Modélisation longitudinales de marqueur du VIH

Rodolphe Thiébaut

Université Victor Segalen – Bordeaux 2

Année 2002 Thèse n°988

THESE pour le

DOCTORAT DE L'UNIVERSITE BORDEAUX 2

Mention : Sciences Biologiques et Médicales

Option : Epidémiologie et Intervention en Santé Publique

présentée et soutenue publiquement

le mardi 17 décembre 2002

par Rodolphe THIEBAUT

né le 3 janvier 1973 à Neuilly sur Seine (Hauts de Seine)

Modélisation longitudinale de marqueurs du VIH

Membres du Jury

Mme Dominique COSTAGLIOLA	**Examinateur**
Mme Hélène JACQMIN-GADDA	**Directeur de Thèse**
M. Michel CHAVANCE	**Rapporteur**
M. Geert MOLENBERGHS	**Rapporteur**
M. Roger SALAMON	**Président**

RESUME Modélisation longitudinale de marqueurs du VIH

L'étude de l'évolution et de la valeur pronostique des marqueurs est très fréquente en épidémiologie. Le taux de lymphocytes T CD4+ et la charge virale plasmatique sont des marqueurs très important de l'infection par le virus de l'immunodéficience humaine (VIH). La modélisation de l'évolution de ces marqueurs présente plusieurs difficultés méthodologiques. D'une part, il s'agit de données répétées incomplètes c'est à dire pouvant être manquantes du fait de la sortie d'étude de certains sujets et de la censure de la charge virale liée à une limite de détection inférieure des techniques de mesure. D'autre part, ces deux marqueurs étant corrélés, il est important de prendre en compte cette information dans le modèle. Nous avons proposé des méthodes basées sur le maximum de vraisemblance pour estimer les paramètres de modèles linéaires mixtes prenant en compte l'ensemble de ces difficultés. Nous avons montré l'impact significatif de ces méthodes biostatistiques sur les estimations et donc nous avons souligné l'importance de leur utilisation dans le cadre des marqueurs du VIH. Pour promouvoir leur diffusion, nous avons présenté des possibilités d'implémentation de certaines des méthodes proposées dans des logiciels statistiques communs.

ABSTRACT Longitudinal modelling of HIV markers

The study of evolution and prognostic value of markers is usual in epidemiology. The CD4+ T lymphocytes cell count and the plasma viral load are useful markers in human immunodeficiency virus infection (HIV). The modelling of those markers evolution raises several methodological difficulties. First of all, these is incomplete data including missing data because of subjects drop-out and censored observation because of a lower quantification limit of assays. Moreover, the two markers are correlated and it is important to take into account this information in the model. We proposed maximum likelihood based methods to estimate the parameters of linear mixed models handling all of these difficulties. We showed the significant impact of these biostatistical methods on the estimations and we underlined their usefulness for HIV markers. To promote their spread, we presented possibilities to implement some of the proposed methods in standard statistical software.

MOTS CLES

Censure, données manquantes, données répétées, infection par le VIH, modélisation, modèles multivariés, modèles conjoints

DISCIPLINE

Doctorat de l'Université Bordeaux 2, Mention : Sciences Biologiques et Médicales

Option : Epidémiologie et Intervention en Santé Publique

LABORATOIRE

Unité INSERM 330 – Equipe Biostatistique - Université Victor Segalen Bordeaux 2

146 rue Léo Saignat 33076 BORDEAUX

Remerciements

A mes filles, à ma femme et à ma mère,

A ma famille,

A mes amis,

A tous les enseignants ayant contribué à ma formation en particulier Geneviève, Daniel et François,

A l’ensemble de mes collègues,

A « Ensemble Contre le SIDA » et leurs généreux donateurs,

A tous ceux qui contribuent aux cohortes APROCO, Aquitaine et à la collaboration CASCADE.

A Madame Dominique Costagliola

Vos connaissances à la fois en biostatistiques et dans le domaine du VIH sont exemplaires. Je vous remercie d'avoir accepté de juger mon travail de thèse.

A Madame Hélène Jacqmin-Gadda

Je ne te remercierai jamais assez pour les connaissances que tu m'as apportées en biostatistiques et en recherche en général. Tu as toujours été très disponible, très pédagogue et jamais avare de ton savoir. Ta gentillesse et ta bonne humeur sont constantes. Sois donc assurée de toute ma reconnaissance.

A Monsieur Michel Chavance

En organisant et en animant la plupart des congrès et ateliers auxquels j'ai participé, vous avez largement contribué à ma formation. Vous avez accepté d'être rapporteur de ma thèse et je vous en remercie.

A Monsieur Geert Molenberghs

La richesse de vos publications et la pédagogie de vos communications sont exemplaires. C'est un honneur que vous me faites en acceptant d'être rapporteur de cette thèse.

A Monsieur Roger Salamon

Vous m'avez encouragé à suivre la voie des biostatistiques et vous m'avez guidé vers la recherche, j'en suis très heureux. Soyez assuré de toute ma gratitude et de mon profond respect.

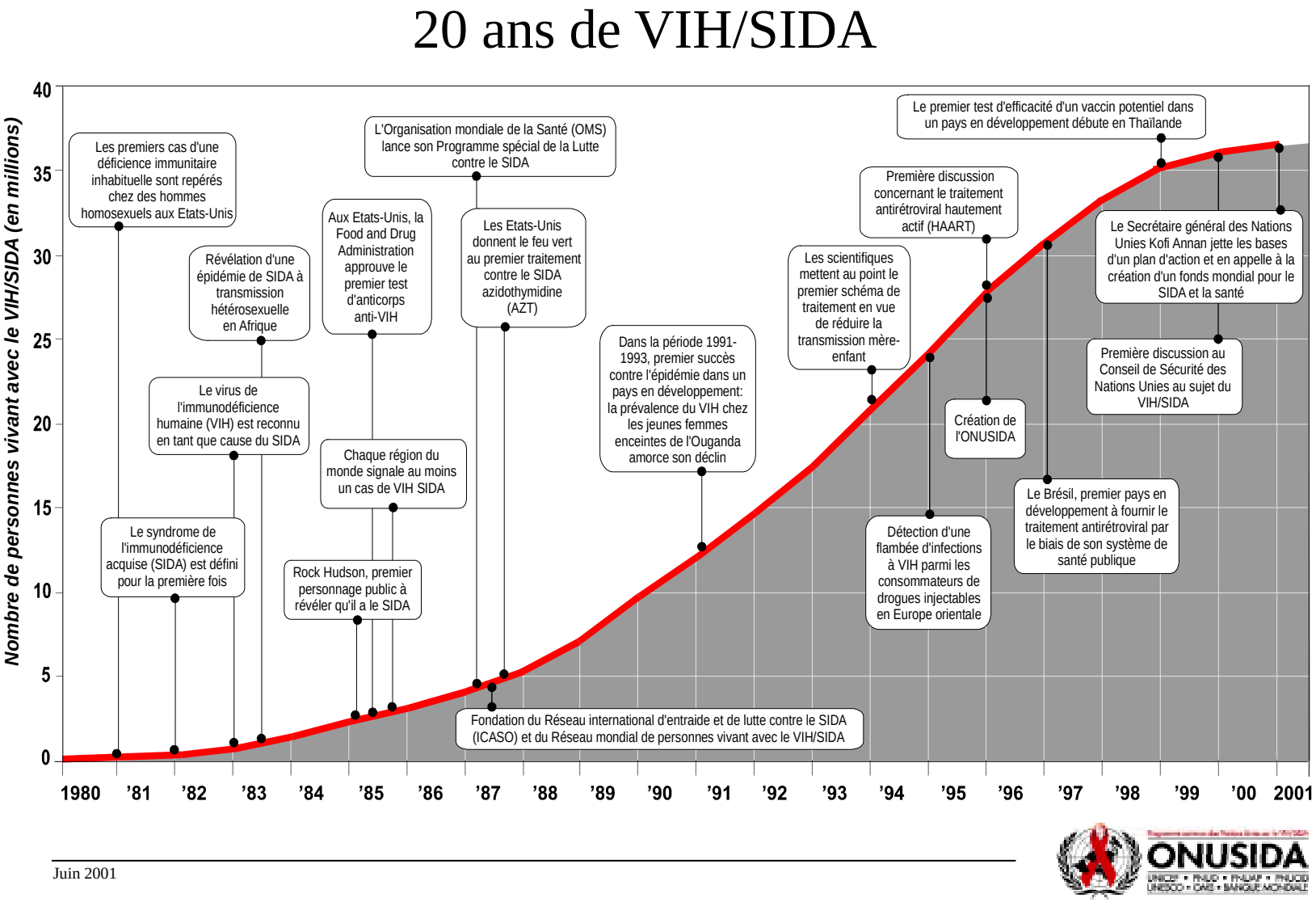
Table des matières

1 Introduction

1.1 Historique de la pandémie du 21ème siècle

Les premiers cas d'infection par le virus de l'immunodéficience humaine (VIH) ont été diagnostiqués au début des années 1980 [1]. Considérée comme une maladie restreinte à des groupes à risque, son impact a été initialement sous-estimé. A partir de 1985, la pandémie du VIH était évidente et devenait la préoccupation de tous. La période de 1986 à 1996 est marquée par la mobilisation croissante des pouvoirs publics pour lutter contre la pandémie (dépistage de l'infection, prévention de la contamination, création de structures spécifiques) mais également par une inquiétude croissante face à l'absence de traitements réellement efficaces et l'importance de la pandémie (Figure 1). A cette époque, la survie à 10 ans était à peu près de 50% et la survie médiane après le diagnostic d'un syndrome d'immunodéficience acquise (SIDA) était de 2 ans [2]. A partir de 1996, les traitements antirétroviraux hautement actifs (HAART pour Highly Active Antiretroviral Therapy) ont été disponibles dans les pays industrialisés modifiant considérablement le pronostique des patients traités. Il s'agissait principalement de trithérapies associant trois antirétroviraux dont deux de la classe des inhibiteurs de l'enzyme transcriptase inverse et un inhibiteur de la protéase. La survie à 10 ans est depuis estimée à près de 80% [2]. Au cours de la même période, la possibilité de quantifier la concentration du virus dans le sang (charge virale) a permis de mettre en évidence une chute de cette quantité au-delà du seuil de détection des techniques de mesure. Cette indétectabilité de la charge virale a engendré des espoirs d'éradication du virus, espoirs très vite déçus par la mise en évidence de réservoirs secondaires viraux. Les traitements antirétroviraux sont depuis et actuellement prescrits à vie engendrant des difficultés d'observance, de résistance du virus au traitement et d'effets secondaires indésirables. Il existe un autre versant de l'épidémie, celui des pays en voie de développement comptant 95% des cas de patients nouvellement infectés. Pour certains de ces pays, l'espérance de vie a reculé d'une vingtaine d'années.

