Analyse longitudinale de la qualité de vie relative à la santé en cancérologie par équation structurelle et modèles mixtes

Myriam Tami[1,4] & Antoine Barbieri[1,3,4] & Caroline Bascoul-Mollevi[3] & Xavier Bry[1,4] & Christian Lavergne[2,4] & David Azria[3] & Sophie Gourgou[3]

[1] *Université de Montpellier*
[2] *Université Paul-Valéry, Montpellier 3*
[3] *Institut régional du Cancer Montpellier (ICM)*
[4] *Institut Montpelliérain Alexander Grothendieck (IMAG)*
myriam.tami,xavier.bry@umontpellier.fr
christian.lavergne@univ-montp3.fr
antoine.barbieri,caroline.mollevi,david.azria,sophie.gourgou@icm.unicancer.fr

Résumé. En oncologie, la qualité de vie relative à la santé (QdV) est un critère secondaire dans les essais cliniques mais son l'analyse longitudinale reste complexe. La QdV est mesurée à travers des questionnaires que remplissent les patients à différentes visites du traitement et au cours du suivi. La structure particulière du questionnaire QLQ-C30 de l'EORTC décompose la QdV en un groupe de dimensions fonctionnelles, un groupe de dimensions symptomatiques et le "Statut global de santé" (GHS, Global Health Status). Par un Modèle à Équation Structurelle (SEM, Structural Equation Model), l'objectif est d'expliquer le GHS par les autres dimensions. Cette modélisation à équation structurelle est réalisée à chaque visite où la variable GHS est expliquée par deux variables latentes. Chaque variable latente résume respectivement le groupe de variables fonctionnelles et le groupe de variables symptomatiques. Puis une maximisation de la vraisemblance par algorithme EM de chacun des modèles, permet d'obtenir à chaque visite une estimation des facteurs. L'analyse longitudinale est alors réalisée par l'intermédiaire d'un modèle linéaire mixte sur la concaténation de la variable GHS et des facteurs estimés à chaque visite. Cette modélisation permet de prendre en compte la variabilité intra-individuelle avec les effets aléatoires et l'influence d'éventuelles covariables tel que le traitement. Nous présenterons une application de cette approche sur des données réelles issues d'un essai clinique en cancérologie.

Mots-clés. Qualité de vie relative à la santé, données oncologiques, analyse longitudinale, modèle à équation structurelle, modèle linéaire mixte.

Abstract. The health-related quality of life data is measured through self-questionnaires filled up at different times. We focused on the oncology data reported through the EORTC questionnaires which decompose the health-related quality of life into several functioning dimensions, several symptomatic dimensions and the Global Health Status. The aim is to

explain the latter, which represents the most general concept, through the other dimensions. First, a similar structural equation model is used at each time, in which the global health status is explained by two latent variables. Each latent variable is a factor which summarizes respectively the functional dimensions and the symptomatic dimensions. This is achieved through the maximization of the likelihood of each structural equation model using the EM algorithm, with the advantage to give an estimation of the subject-specific factors. Then, to consider the longitudinal aspect, the global health status variable and the two factors are concatenated for each visit. The global health status can be then explained by the two factors estimated in the first step and additional explanatory variables using a linear mixed model. This model takes into account the inter-subject variability via specific-subject random effects and other covariates such as the treatment.

Keywords. Health-related quality of life, oncology data, longitudinal analysis, structural equation modeling, mixed models.

1 Introduction

La QdV est un concept subjectif, multidimensionnel et évolutif au cours du temps, ce qui le rend difficile à mesurer et à étudier. Actuellement, les données de QdV sont issues d'auto-questionnaires remplis à différentes visites au cours du traitement et du suivi des patients. Le questionnaire standard en Europe est le "European Organization for Research and Treatment of Cancer Quality of Life Questionnaire - Core Questionnaire" (EORTC QLQ-C30) (cf. Aaronson et al. (1993)). Il décompose la QdV en 15 dimensions : 5 dimensions fonctionnelles, 9 symptomatiques et le "Statut global de santé" (GHS, Global Health Status). Actuellement, l'analyse de la QdV est réalisée dimension par dimension de manière indépendante. Cette analyse présente un inconvénient majeur : la multiplication des tests. Afin de palier à cette limite, nous proposons de construire un modèle à équation structurelle (SEM, Structural Equation Model) à chaque visite où la variable GHS est expliquée par deux variables latentes. Chacune d'elle est un facteur résumant les dimensions fonctionnelles d'une part, et les dimensions symptomatiques d'autre part. La concaténation de la variable GHS et des deux facteurs estimés sur l'ensemble des temps de suivi permettra en plus d'étudier l'aspect longitudinal des données. On peut ainsi avec un modèle linéaire mixte expliquer la variable GHS par les deux facteurs. Ce modèle prendra en compte l'effet individu (effet aléatoire) et l'effet traitement (effet fixe). Cette approche est possible par la maximisation de la vraisemblance de chaque SEM par algorithme EM (cf. Bry et al. (2016), Tami et al. (2014) et Barbieri et al. (2016)). En effet, contrairement aux méthodes classiques de la littérature (cf. Jöreskog, K. G. (1970), Wold, (1966)), celle que nous proposons donne une estimation des facteurs de manière efficace et permet ainsi de les introduire dans un autre modèle (ici un modèle linéaire mixte). Cette technique d'estimation (cf. Bry et al. (2016)) est développée sous le logiciel R et nous l'illustrerons sur des données d'un essai clinique en cancérologie.

