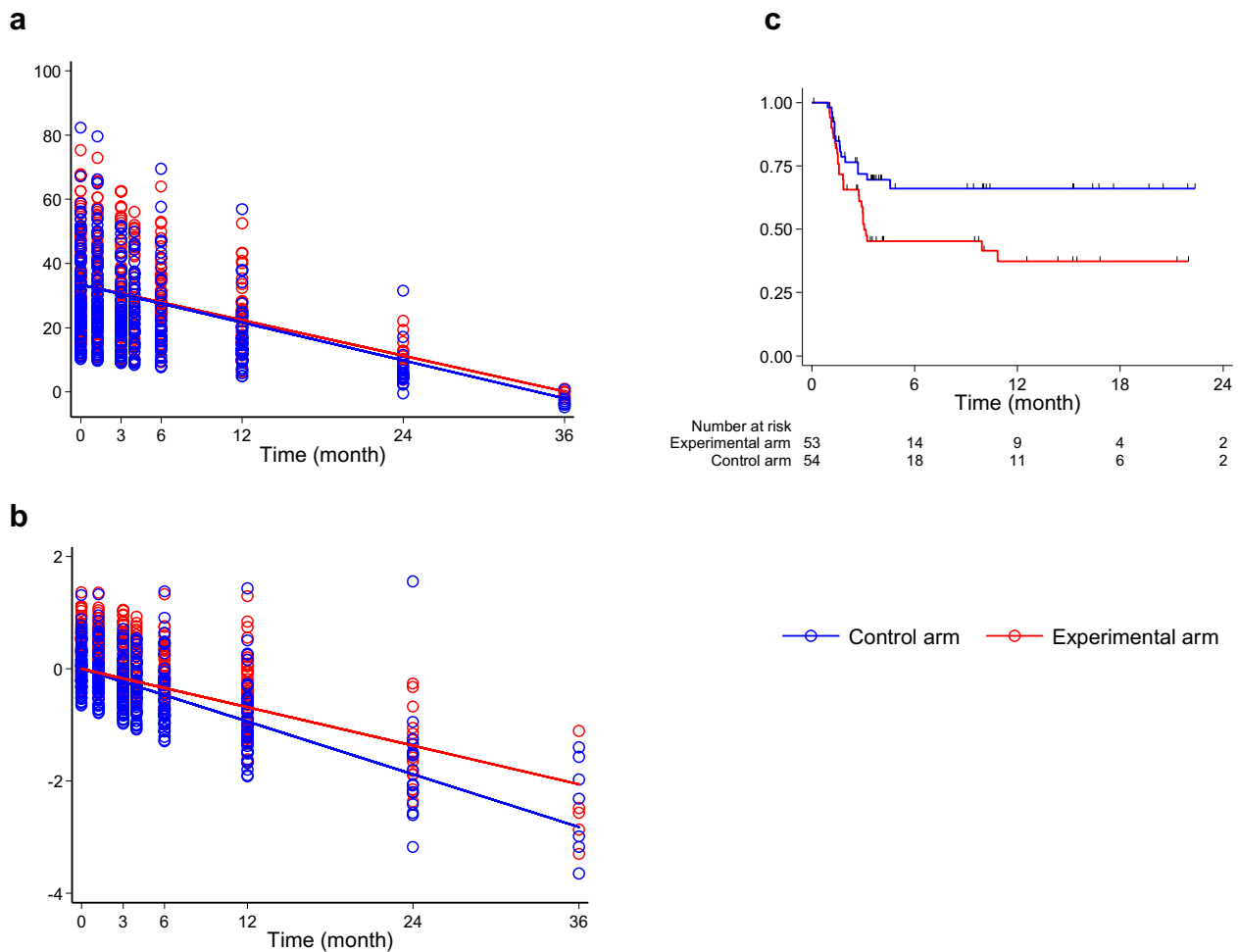


Fig. 2 Graphical outputs for LMM, LPCM and TTD/TUDD. **a, b** Individual (point) and mean (line) predicted values. **c** Kaplan–Meier survival estimate

in the main statistical software packages. Nevertheless, the use of the LPCM could be argued and justified for single-item scales.

