JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

RODOLPHE THIÉBAUT
HÉLÈNE JACQMIN-GADDA

Modélisation longitudinale de données incomplètes : exemple de la charge virale plasmatique du VIH

Journal de la société française de statistique, tome 145, n° 2 (2004), p. 33-47

<http://www.numdam.org/item?id=JSFS_2004__145_2_33_0>

MODÉLISATION LONGITUDINALE DE DONNÉES INCOMPLÈTES : EXEMPLE DE LA CHARGE VIRALE PLASMATIQUE DU VIH

Rodolphe THIÉBAUT * et Hélène JACQMIN-GADDA

RÉSUMÉ

Cet article présente un exemple d'analyse de données longitudinales incomplètes où la variable réponse est susceptible d'être censurée à gauche (du fait d'un seuil de détection de l'outil de mesure) et d'être manquante non aléatoirement du fait de sorties d'étude informatives. L'exemple utilisé est l'évolution de la charge virale plasmatique des patients infectés par le virus de l'immunodéficience humaine (VIH) après la mise en place d'un traitement antirétroviral et chez qui peut survenir un événement clinique (maladie opportuniste, décès) conduisant à la censure informative des données. Un modèle de sélection effets-aléatoires dépendant paramétrique est présenté. Une méthode d'estimation basée sur la maximisation directe de la vraisemblance permet de prendre en compte les données censurées à gauche. Des modèles de mélange sont également évoqués afin de tester l'hypothèse de distribution gaussienne des données censurées. L'application illustre l'impact de la prise en compte des données incomplètes de charge virale. Au total, la prise en compte des données censurées à gauche est facilement mise en œuvre et peut permettre de corriger des biais importants. La prise en compte des données manquantes informatives est difficile de par la définition du mécanisme, du choix du modèle et de sa mise en œuvre. Elle est toutefois souvent nécessaire, au moins en tant qu'analyse de sensibilité, car elle peut avoir un impact majeur sur les résultats.

Mots clés : Modèles mixtes, modèles multivariés, données censurées, données manquantes informatives, VIH.

ABSTRACT

This article shows an example of longitudinal analysis of incomplete data where the response variable can be left-censored (because of the assay detection threshold) and where measurements may be missing because of informative drop-outs. This is illustrated by the change in plasma viral load of patients infected with human immunodeficiency virus (HIV) after the initiation of antiretroviral treatment. Clinical events such as opportunistic diseases and death may occur, leading to an informative censoring of data. A parametric random-effects based selection model is presented. A direct likelihood maximisation allows one to take into account left-censoring of plasma viral load. Mixture models are proposed to test the Gaussian distribution

* Institut National de la Santé et de la Recherche Médicale, Équipe de Biostatistique E0338
INSERM E0338 ISPED, Université Bordeaux 2, Case 11,
146, rue Léo Saignat 33076 Bordeaux Cédex
Tel : (33) 5 57 57 45 21 ; Fax : (33) 5 56 24 00 81
courriel : rodolphe.thiebaut@isped.u-bordeaux2.fr

assumption of left-censored data. The application shows the effect of taking into account left-censoring. Finally, handling for the left-censoring of response variables is easily feasible and allow reducing biases. Dealing with informative drop-out is difficult because of the definition of the drop-out process, the choice of the model and the estimation of model parameters. However, it is most often necessary, at least as a robustness analysis, because it may have a huge effect on results.

Keywords : mixed models, multivariate models, censored data, missing data, HIV infection.

1. Introduction

L'analyse de données longitudinales gaussiennes dans le cadre de mesures répétées est de plus en plus répandue (Verbeke, Molenberghs 2000). Toutefois, la modélisation de données longitudinales peut être compliquée par la présence de données incomplètes. Ces données incomplètes peuvent être des données non observées ou partiellement observées (notamment censurées). Les données non observées peuvent être manquantes par intermittence ou de façon monotone. Soit M le processus d'observation de la variable réponse Y. Au temps de mesure t_{ij}, si $M_{ij} = 0$, on observe la $j^{\text{ème}}$ mesure du sujet i : Y_{ij}. Si $M_{ij} = 1$, la variable réponse n'est pas observée. Les données sont manquantes de façon monotone si, pour tout $k > j$, $P(M_{ik} = 1|M_{ij} = 1) = 1$. Dans ce travail, nous nous limitons aux données manquantes monotones. On note Y_i^o la composante observée et Y_i^m la composante manquante du vecteur Y_i. Dans un schéma d'étude où n_i mesures sont prévues chez chaque sujet i inclus, si $P(M_{ij} = 1|Y_i^o = Y_{i0}, \ldots Y_{ij-1}, Y_i^m = Y_{ij} \ldots Y_{in_i}) \neq P(M_{ij} = 1|Y_i^o)$, le processus d'observation dépend des données manquantes et les données sont dites manquantes non aléatoirement ou encore de façon informative (Rubin 1976; Minini, Chavance 2004). Par exemple, dans le cadre d'études épidémiologiques de cohorte, certains individus peuvent être perdus de vue, générant des données manquantes monotones à partir de la date de sortie d'étude. La raison pour laquelle ces individus ont été perdus de vue peut être associée à la valeur du marqueur après la sortie d'étude. Lorsque les données sont manquantes non aléatoirement, les estimateurs du maximum de vraisemblance sont asymptotiquement biaisés. La modélisation conjointe des données répétées et du processus de données manquantes peut permettre une estimation non biaisée des paramètres du modèle longitudinal (Henderson *et al.* 2000; Minini, Chavance 2004).