Figure 1. Evolution de l'épidémie du VIH (source : ONUSIDA http://www.unaids.org)



1.2 Physiopathologie

Le VIH est un rétrovirus c'est à dire que son matériel génétique est constitué d'ARN (acide ribonucléïque). Pour se multiplier, le virus doit pénétrer à l'intérieur d'une cellule cible. Cette étape nécessite la reconnaissance par l'enveloppe virale (gp110/120) de molécules à la surface de la cellule cible. Ces molécules sont appelées récepteurs et corécepteurs du VIH, les plus connus étant les molécules CD4, CXCR4 (ou fusine) et CCR5 [3, 4]. La seconde étape est la synthèse d'ADN proviral résultant de la copie de l'ARN viral par la transcriptase inverse. C'est cette enzyme qui est la cible des antirétroviraux inhibiteurs de la transcriptase inverse qu'ils soient nucléosidiques ou non nucléosidiques. Par la suite, l'ADN proviral est intégré au génome de la cellule hôte grâce à l'intégrase virale puis le provirus est transcrit en ARN messager par l'ARN polymérase de l'hôte. Cet ARN messager viral migre dans le cytoplasme et permet la synthèse de protéines virales qui seront assemblées en polyprotéïnes et clivées notamment par la protéase virale. Cette enzyme est la cible des antirétroviraux de la classe des inhibiteurs de la protéase. Enfin les nouvelles particules virales vont bourgeonner à la surface de la cellule avant d'être libérées dans le milieu extra-cellulaire.

Les cellules infectées par le VIH sont principalement celles qui expriment à leur surface le récepteur CD4 et l'un des corécepteurs. Il s'agit de la sous-population de lymphocytes T CD4+ helper (ou auxilliaire) mais aussi des monocytes/macrophages ou de cellules de la même origine telles que les cellules dendritiques et les cellules de Langerhans ainsi que les cellules microgliales du cerveau. Dans certaines cellules, les virus sont simplement emprisonnés et ne se répliquent pas. C'est le cas, par exemple, des cellules folliculaires dendritiques présentes dans les centres germinatifs des ganglions lymphatiques [5]. Les lymphocytes T CD4+ infectés représentent plus de 95% des cellules infectées de l'organisme. On distingue parmi ces lymphocytes infectés, les cellules à réplication active (activées par les antigènes ou les cytokines) et les cellules quiescentes. Ces dernières cellules qui représentent moins de 1% des lymphocytes T CD4+ infectés ne produisent pas de virions mais ont déjà été infectées et ont intégré le génome viral sous forme de provirus. Il s'agit le plus souvent de cellules T mémoires n'exprimant à leur surface aucun marqueur d'activation (CD45RO = HLA –, DR –, CD25 –).

Les organes lymphoïdes (ganglions lymphatiques, rate, intestin, thymus) sont atteints dès le stade initial de l'infection ce qui résulte en une activation généralisée et chronique. Il s'agit du site d'élection pour la réplication virale. C'est pourquoi, dans les organes lymphoïdes, la charge virale est dix fois supérieure à celle mesurée dans le sang périphérique [6]. En dépit de

la réponse immunitaire de l'hôte, l'infection VIH persiste du fait de réservoirs viraux et de la réplication constante du virus. Pendant plusieurs années, les lymphocytes T CD4+ détruits par le virus semblent rapidement renouvelés jusqu'à ce que les altérations des organes lymphoïdes centraux (thymus) ne permettent plus leur régénération [7]. Il s'ensuit un déficit quantitatif et qualitatif des lymphocytes T CD4+ évoluant vers un déficit immunitaire profond. Ce déficit immunitaire se traduit par la survenue de pathologies opportunistes dont certaines d'entre elles conduisent à la classification au stade SIDA [8]. Les multithérapies antirétrovirales ont permis la chute de la charge virale dans le compartiment plasmatique et une augmentation du nombre de lymphocytes T CD4+ ainsi qu'une amélioration de leur fonctionnalité [9]. Cependant, la persistance d'un réservoir viral en présence d'un traitement efficace sur la charge virale plasmatique [10] conduit au maintien du traitement à long terme [11].

1.3 Mesure des marqueurs viro-immunologiques

La mesure du taux de lymphocytes T CD4+ totaux circulants (CD4+) est disponible par cytométrie de flux depuis le début de l'épidémie. Bien que les techniques de numération des lymphocytes se soient améliorées, la mesure de CD4+ est connue pour avoir une grande variabilité (coefficient de variation de 15%) avec des fluctuations diurnes [12, 13].

La quantification de la charge virale peut porter sur le virus libre plasmatique (antigénémie, virémie plasmatique par culture, quantification moléculaire de l'ARN viral) ou sur le virus intégré dans les cellules sanguines mononuclées (virémie cellulaire par culture, mesure de l'ADN proviral). Le virus plasmatique reflète essentiellement la multiplication du virus dans le tissu lymphoïde et donc la multiplication active du virus dans l'organisme. L'ADN proviral étant présent dans les cellules quiescentes et dans les cellules productrices de virus, sa signification est moins claire. Toutefois, sous traitement efficace, l'ADN proviral diminue plus lentement que le virus libre mais il diminue quand même de façon significative. La culture cellulaire du VIH-1 est une méthode longue, coûteuse et nécessitant des laboratoires de haute sécurité. Les techniques de quantification moléculaire de l'ARN viral sont disponibles sous forme de trousses agrées depuis 1996, ce qui en a permis une large diffusion. Bien qu'abusif, le terme « charge virale » est utilisé pour cette quantification de l'ARN viral. Le résultat obtenu est rendu en terme de copies/ml correspondant à des molécules d'ARN VIH. Du fait de la variabilité et de l'étendue des valeurs en copie/ml, la charge virale est souvent exprimée sous la forme d'une transformation logarithmique en base 10. Cette transformation permet d'approcher une distribution normale et minimise les variations de

faible amplitude. Les techniques de mesure de la charge virale manquent de sensibilité empêchant de la quantifier au-dessous d'un certain seuil dépendant de la technique utilisée. La sensibilité des techniques s'est améliorée au cours du temps passant de 10 000 copies/ml à 2 copies/ml aujourd'hui. Ainsi, les techniques dites ultrasensibles ont toutes des seuils de détection inférieurs à 200 copies/ml. Cependant, quelle que soit la génération, ces seuils dépendent également de la technique. Par exemple, le seuil de détection de la technique par ADN branché classique (Quantiplex HIV-1 RNA 2.0 ; Chiron, Emeryville, CA, USA) est de 500 copies/ml contre 400 copies/ml pour la technique PCR d'amplification génique (Amplicor HIV-1 1.0 Monitor ; Roche Molecular Systems, Branchburg, NJ, USA). Enfin, même pour des charges virales détectables, la quantification peut être différente selon le test, en particulier selon qu'il est ultrasensible ou non. Par exemple, la technique ultrasensible par ADN branché version 3.0 dont le seuil de détection est de 50 copies/ml tend à trouver des valeurs plus élevées de charge virale par rapport à la technique classique version 2.0 [14, 15]. Enfin, il faut noter qu'il existe une limite de quantification supérieure également dépendante de la technique (autour de 800 000 copies/ml). La variabilité de la charge virale est moins importante que celle des CD4+ (coefficient de variation de 8%) et semble plus lié à des facteurs techniques qu'à des facteurs physiologiques [13]. En général, la variabilité liée à l'erreur de mesure ou aux variations biologiques pouvant être égale à 0,3 $\log_{10}$ copies/ml, on considère une variation de plus de 0,5 $\log_{10}$ copies /ml comme significative.

1.4 Intérêt des marqueurs viro-immunologiques

Depuis l'avènement des HAART, l'incidence des pathologies opportunistes est si faible que l'utilisation de marqueurs de substitution est devenue nécessaire. Un marqueur de substitution est un marqueur qu'on mesure à la place du véritable événement d'intérêt [16]. Prentice [17] a défini plus formellement cette notion de marqueur de substitution comme une variable réponse pour qui le test d'hypothèse nulle d'absence d'effet du groupe de traitement est un test valide de l'hypothèse nulle basée sur le vrai critère à savoir l'événement d'intérêt (progression clinique par exemple). Il propose également une approche de validation du critère de substitution basée sur trois critères :

- ✓ Le marqueur doit être pronostique vis-à-vis de la survenue de l'événement
- ✓ Le marqueur doit pouvoir être affecté par le traitement à évaluer
- ✓ L'effet du traitement sur le marqueur doit mesurer l'ensemble de l'effet du traitement sur la survenue de l'événement d'intérêt.

Le rôle pronostique du comptage des lymphocytes T CD4+ sur la progression clinique est connu depuis le début de l'épidémie. Avant l'ère des HAART, étaient associées à la progression clinique la mesure transversale des CD4+ (souvent à l'entrée de l'étude) [18-20] et l'évolution au cours du suivi [21]. Depuis l'avènement des HAART, la mesure des CD4+ à l'initiation d'un traitement hautement actif [22] et surtout l'évolution des CD4+ en réponse à ce traitement [22, 23] sont parmi les variables les plus pronostiques de la progression clinique. Cependant son utilisation en tant que marqueur de substitution est discuté en particulier du fait du troisième critère [24-29] : une partie seulement des différences de progression clinique en fonction du traitement est expliquée par l'effet du traitement sur les lymphocytes T CD4+. Certains auteurs ont évoqué la difficulté d'utiliser le taux de CD4+ en tant que marqueur de substitution du fait de la grande variabilité des mesures de ce marqueur [12, 13] et donc de l'effet de l'erreur de mesure sur l'estimation du risque [30-32] (voir section 2.3.5). Toutefois, De Gruttola et al. [33] ont utilisé une modélisation conjointe des données longitudinales des CD4+ et de la progression clinique vers le SIDA ou le décès pour prendre en compte cette difficulté. Le bénéfice clinique lié au traitement antirétroviral était faiblement expliqué par l'impact du traitement sur les CD4+ [25].