For the LMM, time was treated as continuous, which implies to make an assumption on the relationship between time and HRQoL. Notice that the linearity assumption considered could be relaxed by including for example a quadratic term or by using splines that would allow a flexible form for the HRQoL trajectories. For the LPCM, time was also treated as continuous. In both the LMM and the LPCM, time could be also treated as a discrete variable.

From a practical point of view, to promote quicker and more systematic analysis of HRQoL data with the three methods in clinical trials in oncology, we developed several commands providing automatic and reliable analyses with the statistical software Stata [16, 17] and R [18]. Moreover, SAS and Stata codes to implement the LPCM are also available from the authors on request.

This article also compares the three methods on real HRQoL data from seven clinical cancer trials and French published prospective databases in adjuvant, advanced or palliative settings, and in different tumor sites. In the majority of cases, we observed concordant results between the LMM and LPCM approaches (significant scales and interpretation of the results, i.e., in favor of the experimental or control arm). However, discordant results were observed between the TTD/TUDD approach and the two others. This issue was not unexpected, since it had been already been raised in two glioblastoma trials [7, 8]. These discordant results are coherent with regard to the different criteria considered: GLMM investigate the change in HRQoL score (LMM) or the latent trait level (LPCM), whereas the TTD/TUDD approach studies the time until the occurrence of a HRQoL score deterioration (whether definitive or not). For TTD/TUDD analyses, the event definition is a major difficulty, especially since the choice of the reference score to

Table 2 Description of the clinical trials or prospective cohort selected

Pathology	Ref	Trial acronym gov number	Group (N = mITT*/ITT°)		Primary endpoint	HRQoL assessment
			Experimental group	Standard group		
Breast cancer (Adjuvant, Phase 2)	[18]	CO-HO-RT NCT00208273	Sequential RT and letrozole N = 70*/75°	Concurrent RT and letrozole N = 73*/74°	Subcutaneous late toxicity (2 years)	Baseline, M3, M6, M9, M12, M15, M18, M21 and M24
Breast cancer (Adjuvant, Phase 3)	[19, 20]	APAD NCT01495650	Adapted physical activity and diet (APAD) N = 72*/72°	Standard Care N = 70*/71°	Subjective fatigue (M18), multidimensional Fatigue Inventory question- naire	Baseline, W18, W27, W54 and W81
Breast cancer (Adjuvant, prospective cohort)	[21]	Response shift	Management with mastec- tomy N = 114*/121°	Management without mas- tectomy N = 195*/208°	Impact of breast cancer management on the change of quality of life, on the components of response shift and health expectations (primary objective)	Baseline, after surgery, M3, M6
Glioblastoma (Advanced, Phase 2)	[25]	Temavir NCT01022918	Bevacizumab and irinotecan N = 45*/60°	Temozolomide N = 45*/60°	Rate of non-progressive disease at 6 months after inclusion	Baseline, just before neoad- juvant treatment (Beva and Irinotecan only), just before RCT, end of RCT, every four weeks until disease progression
Esophageal cancer (Advanced, Phase 2/3)	[23]	PRODIGES/ACCORD17 NCT00861094	N = 119*/134°	Folfox Fluorouracil–cisplatin N = 120*/133°	Progression free survival (Phase 3)	Baseline, final day of RT, D1 of the last CT cycle at W10 or W11, W15, and M6 M12, M24 and M36
Pancreatic cancer (Metastatic, Phase 2/3)	[24]	PRODIGE4/ACCORD11 NCT00112658	Folfirinox N = 163*/171°	Gemzar N = 157*/171°	Overall survival (Phase 3)	Baseline, D15, D30, M2, M4, M6, M8, and M10
Colorectal cancer (Metastatic, Phase 3)	[22]	MIROX NCT00268398	Folfox 7 – Folfiri N = 88*/142°	Folfox 4 N = 83*/142°	2-year disease free survival	Baseline, every 4 cycles

