[Joint modeling of quantitative longitudinal data and censored survival time]

Hélène Jacqmin-Gadda, Rodolphe Thiébaut, Jean-François Dartigues

Revue d'épidémiologie et de santé publique 2004;52(6):502-10

Modélisation conjointe de données longitudinales quantitatives et de délais censurés

Joint Modeling of quantitative longitudinal data and censored survival time

H. Jacqmin-Gadda[1], R. Thiébaut[1], J.F. Dartigues[2]

[1] : INSERM E 0338, Bordeaux, France

[2] : INSERM U593, Bordeaux, France

Correspondance et tirés à part :

Helene Jacqmin-Gadda, INSERM E0338, Université Victor Segalen Bordeaux II, case 11, 146 rue Léo Saignat, 33076 Bordeaux Cedex

Tel : 05 57 57 45 18

Fax : 05 56 24 00 81

Email : helene.jacqmin-gadda@bordeaux.inserm.fr

Titre abrégé : Modèles joints

Abstract

Background : In epidemiology, we are often interested in the association between the evolution of a quantitative variable and the onset of an event. The aim of this paper is to present joint model for the analysis of gaussian repeated data and a survival time. These models allow for example to perform a survival analysis when a time-dependent explanatory variable is measured intermittently, or to study the evolution of a quantitative marker conditionally to an event

Methods : They are constructed by combining a mixed model for repeated Gaussian variables and a survival model which can be parametric or semi-parametric (Cox model).

Results : We discuss the hypotheses underlying the different joint models proposed in the literature and the necessary assumptions for maximum likelihood estimation. The interest of these methods is illustrated on a study of the natural history of dementia in a cohort of elderly.

Keywords □

Longitudinal data. Survival model. Mixed model. Joint model. Random effects

Résumé :

Position du problème : Il est fréquent en épidémiologie de s'intéresser à l'association entre l'évolution d'une variable quantitative et la survenue d'un événement. L'objectif de cet article est de présenter les modèles conjoints pour l'analyse de données gaussiennes répétées et d'un temps de survie. Ces modèles permettent par exemple de réaliser une analyse de survie lorsqu'une variable explicative dépendant du temps est mesurée par intermittence ou d'étudier l'évolution d'un marqueur quantitatif conditionnellement à un événement.

Methods : Ils sont construits en combinant un modèle mixte pour variables gaussiennes répétées et un modèle de survie qui peut être paramétrique ou semi-paramétrique (modèle de Cox).

Results :Nous discutons les hypothèses sous-jacentes aux différents modèles conjoints proposés dans la littérature et les hypothèses nécessaires pour l'estimation de ces modèles par le maximum de vraisemblance. L'intérêt de ces méthodes est illustré sur une étude de l'histoire naturelle de la démence dans une cohorte de personnes âgées.

Mots clés :

Données longitudinales. Modèle de survie. Modèle mixte. Modèle joint. Effets aléatoires.

1 Introduction

Dans les études épidémiologiques prospectives ou les essais cliniques, il est fréquent de recueillir simultanément des données quantitatives répétées (des marqueurs biologiques par exemple) et des délais jusqu'à la survenue d'un événement (décès, pathologies,...). Un objectif important de ces études peut être l'analyse de la relation entre l'évolution des variables quantitatives et la survenue de l'événement. L'étude du risque d'infection opportuniste en fonction de l'évolution des marqueurs du VIH (CD4 et charge virale) constitue un exemple classique. Dans la suite, nous appellerons marqueur, la variable quantitative répétée, qu'il s'agisse ou non d'un marqueur biologique. Par exemple, dans les études du vieillissement cérébral, on s'intéresse également au risque de démence sénile en fonction de l'évolution des performances cognitives mesurées par des tests neuropsychologiques. Ce type d'analyses peut être réalisé à l'aide d'un modèle de survie tel que le modèle à risques proportionnels en considérant le marqueur comme une variable explicative dépendant du temps. Cette approche pose cependant deux problèmes : la variable quantitative est généralement mesurée avec erreur et seulement à des temps discrets. Or, l'estimation d'un modèle à risques proportionnels par la maximisation de la vraisemblance partielle nécessite la connaissance des valeurs de toutes les variables explicatives pour tous les sujets à risque à tous les temps d'événements. Une méthode fréquemment utilisée consiste à imputer la dernière valeur observée (Last Observation Carried Forward : LOCF) mais elle conduit à des estimateurs biaisés [1,2], particulièrement lorsque le délai entre les mesures est long ou variable selon les caractéristiques des sujets (enquêtes d'observation) ou lorsque l'erreur de mesure est importante.