Depuis que la charge virale plasmatique est disponible pour le suivi biologique des patients, son rôle pronostique sur la progression clinique à largement été démontré et ceci indépendamment des CD4+ [34-36]. L'intérêt de l'utilisation à la fois de la charge virale et des CD4+ en tant que marqueurs de substitution a été mis en évidence par la disparition de l'effet du traitement une fois ajusté sur la réponse viro-immunologique [37] ou, dans une autre étude, une augmentation de la part de l'effet du traitement expliqué par l'évolution des deux marqueurs [38]. Cependant, d'autres études ayant évalué l'effet du changement des marqueurs à un temps donné présentent des résultats plus mitigés [39, 40]. Globalement, étant donné l'importance et la taille de l'intervalle de confiance de l'effet du traitement ajusté sur les marqueurs, plusieurs problèmes ont été évoqués. D'une part, les auteurs ont évoqué des difficultés méthodologiques telles que l'indétectabilité de la charge virale [37] ou les sorties d'étude [40]. D'autre part, la toxicité des traitements, les problèmes d'observance des patients au traitement et les résistances virologiques peuvent rendre nécessaire l'utilisation de plusieurs traitements séquentiellement pour obtenir une bonne réponse viro-immunologique. Ce type de difficultés a engendré des débats concernant l'utilisation de critères de substitution combinés de type réponse virologique ou modification du traitement antirétroviral [41].

En pratique, la charge virale plasmatique et les lymphocytes T CD4+ sont utilisés comme critère de jugement dans les essais cliniques randomisés comme pour le suivi habituel des patients. Aujourd'hui, les guides thérapeutiques anglais [42], américains [43, 44] ou français [45] définissent l'initiation et la modification des traitements antirétroviraux selon la valeur et/ou l'évolution de ces marqueurs. L'idée est par exemple de modifier un traitement lorsque le ou les marqueurs atteignent une certaine valeur afin d'éviter la survenue d'un événement clinique [46]. Ainsi, considérant ces marqueurs comme des marqueurs de substitution, l'étude des déterminants de l'évolution de ces marqueurs permet de préciser la prise en charge des patients. Il peut s'agir de l'étude de la réponse viro-immunologique selon le type de traitement dans le cadre d'un essai thérapeutique ou bien de l'étude d'autres facteurs épidémiologiques pouvant influencer cette réponse. Dans les essais cliniques contrôlés randomisés, l'analyse principale est le plus souvent basée sur des critères définis à un temps donné, par exemple le changement moyen de la charge virale à 24 semaines de l'initiation du traitement [47]. Cependant l'analyse de l'ensemble de l'évolution des marqueurs sous traitement peut faire partie des critères secondaires d'un essai clinique [48]. Dans les cohortes observationnelles, les analyses longitudinales de l'évolution des marqueurs viro-immunologiques permettent d'étudier les déterminants de cette évolution soit en dehors d'un traitement soit chez les patients traités. L'intérêt de ces cohortes est de permettre des études incluant des patients « en pratique clinique » avec un suivi qui peut être bien plus long que dans les essais cliniques. Cependant, il existe toujours des biais possibles notamment un biais de sélection par indication lorsqu'on évalue l'efficacité des traitements avec ce type de données [49, 50].

En dehors de la prescription de tout traitement antirétroviral, les études publiées se sont souvent focalisées sur la réponse immunologique [51-54] ou virologique [55-59]. Outre l'accélération de la décroissance des CD4+ et de l'augmentation de la charge virale quelque temps avant le passage au stade SIDA, c'est surtout l'effet du sexe qui a soulevé beaucoup de questions. En effet, de nombreuses études ont rapporté une charge virale en moyenne plus basse chez les femmes par rapport aux hommes pour un taux de CD4+ et un stade clinique donné [60, 61]. De plus, la progression clinique semblant identique dans les deux groupes, certains auteurs ont suggéré de modifier les guides thérapeutiques pour adapter les règles de prescription en fonction du sexe [62].

Chez les patients traités, les facteurs influençant le plus souvent la réponse virologique étaient ceux associés aux modalités de traitement : le type de molécule utilisé [63, 64], les antécédents de traitement antirétroviral avant l'initiation d'un nouveau traitement [65-67], la modification du traitement antirétroviral en cours [63, 65] ou l'observance au traitement [68, 69]. Les valeurs initiales des marqueurs sont inconstamment rapportées comme étant associées à la réponse virologique ultérieure [63, 65, 70].

Quant à la réponse immunologique, elle est le plus souvent associée à la réponse virologique [65, 71-74]. Par exemple, dans une étude récente [74], nous avons montré l'impact des rebonds virologiques sur l'évolution des CD4+ en particulier lorsque ce rebond était au-dessus de 10000 copies/ml.

Les autres facteurs épidémiologiques associés à la réponse viro-immunologique sous traitement antirétroviral tels que le sexe [60, 75], l'utilisation de drogues par voie intraveineuse [76], ou le délai entre la séroconversion et la date d'initiation du traitement [77] ont beaucoup moins été étudiés. Pourtant, ces informations peuvent être très importantes pour la prise en charge des patients. Par exemple, savoir quand débuter un traitement antirétroviral pour la première fois reste une question d'actualité. En effet, bien qu'ayant des traitements efficaces, la question d'initier ces traitements le plus tôt possible se pose dans la mesure où peuvent survenir des effets secondaires indésirables ou des mutations du virus responsables d'une diminution de l'efficacité et du capital des antirétroviraux disponibles pour le patient [78]. Cependant, laisser évoluer la maladie sans intervenir peut conduire à des lésions irréversibles du système immunitaire [79, 80]. Il est donc intéressant de connaître l'effet du délai entre la séroconversion et l'initiation du traitement antirétroviral sur la réponse viro-immunologique.

Au total, les deux marqueurs principaux de l'infection par le VIH, la charge virale plasmatique (ARN VIH) et les lymphocytes T CD4+ (CD4+), bien qu'incomplètement validés en tant que marqueur de substitution, définissent les stratégies de prise en charge des patients. L'évolution de ces marqueurs, en particulier sous traitement antirétroviral, est au centre de l'intérêt de beaucoup d'études récentes dont l'objectif principal est de connaître les déterminants de cette évolution.

1.5 Problèmes méthodologiques

L'étude de l'évolution des marqueurs du VIH présente plusieurs difficultés méthodologiques plus ou moins spécifiques à ce domaine. Tout d'abord, la présence de mesures répétées chez un même sujet conduit à prendre en compte la corrélation entre ces mesures. Proportionnellement au nombre d'études publiées, très peu ont utilisé des modèles adaptés à ce type de données. En fait, beaucoup d'auteurs ont plutôt analysé le délai jusqu'au passage d'un seuil (croissance des CD4>100 cellules/mm^3 par exemple, [77]) à l'aide de technique d'analyse de survie. Ce type d'analyse a l'inconvénient de perdre de l'information et surtout est beaucoup plus sensible à la fréquence de mesure des marqueurs.

Un autre point est que les deux marqueurs d'intérêt sont fondamentalement corrélés : les lymphocytes T CD4+ sont la cible privilégiée du virus mesuré par la charge virale plasmatique. La prise en compte de cette corrélation entre les marqueurs peut permettre non seulement d'améliorer l'ajustement du modèle aux données mais également de mieux comprendre l'effet de déterminants sur l'évolution de l'un ou l'autre marqueur. L'utilisation de tels modèles multivariés dans le cadre de l'infection par le VIH est rare [81-83].

L'analyse de la charge virale sur une échelle continue engendre une autre difficulté : la censure à gauche des mesures dites indétectables. Plusieurs méthodes pour prendre en compte cette difficulté ont été proposées dans le cadre spécifique du VIH mais elles ont été très peu appliquées pour des travaux épidémiologiques. Le plus souvent, les auteurs étudient le délai jusqu'à la survenue d'une mesure de charge virale indétectable.

Enfin, une dernière difficulté est la censure du suivi des patients liée à la sortie d'étude ou la modification d'un traitement antirétroviral, par exemple. En effet, les patients inclus dans des cohortes observationnelles peuvent être perdus de vue pour des raisons très liées à l'évolution clinique et donc viro-immunologique. Ou bien, le suivi peut être censuré à la modification du traitement antirétroviral dans la mesure où l'investigateur ne s'intéresse qu'à la réponse sous le traitement initial. Dans ce cas, si les données manquantes générées sont informatives (voir section 2.3), il est nécessaire de prendre en compte le processus de données manquantes. Ce type d'analyse a déjà été effectué dans le cadre de la modélisation des marqueurs du VIH mais le plus souvent avant l'ère des HAART où la sortie d'étude était essentiellement liée à la progression clinique et en particulier au décès [33, 84, 85].

1.6 Objectifs du travail de thèse

L'objectif principal du travail de thèse présenté ici était de développer des méthodes pour modéliser au mieux l'évolution longitudinale des lymphocytes T CD4+ et de la charge virale plasmatique en prenant en compte les difficultés méthodologiques présentées ci-dessus.
Les objectifs secondaires étaient d'appliquer ces méthodes afin de répondre à des questions épidémiologiques concrètes et de rendre ces méthodes applicables par le plus grand nombre.

Dans la partie suivante (section 2), les méthodes d'analyses des données longitudinales sont présentées dans le cadre général (2.1), puis leur extension pour des modèles multivariés (2.2). Ensuite, on présente des méthodes de modélisation de données longitudinales incomplètes soit parce qu'elles sont manquantes du fait de la sortie d'étude (2.3), soit parce qu'elles sont censurées à gauche du fait d'un manque de sensibilité des techniques de mesure (2.4). Dans la section 3, est présenté un article original sur l'utilisation du logiciel SAS® pour estimer les paramètres d'un modèle pour des données longitudinales bivariées (3.2). Dans la section 4, la première partie est consacrée à nos travaux antérieurs sur la prise en compte de la censure à gauche de la charge virale et son utilisation en tant que marqueur pronostique (4.1). Puis, on présente l'extension de la modélisation bivariée pour prendre en compte la censure selon deux méthodes : la première programmée en Fortran a fait l'objet d'une publication (4.3), la seconde utilise le logiciel SAS® (4.4). La première partie de la cinquième section (5.1) est consacrée à la prise en compte des données manquantes informatives dans le cadre du VIH. Le modèle bivarié des deux marqueurs, prenant en compte la censure de la charge virale et les données manquantes informatives, est présenté dans la section 5.2 sous forme d'un article original soumis. Une discussion générale se trouve en section 6.