D day, W week, M month, CT chemotherapy, RT radiotherapy, RCT radiochemotherapy

Table 3 Evaluation of the three longitudinal methods on the selected databases

	TTD (adjuvant)/TUDD (advanced or metastatic)				
	LMM	LPCM	Baseline (10-point MCID)	Baseline + Death (10-point MCID)	Best previous score (10-point MCID)
Adjuvant situation					
CO-HO-RT	-	-	-	-	-
Time effect	-	CO, DY	-	-	-
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	CO-, DY- (2)	CO-, DY- (2)	FI- (1)	-	FI- (1)
APAD					
Time effect	GHS/QoL, PF, RF, SF, FI	GHS/QoL, PF, RF, SF, FI	-	-	-
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	GHS/QoL+, PF+, RF+, SF+, FI+(5)	GHS/QoL-, PF+, RF+, SF+, FI+(5)	GHS/QoL+, PF+, RF+(3)	-	GHS/QoL+, PF+, RF+, FI+(4)
RESPONSE SHIFT					
Time effect	CF, PA, DI	CF, NV, PA, DI	-	-	-
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	CF+, PA+, DI- (3)	CF+, NV-, PA+, DI- (4)	EF+(1)	-	DI- (1)
Advanced situation					
TEMAVIR	-	-	-	-	-
Time effect	PF, DY, SL, CO, FI	PF, CO	-	-	-
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	CO+(1)	CO+(1)	DI+, FI+(2)	FI+(1)	DI+, FI+(2)
PRODIGES/ACCORD17					
Time effect	GHS/QoL, RF, EF, FA, NV, PA, SL, AP, CO	EF, NV, CO	-	-	-
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	PA- (1)	-	AP- (1)	-	AP- (1)
Metastatic situation					
PRODIGE4/ACCORD11	-	-	-	-	-
Time effect	GHS/QoL, EF, FA, NV, PA, SL, AP, CO, DI	EF, PA, SL, AP, CO, DI	-	-	-
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	DI- (1)	DI-, PA+(2)	GHS/QoL+, PF+, RF+, EF+, CF+, SF+, FA+, NV+, PA+, DY+, SL+, AP+, CO+(12)	GHS/QoL+, RF+, EF+, CF+, SF+, FA+, NV+, PA+, DY+, SL+, AP+, CO+(12)	GHS/QoL+, RF+, EF+, CF+, SF+, PA+, SL+, AP+, CO+(9)
MIROX					
Time effect	PF, EF, NV, DY	RF, SF, FA, NV, AP, DI	-	-	-

Table 3 (continued)

	TTD (<i>adjuvant</i>)/TUDD (<i>advanced or metastatic</i>)		
	LMM	LPCM	Best previous score (10-point MCID)
Group-by-time interaction effect (LMM/LPCM) or group effect (TTD/TUDD)	FA + (1)	-	Baseline (10-point MCID) PA + (1) GHS/QoL +, PF +, PA + (3) PA + (1)

PF physical functioning, *RF* role functioning, *EF* emotional functioning, *CF* cognitive functioning, *SF* social functioning, *FA* fatigue, *NV* nausea and vomiting, *PA* pain, *DY* Dyspnea, *SL* insomnia, *AP* appetite loss, *CO* constipation, *DI* diarrhea, *FI* financial difficulties, *GHS/QoL* global health status