Les données observées partiellement peuvent être des données censurées c'est-à-dire dont on sait que la valeur est au-dessus ou au-dessous d'un seuil. Ce phénomène peut survenir lorsqu'un marqueur est quantifié avec une méthode dont la sensibilité n'est pas parfaite (avec un seuil de détection) si bien que certaines valeurs sont censurées à gauche. On note Y_i^o la composante observée et Y_i^c la composante de mesures censurées du vecteur Y_i. Sur le nombre total de mesures n_i pour un sujet i : n_i^o données sont complètement observées avec Y_{ij} et n_i^c données sont censurées à gauche, c'est-à-dire que l'on sait seulement qu'elles sont inférieures à Y_{ij}^c, le seuil de détection de la technique de mesure qui peut varier d'un sujet à l'autre et en fonction du temps. En présence de

telles données, les analyses ne tenant pas compte des données censurées (en imputant la moitié de la valeur du seuil, *i.e.* $Y_{ij} = \frac{Y_{ij}^c}{2}$, par exemple) génèrent des estimations biaisées (Jacqmin-Gadda *et al.* 2000).

L'objectif de cet article est de présenter une méthode d'analyse des données longitudinales gaussiennes tenant compte de ces deux types d'observation incomplète (données manquantes informatives et données indétectables). On prendra l'exemple d'une étude de cohorte de patients infectés par le virus de l'immunodéficience humaine (VIH) chez qui on a mesuré la charge virale plasmatique du VIH de manière répétée après la mise en place d'un traitement antirétroviral. En fait, la charge virale plasmatique représente la quantité de virus dans le sang circulant et fait partie des principaux marqueurs utilisés pour la décision de mise en place d'un traitement et pour le suivi de la réponse au traitement (Delfraissy 2002). L'analyse de la réponse au traitement nécessite la prise en compte de données répétées qui sont censurées à gauche du fait de la limite de détection des techniques de dosage (charges virales indétectables) et qui peuvent être manquantes du fait notamment de la survenue d'un événement clinique comme dans l'exemple utilisé dans cet article (Thiébaut *et al.* 2003). Dans la suite de cet article, on parlera de données indétectables pour les données censurées à gauche afin de faciliter la lecture.

2. Modèles pour données longitudinales gaussiennes incomplètes

2.1. Modèle linéaire mixte

Le modèle pour données longitudinales considéré est le modèle classique à effets mixtes décrit par Laird et Ware (Laird, Ware 1982) :

$$\begin{aligned} Y_i &= X_i\beta + Z_i\alpha_i + e_i,\ \alpha_I \sim N(0, G), \\ e_i &\sim N(0, \sigma^2 I_{n_i}) \text{ et } \alpha_i \text{ indépendant de } e_i \end{aligned} \tag{1}$$

avec X_i de dimension $(n_i \times p)$, Z_i de dimension $(n_i \times q)$, $p \geqslant q$ le nombre de variables explicatives, Z_i étant le plus souvent une sous-matrice de X_i. Les effets aléatoires individu-dépendants permettent de prendre en compte la corrélation des mesures répétées chez un même individu.

Les paramètres de ces modèles peuvent être estimés sans biais par maximisation de vraisemblance lorsque les données observées ne sont pas censurées à gauche et lorsque le mécanisme de données manquantes est aléatoire. Dans ce cas, les procédures classiques de logiciels standards sont utilisables, par exemple proc MIXED de SAS® (Littell *et al.* 1996) ou NLME de S-PLUS® (Pinheiro, Bates 2000).

2.2. Modèle linéaire mixte pour données indétectables

En présence de données indétectables, les paramètres du modèle mixte peuvent être estimés asymptotiquement sans biais grâce à la maximisation d'une vraisemblance prenant en compte la contribution des données indétectables en tant que telles. Deux formulations de la vraisemblance ont déjà été proposées. La première est conditionnelle aux données observées (Hughes 1999; Jacqmin-Gadda *et al.* 2000) :

$$L(\theta) = \prod_{i=1}^{N} f_{Y_i^o|\theta}(Y_i^o|\theta) \int_{H_1} \int_{H_2} \dots \int_{H_{n_i^c}} f_{Y_i^c|Y_i^o,\theta}(u|Y_i^o,\theta)\, \mathrm{d}u_1\, \mathrm{d}u_2 \dots \mathrm{d}u_{n_i^c} \quad (2)$$

avec θ le vecteur de paramètres du modèle à estimer, $H_d =]-\infty, Y_{id}^c]$ et $d = 1, \dots, n_i^c$. $f_{Y_i^o|\theta}$ est la densité multivariée normale du vecteur Y_i^o et $f_{Y_i^c|Y_i^o,\theta}$ est la densité multivariée normale du vecteur Y_i^c conditionnelle aux mesures observées Y_i^o.