Pour s'affranchir de l'erreur de mesure et obtenir des estimations des valeurs individuelles

du marqueur à tous les temps, on peut estimer les paramètres d'un modèle mixte sur les données quantitatives répétées et utiliser les prédictions individuelles issues de ce modèle comme variable dépendant du temps dans le modèle de survie. Nous détaillerons plus loin l'intérêt et les limites de cette stratégie en deux étapes qui permet de réduire les biais sur les paramètres du modèle de survie. L'estimation simultanée des paramètres du modèle de survie pour le délai jusqu'à l'événement et des paramètres du modèle mixte pour l'évolution du marqueur, par maximisation de la vraisemblance conjointe des deux variables en tenant compte de leur dépendance [3], est cependant la méthode la plus satisfaisante car elle conduit à des estimateurs non biaisés si les hypothèses paramétriques sont vérifiées. Les modèles conjoints sont également utiles pour étudier la qualité de marqueurs biologiques en tant que marqueur de substitution dans un essai thérapeutique [4], pour améliorer l'estimation de la survie marginale en tenant compte d'une variable auxiliaire (le marqueur) ou pour prédire le délai de survenue de l'événement en fonction de l'évolution du marqueur.

Dans les exemples précédents, l'objectif principal est l'estimation des paramètres d'un modèle de survie, mais la modélisation conjointe peut également avoir pour objectif majeur l'analyse de l'évolution du marqueur. Un domaine d'application important est ainsi l'étude de l'évolution du marqueur en présence de sorties d'étude informatives. En effet, dans les études longitudinales, lorsque la probabilité de sortie d'étude dépend des valeurs manquantes du marqueur les données manquantes sont dites informatives ou non ignorables [5]. Dans ce cas, les sorties d'études informatives induisent un biais de sélection qui affectent les estimateurs des paramètres du modèle d'évolution du marqueur. Un modèle conjoint pour l'évolution du marqueur et le temps de sortie d'étude (considéré comme un événement, lui-même censuré par la fin programmée de l'étude) permet de tenir compte de la probabilité de sortie d'étude et donc de corriger ces bais si le modèle est

correcte (ce qui est difficile à vérifier en pratique). Dans ce contexte, ces modèles conjoints sont appelés modèles de sélection car ils décrivent le processus de sélection des données [6]. Plus généralement, la modélisation conjointe permet d'étudier l'évolution d'un marqueur conditionnellement à la survenue d'un événement.

L'objectif de cet article est de présenter les principes de la modélisation conjointe de données longitudinales gaussiennes et d'un délai censuré en décrivant les principaux modèles proposés dans la littérature. L'intérêt et les limites de ces approches sont exposés en insistant sur les hypothèses sous-jacentes aux différents modèles et aux méthodes d'estimation. La section 2 est consacrée à la description des modèles et la section 3 aux méthodes d'estimation. Nous illustrons ensuite l'intérêt de ces modèles sur une étude de la détérioration cognitive dans la phase prédiagnostique de la démence sénile. L'objectif principal de cette analyse est la description de l'évolution d'un test de mémoire visuelle conditionnellement à l'âge de survenue de la démence.

2 Modèles conjoints

2.1 Modèle pour données quantitatives longitudinales

Le modèle classique pour l'évolution d'une variable quantitative gaussienne répétée est le modèle linéaire mixte [7]. Il est donc naturellement utilisé pour la modélisation de l'évolution du marqueur dans les modèles conjoints [3,8-13]. Soit $Y_i^*(t)$ la vraie valeur du marqueur pour le sujet i au temps t et $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ le vecteur des réponses observées (avec erreur) du sujet i aux temps $t_i = (t_{i1}, t_{i2}, ..., t_{in_i})'$: $Y_{ij} = Y_i^*(t_{ij}) + e_{ij}$ avec $e_{ij} \sim N(0, \sigma_e^2)$ et les erreurs e_{ij} sont indépendantes. On note Z_{ij} un vecteur de variables explicatives incluant le temps, éventuellement des variables dépendant du temps dont les

valeurs sont connues en tous temps et des interactions avec le temps ; β est le vecteur d'effets fixes correspondant. Le modèle le plus couramment utilisé est le modèle à ordonnée à l'origine (α_{0i}) et pente (α_{1i}) aléatoires qui s'écrit :

$$Y_i^*(t_{ij}) = Z_{ij}\beta + \alpha_{0i} + \alpha_{1i}t_{ij} \text{ avec } \alpha_i = (\alpha_{0i}, \alpha_{1i})' \sim N(0, G) \quad (1)$$