2 Analyse de la QdV par une approche en deux étapes

2.1 Première étape : Analyse transversale

Le SEM que nous proposons est issu de la décomposition du questionnaire QLQ-C30 qui distingue un groupe de dimensions fonctionnelles, un groupe de dimensions symptomatiques et une dimension GHS. La structure de ce questionnaire suggère une modélisation à deux facteurs f^1 et f^2 qui quantifient et résument respectivement les statuts fonctionnels et symptomatiques pour tous les individus. f^1 et f^2 sont respectivement liés au groupe de variables observables fonctionnelles X^1 et symptomatiques X^2. Ces liaisons sont formalisées par des équations dites de mesures (cf. (**??**), (**??**)). Les facteurs f^1 et f^2 sont aussi liés à la variable GHS notée y. y est expliquée par les facteurs et ces liaisons entre les facteurs sont quant à elles formalisées par une équation structurelle (cf.(**??**)). La concaténation des équations de mesures et de l'équation structurelle forment un système d'équations qui aboutit au modèle structurel que nous construisons à chaque temps de suivi (cf. (**??**)). Ces relations de dépendance peuvent être chacune enrichie par une dépendance supplémentaire aux covariables T, T^1 et T^2 tel que le traitement par exemple. En outre, l'ensemble des variables observables sont quantitatives. En effet, les données résultant des questionnaires subissent un pré-traitement selon une procédure de scoring (cf. Fayers et al. (2001)) permettant de conjecturer la nature quantitative des variables observables.

2.1.1 Formulation du modèle à équation structurelle établi à chaque visite et ses notations

Les données sont organisées sous forme de groupes de variables observables décrivant les mêmes n individus :
$X^m = \{x_i^{j,m}\}$; $i \in [\![1, n]\!]$, $j \in [\![1, q_m]\!]$, $m \in [\![1, 2]\!]$ la m-ième matrice explicative de dimension $n \times q_m$ constituée des variables $x^{1,m}, \ldots, x^{q_m,m}$. La valeur de la variable $x^{j,m}$ pour l'individu i est notée $x_i^{j,m}$.
$y = \{y_i\}$; $i \in [\![1, n]\!]$, est le vecteur de longueur n représentant la variable observable dépendante des facteurs f^1 et f^2.
T (resp. T^1, T^2) de dimension $n \times r_T$ (resp. $n \times r_1$, $n \times r_2$) sont les matrices de covariables. Introduisons également d (resp. D^m) un vecteur $r_T \times 1$ (resp. une matrice $r_m \times q_m$) de coefficients pondérateurs, a^m un vecteur $q_m \times 1$ de coefficients pondérateurs, et ε^y (resp. ε^m) un vecteur des erreurs $n \times 1$ (resp. une matrice des erreurs $n \times q_m$) associées à y (resp. X^m). Le modèle peut alors être formulé ainsi :

$$\begin{cases} X^1 = T^1 D^1 + f^1 a^{1'} + \varepsilon^1 & \text{(1a)} \\ X^2 = T^2 D^2 + f^2 a^{2'} + \varepsilon^2 & \text{(1b)} \\ y \ = Td + f^1 c^1 + f^2 c^2 + \varepsilon^y & \text{(1c)} \end{cases}$$