La seconde formulation est conditionnelle aux effets aléatoires (Lyles *et al.* 2000) :

$$L(\theta) = \prod_{i=1}^{N} \left[\int_{R^q} \left\{ \prod_{j=1}^{n_{io}} f_{Y_{ij}^o|\alpha_i}(Y_{ij}^o|\alpha_i = u) \right\} \left\{ \prod_{j=1}^{n_{ic}} F_{Y_{ij}^c|\alpha_i}(Y_{ij}^c|\alpha_i = u) \right\} f_{\alpha_i}(u)\, \mathrm{d}u \right] \quad (3)$$

avec $f_{Y_{ij}^o|\gamma_i}$ densité conditionnelle univariée normale et $f_{Y_{ij}^c|\alpha_i}$ est la fonction de répartition de la loi univariée normale des mesures indétectables conditionnellement aux effets aléatoires. Dans la mesure où les effets aléatoires (α_i) sont supposés indépendants de l'erreur de mesure (e_i), chaque observation Y_{ij} est indépendante conditionnellement aux effets aléatoires. Chaque observation (mesure) est donc traitée indépendamment et selon son statut (observée ou censurée). Les vraisemblances (2) et (3) sont directement maximisables avec des logiciels disponibles (Thiébaut, Jacqmin-Gadda 2004). Le choix de la formulation peut dépendre notamment des difficultés numériques engendrées par le calcul des intégrales multiples. Ainsi, face à un modèle comprenant beaucoup d'effets aléatoires (plus de trois en pratique), la première formulation pourra être plus adaptée. En revanche, si de nombreuses mesures sont censurées (plus de dix en pratique), il sera préférable d'utiliser la formulation (3) intégrant sur les effets aléatoires.

Une hypothèse majeure de cette méthode est que l'ensemble des mesures Y_i (indétectables ou non) est issu d'une même distribution gaussienne. On peut cependant imaginer que les mesures censurées sont issues du mélange de distributions (Moulton, Halsey 1995). Par exemple, il peut s'agir d'un mélange d'une distribution normale et d'une distribution de Dirac (Berk, Lachenbruch 2002). Dans ce cas, on définit une variable aléatoire D suivant une loi de Bernoulli de paramètre τ avec $P(D = 1) = \tau$. Si $D = 1$ alors Y suit le modèle (1) et si $D = 0$, Y prend une valeur fixée avec une probabilité de 1. Par

exemple, on peut imaginer qu'une partie des données indétectables de charge virale correspondent à des échantillons où il n'y a plus de virus ($Y_{ij} = 0$). Sous ce modèle, la vraisemblance s'écrit :

$$L(\theta) = \prod_{i=1}^{N} \left[\int_{R^q} \left\{ \prod_{j=1}^{n_{io}} \tau f_{Y_{ij}^o|\alpha_i,n_i}(Y_{ij}^o|u) \right\} \left\{ \prod_{j=n_{io}+1}^{n_{ic}} (1-\tau) + \tau F_{Y_{ij}^c|\alpha_i,n_i}(Y_{ij}^c|u) \right\} f_{\alpha_i,n_i}(u)\,\mathrm{d}u \right]$$

Les paramètres de cette vraisemblance sont estimables à l'aide de la procédure NLMIXED de SAS® Un exemple de code est donné par Berk et Lachenbruch (Berk, Lachenbruch 2002).

2.3. Modèle de sélection effets-aléatoires dépendant

En présence de données manquantes informatives, la modélisation conjointe du processus d'observation et du marqueur permet potentiellement de réduire le biais sur les estimations. Par exemple, dans le cadre de données manquantes liées à une sortie d'étude définitive, on modélise conjointement l'évolution du marqueur par un modèle mixte et le délai jusqu'à la sortie d'étude informative ou la fin de l'étude. On considère comme événement une sortie d'étude informative ($\delta_i = 1$) avec T_i^o le temps ou une transformation du temps jusqu'à la sortie d'étude. On note C_i le temps jusqu'à la fin de l'étude ou jusqu'à la sortie d'étude si elle est considérée comme non informative ($\delta_i = 0$). On observe donc le couple $(T_i = \min\lfloor T_i^o, C_i \rfloor, \delta_i)$. Il faut noter que, dans ce modèle de survie destiné à prendre en compte une censure informative des données longitudinales, on considère la censure du temps de sortie d'étude comme non informative.

Plusieurs types de modèles conjoints ont été proposés dans la littérature (Little 1995; Minini, Chavance 2004), notamment les modèles de sélection variable réponse dépendants ou effets aléatoires dépendants. Nous présentons dans cet article un modèle de sélection effets aléatoires dépendants. Le choix de ce modèle a été guidé par l'application présentée dans la section 3. En effet, les sorties d'étude informatives liées à une progression clinique pouvaient survenir à tout moment. Pour utiliser un modèle de mélange, il faudrait arbitrairement regrouper les dates de sortie d'étude (par exemple par semestre). Un modèle de sélection nous semblait donc plus adapté. Le choix d'un modèle effets aléatoires dépendant semblait pertinent dans l'application car la progression clinique (définissant la sortie d'étude informative) était plus clairement associée à l'évolution sous-jacente du marqueur plutôt qu'à la valeur du marqueur à un temps déterminé (De Gruttola, Tu 1994). De plus, le choix d'un modèle effets aléatoires dépendant plutôt que variable réponse dépendant se justifiait par la sensibilité plus importante à la distribution des Y dans ce dernier type de modèle (Kenward 1998; Jacqmin-Gadda, Thiébaut 2004).