A + (-) sign indicates an improvement (deterioration) in HRQoL in the experimental group compared to the standard group

determine the deterioration (for example, baseline or best previous score) and the possible inclusion of death in the event definition (as recommended in the palliative setting) could produce discordant results. Besides, in the case where death is included in the event definition, it is necessary to remain vigilant with regard to death occurrence in relation to the questionnaire collection time: with too many early or late deaths, TUDD analysis could coincide with overall survival analysis. On the other hand, if death is not included in the event definition, it may introduce an informative censoring and bias the results. At last, our study presented a limit: the choice of a ten-point responder threshold, for all scales, to qualify an individual deterioration in the TTD/TUDD approaches. Indeed, for single-item scale, for example, a ten-point difference is achievable with a movement of only one response level and for others, it would require more than this. The choice of ten points cannot be clinically meaningful across all subscales. In fact, the EORTC is currently working on the definition of MID for group-level as well as responder threshold per EORTC questionnaire and cancer sites [29]. These new recommendations could then be used and will be better adapted according to the type of analysis.

Overall, these methods seem to provide complementary interpretations and information. The LMM and LPCM are the most powerful methods on simulated data, while the TUDD/TDD approach gives more clinically understandable results. Thus, one method does not outperform the others and we would recommend combining the LMM and TTD/TUDD approaches, except for single-item scales, for longitudinal analysis of HRQoL data (if HRQoL is a secondary endpoint) in cancer clinical trials.

This statement has been supported by the glioblastoma trials previously cited [7, 8]. In fact, a secondary paper on HRQoL data has been published for AVAglio trial using both LMM and TTD approaches [30], which enabled the comparison of the results with the RTOG0825 trial [8]. For single-item scales, the TTD/TUDD approach is not appropriate, and LPCM is more appropriate than LMM and seems easy to implement because, in this case, LPCM is only a classical GLMM for ordinal data.

An article fully dedicated to reporting HRQoL data should be systematically proposed after the first publication of the trial's results and should be written according to the recommendations of Calvert et al. [31]. In particular, statistical approaches for dealing with missing data (such as sensitivity analysis with joint modeling) and type I error adjustment must be explicitly detailed. Moreover, it is highly recommended and appreciated when the HRQoL results are published relatively soon after the main paper comes out.

Our work and views are consistent with the objective of the SISAQOL Consortium [32, 33] to propose recommendations for standardizing analyses of patient-reported outcome data in cancer clinical trials. Indeed, before planning clinical

cancer trials with HRQoL as a primary/co-primary endpoint, it is essential to harmonize the methodology for HRQoL analysis and the reporting of the results. This seems the most reliable way to obtain comparative results between trials in order to make assumptions to plan future clinical trials.

Conclusion

In conclusion, these results pledge for the recommendation to use both longitudinal methods LMM and TTD/TUDD (except for single-item scales) in HRQoL-specific publications to move towards becoming a consensus. The choice of the method should be also guided by the clinical objective, depending on whether the objective is to show a difference in the evolution of the mean score over time (LMM, LPCM) or a difference in the risk of HRQoL deterioration over time (TTD/TUDD).

Standardization of the longitudinal analysis of HRQoL is an essential step towards confirming its position as a primary or co-primary endpoint in cancer clinical trials, ultimately leading to a change in clinical practice in light of HRQoL data.

Acknowledgements The authors thank Fiona Ecartot (EA3920, University Hospital Besançon, University of Franche-Comté, Besançon, France) for editorial assistance.

Funding This work was supported by grants from French Public Health Research Institute (<https://IRESP.net>) and SIRIC Montpellier Cancer [Grant INCa_Inserm_DGOS_12553].

Compliance with ethical standards

Conflicts of interest The authors have declared no conflicts of interest.