Ce modèle suppose que l'évolution de chaque sujet est linéaire avec une pente et une ordonnée à l'origine différentes pour chaque individu. Il se généralise sans difficulté au cas où la composante aléatoire est un polynôme du temps ou une combinaison linéaire de fonctions quelconques du temps. Même sous cette forme plus générale, ce modèle peut paraître trop rigide puisqu'il suppose que la trajectoire d'un individu est totalement déterminée par les effets aléatoires qui ne varient pas au cours du temps. Certains auteurs ont donc proposé d'inclure un processus stochastique $U_i(t)$ dans le modèle de $Y_i^*(t)$, qui représente les fluctuations individuelles à court terme autour de la tendance individuelle à long terme décrite par les effets aléatoires [9,10]:

$$Y_i^*(t_{ij}) = Z_{ij}\beta + \alpha_{0i} + \alpha_{1i}t_{ij} + U_i(t) \quad (2)$$

où $U_i(t)$ est un processus gaussien. Il s'agit fréquemment d'un processus autoregressif [9] tel que $Cov(U_i(t), U_i(s)) = \sigma_U^2 \exp\left(-|t-s|^{\upsilon}\right)$ mais une structure de covariance plus générale peut-être envisagée [10]. Ce modèle est plus souple puisqu'il tient compte des fluctuations à court terme, dues par exemple aux variations de l'état de santé du sujet, mais il est peu utilisé dans les modèles conjoints car il pose plus de problèmes d'estimation (cf section 3). Dans les modèles (1) et (2), $Z_{ij}\beta$ représente la moyenne pour la population, et nous noterons $W_i(t_{ij})$, la déviation individuelle par rapport à la moyenne de la population : $W_i(t_{ij}) = \alpha_{0i} + \alpha_{1i}t_{ij}$ ou $W_i(t_{ij}) = \alpha_{0i} + \alpha_{1i}t_{ij} + U_i(t)$ selon le modèle.

2.2 Modèle pour la durée de vie

Le délai de survenue de l'événement peut suivre un modèle paramétrique [11-13], ce qui simplifie l'estimation des paramètres, ou un modèle semi-paramétrique des risques proportionnels [8-10]. Nous présenterons un modèle log-normal (ou modèle de vie accélérée) dans l'application (section 4) et nous nous centrerons ici sur le modèle de Cox, plus fréquemment utilisé. Notons T_i le temps de survenue de l'événement et C_i le temps de censure. On observe $X_i = \min[T_i, C_i]$, l'indicateur de survenue de l'événement $\delta_i = 1_{[T_i < C_i]}$ et éventuellement un vecteur de variables explicatives ζ_i. Le risque instantané peut s'écrire :

$$\lambda_i(t) = \lambda_0(t)\exp\left(Y_i^*(t)\gamma + \zeta_i'\eta\right) \quad (3)$$

où η est le vecteur de paramètres fixes correspondants aux variables explicatives ζ_i, et γ est le paramètre qui mesure l'association entre le marqueur et le risque de l'événement. Outre la forme paramétrique ou non du modèle, un point essentiel de la modélisation conjointe est la forme de la dépendance entre le modèle de survie et le modèle mixte. Dans l'expression ci-dessus, le risque instantané de survenue de l'événement dépend de la « vraie » valeur courante du marqueur (« vraie » signifiant « débarrassée de l'erreur de mesure ») mais il peut dépendre seulement des effets aléatoires (ordonnée à l'origine et pente individuelles par exemple) :

$$\lambda_i(t) = \lambda_0(t)\exp\left(\alpha_i'\gamma + \zeta_i'\eta\right), \quad (4)$$

ou de la déviation individuelle courante par rapport à la moyenne de la population $W_i(t)$:

$$\lambda_i(t) = \lambda_0(t)\exp\left(W_i(t)\gamma + \zeta_i'\eta\right). \quad (5)$$

Des fonctions plus complexes de l'historique du processus Y peuvent également être envisagées, de même qu'une dépendance simultanée sur différents aspects de la

trajectoire : valeur courante et pente individuelle par exemple. Selon le cas, γ est un vecteur ou un scalaire. Si $W_i(t)$ inclut un processus stochastique, les modèles conjoints associant (2) et (3) ou (5), font l'hypothèse que le risque de l'événement dépend des fluctuations du marqueur à court terme et pas seulement de la tendance à long terme d'évolution du patient. Au contraire, si le modèle mixte n'inclut pas de processus stochastique (modèle (1)) ou si le modèle de survie n'en dépend pas (modèle associant (2) et (4)), le risque de survenue de l'événement ne dépend que de l'évolution à long terme du sujet.