Le modèle conjoint s'écrit :

$$\begin{cases} Y_i &= X_i\beta + Z_i\alpha_i + e_i \\ T_i^o &= \mu_{T^o} + \varepsilon_i \end{cases} \tag{4}$$

$$\text{avec } \begin{pmatrix} \boldsymbol{\alpha_i} \\ T_i^o \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mathbf{0} \\ \mu_{T^o} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{B} \\ \boldsymbol{B}^T & \sigma^2_{T^o} \end{pmatrix} \right\}$$

Le lien entre l'évolution du marqueur et le temps de sortie d'étude est donc modélisé *via* la matrice de covariance B. Si les termes de covariance de B sont significativement différents de 0 (vérifier avec un test de Wald, par exemple) alors le marqueur Y et le temps de sortie d'étude sont associés *via* les effets aléatoires. Par ailleurs, on peut également interpréter cette covariance comme l'effet de l'évolution du marqueur sur le temps de participation à l'étude qui peut être lié à la survenue d'un événement clinique par exemple (De Gruttola, Tu 1994).

La vraisemblance du modèle (4) s'écrit :

$$L(\theta) = \prod_{i=1}^{N} \left[\int_{R^q} \left\{ \prod_{j=1}^{n_i} f_{Y_{ij}|\alpha_i}(Y_{ij}|u) \right\} \left\{ f_{T_i^o|\alpha_i}(T_i|u) \right\}^{\delta_i} \right.$$
$$\left. \left\{ F_{T_i^o|\alpha_i}(T_i|u) \right\}^{1-\delta_i} f_{\alpha_i}(u)\, \mathrm{d}u \right]$$

Elle est donc composée de deux parties : la contribution de chaque mesure du marqueur $f_{Y_{ij}|\alpha_i}$ et la contribution du temps jusqu'à la sortie d'étude de chaque individu. Si la sortie d'étude était considérée comme informative (donc comme un événement) alors $\delta_i = 1$ et la contribution pour le sujet i était $f_{T_i^o|\alpha_i}$, la densité conditionnelle d'une loi univariée normale. Si la sortie d'étude était considérée comme non informative, la contribution pour le sujet i était $F_{T_i^o|\alpha_i}$, la fonction de répartition de T_i^o conditionnelle aux effets aléatoires.

Les paramètres de ce modèle peuvent être estimés par maximisation directe de cette vraisemblance. Un programme sous SAS/IML a été proposé pour un modèle à intercept et pente aléatoires (Lyles *et al.* 2000). Lyles *et al.* utilisent une technique de quadrature gaussienne pour calculer l'intégrale et maximisent la vraisemblance à l'aide d'un algorithme de Newton-Raphson. Nous avons développé un programme Fortran 90 capable d'estimer les paramètres de modèles linéaires plus généraux. La vraisemblance est reparamétrée (décomposition de Cholesky) pour assurer une contrainte de positivité sur les paramètres de variance. L'intégrale multiple est calculée par une méthode de Monte Carlo ($n = 2\,000$ simulations). La maximisation est réalisée grâce à un algorithme de Marquardt. La convergence est considérée atteinte lorsque les trois conditions suivantes sont remplies :

- variation relative des paramètres $ca = \max \left| \frac{\theta^k - \theta^{k-1}}{\theta^k} \right| < 10^{-4}$ où θ^k est le vecteur de paramètres à estimer à l'itération k,
- variation relative de la vraisemblance $cb = \left| \frac{\ell^k - \ell^{k-1}}{\ell^k} \right| < 10^{-4}$ où ℓ^k est la vraisemblance à l'itération k,
- changement prédit de la fonction de vraisemblance $dd = (g^k)^T (H^k)^{-1} g^k < 10^{-4}$ où g^k est le gradient et H^k l'Hessienne.

Outre les hypothèses du modèle linéaire mixte, ce modèle paramétrique comporte également une hypothèse de distribution normale du temps (ou d'une transformation de ce temps). On peut vérifier l'adéquation de l'hypothèse de normalité avec les données en comparant les courbes de risques et/ou de survie estimées à l'aide du modèle normal avec celle obtenue non paramétriquement (avec l'estimateur de Kaplan-Meier, par exemple).

3. Application

3.1. Objectif

L'avènement des traitements antirétroviraux dits hautement actifs a permis une réduction majeure de l'incidence des maladies opportunistes. Ainsi, la décision d'initiation d'un traitement et son évaluation utilisent notamment la quantification de la charge virale plasmatique en substitution des événements cliniques. En effet, la charge virale plasmatique s'est avérée un bon marqueur de pronostic clinique (Mellors *et al.* 1997) mais sa valeur en tant que marqueur de substitution a été discutée (Albert *et al.* 1998). Disponibles depuis 1996, les techniques de quantification se sont considérablement améliorées avec des seuils de quantification passant de 10 000 copies/mL à quelques copies/mL. Cependant, dans la plupart des analyses effectuées, la charge virale était analysée en tant que variable catégorielle plutôt que continue en individualisant la catégorie « indétectable ». Bien que les techniques de quantification se soient améliorées, les données indétectables sont souvent nombreuses car les traitements hautement actifs engendrent une chute importante de la charge virale le plus souvent en dessous du seuil de détection. De plus, les analyses effectuées avec des données issues de cohortes observationnelles utilisent souvent les anciennes mesures effectuées avec des seuils assez élevés. L'objectif de l'application est d'illustrer les conséquences de la prise en compte des données incomplètes de charge virale du VIH dans le cadre de l'étude de l'efficacité des traitements antirétroviraux.

3.2. Données

Les données utilisées pour cette application proviennent d'une étude de cohorte déjà publiée (Thiébaut *et al.* 2003). Il s'agit d'un échantillon de 551 patients infectés par le VIH-1 chez qui un traitement antirétroviral hautement actif a été initié alors qu'ils n'avaient jamais reçu de traitement auparavant.