References

- Osoba, D. (2011). Health-related quality of life and cancer clinical trials. *Therapeutic Advances in Medical Oncology*, 3(2), 57–71.
- Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (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. *Journal of National Cancer Institute*, 85(5), 365–376.
- Fayers, P. M., Aaronson, N. K., Bjordal, K., Groenvold, M., Curran, D., & Bottomley, A. on behalf of the E. Q. of L. Group. (2001). EORTC QLQ-C30 Scoring Manual (3rd edition).
- Bonnetain, F., Fiteni, F., Efficace, F., & Anota, A. (2016). Statistical challenges in the analysis of health-related quality of life in cancer clinical trials. *Journal of Clinical Oncology*, 34(16), 1953–1956.
- Kiebert, G. M., Curran, D., & Aaronson, N. K. (1998). Quality of life as an endpoint in EORTC clinical trials european organization for research and treatment for cancer. *Statistics in Medicine*, 17(57), 561–569.
- Hamel, J.-F., Saulnier, P., Pe, M., Zikos, E., Musoro, J., Coens, C., et al. (2017). A systematic review of the quality of statistical methods employed for analysing quality of life data in cancer randomised controlled trials. *European Journal of Cancer*, 83, 166–176.
- Chinot, O. L., Wick, W., Mason, W., Henriksson, R., Saran, F., Nishikawa, R., et al. (2014). Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *New England Journal of Medicine*, 370(8), 709–722.
- Gilbert, M. R., Dignam, J. J., Armstrong, T. S., Wefel, J. S., Blumenthal, D. T., Vogelbaum, M. A., et al. (2014). A randomized trial of bevacizumab for newly diagnosed glioblastoma. *New England Journal of Medicine*, 370(8), 699–708.
- Fairclough, D. L. (2002). Design and Analysis of Quality of Life Studies in Clinical Trials. (Boca Raton: Chapman and Hall, Ed.).
- Anota, A., Hamidou, Z., Paget-Bailly, S., Chibaudel, B., Bascoul-Mollevi, C., Auquier, P., et al. (2015). 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*, 24(1), 5–18.
- Barbieri, A., Peyhardi, J., Conroy, T., Gourgou, S., Lavergne, C., Mollevi, C. (2017). Item response models for the longitudinal analysis of health-related quality of life in cancer clinical trials. *BMC Medical Research Methodology*, 17(1).
- Barbieri, A., Anota, A., Conroy, T., Gourgou-Bourgade, S., Juzyna, B., Bonnetain, F., et al. (2016). Applying the longitudinal model from item response theory to assess health-related quality of life in the PRODIGE 4/ACCORD 11 randomized trial. *Medical Decision Making*, 36(5), 615–628.
- Anota, A., Barbieri, A., Savina, M., Pam, A., Gourgou-Bourgade, S., Bonnetain, F., et al. (2014). 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(1), 1326.
- Charton, E., Cuer, B., Cottone, F., Efficace, F., Touraine, C., Hamidou, Z., et al. (2020). Time to deterioration in cancer randomized clinical trials for patient-reported outcomes data: a systematic review. *Quality of Life Research*, 29(4), 867–878.
- Van der Linden, W. J., Hambleton, R. K. (1997). Handbook of Modern Item Response Theory (Vol. Springer Verlag, New York).
- Bascoul-Mollevi Caroline, Castan Florence, Azria David, Gourgou-Bourgade Sophie. (2015). EORTC QLQ-C30 descriptive analysis with the qlqc30 command. The Stata Journal, pp. 1060–1074. (<https://www.stata-journal.com/article.html?article=dm0084>).
- Bascoul-Mollevi, C., Savina, M., Anota, A., Barbieri, A., Azria, D., Bonnetain, F., et al. (2018). Longitudinal health-related quality of life analysis in oncology with time to event approaches, the STATA command qlqc30_TTD. *Computer Methods and Programs in Biomedicine*, 158, 153–159.
- Anota, A., Savina, M., Bascoul-Mollevi, C., Bonnetain, F. (2017). QoLR: An R package for the longitudinal analysis of health-related quality of life in oncology. *Journal of Statistical Software*, p. Vol 77: Issue 12. (<https://www.jstatsoft.org/article/view/v077i12>).
- Troxel, A. B., Fairclough, D. L., Curran, D., & Hahn, E. A. (1998). Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine*, 17(5–7), 653–666.
- Oza, A. M., Matulonis, U. A., Malander, S., Hudgens, S., Sehoul, J., del Campo, J. M., et al. (2018). Quality of life in patients with recurrent ovarian cancer treated with niraparib versus placebo (ENGOT-OV16/NOVA): Results from a