2 Méthodes d'analyse de données longitudinales

L'étude de données longitudinales (répétées) peut être très puissante, car elle permet d'analyser les modifications au sein d'un individu au cours du temps. La plupart des techniques statistiques standards font l'hypothèse que chaque observation est indépendante des autres observations. Cependant, cette hypothèse n'est pas appropriée lorsqu'il s'agit d'observations répétées chez un même sujet, dans la mesure où ces observations tendent à être corrélées entre elles. Si on prend deux observations chez un même individu, elles ont tendance à avoir une valeur plus proche que deux observations prises chez deux individus distincts. Ainsi, l'utilisation de techniques statistiques classiques pour analyser des données répétées conduit à des inférences incorrectes, notamment en ce qui concerne les écart-types des estimations qui sont soit trop petits soit trop larges selon la variable d'intérêt [86]. Si on s'intéresse à l'effet de variables constantes au cours du temps chez un même sujet, une mesure répétée chez un individu apporte moins d'information qu'une observation chez un nouvel individu. De même, si on utilise des techniques classiques d'analyse statistique sur un échantillon avec mesures répétées pour étudier la taille moyenne dans une population d'adulte, on sous-estime la variance de la taille moyenne de cette population [87]. Si on s'intéresse aux variables associées aux observations individuelles, par exemple le temps dans une enquête longitudinale, la répétition des observations chez un même sujet mets en évidence les changements intra-individuels de la variable réponse. Ce schéma d'étude permet donc de répondre à la question en minimisant le bruit issu de la variabilité inter-individuelle. Dans ces circonstances, ignorer la structure de corrélation des données engendre une perte d'information qui se traduit par des estimations trop larges des écart-types.

Le type de modèle utilisé pour prendre en compte la corrélation des observations dépend du type de données et de l'objectif de l'analyse [86]. Le fait qu'on s'intéresse explicitement à la variabilité inter-individuelle justifie l'utilisation de modèles à effets aléatoires (ou modèles mixtes) au lieu de modèles marginaux. De plus, étant donné que l'on s'intéresse à des marqueurs biologiques quantitatifs continus, on considère uniquement les modèles pour données longitudinales gaussiennes.

2.1 Modèle linéaire mixte pour données longitudinales gaussiennes

2.1.1 Formulation générale

On parle de modèle linéaire mixte car ce modèle est composé de deux parties :

- les effets fixes (β) identiques pour tous les sujets de la population, représentant la tendance moyenne pour la population,
- les effets aléatoires (γ_i) représentant l'écart de chaque individu i par rapport à la tendance de la population.

Ces effets aléatoires permettent d'expliciter les différences entre les individus (et donc les corrélations au sein de chaque individu) sans observer les déterminants de cette variabilité interindividuelle. Si on était capable de mesurer toutes les variables expliquant les différences entre les individus, l'ensemble de la variabilité entre les observations de la variable à expliquer (Y) pourrait être modélisé à l'aide d'effets fixes et d'une erreur de mesure (ε) sans effet aléatoire. Soit $Y_i = \left(Y_{i1},...,Y_{in_i}\right)$ le vecteur réponse de dimension n_i pour le sujet i d'un échantillon de N sujets, le modèle s'écrit [88]:

$$\begin{cases} Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i \\ \gamma_i \sim N(0, G) \\ \varepsilon_i \sim N(0, \Sigma_i) \\ \gamma_i \perp \varepsilon_i, \forall\, i \end{cases} \quad (1)$$

avec X_i la matrice de variables explicatives de dimension $n_i \times p$ (p étant le nombre de variables explicatives et donc le nombre d'effets fixes β), Z_i une sous-matrice de X_i de dimension $n_i \times q$ (avec $q \leq p$ le nombre d'effets aléatoires γ_i). De plus, effets aléatoires et erreurs sont indépendants d'un sujet à l'autre et les erreurs sont indépendantes d'une mesure à l'autre chez un même sujet :

$$\begin{cases} \gamma_i \perp \gamma_j, \forall\, i, j = 1,...N \\ \varepsilon_i \perp \varepsilon_j, \forall\, i, j = 1,...N \\ \varepsilon_{ij} \perp \varepsilon_{ik}, \forall\, j, k = 1,...n_i \end{cases}$$

Les effets aléatoires γ_i ne sont pas estimables comme les paramètres fixes β sous peine d'estimer N paramètres supplémentaires. On suppose en fait que l'ensemble des effets aléatoires γ_i suit une loi normale centrée de matrice de variance-covariance G. Ceci revient donc à estimer au minimum (G diagonale) q paramètres de variance et au maximum (G non

structurée) $\frac{q\times(q+1)}{2}$ pour q effets aléatoires. On suppose également que l'erreur de mesure est indépendante pour chaque observation suivant une loi normale centrée. La matrice de variance de l'erreur de mesure ε_i pour l'ensemble des observations d'un sujet i est donc la matrice diagonale $\sum_i = \sigma_\varepsilon^2 I_{n_i}$.

Le type de modèle qui vient d'être décrit présente donc deux sources de variabilité :

- la variabilité inter-individuelle modélisée via les effets aléatoires suivant l'idée que les individus représentent un échantillon aléatoire d'une population,
- la variabilité intra-individuelle modélisée par l'erreur de mesure représentant une erreur aléatoire engendrée par la technique de mesure indépendante pour chaque mesure de chaque sujet.

On peut concevoir que la seule erreur de mesure ne puisse pas modéliser l'ensemble de la variabilité intra-individuelle. Une part de l'évolution des mesures observées chez un même individu peut correspondre à la réalisation d'un processus stochastique dépendant du temps propre à chaque individu. Par exemple, les mesures répétées de CD4+ représentent l'évolution du système immunitaire de chaque individu à laquelle est ajoutée une erreur de mesure. Ce type de variation stochastique engendre une corrélation entre les mesures d'un même individu qui dépend de l'écart entre les mesures. Classiquement, cette corrélation diminue avec le temps. Pour prendre en compte cette corrélation dite sérielle (ou autocorrélation), il a été proposé [89, 90] une extension du modèle de Laird et Ware [88] en incluant un processus autorégressif d'ordre 1. Cependant, la corrélation sérielle et la corrélation des effets aléatoires ne sont pas toujours distinguables selon l'information disponible. En particulier, ceci peut survenir lorsque le nombre d'observations par sujet est à peu près identique et peu important [91, 92].

Le modèle incluant des effets aléatoires, un terme d'auto-corrélation et une erreur de mesure s'écrit :

$$Y_i = X_i\beta + Z_i\gamma_i + W_i + \varepsilon_i \text{ avec } \begin{cases} \varepsilon_i \sim N(0,\Sigma_i) \\ W_i \sim N(0,R_i) \\ \gamma_i \sim N(0,G) \end{cases}$$

W_i est le vecteur des réalisations d'un processus autorégressif d'ordre 1, $w_i(t)$, dont la covariance entre deux temps s et t est $R_i(w_i(s), w_i(t)) = \sigma_w^2 \times e^{\alpha|t-s|}$. σ_w^2 est la variance du processus et α est le paramètre de corrélation du processus.

D'autres processus peuvent être utilisés tels que le processus d'Ornstein-Uhlenbeck intégré [93]. La covariance et la variance de ce processus sont définies comme suit :

$$R_i(w_i(s), w_i(t)) = \frac{\sigma_w^2}{2\alpha^3} \times \left[2\alpha \min(s,t) + e^{-\alpha t} + e^{-\alpha s} - 1 - e^{-\alpha |t-s|}\right]$$

et $\text{var}(w_i(t)) = \frac{\sigma_w^2}{\alpha^3}\left[\alpha t + e^{-\alpha t} - 1\right]$

Les deux paramètres du processus σ_w^2 et α contrôlent la variabilité et le degré de constance de la tendance (concept de 'derivative tracking'), c'est à dire comment un individu tend à maintenir la même pente sur une longue période de temps. Quand σ_w^2 et α sont petits, le processus tend à garder la même trajectoire sur une longue période comme dans le cas d'un modèle à effets aléatoires. Quand σ_w^2 et α sont grands, le processus a très peu de mémoire. D'ailleurs, le mouvement brownien est un cas particulier de ce processus lorsque $\alpha \to \infty$ et $\frac{\sigma_w^2}{\alpha^2}$ est constant, puisque $R_i(w_i(s), w_i(t)) \to \frac{\sigma_w^2}{\alpha^2}\min(s,t)$.

2.1.2 Estimation

On distingue deux formulations du modèle linéaire mixte : la formulation hiérarchique et la formulation marginale. La formulation princeps est la formulation hiérarchique où le modèle (1) est défini par les distributions $f_{Y_i|\gamma_i}(.|.)$ et $f_{\gamma_i}(.)$. Dans une première étape [88], pour chaque individu i, β et γ_i sont fixés. Conditionnellement aux effets aléatoires, Y_i est distribué normalement de moyenne le vecteur $X_i\beta + Z_i\gamma_i$ et de variance Σ_i. Dans une seconde étape, seul β est fixé, les effets aléatoires sont supposés suivre une loi normale d'espérance nulle et de matrice de covariance G. En utilisant cette formulation hiérarchique du modèle (1), les paramètres peuvent être estimés par les techniques Bayésiennes [88, 94].

La fonction de densité marginale de Y_i est donné par :

$$f_{Y_i}(y_i) = \int_{R^q} f_{Y_i|\gamma_i}(y_i | \gamma_i = u) f_{\gamma_i}(u)\, du$$

Il s'agit de la fonction de densité d'une loi multivariée normale de moyenne le vecteur $X_i\beta$ et de matrice de covariance $V_i = Z_i G Z_i^T + \Sigma_i$ (du fait de l'hypothèse d'indépendance entre les effets aléatoires et l'erreur de mesure). C'est cette formulation du modèle qui est classiquement utilisée pour estimer les paramètres. Pourtant, la structure hiérarchique du

modèle initial n'est alors pas prise en compte. Les inférences basées sur le modèle marginal ne prennent pas explicitement en compte le fait que les effets aléatoires représentent l'hétérogénéité entre les sujets. Autrement dit, le modèle marginal n'est pas équivalent au modèle hiérarchique original. Par exemple, tant que les matrices de variance-covariance $V_i = Z_i G Z_i^T + \Sigma_i$ sont définies positives (toutes les valeurs propres sont positives), un modèle marginal valide peut être obtenu. Autrement dit, selon la forme de Z_i, V_i peut être définie positive alors que G ne l'est pas. Ainsi, une variance d'effet aléatoire négative pourra être compatible avec un modèle marginal mais pas avec le modèle hiérarchique initial. Cependant, tout en adoptant une formulation marginale, on peut contraindre G à être défini positive (en utilisant une décomposition de Cholesky, par exemple).