3 Estimation des modèles conjoints

3.1 L'approche en deux étapes

Si l'objectif est l'étude du risque de survenue de l'événement en fonction du marqueur, c'est à dire une analyse de survie avec une variable explicative dépendant du temps, on peut réaliser l'analyse en deux étapes. Dans un premier temps le modèle mixte est estimé sur les données longitudinales et les prédictions individuelles de la valeur du marqueur $Y_i^*(t)$ pour chaque sujet à chaque temps d'événement sont calculées en utilisant les estimateurs bayésiens empiriques. Le modèle de survie est ensuite estimé en utilisant ces prédictions comme variables explicatives dépendant du temps. Cette méthode d'imputation réduit nettement le biais par rapport aux méthodes d'imputations naïves telles que LOCF tout en étant réalisable avec les logiciels courants [1,14,15].

Plus précisément, deux procédures peuvent être adoptées : soit le modèle mixte n'est estimé qu'une fois sur l'ensemble des mesures du marqueur recueillies sur tous les sujets jusqu'à leur sortie d'étude ou jusqu'à la survenue de l'événement [14], soit il est réestimé pour chaque temps d'événement t_k en utilisant exclusivement les mesures effectuées avant t_k sur les sujets à risque en t_k [15]. Cette seconde stratégie évite des biais potentiels dus à

l'utilisation de données postérieures à t_k pour estimer $Y_i^*(t_k)$, mais elle est plus lourde et présente plusieurs inconvénients : pour les premiers temps d'événements les modèles sont estimés sur très peu de données ; les paramètres du modèle utilisé pour les prédictions changent à chaque temps ce qui est peu cohérent ; enfin, si l'événement est lié au marqueur, les échantillons à risque (qui par définition exclus les sujets qui ont déjà connus l'événement) sont de plus en plus sélectionnés au cours du temps et l'hypothèse de normalité des effets aléatoires n'est clairement pas vérifiée pour l'ensemble des modèles estimés. Sachant que les prédictions individuelles obtenues par l'estimateur bayésien empirique dépendent fortement de l'hypothèse de normalité des effets aléatoires, ce dernier point explique en grande partie la persistance d'un biais mis en évidence par les simulations [16]. Bycott et Taylor [17] proposent des procédures intermédiaires dans lesquelles une partie seulement des données postérieures au temps d'événement considéré t_k est utilisée pour prédire $Y_i^*(t_k)$.

Ces procédures d'imputation utilisant le modèle mixte ne sont d'autre part pas totalement satisfaisante car les estimateurs des variances des paramètres du modèle de survie ne tiennent pas compte de l'incertitude sur l'estimation de $Y_i^*(t_k)$. Une étude par simulation suggère cependant que l'impact de ce problème est mineur [16]. Par contre, si l'événement est associé à la valeur courante du marqueur ou aux effets aléatoires, il induit une sortie d'étude informative [5] car les mesures postérieures à l'événement ne sont pas utilisées et conduit à des estimations biaisées des paramètres du modèle mixte. Enfin, cette approche ne répond pas à l'ensemble des objectifs des modèles conjoints : elle permet l'estimation d'un modèle de survie avec une variable explicative dépendant du temps mesurée par intermittence mais, si l'analyse est centrée sur l'évolution du marqueur, la maximisation de la vraisemblance conjointe est nécessaire.

3.2 La vraisemblance conjointe

Les paramètres du modèle mixte et du modèle de survie peuvent être estimés simultanément par maximisation de la vraisemblance conjointe. Nous supposons que la censure du temps de survenue de l'événement et le calendrier des mesures du marqueur sont non-informatifs. Plus précisément, pour un sujet dans l'étude au temps t, la probabilité d'être censuré en t et la probabilité que le marqueur soit mesuré en t sont conditionnellement indépendantes du temps de l'événement, de la valeur courante du marqueur et des effets aléatoires connaissant les valeurs passées observées du marqueur.