- double-blind, phase 3, randomised controlled trial. *The Lancet Oncology*, 19(8), 1117–1125.
21. Azria, D., Belkacemi, Y., Romieu, G., Gourgu, S., Gutowski, M., Zaman, K., et al. (2010). Concurrent or sequential adjuvant letrozole and radiotherapy after conservative surgery for early-stage breast cancer (CO-HO-RT): a phase 2 randomised trial. *The lancet oncology*, 11(3), 258–265.
 22. Carayol, M., Romieu, G., Bleuse, J. P., Senesse, P., Gourgu-Bourgade, S., Sari, C., et al. (2013). Adapted physical activity and diet (APAD) during adjuvant breast cancer therapy: Design and implementation of a prospective randomized controlled trial. *Contemporary Clinical Trials*, 36, 531–543.
 23. Carayol, M., Ninot, G., Senesse, P., Bleuse, J.-P., Gourgu, S., Sancho-Garnier, H., et al. (2019). Short- and long-term impact of adapted physical activity and diet counseling during adjuvant breast cancer therapy: The “APAD1” randomized controlled trial. *BMC cancer*, 19(1), 737.
 24. Dabakuyo, T. S., Guillemin, F., Conroy, T., Velten, M., Jolly, D., Mercier, M., et al. (2013). Response shift effects on measuring post-operative quality of life among breast cancer patients: A multicenter cohort study. *Quality of life research*, 22(1), 1–11.
 25. Hebbar, M., Chibaudel, B., Andre, T., Mineur, L., Smith, D., Louvet, C., et al. (2015). FOLFOX4 versus sequential dose-dense FOLFOX7 followed by FOLFIRI in patients with resectable metastatic colorectal cancer (MIROX): A pragmatic approach to chemotherapy timing with perioperative or post-operative chemotherapy from an open-label, randomized phase III trial. *Annals of Oncology*, 26, 1040.
 26. Conroy, T., Galais, M.-P., Raoul, J.-L., Bouché, O., Gourgu-Bourgade, S., Douillard, J.-Y., et al. (2014). Definitive chemoradiotherapy with FOLFOX versus fluorouracil and cisplatin in patients with oesophageal cancer (PRODIGE5/ACCORD17): Final results of a randomised, phase 2/3 trial. *The lancet oncology*, 15(3), 305–314.
 27. Conroy, T., Desseigne, F., Ychou, M., Bouché, O., Guimbaud, R., Bécouarn, Y., et al. (2011). FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *The New England journal of medicine*, 364(19), 1817–1825.
 28. Chauffert, B., Feuvret, L., Bonnetain, F., Taillandier, L., Frappaz, D., Taillia, H., et al. (2014). Randomized phase II trial of irinotecan and bevacizumab as neo-adjuvant and adjuvant to temozolomide-based chemoradiation compared with temozolomide-chemoradiation for unresectable glioblastoma: Final results of the TEMAVIR study from ANOCEF†. *Annals of Oncology*, 25(7), 1442–1447.
 29. Musoro, Z. J., Hamel, J.-F., Ediebah, D. E., Cocks, K., King, M. T., Groenvold, M., et al. (2018). Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-analysis protocol. *British Medical Journal Open*, 8(1), e019117.
 30. Taphoorn, M. J. B., Henriksson, R., Bottomley, A., Cloughesy, T., Wick, W., Mason, W. P., et al. (2015). Health-related quality of life in a randomized phase iii study of bevacizumab, temozolomide, and radiotherapy in newly diagnosed glioblastoma. *Journal of Clinical Oncology*, 33(19), 2166–2175.
 31. Calvert, M., Blazeby, J., Altman, D. G., Revicki, D. A., Moher, D., Brundage, M. D., et al. (2013). Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *JAMA*, 309(8), 814–822.
 32. Bottomley, A., Pe, M., Sloan, J., Basch, E., Bonnetain, F., Calvert, M., et al. (2016). Analysing data from patient-reported outcome and quality of life endpoints for cancer clinical trials: a start in setting international standards. *The Lancet. Oncology*, 17(11), e510–e514.
 33. Bottomley, A., Pe, M., Sloan, J., Basch, E., Bonnetain, F., Calvert, M., Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium. (2018). Moving forward toward standardizing analysis of quality of life data in randomized cancer clinical trials. *Clinical Trials*, 1740774518795637.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Caroline Bascoul-Molle^{1,2,3}  · Antoine Barbieri⁴ · Céline Bourcier^{5,2} · Thierry Conroy^{6,7} · Bruno Chauffert⁸ · Mohamed Hebbar⁹ · William Jacot^{10,2} · Beata Juzyna¹¹ · Hélène De Forges¹² · Sophie Gourgu^{1,3} · Franck Bonnetain^{3,13,14} · Célia Touraine^{1,3} · Amélie Anota^{3,13,14}