Dans cette cohorte, les patients étaient suivis selon les pratiques cliniques usuelles c'est-à-dire tous les 3 à 6 mois. Au cours d'une durée de suivi médiane de 33 mois (Etendue inter-quartile : 22-48), 5 331 mesures de charge virale ont été effectuées soit 4 mesures par sujet en médiane dont 56 % étaient indétectables, la plupart au seuil de 2.7 $\log_{10}$ copies/mL. Ces seuils étaient variables du fait de l'évolution des techniques au cours du temps et de l'utilisation de techniques différentes selon les centres.

Le suivi des patients a été censuré au 31 décembre 2001 ou au dernier suivi si celui-ci était antérieur à cette date ou en cas de survenue d'un événement clinique. Un événement clinique (pathologie classant SIDA ou décès) est survenu chez 60 patients. La censure du suivi à la survenue d'un événement clinique est justifiée par la modification majeure et variable de l'évolution de la charge virale plasmatique qui s'en suit. On peut donc supposer que les données manquantes à la suite d'un événement clinique sont des données manquantes informatives.

3.3. Modèle

La variable dépendante Y_{ij} est le logarithme en base décimale de la charge virale plasmatique (ARN VIH). Cette transformation améliore la normalité et l'homoscédasticité des résidus.

Étant donné l'évolution observée de la charge virale plasmatique après la mise en place du traitement (figure 1), un modèle linéaire par morceau a été proposé :

$$\begin{aligned} Y_{ij} =& \beta_0 + \beta_1 \inf(t_{ij}, 2) + \beta_2 (t_{ij} - 2) I_{t_{ij} \geqslant 2} \\ &+ \alpha_{0i} + \alpha_{1i} \inf(t_{ij}, 2) + \alpha_2 (t_{ij} - 2) I_{t_{ij} \geqslant 2} + e_i \end{aligned} \tag{5}$$

Le temps de changement de pente a été fixé à la même valeur pour tous les sujets car il était très dépendant du schéma d'étude et finalement peu variable d'un individu à l'autre. Il a été estimé par profil de vraisemblance en comparant la vraisemblance des modèles pour des valeurs de temps de changement de pente allant de 1 à 6 mois. Un changement de pente à 2 mois conduisait à la meilleure vraisemblance. L'inclusion d'effets aléatoires sur l'intercept (α_{0i}), la première pente (α_{1i}) et la seconde pente (α_{2i}) permet que le niveau initial et les pentes soient propres à chaque individu. β_0, β_1 et β_2 représentent le niveau moyen initial et les pentes moyennes pour la population d'étude.

Le modèle conjoint correspondant est le suivant :

$$\begin{pmatrix} \alpha_{0i} \\ \alpha_{1i} \\ \alpha_{2i} \\ T_i^o \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mu_{T^o} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\alpha_0} & & & \\ \sigma_{\alpha_0\alpha_1} & \sigma^2_{\alpha_1} & & \\ \sigma_{\alpha_0\alpha_2} & \sigma_{\alpha_1\alpha_2} & \sigma^2_{\alpha_2} & \\ \sigma_{\alpha_0 T^o} & \sigma_{\alpha_1 T^o} & \sigma_{\alpha_2 T^o} & \sigma^2_{T^o} \end{pmatrix} \right\} \tag{6}$$

T_i^o est le logarithme du délai en jour entre l'initiation du traitement antirétroviral et la survenue d'un événement clinique. Ce délai peut être censuré par le délai jusqu'à la date de point ou le dernier suivi (C_i). Cette transformation a été choisie car elle conduisait à des estimations proches des estimations non paramétriques (figure 2).

Les paramètres du modèle (5) ont été estimés sans prendre en compte les données indétectables de charge virale (en imputant la moitié de la valeur du seuil pour les charges virales indétectables) et en les prenant en compte à l'aide de la formulation de la vraisemblance conditionnelle aux effets aléatoires (3). Afin de vérifier l'hypothèse de normalité des données censurées, une troisième analyse a été réalisée en utilisant un modèle de mélange en supposant que les valeurs indétectables étaient issues d'un mélange d'une distribution de Dirac avec une probabilité τ et de la distribution gaussienne définie par (5) et (6) avec une probabilité $1-\tau$:

$$logit\ (\tau) = \gamma_0 + \eta_i \qquad \text{et} \qquad \eta_i \sim N(0, \sigma_n^2)$$

Enfin, les paramètres du modèle conjoint ont été estimés en prenant en compte les donnees indétectables de charge virale et les sorties d'étude informatives.

3.4. Résultats

Les estimations des paramètres des modèles sont présentées dans le tableau 1. Les courbes d'évolution de la charge virale prédites selon le modèle sont présentées dans la figure 1. Les estimations obtenues à partir d'un modèle mixte sans prendre en compte les données indétectables de charge virale sont très proches des observations où les données indétectables ont été remplacées par la moitié de la valeur du seuil. On peut noter toutefois que les estimations à partir de 21 mois tendent à être moins optimistes que les observations (charge virale plus élevée). Cela peut être lié aux données manquantes : la moyenne des observations est calculée sur 270/551 patients. Si la non observation d'une partie de ces données est associée aux valeurs antérieures de la charge virale, alors le biais des moyennes estimées sur les données observées est partiellement corrigé par le modèle mixte.