où nous imposons que les éléments de la première colonne des matrices de covariables T et T^m soient fixées à 1. Ainsi d et la première ligne de chaque matrice D^m correspondent aux paramètres de moyenne. On y adjoint sous contraintes d'identifiablité, les hypothèses suivantes :
Les $i \in [\![1, n]\!]$ observations sont indépendantes ; $\varepsilon_i^y \sim \mathcal{N}(0,\ \sigma_y^2)$; $\forall m \in [\![1, 2]\!]$: $\varepsilon_i^m \sim \mathcal{N}(0,\ \psi_m)$, où $\psi_m = diag(\sigma_m^2)$; ε^y and ε^m sont indépendant $\forall m \in [\![1, 2]\!]$; $\forall m \in [\![1, 2]\!]$: $f^m \sim \mathcal{N}(0, I_n)$ avec f^1, f^2 indépendant et ε^y, ε^m, f^m, $\forall m \in [\![1, 2]\!]$ sont mutuellement indépendant. Nous faisons aussi les hypothèses que $\forall m \in [\![1, 2]\!]$ les variables observées X^m (par exemple $x_j^m, j \in [\![1, q_p]\!]$) dépendent linéairement du facteur f^m et des covariables T^m, conditionnellement auxquelles ils sont indépendant ; y dépend linéairement des facteurs f^1 et f^2 et de la covariable T.
Ce SEM est établi pour chaque temps de suivi et pour i une observation, le modèle peut être formulé selon le système d'équation suivant :

$$\begin{cases} x_i^{1\prime} &= t_i^{1\prime} D^1 + f_i^1 a^{1\prime} + {\varepsilon_i}^{1\prime} \\ x_i^{2\prime} &= t_i^{2\prime} D^2 + f_i^2 a^{2\prime} + {\varepsilon_i}^{2\prime} \\ y_i' &= t_i' d + f_i^1 c^1 + f_i^2 c^2 + {\varepsilon_i}^{y\prime} \end{cases} \tag{2}$$

La figure **??** représente son diagramme.

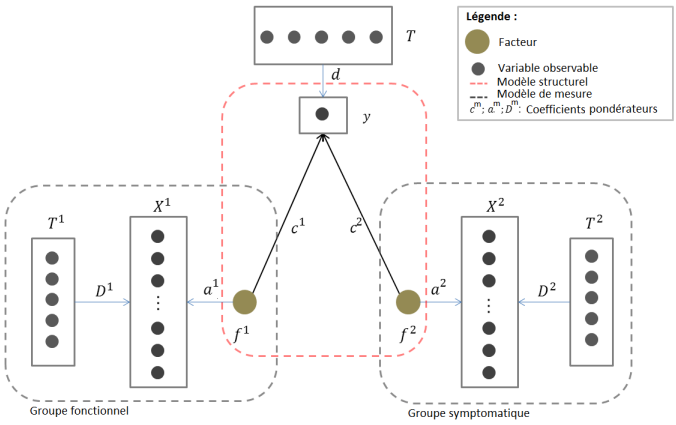

Figure 1: Diagramme du modèle structurel établi par la décomposition du QLQ-C30 à chaque temps de suivi.

2.1.2 Estimation par algorithme EM

L'algorithme EM de Dempster et al. (1977) est une procédure générale pour maximiser la vraisemblance. Il est souvent utilisé dans le cas de problèmes à données manquantes. Dans le cadre des SEMs à variables latentes, une approche proposée par Bry et al (2016) et Tami et al. (2014) utilise l'outil EM en considérant les facteurs comme des données manquantes pour maximiser la vraisemblance complète. Ainsi cette approche a l'avantage

d'estimer les facteurs $\widetilde{f_i^m}$ pour chaque individus i en plus des paramètres θ du modèle, où $\theta = \{d, D^1, D^2, a^1, a^2, c^1, c^2, \sigma_y^2, \sigma_1^2, \sigma_2^2\}$.
Dans le cadre de l'analyse transversale, le modèle (**??**) est établi à chaque visite v. Puis par la méthode d'estimation fondée sur EM, à chaque visite v, les estimations facteurs $\widetilde{f}_{iv} = \left(\widetilde{f}_{iv}^1, \widetilde{f}_{iv}^2\right)$ en plus des estimations des paramètres sont obtenus. Alors, la concaténation de la variable GHS et de ces estimations des deux facteurs sur l'ensemble des temps de suivi permet de passer à l'étude longitudinal des données. Ainsi, on pallie à l'inconvénient de la multiplicité des tests : au lieu d'analyser la QdV dimension par dimension de manière indépendante, on ne se focalise que sur les deux facteurs estimés.