Antoine Barbieri
antoine.barbieri@u-bordeaux.fr

Céline Bourcier
Celine.Bourcier@icm.unicancer.fr

Thierry Conroy
t.conroy@nancy.unicancer.fr

Bruno Chauffert
chauffert.bruno@chu-amiens.fr

Mohamed Hebbar
mohamed.hebbar@chru-lille.fr

William Jacot
William.Jacot@icm.unicancer.fr

Beata Juzyna
b-juzyna@unicancer.fr

Hélène De Forges
Helene.deForges@icm.unicancer.fr

Sophie Gourgu
Sophie.Gourgu@icm.unicancer.fr

Franck Bonnetain
franck.bonnetain@univ-fcomte.fr

Célia Touraine
Celia.Touraine@icm.unicancer.fr

Amélie Anota
aanota@chu-besancon.fr

¹ Biometrics Unit – CTD INCa, Institut du Cancer Montpellier, Univ. Montpellier, Montpellier, France

² Institut de Recherche en Cancérologie de Montpellier Inserm U1194, University Montpellier, 208 rue des Apothicaire, Montpellier Cedex 5, 34298 Montpellier, France

- ³ National Platform Quality of Life and Cancer, Montpellier, France
- ⁴ Inserm UMR 1219, University Bordeaux, Bordeaux, France
- ⁵ Department of Radiation Oncology, Institut du Cancer Montpellier, University Montpellier, Montpellier, France
- ⁶ Medical Oncology Department, Institut de Cancérologie de Lorraine, Vandœuvre-lès-Nancy, France
- ⁷ Lorraine University, APEMAC, Team MICS, Nancy, France
- ⁸ Medical Oncology Department, Amiens University Hospital, Amiens, France
- ⁹ Department of Medical Oncology, University Hospital, Lille, France
- ¹⁰ Department of Medical Oncology, Institut du Cancer Montpellier, University Montpellier, Montpellier, France
- ¹¹ R&D Unicancer, Paris, France
- ¹² Clinical Research and Innovation Department, Institut du Cancer Montpellier, University Montpellier, Montpellier, France
- ¹³ Methodology and Quality of Life in Oncology Unit, University Hospital of Besançon, Besançon, France
- ¹⁴ UMR1098, Interactions Hôte-Greffon-Tumeur/Ingénierie Cellulaire Et Génique, Bourgogne Franche-Comté University, Inserm, EFS BFC, Fédération Hospitalo-Universitaire INCREASE, Besançon, France