Dans l'approche classique, l'estimation du vecteur de paramètres du modèle $\theta = \left(\beta^T, \alpha^T\right)^T$ avec α le vecteur de tous les paramètres de variance-covariance de G et de Σ_i, est obtenue par maximisation de la vraisemblance marginale :

$$L(\theta) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} \left|V_i(\alpha)\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (Y_i - X_i\beta)^T V_i^{-1}(\alpha)(Y_i - X_i\beta) \right) \right\} \qquad (2)$$

Conditionnellement à α (c'est à dire en supposant α connu), l'estimateur de maximum de vraisemblance de β est :

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\alpha) X_i \right)^{-1} \sum_{i=1}^{N} X_i^T V_i^{-1}(\alpha)\, y_i \qquad (3)$$

Si α est inconnu, mais qu'une estimation est disponible ($\hat{\alpha}$) on peut définir $\hat{V}_i = V_i(\hat{\alpha})$ et estimer β en remplaçant V_i^{-1} par $\hat{V}_i^{-1}$ dans l'expression (3). α est estimé par la méthode du maximum de vraisemblance ou par la méthode du maximum de vraisemblance restreinte (REML pour Restricted Maximum Likelihood). L'estimateur du maximum de vraisemblance de α est obtenu par maximisation de (2) après avoir remplacé β par (3). La méthode du maximum de vraisemblance restreinte est utilisée pour corriger le biais sur l'estimateur des paramètres de covariance dû au fait que les estimateurs du maximum de vraisemblance ne tiennent pas compte de la perte de degré de liberté engendrée par l'estimation des effets fixes. En général, les deux méthodes sont équivalentes mais elles tendent à différer quand le nombre p d'effets fixes est grand et/ou l'échantillon est de petite taille. Il faut noter que les inférences basées sur le test du rapport de vraisemblance ne sont pas valides lorsque la méthode REML

est utilisée. Les travaux de cette thèse ont toujours porté sur des grands échantillons, ce qui a permis de préférer les estimations par maximum de vraisemblance.

Dans l'article princeps de Laird et Ware [88], l'algorithme proposé pour estimer les paramètres du modèle par maximisation de la vraisemblance était l'EM algorithme (traitant les effets aléatoires comme des données manquantes). Pourtant, les auteurs rapportaient déjà une convergence lente pour l'estimation des paramètres de covariance, en particulier lorsque l'estimateur du maximum de vraisemblance est proche des limites de l'espace de définition des paramètres (variances des effets aléatoires proches de 0). Plus tard, Lindström et Bates [95] ont montré que l'algorithme de Newton-Raphson était plus rapide dans ce cadre. C'est d'ailleurs ce dernier algorithme qui est implémenté dans la procédure MIXED du logiciel SAS® [96].

2.1.3 Prédictions individuelles

L'estimation des γ_i peut être intéressante pour repérer des individus particuliers dans la mesure où les effets aléatoires représentent les écarts entre les profils individuels et le profil moyen. De plus, l'estimation des effets aléatoires est également nécessaire pour prédire les valeurs individuelles de la variable réponse Y_i.

En utilisant la théorie bayésienne [97], les γ_i sont estimés par la moyenne de la distribution a posteriori $f_{\gamma_i|Y_i}(.|.)$. On a :

$$\hat{\gamma}_i = E\left[\gamma_i | Y_i = y_i\right] = GZ_i^T V_i^{-1}(\alpha)(y_i - X_i\beta)$$

Cet estimateur correspond aussi à la solution du système linéaire d'équation proposé par Henderson [98]. Cet estimateur est appelé estimateur bayésien empirique (« Empirical Bayes ») car les paramètres inconnus α et β sont remplacés par leurs estimateurs du maximum de vraisemblance conduisant à une sous estimation de la véritable variabilité de γ_i.

Une simple étude de la distribution des estimations bayésiennes empiriques par un histogramme ne permet pas la vérification de la normalité de la distribution des γ_i du fait de cette sous-estimation de la variabilité et de la sensibilité des estimations bayésiennes empiriques à l'hypothèse de normalité. Verbecke et al. [94, 99] proposent de détecter la non-normalité des effets aléatoires par une comparaison des résultats sous l'hypothèse de normalité avec des résultats issus d'un modèle avec des hypothèses plus souples sur les effets aléatoires.

2.2 Modèles multivariés de données longitudinales

Dans la suite, on appellera données longitudinales univariées des données générées lorsqu'une même variable est mesurée de façon répétée chez un même sujet. Toute approche pour analyser ce type de données doit prendre en compte la corrélation des observations chez un même sujet. Dans le cas de données longitudinales multivariées, plusieurs variables sont mesurées de façon répétée chez un même sujet. Si on est intéressé par la relation entre ces différentes variables, il est nécessaire de prendre en compte la corrélation entre ces variables aux même temps de mesure ou à des temps différents. L'utilisation de modèle multivarié pour données longitudinales permet d'évaluer l'association entre plusieurs marqueurs sans imposer la régression d'un marqueur (variable dépendante) sur un autre (variable indépendante). En épidémiologie, des exemples de ce type de données sont les pressions artérielles systoliques et diastoliques [100, 101], les résultats de différents tests psychométriques [102], la filtration glomérulaire et la créatinine sérique [103], l'indice de masse corporelle et l'insulinémie [104]. L'infection par le VIH n'est pas en reste puisque des modèles ont été présentés pour la modélisation conjointe des lymphocytes T CD4+ et CD8+ [81], des lymphocytes T CD4+ et de la beta-2-microglobuline [82] et de la charge virale avec les CD4+ [83, 105].

Les modèles multivariés proposés diffèrent selon leur structure de covariance et la méthode d'estimation des paramètres utilisée. Ils peuvent comprendre des effets aléatoires, une erreur autocorrélée ou les deux. Pour un modèle bivarié, le vecteur des variables réponses peut être partitionné en deux selon la variable considérée : $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$ avec Y_i^k le vecteur n_i^k de mesures du marqueur k (k=1, 2).

En reprenant les notations de la section 2.1.1, le <u>modèle bivarié à effets aléatoires</u> s'écrit :

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i \text{ avec } \begin{cases} \varepsilon_i \sim N(0, \Sigma_i) \\ \gamma_i \sim N(0, G) \end{cases}$$

et $\beta = \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix}$, $X_i = \begin{bmatrix} X_i^1 & 0 \\ 0 & X_i^2 \end{bmatrix}$, $\gamma_i = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix}$, $Z_i = \begin{bmatrix} Z_i^1 & 0 \\ 0 & Z_i^2 \end{bmatrix}$, $G = \begin{bmatrix} G^1 & G^{12} \\ G^{21} & G^2 \end{bmatrix}$, $\varepsilon_i = \begin{bmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \end{bmatrix}$,

avec X_i^k une matrice de variable explicative de dimension $n_i \times p^k$ pour le marqueur k, β^k le vecteur d'effets fixes de dimension p^k, Z_i^k une matrice $n_i \times q^k$ de variables explicatives pour les effets aléatoires γ_i^k de dimension q^k sachant que $q^k \leq p^k$.

La matrice de variance-covariance des effets aléatoire G peut être partitionnée en 4 parties : G^1 la matrice de variance-covariance des effets aléatoires pour le premier marqueur, G^2 pour le second marqueur et $G^{12} = G^{21^T}$ la matrice de covariance entre les effets aléatoires des deux marqueurs. Les erreurs de mesure sont le plus souvent considérées comme indépendantes à chaque mesure et entre les deux marqueurs. Ainsi, Σ_i est une matrice diagonale composée de deux éléments $\sigma^2_{\varepsilon^1}$ et $\sigma^2_{\varepsilon^2}$ représentant la variance des erreurs de mesure de chaque marqueur. Cependant, lorsque les deux marqueurs sont mesurés au même moment j chez le même sujet i, Schluchter [103] propose d'estimer une covariance entre ces deux erreurs. Autrement dit, $\text{cov}(\varepsilon^1_{ij}, \varepsilon^2_{ij}) = \sigma_{\varepsilon^1\varepsilon^2}$ et $\text{cov}(\varepsilon^1_{ij}, \varepsilon^2_{i'j'}) = 0$ si $i \neq i'$ ou $j \neq j'$. L'intérêt du modèle bivarié porte sur la sous-matrice $G^{12} = G^{21^T}$. Si $G^{12} = G^{21^T} = 0$ et $\text{cov}(\varepsilon^1_{ij}, \varepsilon^2_{ij}) = 0$ alors le modèle bivarié est équivalent à deux modèles univariés. Les méthodes d'estimation des paramètres du modèle bivarié sont le plus souvent basées sur le maximum de vraisemblance restreint ou non via un algorithme EM [81, 83], un Fisher scoring [103] ou un algorithme de Newton-Raphson [94]. L'interprétation de la relation entre les deux marqueurs dépend de la structure des effets aléatoires. De plus, le lien entre des variables explicatives et les marqueurs peut être précisé par l'utilisation d'une variable latente. Par exemple, Roy et Lin [106] propose un modèle à variable latente dans lequel les marqueurs sont associés à la variable latente, laquelle est une fonction de variables explicatives. L'intérêt de ce type d'approche est de modéliser une variable latente (représentant par exemple le système immunitaire) dont plusieurs aspects sont mesurés par les marqueurs observés. On peut ainsi tester l'effet global des variables explicatives sur la variable latente.

Parmi les structures d'erreur autocorrélée proposée, une des plus classiques est probablement le processus autorégressif d'ordre 1 en temps continu ou processus d'Ornstein-Uhlenbeck (OU) qui est un processus stochastique gaussien markovien avec une covariance entre deux temps s et t définie comme suit :

$R_i(w_i(s), w_i(t)) = C \times e^{B|t-s|}$ et $0 \leq s \leq t$

où B et C sont des matrices de dimension $k \times k$, k étant le nombre de marqueurs. Ces matrices sont définies telles que :

- ✓ les valeurs propres de B ont une partie réelle négative
- ✓ C et $D = -(CB + B^T C)$ sont symétriques définies positives

Ce processus a une espérance infinitésimale $E(dW(t)) = B^T W(t)$ et une matrice de covariance D. La matrice de covariance du processus lorsque $t = s$ est C [82].

Plusieurs extensions de ce processus ont été proposées dont le processus d'Ornstein-Uhlenbeck intégré (IOU) [82] ou le modèle MVRDEC pour « Multivariate Regression with Damped Exponential Correlation » [101].