La prise en compte des données indétectables de charge virale met en évidence la sous-estimation de la décroissance de la charge virale déjà illustrée ailleurs (Hughes 1999; Jacqmin-Gadda *et al.* 2000). Ainsi, on retrouve une pente à long-terme dont la décroissance est plus forte ($\widehat{\beta}_2 = -0.18$ versus $0.020\log_{10}$ copies/mL/an) et dont la variabilité individuelle est beaucoup plus importante ($\widehat{\alpha}_2 = 0.36$ vs. 0.081). De plus, on retrouve une sous-estimation de la variance des estimations qui est classique avec les méthodes d'imputation simple.

Les estimations des paramètres fixes et de covariance diffèrent peu que l'on traite les données indétectables en considérant une loi normale ou un mélange de loi normale et de Dirac (tableau 1). La probabilité d'appartenir à une loi normale est estimée à $\widehat{\tau} = 0.999$. Autrement dit, il n'y avait pas d'argument pour supposer que les données censurées ne provenaient pas entièrement d'une loi normale.

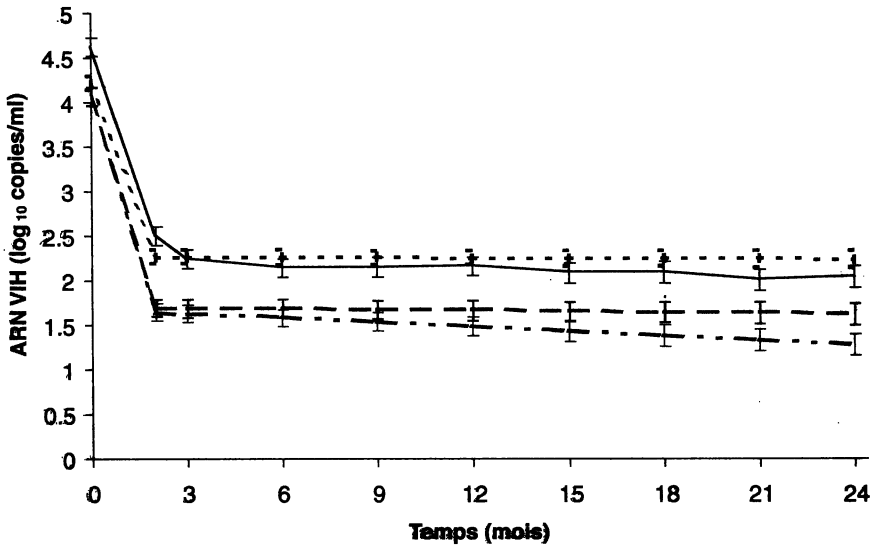

FIG 1. — Évolution de la charge virale moyenne observée et prédite après l'initiation d'un traitement antirétroviral hautement actif. Les charges virales indétectables des données observées ont été remplacées par la moitié de la valeur du seuil (—). Les prédictions sont issues d'un modèle mixte brut, données indétectables étant remplacés 1/2 seuil (····), d'un modèle mixte prenant en compte la censure de la charge virale (— · · —), d'un modèle conjoint prenant en compte la censure de la charge virale et les sorties d'étude informatives (– – – –). Cohorte Aquitaine ($N = 551$).

Les estimations prenant en compte la censure à gauche à partir d'un simple modèle mixte différaient de celles issues du modèle conjoint dans la deuxième partie de l'évolution (après 2 mois). L'écart sur les estimations moyennes des deuxièmes pentes ($\widehat{\beta}_2 = -0.18$ versus $0.037 \log_{10}$ copies/mL/an) se traduisait par des estimations de la charge virale moyenne à un temps donné qui différaient significativement à partir de 18 mois. Lorsque la sortie d'étude informative des sujets ayant présenté un événement clinique était prise en compte, l'évolution de la charge virale était moins optimiste. En effet, la corrélation estimée entre l'évolution de la charge virale (essentiellement la deuxième pente après deux mois) et le temps de survenue d'un événement clinique était significativement négative (tableau 2). Ainsi, plus la charge virale était décroissante, en particulier après 2 mois, c'est-à-dire meilleure était la réponse au traitement et plus le délai de survenu d'un événement clinique était long. De ce fait, les patients ayant une moins bonne réponse virologique, c'est-à-dire une charge virale moins décroissante que les autres, avaient tendance à sortir plus tôt de l'étude en raison de la survenue d'un événement clinique. Par conséquent, ces patients contribuaient moins à l'estimation de la tendance moyenne de la charge virale, engendrant une estimation trop optimiste de l'évolution de la charge virale (figure 2).

Tableau 1. Estimations des paramètres du modèle mixte (intercept, première pente en mois, seconde pente en année) avec et sans prise en compte de la censure (imputation simple de la moitié du seuil de censure dans ce dernier cas) et du modèle conjoint prenant en compte la censure à gauche de la charge virale et les sorties d'étude potentiellement informatives liées à la survenue d'un événement clinique. Cohorte Aquitaine (N=551).