Si le modèle mixte est de la forme (1), en exploitant l'indépendance du marqueur Y_i et du temps de l'événement T_i conditionnellement aux effets aléatoires, la vraisemblance conjointe des données (Y_i, X_i, δ_i) peut s'écrire :

$$L(\theta) = \prod_{i=1}^{N} \int_{R^q} f_{Y_i|\alpha_i}(y_i|u)\left\{f_{T_i|\alpha_i}(x_i|u)\right\}^{\delta_i}\left\{1 - F_{T_i|\gamma_i}(x_i|u)\right\}^{1-\delta_i} f_{\alpha_i}(u)\, du \qquad (6)$$

Le terme $f_{Y_i|\alpha_i}$ est le produit de densités gaussiennes univariées car, sachant α_i, les Y_{ij} sont indépendants. Le terme suivant est soit la densité du temps de l'événement pour les sujets ayant subi l'événement, soit la fonction de survie pour les sujets censurés. L'intégrale sur les effets aléatoires n'a généralement pas de solution analytique et doit être calculée par quadrature gaussienne ou Monte Carlo. Lorsque le modèle de survie est paramétrique, $L(\theta)$ peut-être maximisée directement par un algorithme de type Newton-Raphson [13] ou par l'algorithme EM [11,12]. Avec un modèle semi-paramétrique, le nombre de paramètres augmente car la fonction de risque de base $\lambda_0(t)$ doit être estimée et les différents auteurs utilisent un algorithme EM ou une approche bayésienne par MCMC (Monte Carlo Markov Chain) [8-10]. Lorsque le modèle mixte et le modèle de survie dépendent d'un processus gaussien, la vraisemblance peut être développée conditionnellement à $W_i(t)$ mais les

difficultés numériques croissent considérablement, même pour un modèle paramétrique, car la dimension de l'intégrale est alors le nombre de mesures du marqueur. Pour un modèle des risque proportionnels, Henderson *et al* [9] ont proposé un algorithme MCEM tandis que Wang et Taylor [10] ont choisi un algorithme de type MCMC en supposant le risque $\lambda_0(t)$ et le processus $U_i(t)$ constants par morceaux.

4 Exemple : Histoire naturelle de la démence sénile

Cette étude a pour objectif la description de l'évolution cognitive en phase prédiagnostique de la démence sénile et la comparaison avec l'évolution des sujets non déments. Nous souhaitons en particulier identifier le moment où le déclin des sujets qui vont devenir déments commence à se distinguer du vieillissement cognitif normal et estimer le délai entre cette accélération du déclin cognitif et le diagnostic de démence.

Les données sont issues de la cohorte de personnes âgées Paquid [18] qui comportait 3777 sujets de 65 ans et plus à l'inclusion en 1988-1989. Le niveau cognitif est mesuré par le test de mémoire visuelle de Benton qui a été proposé aux participants lors de la visite initiale puis à l'occasion de chaque suivi (1 an, 3 ans, 5 ans, 8 ans et 10 ans après la visite initiale). A chaque suivi, la présence ou l'absence d'une démence a été recherchée cliniquement. L'échantillon d'analyse est constitué de 2960 sujets non déments à l'inclusion et revus au moins une fois au cours du suivi, parmi lesquels 437 ont été diagnostiqués déments entre T1 et T10. Les mesures du Benton postérieures à la visite de diagnostic de la démence sont ignorées ; de même, les scores recueillis lors de la visite initiale ne sont pas utilisés car un effet de primo-passation précédemment mis en évidence aurait compliqué l'analyse [18]. L'âge de démence est estimée par la moyenne entre l'âge à la visite de diagnostic et l'âge à la visite précédente. L'âge de démence médian observé parmi les 437 sujets diagnostiqués déments au cours du suivi est de 85 ans.

Une première analyse a été réalisée sans modéliser conjointement la survenue de la démence. Afin d'identifier la période d'accélération du déclin cognitif avant le diagnostic de démence, nous avons fait l'hypothèse que l'évolution du score au test de Benton pour les déments présente deux phases : une première phase de déclin linéaire lent qui traduit le vieillissement cognitif normal, suivie, quelques années avant le diagnostic, par une phase de déclin plus prononcé et non linéaire. Ce type d'approche a été utilisé précédemment pour étudier le délai entre l'accélération du déclin cognitif et la démence [19] ou le décès [20]. Le modèle retenu ici est un modèle à effets mixtes linéaire dans la première phase et quadratique dans la seconde phase. Si on note Y_{ij} le score du sujet i à l'âge t_{ij}, D_i la variable indicatrice de démence et T_i l'âge de démence; ce modèle peut s'écrire :