Le processus IOU multivarié est un processus gaussien d'espérance nulle et de fonction de covariance R [107] :

$$R_i(w_i(s), w_i(t)) = \left(e^{B^T s} - I - B^T s\right)B^{T-2}C + CB^{-2}\left(e^{Bt} - e^{B(t-s)} - Bs\right), \quad 0 \le s \le t$$

Le mouvement brownien correspond à la limite du processus IOU quand les valeurs propres de B tendent vers moins l'infini et quand CB^{-1} est une matrice constante. Pour estimer les paramètres du IOU, Sy et al. [82] utilisent un algorithme EM.

Le modèle MVRDEC est défini par Carey et Rosner [101] comme suit :

$Y_i = X_i\beta + \varepsilon_i$ avec $\varepsilon_i \sim N(0, \Sigma_i)$

$\sum_i$ est une matrice partitionnée en quatre sous-matrices de dimension $n_i \times n_i$ dans le cas d'un modèle bivarié $\sum_i = \begin{bmatrix} \sum_i^1 & \sum_i^{21} \\ \sum_i^{12} & \sum_i^2 \end{bmatrix}$. $\sum_i^1$ et $\sum_i^2$ sont les matrices de covariance spécifiques à chaque marqueur k avec $\text{cov}\left(Y_i^k(t), Y_i^k(s)\right) = \sigma_k^2 \alpha_k^{|t-s|^{\theta_k}}, k = 1,2, \quad 0 \le \alpha_k \le 1, \quad \theta_k \ge 0$.

Quand $\theta_k = 1$ la structure de covariance correspond à un processus auto-régressif d'ordre 1, quand $\theta_k \to \infty$ la structure de covariance tend vers un processus de moyenne mobile d'ordre 1. $\sum_i^{12}$ contient les estimations des covariances entre les marqueurs avec :

$$\text{cov}\left(Y_i^1(t), Y_i^2(s)\right) = \sigma_1 \sigma_2 \alpha_{12}^{|t-s|^{\theta_{12}}}, \quad -1 \le \alpha_{12} \le 1, \quad \theta_{12} \ge 0$$

Le paramètre α_{12} représente la corrélation simultanée $(t = s)$ entre les deux marqueurs. La vitesse de décroissance de cette corrélation en fonction du temps dépend de θ. Carey et Rosner [101] proposent une maximisation classique de vraisemblance par un algorithme de Newton-Raphson pour obtenir une estimation des paramètres du modèle.

Un point intéressant de ces modèles est l'interprétation de la relation entre les marqueurs. Pour le processus auto-régressif d'ordre 1, si B est une matrice triangulaire alors un des marqueurs influence l'autre mais le contraire n'est pas vrai [82, 90]. Ce même type de raisonnement peu être réalisé en étudiant les corrélations entre les marqueurs à des temps différents. Par exemple dans l'article de Sy et al. [82], pour deux temps $s < t$, ils trouvent $corr(W_2(s), W_1(t))$

très supérieur à $corr(W_1(s), W_2(t))$, ce qui indique que W_2 affecte plus W_1 que W_1 n'affecte W_2.

La spécification de la matrice de covariance joue un rôle important dans la modélisation longitudinale et peut avoir un impact non négligeable notamment sur les prédictions [108]. Dans le cadre des modèles multivariés, une mauvaise spécification de la matrice de covariance d'un marqueur peut influencer non seulement les résultats concernant ce marqueur mais également ceux des autres marqueurs.

2.3 Traitement des données manquantes informatives

Dans un modèle de régression, les données de la variable à expliquer ou des variables explicatives peuvent être incomplètes. On considérera uniquement les données incomplètes pour la variable à expliquer. Ces données peuvent être manquantes ou censurées. Les données manquantes peuvent être manquantes par intermittence (un patient ne se présente pas à une visite mais est revu à la visite suivante) ou de façon définitive (un patient perdu de vue, par exemple). C'est ce dernier type de données manquantes, appelé monotone, que nous considérerons par la suite.

2.3.1 Taxonomie

En présence de données manquantes, la variable réponse Y_i peut être séparée en deux sous-vecteurs: Y_i^o pour les données observées et Y_i^m pour les données manquantes. On observe également une variable D_i générée par le processus de données manquantes ayant pour paramètres ψ. Dans le cadre d'un schéma d'étude comportant des visites définies, D_i peut être discret (au même titre que le temps de chaque mesure). Il indique alors à quelle visite est survenue la sortie d'étude ou bien, quand le sujet n'est pas sorti prématurément de l'étude, il vaut $D_i = n_i + 1$. Pour les cohortes basées sur des pratiques cliniques où les visites ne sont pas fixes, il est plus adéquat de définir D_i comme un temps continu jusqu'à la sortie d'étude (prématurée ou non). Enfin, on observe également une matrice de variables explicatives X_i. On rappelle que $\theta = (\beta^T, \alpha^T)^T$ est le vecteur de paramètres du modèle linéaire mixte (1).

La densité conjointe de Y_i et D_i peut se factoriser comme suit :

$$f(y_i,d_i|X_i,\theta,\psi)=f(y_i|X_i,\theta)f(d_i|y_i,X_i,\psi) \quad (4)$$

Le premier facteur correspond à la densité marginale de Y_i et le second facteur correspond à la sélection des individus selon les groupes "observé" ou "manquant". C'est pour cela que cette factorisation est la base des modèles dits de sélection.

Une autre factorisation possible conduit aux modèles dits de mélange (pattern-mixture models) [109] :

$$f(y_i,d_i|X_i,\theta,\psi)=f(y_i|d_i,X_i,\theta)f(d_i|X_i,\psi)$$

La réponse est cette fois étudiée conditionnellement à la date de sortie d'étude. En d'autres termes, la population est stratifiée selon la sortie d'étude ce qui implique que le modèle pour l'ensemble de la population est un mélange des modèles pour les différents temps de sortie d'étude.

Lorsqu'on s'intéresse aux données manquantes concernant la variable à expliquer dans un modèle de régression, il est important de définir dans quel but on désire tenir compte de ces données manquantes [110, 111]. Si l'intérêt porte sur la réponse complète (observée ou non), alors il est utile d'obtenir une description marginale de la réponse, c'est à dire $f(y_i|X_i,\theta)$. Si, au contraire, cela n'a pas de sens de se demander quelle réponse aurait eu un sujet s'il était resté dans l'étude, alors on préférera une description de la réponse conditionnelle au fait que le sujet soit resté dans l'étude, c'est à dire $f(y_i|d_i,X_i,\theta)$. Par exemple, la représentation marginale pourrait ne pas être adaptée quand la nature de la sortie d'étude _ le décès, par exemple _ rend impossible l'existence d'une réponse. Dans un essai thérapeutique, les sujets peuvent arrêter de participer du fait de la survenue d'un effet secondaire indésirable. Quand l'essai est explicatif (ou quand il y a des objectifs secondaires explicatifs au sein d'un essai pragmatique), on s'intéresse plutôt aux propriétés pharmacologiques du traitement [112]. Dans ce cas, la réponse conditionnelle au fait que le sujet soit rester dans l'étude peut sembler être la réponse d'intérêt. Cependant, la réponse marginale peut aussi être utile [110]. En effet, l'investigateur peut, par exemple, être intéressé par la réponse au traitement si les patients n'étaient pas sortis de l'étude pour cause d'effets secondaires indésirables. L'intérêt de ce type

d'étude est évident notamment si on pense améliorer ultérieurement la tolérance du traitement sans modifier son efficacité.

Par la suite, on s'intéressera spécifiquement aux modèles dits de sélection dans la mesure où dans notre travail nous avons utilisé ce type de modèle notamment du fait de l'intérêt porté à la réponse marginale (confère section 2.3). Cependant, il faut noter que la réponse marginale peut également être obtenue à partir d'un modèle de mélange.

Little et Rubin [113] ont proposé un cadre conceptuel pour définir le type de données manquantes à partir de la factorisation (4). Le second facteur de l'expression (4) peut être explicité selon le statut des données :

$$f\left(d_i|y_i,X_i,\psi\right)=f\left(d_i\left|y_i^o,y_i^m,X_i,\psi\right.\right) \qquad (5)$$

Le mécanisme du processus de données manquantes est défini comme :

- ✓ complètement aléatoire (MCAR pour Missing Completely At Random) si la probabilité de réponse est indépendante de la variable étudiée qu'elle soit observée ou manquante, l'expression (5) se simplifie : $f\left(d_i|y_i,X_i,\psi\right)=f\left(d_i|X_i,\psi\right)$.
- ✓ Aléatoire (MAR pour Missing At Random) si conditionnellement aux données observées, le processus est indépendant des données manquantes, c'est-à-dire $f\left(d_i|y_i,X_i,\psi\right)=f\left(d_i\left|y_i^o,X_i,\psi\right.\right)$. La probabilité qu'une observation soit incomplète dépend uniquement des valeurs observées.
- ✓ Informatif (ou non aléatoire ou MNAR pour Missing Not At Random) si le processus de réponse dépend des données non observées, $f\left(d_i|y_i,X_i,\psi\right)=f\left(d_i\left|y_i^o,y_i^m,X_i,\psi\right.\right)$

Dans le cas particulier des perdus de vue dans les études longitudinales, Diggle et Kenward [114] ont suivi cette taxonomie pour définir le processus de sortie d'étude comme :

- ✓ complètement aléatoire (CRD pour completely random drop-out) où les processus de sortie d'étude et de mesure sont complètement indépendants.
- ✓ Aléatoire (RD pour random drop-out) où le processus de sortie d'étude dépends des mesures déjà observées, c'est à dire précédent la sortie d'étude.
- ✓ Informatif (ID pour informative drop-out) où le processus dépend des mesures non observées c'est à dire qui auraient été observée si le sujet n'était pas perdu de vue.

On préfère parler de sortie d'étude informative plutôt que de censure informative [115] car les observations même du processus de sortie d'étude D_i peuvent être censurées [116]. Dans ce cas, D_i est censuré par un temps de censure C_i : la fin de l'étude par exemple, ou une autre cause de sortie d'étude supposée non informative. On observe $d_i^o = \min\{d_i, c_i\}$ et un indicateur de censure $\delta_i = I_{\{d_i \le c_i\}}$.