Modèle	Effets fixes			Variances des effets aléatoires			Variance résiduelle
	$\hat{\beta}_0\ (\hat{\sigma}_{\hat{\beta}_0})$	$\hat{\beta}_1\ (\hat{\sigma}_{\hat{\beta}_1})$	$\hat{\beta}_2\ (\hat{\sigma}_{\hat{\beta}_2})$	$\hat{\sigma}^2_{\alpha_0}\ (\hat{\sigma}_{\hat{\sigma}_{\alpha_0}})$	$\hat{\sigma}^2_{\alpha_1}\ (\hat{\sigma}_{\hat{\sigma}_{\alpha_1}})$	$\hat{\sigma}^2_{\alpha_2}\ (\hat{\sigma}_{\hat{\sigma}_{\alpha_2}})$	$\hat{\sigma}^2_{\varepsilon}\ (\hat{\sigma}_{\hat{\sigma}_{\varepsilon}})$
Modèle mixte sans prise en compte de la censure	4.22 (0.040)	-0.98 (0.024)	-0.020 (0.018)	0.55 (0.033)	0.19 (0.016)	0.081 (0.02)	0.42 (0.0074)
Modèle mixte prenant en compte la censure (loi normale)	4.079 (0.046)	-1.21 (0.036)	-0.18 (0.044)	0.40 (0.065)	0.31 (0.039)	0.36 (0.057)	1.00 (0.037)
Modèle mixte prenant en compte la censure (mélange)	4.064 (0.047)	-1.26 (0.039)	-0.21 (0.050)	0.38 (0.065)	0.36 (0.043)	0.51 (0.077)	1.06 (0.038)
Modèle conjoint prenant en compte la censure et les données manquantes informatives	4.076 (0.045)	-1.19 (0.034)	-0.037 (0.039)	0.40 (0.072)	0.27 (0.041)	0.69 (0.032)	1.00 (0.018)

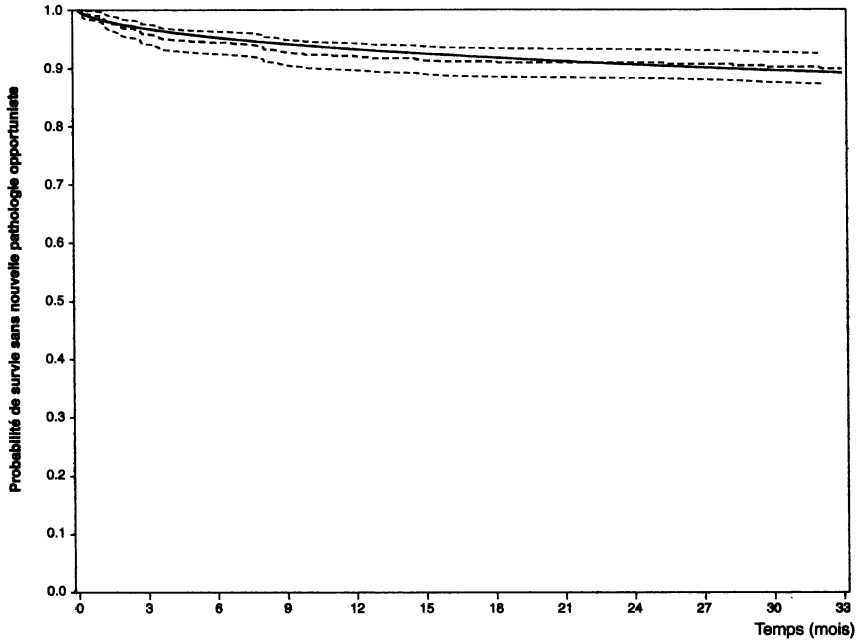

FIG 2. — Probabilité d'être suivi sans survenue d'infection opportuniste estimée selon un modèle log-normal (trait plein) et selon l'estimateur non paramétrique de Kaplan-Meier (traits pointillés avec l'intervalle de confiance à 95 %). Cohorte Aquitaine ($N = 551$).

TABLEAU 2. – Matrice de corrélation des effets aléatoires et du temps de sortie d'étude informative estimée par un modèle conjoint incluant un modèle linéaire mixte pour la charge virale plasmatique et un modèle log-normal pour le temps de sortie d'étude informative. Cohorte Aquitaine ($N = 551$, 60 événements cliniques ayant généré une sortie d'étude)

	Intercept	Première pente	Deuxième pente	Temps de sortie d'étude informative
Intercept	1	−0.16	0.12	−0.41
Première pente	−0.16	1	0.048	−0.043
Deuxième pente	0.12	0.048	1	−0.96
Temps de sortie	-0.41	−0.043	−0.96	1

4. Discussion

L'exemple de la charge virale plasmatique du VIH permet d'illustrer l'importance de la prise en compte des données incomplètes dans le cadre des modèles pour données longitudinales. Si les modèles mixtes sont très utilisés du fait de leur souplesse face à des données déséquilibrées, ils sont toutefois sensibles à

certains types de données incomplètes tels que les données censurées à gauche (indétectables) ou les données manquantes informatives.

En ce qui concerne la censure à gauche dans l'exemple de la charge virale du VIH, il semble évident que sa prise en compte est nécessaire étant donné l'impact potentiel sur les estimations. De plus, la maximisation de la vraisemblance complète est facilement mise en œuvre (Thiébaut, Jacqmin-Gadda 2004) et un modèle de mélange permet de vérifier l'adéquation de l'hypothèse de distribution pour les données censurées (Berk, Lachenbruch 2002). Bien entendu, la présence de nombreuses données censurées à des seuils élevés peut rendre le poids des hypothèses paramétriques si important que l'intérêt de telles analyses peut être remis en cause.