$$Y_{ij} = (\beta_0 + u_{oi}) + (\beta_1 + u_{1i})(t_{ij} - 65) + (\beta_2 + u_{2i})(t_{ij} - (T_i - \theta_i))^{+2} D_i + e_{ij} \qquad (7)$$

où $(t_{ij} - (T_i - \theta_i))^{+2} = (t_{ij} - (T_i - \theta_i))^2$ si $t_{ij} > (T_i - \theta_i)$ et 0 sinon. Le délai entre l'accélération du déclin cognitif et la démence, noté θ_i, peut être supposé identique pour tous les sujets ($\theta_i = \theta$ pour tout i) et estimé par profil de vraisemblance [19,20], mais dans le modèle proposé ici, θ_i est aléatoire, spécifique à chaque individu, et distribuée normalement d'espérance μ_θ et variance σ_θ^2. De même, le vecteur $u_i = (u_{0i}, u_{1i}, u_{2i})'$ d'effets aléatoires suit une distribution trivariée normale.

Les estimations des paramètres du modèle (7) ajusté sur le niveau d'études sont présentés dans la table 1. Le score au test de Benton des sujets qui n'ont pas le certificat d'études primaires est en moyenne inférieur de 1.8 points (intervalle de confiance à 95 % [-2.0 - -1.6]) au score des sujets de niveaux d'études plus élevés. Parmi les sujets normaux et parmi les déments avant la phase d'accélération du déclin, le score au test de Benton décroit d'environ 0.10 points par an (IC à 95 % [-0.11 - -0.09]). Chez les déments, le délai moyen entre l'accélération du déclin et la démence est estimé à 7.7 ans (erreur

standard, E.S : 1.13) et son écart-type à 1.8 ans (E.S. : 0.58).

Dans le modèle précédent, seuls les sujets diagnostiqués déments au cour du suivi connaissent une accélération du déclin. Ceci constitue une limite importante car certains des sujets considérés comme non déments peuvent être en phase prédiagnostique de la démence. En particulier, dans notre application, le groupe de non-déments inclut probablement une proportion non négligeable de sujets en phase prédémentielle car il comprend tous les sujets qui n'ont pas été diagnostiqués déments avant la fin de leur suivi, quelle que soit la durée du suivi, et il a été précédemment montré dans l'étude Paquid que la sortie d'étude et les non-réponses aux tests psychométriques étaient associés à un déclin cognitif plus marqué et un risque de démence élevé [18,21]. Si l'on restreint le groupe des non-déments aux sujets vus non-déments à T10, le déclin estimé pour les normaux est effectivement plus faible (-0.087 points /an , IC à 95 % [-0.10 - -0.075]).et le délai moyen entre l'accélération du déclin et la démence est plus long (9.1 ans, ES : 1.3) car l'écart entre les normaux et les déments est plus grand. Cette analyse n'est cependant pas satisfaisante car l'échantillon des sujets non-déments vus à T10 constitue un échantillon très sélectionné de l'ensemble des non-déments et il inclut probablement toujours une fraction de sujets en phase prédiagnostique de la démence.

Une seconde analyse a donc été réalisée en utilisant l'ensemble des données disponibles et un modèle conjoint pour l'évolution du score cognitif et la survenue de la démence. Dans ce modèle, tous les sujets, qu'ils aient ou non été observés déments au cours du suivi, peuvent présenter une évolution en deux phases (linéaire puis quadratique) et le logarithme de l'âge au changement de phase, notée τ_i, suit une distribution lognormale. L'âge de survenue de la démence suit une distribution log-normale dont l'espérance dépend de l'âge à l'accélération du déclin cognitif τ_i. Le modèle conjoint s'écrit donc :

$$Y_{ij} = (\beta_0 + u_{oi}) + (\beta_1 + u_{oi})(t_{ij} - 65) + (\beta_2 + u_{2i})(t_{ij} - \tau_i)^{+2} + e_{ij} \quad (8)$$

$$\log(T_{di}) = \eta + \gamma \log(\tau_i) + \varepsilon_i \text{ avec } \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad (9)$$

et $\log(\tau_i) \sim N(\mu_\tau, \sigma_\tau^2)$.

L'hypothèse d'une distribution log-normale a été validée en comparant la distribution marginale de l'âge de démence estimée par ce modèle à une estimation non-paramétrique. Le modèle conjoint définit par (8) et (9) permet d'une part d'estimer la distribution de l'âge de démence sachant l'âge à l'accélération du déclin cognitif, et notamment la médiane de cette distribution par $Med(T_i|\tau_i) = \exp(\eta)\tau_i^{\gamma}$, et d'autre part d'estimer l'espérance du score au Benton sachant l'âge de démence qui constitue l'objectif principal de cette étude.