Par la suite, du fait de l'utilisation des modèles à effets aléatoires, la définition de données manquantes informatives a été affinée selon que le processus de données manquantes dépend effectivement des valeurs non observées de la variable réponse ou des effets aléatoires non observés [116, 117].
On définira donc un processus de données manquantes :

- ✓ Variable réponse-dépendant si $f(d_i|y_i,\gamma_i,X_i,\psi) = f(d_i|y_i^o,y_i^m,X_i,\psi)$
- ✓ Effets aléatoires-dépendant si $f(d_i|y_i,\gamma_i,X_i,\psi) = f(d_i|y_i^o,\gamma_i,X_i,\psi)$

Un cas particulier se présente lorsque la variable à expliquer Y_i et le processus de données manquantes D_i sont tous les deux associés à une variable explicative X_i. Si conditionnellement à X_i, c'est à dire ajusté sur cette variable explicative, Y_i et D_i sont indépendants, on a $f(y_i,d_i|X_i,\theta,\psi) = f(y_i|X_i,\theta)f(d_i|X_i,\psi)$ et les données manquantes peuvent être considérées comme manquantes complètement au hasard (MCAR). Au contraire, si on ignore l'effet de X_i, on se retrouve dans un contexte MAR voir MNAR. Cette situation est identifiée par Little [117] comme une sortie d'étude covariable-dépendante, réservant le terme MCAR lorsque le processus est indépendant des données observées de la réponse et des covariables, c'est à dire que l'expression (5) se résume à $f(d_i|y_i,X_i,\psi) = f(d_i|\psi)$.

Avant de présenter les modèles de sélection pour la modélisation conjointe du processus de données manquantes et des données répétées en section 2.3.4, des techniques simples (2.3.2) ou des méthodes usuelles basées sur la maximisation de vraisemblance (2.3.3) sont évoquées.

2.3.2 Méthodes simples d'analyse de la réponse en présence de données manquantes non informatives

Pour traiter les données manquantes non informatives, plusieurs méthodes simples sont classiquement utilisées.

La modélisation des données complètes utilise uniquement les sujets pour lesquels toutes les mesures ont été observées. Cette méthode présente l'avantage d'être simple et elle peut être mise en œuvre avec des logiciels standards. En revanche, non seulement elle entraîne une perte de puissance mais en plus elle n'est valide que dans le cas MCAR. Dans le cas contraire, elle conduit à des estimations biaisées.

Une autre méthode souvent utilisée est l'imputation simple. Il s'agit de remplacer les données manquantes par des prédictions effectuées à partir d'informations issues d'un même sujet (dernière observation observée ou LOCF pour Last Observation Carried Forward), des autres sujets (imputation de la moyenne, par exemple) ou des deux (imputation conditionnelle ou méthode de Buck, imputation "Hot Deck"). Dans tous les cas ce type d'approche peut poser deux problèmes. Tout d'abord le modèle d'imputation peut être faux. Par exemple, la méthode LOCF fait l'hypothèse que le marqueur reste à la même valeur après la sortie d'étude du sujet, ce qui est souvent peu probable. Cependant, le modèle d'imputation basé sur les observations déjà effectuées peut être correct en particulier en cas de données manquantes aléatoirement (MAR). L'autre problème majeur est que l'incertitude due à l'incomplétude des données est masquée ce qui conduit à une sous-estimation des variances des paramètres.

Pour corriger ce type de difficulté, les méthodes d'imputation multiple permettent de remplacer chaque donnée manquante par plusieurs valeurs [118]. Le principe est de produire $k = 1,...M$ échantillons du vecteur Y_i^m à partir d'un modèle pour la distribution conditionnelle $f\left(y_i^m \middle| y_i^o\right)$. Ensuite, le vecteur de paramètres d'intérêt θ est estimé séparément avec chacun des k échantillons de données complétées $\left(Y_i^o, Y_i^{mk}\right)$. Enfin, les M inférences sont combinées et la variance des estimations finale de θ est calculée en prenant en compte la variance intra et inter-imputations.

2.3.3 Méthodes basées sur la vraisemblance et ignorabilité du processus de données manquantes

On peut également utiliser des estimations basées sur la vraisemblance. Les inférences étant basées sur les données observées $f\left(y_i^o, d_i\right)$, on écrit la vraisemblance pour ces données :

$$L\left(\theta,\psi \middle| X_i, y_i^o, d_i\right) \propto f\left(y_i^o, d_i \middle| X_i, \theta, \psi\right)$$

avec

$$f\left(y_i^o, d_i \middle| X_i, \theta, \psi\right) = \int f\left(y_i^o, y_i^m, d_i \middle| X_i, \theta, \psi\right) dy_i^m = \int f\left(y_i^o, y_i^m \middle| X_i, \theta\right) f\left(d_i \middle| y_i^o, y_i^m, X_i, \psi\right) dy_i^m \qquad \textbf{(6)}$$

Lorsqu'on utilise un modèle à effets aléatoires, les distributions de Y_i et D_i peuvent également dépendre des effets aléatoires non observés. Sachant $\theta = \left(\beta^T, \alpha^T\right)^T$, on a :

$$f\left(y_i^o, d_i \middle| X_i, \theta, \psi\right) = \iint f\left(y_i^o, y_i^m \middle| \gamma_i, X_i, \theta\right) f\left(d_i \middle| y_i^o, y_i^m, \gamma_i, X_i, \psi\right) f\left(\gamma_i \middle| \alpha\right) d\gamma_i dy_i^m \qquad \textbf{(7)}$$

Sous l'hypothèse d'un processus MAR (section 2.3.1) :

$$f\left(d_i \middle| y_i^o, y_i^m, \gamma_i, X_i, \psi\right) = f\left(d_i \middle| y_i^o, X_i, \psi\right)$$

d'où à partir de (7) :

$$f\left(y_i^o, d_i \middle| X_i, \theta, \psi\right) = \iint f\left(y_i^o, y_i^m \middle| \gamma_i, X_i, \theta\right) f\left(d_i \middle| y_i^o, X_i, \psi\right) f\left(\gamma_i \middle| \alpha\right) d\gamma_i dy_i^m$$

d'où $f\left(y_i^o, d_i \middle| X_i, \theta, \psi\right) = f\left(y_i^o \middle| X_i, \theta\right) f\left(d_i \middle| y_i^o, X_i, \psi\right)$

Donc la log-vraisemblance peut se décomposer comme suit :

$$\ell\left(\theta, \psi, y_i^o, d_i\right) = \ell_1\left(\theta; y_i^o\right) + \ell_2\left(\psi; d_i, y_i^o\right)$$

Si ψ et θ sont distincts (c'est à dire que la connaissance de l'un n'entraîne pas de contraintes sur les valeurs de l'autre), l'ensemble de l'information concernant les paramètres modélisant la variable réponse est comprise dans le terme ℓ_1. Dans ce cas, le processus de données manquantes est dit ignorable et les méthodes de maximisation de la vraisemblance basée uniquement sur ℓ_1 conduisent à des inférences valides sur le vecteur de paramètres θ.

Ce résultat est important mais pour s'assurer que le processus de données manquantes est aléatoire (MAR) et pas informatif (MNAR), il est nécessaire de savoir s'il dépend ou non des observations manquantes. En pratique, ceci n'est pas réalisable puisque par définition les observations manquantes sont inconnues. Cependant, certains schémas d'étude peuvent générer des données manquantes selon un processus MAR. Par exemple, c'est le cas lorsque la

variable réponse n'est pas mesurée pour certaines valeurs d'une variable explicative. Il est possible de tester l'hypothèse de données MCAR versus MAR [119, 120]. Dans la situation où les données manquantes sont MCAR, une analyse sur données complètes ou des méthodes de type équations d'estimation généralisées [121] peuvent être utilisées.

Au total, la condition d'ignorabilité du processus de données manquantes permet l'utilisation des techniques classiques de maximisation de vraisemblance pour estimer les paramètres de la distribution marginale de la variable réponse. Ces techniques sont disponibles dans des logiciels classiques tels que S-PLUS® (fonction lme) ou SAS® (procédure MIXED). Cependant, sauf cas particuliers, l'hypothèse MAR n'étant pas vérifiable, des analyses de sensibilité sont souvent nécessaires [94].

2.3.4 Modèles de sélection pour la modélisation conjointe de la variable réponse et du processus de données manquantes

Les modèles de sélection destinés à modéliser conjointement les processus de mesures et de données manquantes peuvent être séparées en deux catégories : les modèles variable-réponse-dépendants et les modèles effets-aléatoires-dépendants [116, 117].

Dans les modèles variable réponse-dépendants, le processus de données manquantes dépend de Y_i^o et de Y_i^m. Ces modèles nécessitent que Y_i soit mesuré à des temps discrets, prédéfinis et que les données manquantes soient monotones.
Dans cette classe de modèle, le plus classique est sans doute celui de Diggle et Kenward [114]. La probabilité de sortie d'étude au temps t_{ij} est modélisée par un modèle logistique dont les variables explicatives sont les valeurs de Y du sujet i au temps t_{ij}, et éventuellement au temps précédent t_{ij-1} ainsi qu'un vecteur de variables explicatives K_{ij} qui peut être différent de X_{ij} : $\text{logit}\left\{P\left(d_{ij}=t_{ij}\right)\right\}=\varphi_0+\varphi_1 y_{ij}+\varphi_2 y_{ij-1}+\phi K_{ij}$
On peut noter que $\varphi_1=0$ signifie que la sortie d'étude est aléatoire (MAR) et si $\varphi_1=0$ et $\varphi_2=0$ alors la sortie d'étude est complètement aléatoire (MCAR). La maximisation de la vraisemblance de leur modèle utilise l'algorithme simplex de Nelder et Mead et nécessite une intégration numérique. Parmi les hypothèses du modèle, on retient le lien logistique entre la probabilité de sortie et la variable réponse, la normalité de la distribution de la variable

réponse [122] et le caractère monotone des données manquantes. Cependant le modèle a été étendu à des données manquantes par intermittence [123].

Il existe des situations dans lesquelles la sortie d'étude est associée à une tendance au cours du temps plutôt qu'à une valeur donnée de la variable réponse. Dans ce cas, les modèles effets-aléatoires-dépendants semblent plus adaptés. Wu et Caroll [115] ont proposé un modèle probit liant la probabilité de sortie d'étude à l'intercept (γ_{oi}) et la pente (γ_{1i}) individuels de chaque patient. Considérant D_i comme un temps de sortie d'étude continu et potentiellement censuré, la régression probit proposée est :

$\Phi^{-1}\left\{P(d_i \le t_{ij})\right\} = \varphi_{0j} + \varphi_1\gamma_{0i} + \varphi_2\gamma_{1i}$ où Φ est une fonction de répartition univariée normale.