La prise en compte des données manquantes potentiellement informatives est beaucoup plus délicate pour plusieurs raisons. Tout d'abord, si ces données manquantes sont générées par des sorties d'étude, celles-ci peuvent être d'origine différente. Dans l'application du présent article, on considérait les sorties d'étude comme informatives uniquement lorsqu'elles étaient associées à la survenue d'une pathologie opportuniste ou du décès. Cette cause était potentiellement informative du fait du rôle pronostique connu de la charge virale (Mellors *et al.* 1997). Cependant, parmi les autres causes de sortie d'étude (fin de l'étude, patients perdus de vue) qui étaient classées comme non informatives, on peut imaginer que certaines étaient en fait informatives. Outre la définition de la sortie d'étude informative, le choix du modèle conjoint pour prendre en compte les données manquantes informatives peut également être discuté (Jacqmin-Gadda, Thiébaut 2004). En particulier, dans le présent article, on propose un modèle entièrement paramétrique dont on peut supposer que les hypothèses ne reflètent pas l'association réelle entre l'évolution du marqueur et le mécanisme de sortie d'étude. Une autre difficulté non négligeable est la mise en œuvre de ces modèles nécessitant des calculs numériques complexes. Cependant, dans certains cas, il est possible d'utiliser des logiciels de calcul statistique classiques (Guo, Carlin 2004). L'accessibilité de ces différents modèles devrait faciliter leur comparaison afin de vérifier la robustesse des résultats aux données manquantes.

Remerciements : Les auteurs remercient l'ensemble des patients participants à la Cohorte Aquitaine ainsi que l'ensemble des cliniciens, techniciens et méthodologistes gérant cette cohorte. La Cohorte Aquitaine est financée en partie par l'Agence Nationale de Recherches sur le SIDA (ANRS, Action Coordonnées n°7, Cohortes).

Références

ALBERT J. M., IOANNIDIS J. P. A., REICHELDERFER P., CONWAY B., COOMBS R. W., CRANE L., DEMASI R., DIXON D. O., FLANDRE P., HUGHES M. D., KALISH L. A., LARNTZ K., LIN D. Y., MARSCHNER I. C., MUNOZ A., MURRAY J., NEATON J., PETTINELLI C., RIDA W., TAYLOR J. M. G. and WELLES S. L. (1998). Statistical issues for HIV surrogate endpoints : Point/counterpoint. *Statistics in Medicine*, **17**, 2435-62.

BERK K. N. and LACHENBRUCH P. A. (2002). Repeated measures with zeros. *Statistical Methods in Medical Research*, **11**, 303-16.

DE GRUTTOLA V. and TU X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003-14.

DELFRAISSY J. F. (2002). *Prise en charge des personnes infectées par le VIH. Recommandations du groupe d'experts.* Paris, Médecine-Sciences Flammarion.

GUO X. and CARLIN P. (2004). Separate and joint modelling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, 1-9.

HENDERSON R., DIGGLE P. and DOBSON A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465-80.

HUGHES J. P. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, **55**, 625-9.

JACQMIN-GADDA H. and THIÉBAUT R. (2004). Modèles de sélection pour données longitudinales gaussiennes : Application à l'étude du vieillissement cognitif. *Journal de la Société Française de Statistique*, (sous presse).

JACQMIN-GADDA H., THIÉBAUT R., CHÊNE G. and COMMENGES D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, **1**, 355-68.

KENWARD M. G. (1998). Selection models for repeated measurements with non-random dropout : an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723-32.

LAIRD N. M. and WARE J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-74.

LITTELL R. C., MILLIKEN G. A., STROUP W. W. and WOLFINGER R. D. (1996). *SAS System for Mixed Models.* Cary, NC, SAS Institute.

LITTLE R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112-21.

LYLES R. H., LYLES C. M. and TAYLOR D. J. (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. *Journal of the Royal Statistical Society : Series C*, **49**, 485-97.

MELLORS J. W., MUOZ A., GIORGI J. V., MARGOLICK J. B., TASSONI C. J., GUPTA P., KINGSLEY L. A., TODD J. A., SAAH A. J., DETELS R., PHAIR J. P. and RINALDO C. R., Jr. (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine*, **126**, 946-54.

MININI P. and CHAVANCE M. (2004). Observations longitudinales incomplètes : de la modélisation des observations disponibles à l'analyse de sensibilité. *Journal de la Société Française de Statistique*, **145**, 2, 5-18.

MOULTON L. H. and HALSEY N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, **51**, 1570-8.

PINHEIRO J. C. and BATES D. M. (2000). Mixed-Effects Models in S and S-PLUS. New York, Springer-Verlag.

RUBIN D. (1976). Inference and missing data. *Biometrika*, **63**, 581-92.

THIÉBAUT R., CHÊNE G., JACQMIN-GADDA H., MORLAT P., MERCIÉ P., DUPON M., NEAU D., RAMAROSON H., DABIS F. and SALAMON R. (2003). Time updated CD4+ T Lymphocyte count and HIV RNA as major markers of disease progression in naive HIV-1 infected patients treated with an highly active antiretroviral therapy. The Aquitaine Cohort, 1996-2001. *Journal of Acquired Immune Deficiency Syndromes*, **33**, 380-6.

THIÉBAUT R. and JACQMIN-GADDA H. (2004). Mixed models for longitudinal left-censored repeated measures. *Computer Methods and Programs in Biomedicine*, **74**, 255-60.

VERBEKE G. and MOLENBERGHS G. (2000). *Linear mixed models for longitudinal data*. New York, Springer.