Les paramètres de ce modèle ajusté sur le niveau d'études ont été estimés par maximisation de la vraisemblance (6) modifiée pour tenir compte de la troncature à gauche des données car les sujets doivent être non-déments à l'âge d'entrée dans la cohorte (Table 1). L'âge médian de démence et l'âge médian à l'accélération du déclin cognitif sont estimés respectivement à 91,6 ans et 84,1 ans. Le délai estimé entre l'accélération du déclin cognitif et la démence diminue avec l'âge : le délai médian estimé est de 10,8 ans lorsque l'accélération du déclin cognitif survient à 65 ans et de 7,3 ans lorsqu'il survient à 85 ans. Le calcul de la durée médiane de la phase de déclin accélérée pour un sujet dont l'âge de démence est 85 ans (âge médian de démence observé) donne un résultat très proche de la valeur estimée dans la première analyse (7,8 ans contre 7,7 ans). La figure 1 représente l'évolution moyenne estimée du score au test de Benton dans la période précédant la démence pour un âge de démence égal à 70 ans ou 90 ans.

L'hypothèse d'indépendance entre l'âge de démence et l'âge de censure conditionnellement aux scores observés au test de Benton constitue une limite de ce

modèle car nous n'avons pas tenu compte des décès. Les sujets décédés sont considérés ici comme censurés après leur dernière visite bien qu'une précédente analyse des données de Paquid ait montré que le risque de démence est plus élevé entre la dernière visite et le décès [22]. Un modèle conjoint trivariée pour le déclin cognitif, l'âge de démence et l'âge au décès permettrait de corriger d'éventuels biais liés aux décès et d'estimer l'évolution cognitive sachant que le sujet est vivant.

5 Discussion

Nous avons tenté de mettre en évidence la richesse de la modélisation conjointe de délais censurés et de données longitudinales gaussiennes pour l'analyse de données épidémiologiques. Nous nous sommes restreints au cas d'un marqueur et d'un événement mais des modèles plus généraux ont été proposés pour plusieurs marqueurs et/ou plusieurs événements [13,23].

Ces modèles requièrent cependant des hypothèses paramétriques qui sont difficiles à vérifier et peu d'études ont porté sur la robustesse aux différentes hypothèses. Lorsque le modèle de survie est paramétrique, il est généralement possible d'estimer la survie marginale et de la comparer à des estimations obtenues par des méthodes non paramétriques. Une évaluation succincte de l'ajustement du modèle mixte est également possible en comparant les moyennes observées à chaque temps aux moyennes des prédictions des sujets ayant une mesure au temps considéré. Par contre, la représentation de l'histogramme des estimateurs empiriques bayésiens des effets aléatoires apporte peu d'information concernant la validité de l'hypothèse de normalité des effets aléatoires car ces estimateurs sont très dépendants de la loi a priori retenue. Une étude par simulation a cependant montré la robustesse des estimations des effets fixes à cette hypothèse [24]. Enfin, la forme de la dépendance entre les deux variables est un élément central du modèle

particulièrement difficile à évaluer.

Par ailleurs, bien qu'utile pour analyser les données incomplètes, l'estimation des modèles conjoints nécessite des hypothèses concernant les données manquantes [3]. Le temps de censure doit par exemple être indépendant du temps de l'événement conditionnellement aux valeurs observées du marqueur, et les données manquantes sur le marqueur doivent être ignorables si elles ne sont pas dues à l'événement. Les hypothèses concernant les données manquantes ne sont pas vérifiables en pratique mais elles sont généralement moins fortes que celles qui sont nécessaires pour analyser séparément le risque de l'événement et l'évolution du marqueur. Par exemple, lors de l'estimation d'un modèle mixte classique sur les mesures répétées du marqueur, les sorties d'études liées à la survenue de l'événement introduiront un biais si elles sont informatives.

Les modèles conjoints sont intéressants pour l'étude de l'évolution de marqueurs et de la survenue d'événements et notamment dans trois types de problèmes fréquemment rencontrés dans les enquêtes prospectives : l'analyse de l'évolution d'un marqueur en tenant compte des sorties d'étude informatives, les analyses de survie avec une variable explicative dépendant du temps mesurée par intermittence et l'étude de l'histoire naturelle d'une pathologie. Leur usage reste cependant encore limité par la complexité des procédures d'estimation et l'absence de logiciel, et par le peu d'outils disponibles pour évaluer les hypothèses de ces modèles. L'estimation en 2 étapes est un compromis intéressant pour les analyses de survie avec variables explicatives dépendant du temps.