2.2 Seconde étape : Analyse longitudinale par modèle linéaire mixte

L'analyse longitudinale est réalisée via un modèle linéaire mixte. Le but est d'étudier l'évolution de la QdV au cours du temps de suivi représentée par les mesures répétées de la variable GHS notée y_{iv} lors de la visite v pour un individu i. Ceux-ci sont décrits par un modèle linéaire mixte standard :

$$y_{iv} \quad = \quad \alpha + \boldsymbol{x}_{iv}'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\xi}_i + \varepsilon_{iv}, \tag{3}$$

où, α est l'intercept ; $\boldsymbol{\beta}$ le vecteur des effets fixes ; $\boldsymbol{x}_{iv}$ le vecteur "design" contenant les estimations des facteurs obtenus par la procédure précédente $\widetilde{f}_{iv} = \left(\widetilde{f}_{iv}^1, \widetilde{f}_{iv}^2\right)$ et d'autres covariables ; $\boldsymbol{\xi_i}$ le vecteur des effets aléatoires individuels tel que $\xi_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ où Σ est la matrice de covariance et $\boldsymbol{u}_i$ le vecteur "design" et $\varepsilon_{iv} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ sont les termes d'erreurs. Les deux principaux avantages de cet outils vont être de prendre en compte la variabilité induite par les données répétées dans le temps pour un même patient i et de quantifier la part d'information apportée par les variables explicatives.

3 Application

Une application sur des données réelles de QdV issues de l'essai clinique CO-HO-RT (Azria et al. (2010)) sera présentée et complétée par son analyse longitudinale. Le contexte est celui d'un étude de phase 2 randomisée évaluant les toxicités cutanées d'un traitement par radiothérapie-létrozole concomitant ou radiothérapie suivie par létrozole en situation adjuvante de cancer du sein. Le nombre d'observations par visites (questionnaires entièrement remplis) pour un total de 121 patientes restantes est variable au cours du temps de suivi. Le tableau suivant récapitule le nombre d'observations i disponibles en fonctions des 8 visites v.

Visite v (mois)	0	3	6	12	15	18	21	24
n_v	113	106	102	100	102	84	91	90

Dans le cadre de ces données, nous avons étudié l'effet du type de traitement sur l'évolution de la QdV au cours du temps de suivi des patientes. Pour ce faire, nous avons dans un premier temps établi des SEMs avec covariables et procédé à leurs estimations. Nous avons ensuite à chaque visite comparés les facteurs estimés $\widetilde{f}^1_{iv}$ entre eux et procédé de même pour les estimations $\widetilde{f}^2_{iv}$. Les résultats obtenus sont qu'aucun des deux types de traitements reçus par les patients n'a d'effet significatif sur leur QdV.
Lors de la seconde étape, nous avons également réalisé une procédure de sélection de modèles au sens du BIC et celui retenu est le suivant :

$$y_{iv} \quad = \quad \alpha + \beta_1 \widetilde{f^1_{iv}} + \beta_2 \widetilde{f^2_{iv}} + \xi_i + \varepsilon_{iv}$$

À partir de ce modèle nous nous sommes intéressé à la part d'information portée par les différents éléments du modèle. Le résultat est que le facteur fonctionnel $\widetilde{f^1_{iv}}$ explique deux fois plus le GHS que le facteur symptomatique $\widetilde{f^2_{iv}}$. Mais cela est à pondérer par la forte corrélation existante entre les deux facteurs : la présence des deux facteurs dans le modèle a du sens.

Bibliographie

[1] Aaronson, N K et Ahmedzai, S et Bergman, B et Bullinger, M et Cull, A et Duez, N J et Filiberti, A et Flechtner, H et Fleishman, S B et de Haes, J C (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 the National Cancer Institute*.
[2] Bry, X. et Lavergne C. et Tami M. (2016), EM estimation of a Structural Equation Model, *pré-publication*, Montpellier, France.
[3] Tami M. et Bry, X. et Lavergne C. (2014), Estimation par algorithme EM pour modèles à facteurs et à équations structurelles, *JDS 2014*, Rennes, France.
[4] Barbieri A. et Tami M. et Bry X., Azria D. et Gourgou S. et Bascoul-Mollevi C. et Lavergne C. (2016), EM algorithm estimation of a structural equation model for the longitudinal study of the quality of life, *pré-publication*, Montpellier, France.
[5] Jöreskog, K. G. (1970), A general method for analysis of covariance structures, *Biometrika*.
[6] Wold H. (1966), Estimation of Principal Components and Related Models by Iterative Least squares, *Academic Press*.
[7] Fayers, P. M. et Aaronson, N. K. et Bjordal, K. et Groenvold, M. et Curran, D. et Bottomley, A. (2001), *EORTC QLQ-C30 Scoring Manual* (3rd edition).
[8] Dempster, A. P. et Laird, N. M. et Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B.
[9] Azria D., Belkacemi Y., Romieu G et al. (2010), Concurrent or sequential adjuvant letrozole and radiotherapy after censervative surgery for early-stage breast cancer (CO-HO-RT): a phase 2 randomised trial, *The Lancet Oncology*.