Références

1. Prentice R. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 1982; 69:331-342.
2. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. Stat Med. 1993; 12:835-842.
3. Tsiatis A, Davidian M. Joint modeling of longitudinal and time-to-event data: An overview. Stat Sinica 2004; sous presse.
4. Taylor JM, Wang Y. Surrogate markers and joint models for longitudinal and survival data. Control Clin Trials 2002; 23: 626-34.
5. Little RJA, Rubin DB. Statistical analysis with missing data. New-York : Wiley, 1987.
6. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995; 90:1112-1121.
7. Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982; 38: 963-974.
8. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics 1997; 53:330-339.
9. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics 2000; 1:465-480.
10. Wang Y, Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. J Am Stat Assoc 2001; 96:895-905.
11. De Gruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics 1994; 50:1003-1014.

12. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. Stat Med 1992; 11:1861-70.

13. Thiébaut R, Jacqmin-Gadda H, Babiker A, Commenges D and the CASCADE Collaboration. Joint modelling of bivariate longitudinal data with informative drop out and left censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. Stat Med 2004, sous presse.

14. Thiébaut R, Chêne G, Jacqmin-Gadda H *et al*. Time updated CD4+ T lymphocyte count and HIV RNA as prognostic factors in HIV-1 infected patients naive of antiretrovirals and newly treated with an highly active antiretroviral therapy. J Acquir Immune Defic Syndr 2003; 33 :380-386.

15. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error : Application to survival and CD4 counts in patients with AIDS. J Am Stat Assoc 1995; 90:27-37

16. Dafni UG, Tsiatis AA. Evaluating surrogate markers of clinical outcome measured with error. Biometrics 1998; 54: 1445-1462.

17. Bycott P, Taylor JMG. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. Stat Med 1998; 17: 2061-2077.

18. Jacqmin-Gadda H, Fabrigoule C, Commenges D, Dartigues JF. A five-year longitudinal study of Mini-Mental State Examination in normal aging. Am J Epidemiol 1997; 145: 498-506.

19. Hall CB, Lipton RB, Sliwinski M, Stewart WF. A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. Stat. Med. 2000; 19: 1555-1566.

20. Wilson RS, Beckett LA, Bienias JL, Evans MD, Bennett MD. Terminal decline in cognitive function. Neurology 2003; 60 : 1782-1787.

21. Dartigues JF, Commenges D, Letenneur D, *et al*. Cognitive predictors of dementia in elderly community residents. Neuroepidemiology 1997; 16: 29-39.

22. Joly P, Commenges D, Helmer C, Letenneur L. A Penalized likelihood approach for an Illness-Death model with interval-censored data: Application to age-specific incidence of dementia. Biostatistics 2002;3:433-43.

23. Xu J, Zeger SL. The evaluation of multiple surrogate endpoints. Biometrics 2001; 57: 81-87.

24. Song X, Davidian M, Tsiatis A. A semi-parametric approach to joint modelling of longitudinal and time-to-event data. Biometrics 2002; 58:742-753.

Table I : Estimation des parametres du modèle mixte à changement de pente aléatoire et du modèle conjoint pour l'évolution du score au test de Benton et la survenue de la démence (Paquid, N=2960)

	Modèle mixte (6)		Modèle conjoint (7 et 8)	
Paramètre	Estimation	(Erreur standard)	Estimation	(Erreur standard)
Intercept	12.6	(0.07)	12.4	(0.08)
Niveau d'études[a]	-1.8	(0.08)	-1.8	(0.08)
β_1	-0.10	(0.005)	-0.07	(0.007)
β_2	-0.037	(0.010)	-0.038	(0.007)
μ_θ	7.7	(1.13)	-	-
σ_θ	1.8	(0.58)	-	-
μ_τ	-	-	4.43	(0.014)
σ_τ	-	-	0.14	(0.007)
η	-	-	1.25	(0.16)
γ	-	-	0.74	(0.036)

[a] « Pas d'études » ou « pas de certificat d'études primaires (CEP) » versus « au moins le CEP »

Figure 1 : Moyenne estimée du score au test de Benton sachant l'âge de démence dans les années précédant le diagnostic pour des sujets ayant un niveau d'études supérieur au Certificat d'études primaires

_____ : âge de démence = 70 ans

-------: âge de démence = 90 ans