D'après ce modèle, la sortie d'étude est informative si $(\varphi_1, \varphi_2) \neq (0,0)$. Les estimations des paramètres du modèle sont obtenues à l'aide des estimateurs du pseudo-maximum de vraisemblance (PMLE) [124] et de l'algorithme de Newton-Raphson. Par la suite, Wu et Bailey [125] et Mori et al. [126] ont proposé d'autres estimateurs basés sur un modèle linéaire conditionnel. Cette dernière approche est limitée par le fait que les patients sont supposés être potentiellement suivis la même longueur de temps ce qui signifie qu'ils sont tous inclus au même moment. De plus les approches par modèle conditionnel comme par PMLE utilisent des estimations des pentes par les moindres carrés ce qui rend impossible le calcul d'une pente avec un seul point.

Schluchter [127] et DeGruttola et Tu [33] ont proposé un modèle où (γ_i, D_i) a une distribution multivariée normale et les paramètres basés sur le maximum de vraisemblance sont obtenus avec un algorithme EM. Bien que les deux modèles soient équivalents, les motivations dans les deux approches étaient différentes. Schluchter [127] désirait corriger le biais des paramètres du modèle linéaire en présence de sortie informative. Au contraire, DeGruttola et Tu [33] étaient intéressés par l'effet de la pente individuelle sur D_i, qui était un temps de survie dans leur cas.

Par la suite, les modèles conjoints ont été améliorés par l'utilisation de modèles plus souples. Par exemple, le modèle de survie ou de sortie d'étude utilisé peut être un modèle de Weibull [128] ou un modèle de Cox [129, 130] avec éventuellement un terme de fragilité [131]; le modèle longitudinal pour le marqueur peut inclure un processus stochastique en plus des

effets aléatoires [132, 133]. Les méthodes d'estimations utilisées sont variées incluant le MCMC [129], l'EM [130], le simplex [128] ou une conjonction de ces deux derniers [131].

2.3.5 Modèles conjoints pour l'étude de l'effet d'un marqueur sur la survenue d'un événement

De manière générale, les méthodes d'analyse de données longitudinales comportant des sorties d'études informatives consistent à modéliser conjointement l'évolution longitudinale d'un marqueur et le temps de survenue d'un événement engendrant la sortie d'étude. Ces modèles conjoints peuvent donc également être utilisés pour étudier l'effet de l'évolution d'un marqueur sur la survenue de tout type d'événement.

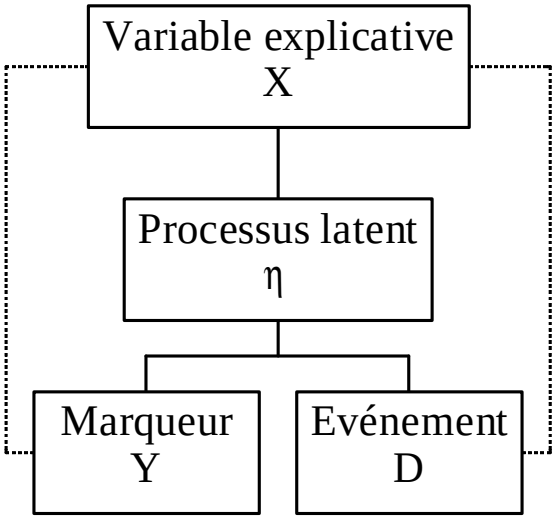
L'approche la plus simple utilisée pour étudier l'effet d'un marqueur sur la progression clinique utilise directement les mesures observées du marqueur en tant que variable dépendante du temps dans un modèle de survie. Cette approche a comme inconvénient de considérer la valeur du marqueur constante entre deux mesures et de ne pas tenir compte de l'erreur de mesure faite sur le marqueur. L'estimation de l'effet du marqueur sur la progression clinique est alors biaisée vers 0 (pas de relation) [32, 134]. De plus, à l'occasion de l'étude d'un marqueur de substitution (voir section 1.4), une approche naïve ne prenant pas en compte l'erreur de mesure sur le marqueur engendre un effet résiduel du traitement ajusté sur le marqueur faisant conclure que le marqueur est un mauvais marqueur de substitution [31].

Une approche en deux étapes a été proposée [135]. Le principe est d'utiliser les estimations bayésiennes empiriques issues d'un modèle linéaire mixte (première étape) pour étudier l'effet des marqueurs sur la progression clinique au sein d'un modèle de Cox en tant que variables dépendantes du temps (deuxième étape). Ce type d'approche permet de réduire le biais sur les estimations du modèle de Cox lié à l'erreur de mesure sur la variable explicative dépendante du temps. Le modèle linéaire mixte utilisé dans la première étape peut comprendre des effets aléatoires [31, 135] éventuellement associé à un mouvement brownien [136] ou plus généralement un processus d'Ornstein-Uhlenbeck intégré [137]. Le problème de ce type d'approche est que la variabilité engendrée par l'estimation des paramètres du modèle mixte dans la première étape n'est pas prise en compte. Self et Pawitan [138] ont proposé un estimateur de la variance pour $\hat{\phi}$ prenant en compte cette variabilité supplémentaire. D'autres

corrections de l'erreur de mesure ont été proposées avec notamment un lissage du marqueur par des moyennes mobiles [30] ou une méthode d'imputation multiple [139].

L'autre solution est de modéliser conjointement le marqueur et le temps de survie comme rapporté dans la section précédente 2.3.4. Les modèles effets-aléatoires-dépendants peuvent également être assimilés à des modèles à variable latente, l'effet aléatoire joue le rôle d'une variable latente [33, 133]. L'idée est qu'un processus stochastique sous-jacent, non observé, η représente, par exemple, l'état de santé de l'individu i. Dans ce cas D_i est plutôt un délai jusqu'à la survenue d'un événement clinique. Le marqueur Y_i représente une mesure imparfaite de cet état de santé. On suppose donc que (a) D_i et Y_i sont indépendants conditionnellement à η, (b) X_i peut affecter D_i soit par l'intermédiaire de η soit directement et (c) X_i peut influencer Y_i ou D_i par l'intermédiaire de η. Ceci est résumé par le schéma causal présenté figure 2.

Figure 2. Schéma causal de l'influence d'une variable explicative sur un marqueur et un événement par l'intermédiaire d'un processus latent



Un des critères principaux pour que Y_i soit considéré comme un bon marqueur de substitution du véritable événement est que l'effet du traitement X_i sur la survenue d'un événement clinique D_i passe entièrement par Y_i. Il faut donc que $[D_i|X_i, Y_i] = [D_i|Y_i]$ ce qui est testable

par cette approche. Cette modélisation conjointe est donc très utile pour l'étude de la qualité des biomarqueurs du VIH en tant que marqueur de substitution (voir sections 1.4).

Les développements les plus récents des modèles conjoints ont porté sur l'extension de ces modèles à l'étude simultanée de plusieurs marqueurs [140] ou de plusieurs temps de survie [141]. Dans ces deux articles le lien entre les marqueurs et la survenue d'un événement est un ou plusieurs processus latents dépendants du temps. L'approche bayésienne de Xu et Zeger [133, 140] pour estimer les paramètres du modèle utilisent un algorithme MCMC avec des distributions a priori non informatives. Huang et al. [141] utilise un algorithme de Newton-Raphson en calculant les dérivées secondes par différentiation automatique. Toutefois, il faut noter que leur vraisemblance et le calcul des intégrales sont simplifiés par l'utilisation de variables latentes et de marqueurs binaires. Ainsi, bien qu'il soit un peu hors sujet du fait d'une variable réponse binaire plutôt que continue gaussienne, nous citons cet article du fait de son caractère novateur.

2.4 Mesures censurées

Dans les modèles de régression, les données sur la variable à expliquer peuvent être incomplètes parce que manquantes ou censurées. Dans le contexte des données longitudinales gaussiennes, la censure est générée par le manque de sensibilité des techniques de mesure. Il faut donc distinguer ce type de censure avec celle d'un temps de survie où le décès n'a pas été observé bien que certaines techniques statistiques puissent être utiles dans les deux cas.

2.4.1 Définitions

Une variable est dite censurée quand sa valeur exacte est inconnue mais que l'on sait seulement si elle est inférieure ou supérieure à une valeur. Dans une analyse de survie, il s'agit d'un délai entre l'entrée dans l'étude et la survenue d'un événement. Ce délai est censuré parce qu'il n'a pas été possible d'observer l'événement du fait de la fin de l'étude (censure à droite) ou du fait de sa survenue entre deux périodes d'observation (censure à gauche et à droite, c'est à dire par intervalle). Il peut également s'agir de concentrations qui ne peuvent être déterminées précisément à partir d'un seuil donné du fait de limitations techniques. On peut citer par exemple les concentrations de polluants [142], d'anticorps [143] ou de la charge virale plasmatique du VIH (voir section 1.3). Ces observations peuvent être censurées à gauche s'il existe un seuil de détection inférieur ou elles peuvent être censurées à droite s'il existe un seuil de détection supérieur. Classiquement, on distingue les censures aléatoires et non aléatoires. Parmi ces dernières, les censures de type I surviennent lorsque le seuil de censure C est fixé a priori pour toutes les observations et que le nombre d'observations censurées varient. En analyse de survie, un protocole ou la durée de suivi est fixée à l'avance répondrait à cette définition (censure à droite de type I). Le plus souvent, les censures liées à une limite de quantification sont de type I puisqu'on connaît le seuil de détectabilité de la technique. Les censures de type II surviennent lorsque le nombre d'événements est connu à l'avance mais que le point de censure peut être variable. Un exemple de censure à droite de type II est une étude de fiabilité qui est arrêtée lorsqu'un nombre défini d'événements est observé.

Par ailleurs, la censure peut être aléatoire ou non. Par exemple, dans une analyse de survie, lorsque l'entrée dans l'étude dépend de la survenue d'un événement et que la date de fin d'étude est fixée, le délai entre l'entrée et la fin de l'étude est aléatoire. Dans le contexte de la quantification d'une variable biologique, si l'intérêt porte sur l'évolution de cette quantité

entre deux points et que la valeur au deuxième temps est potentiellement censurée alors la différence entre la valeur initiale et la valeur au deuxième temps sera censurée de façon aléatoire. Par exemple, la différence entre la charge virale plasmatique du VIH à la mise en place d'un traitement antirétroviral et à la date de point (6 mois, par exemple) est un critère classique dans les essais thérapeutiques.

2.4.2 Méthodes d'analyse de données censurées

De nombreuses méthodes d'estimation de la distribution de données censurées de type I existent dans la littérature allant de l'imputation simple (de la valeur ou de la moitié de la valeur du seuil de censure) à des techniques de régression ou de maximisation de vraisemblance [142]. Pour la charge virale plasmatique, Lynn [144] montre les biais engendrés par l'imputation simple de la valeur ou de la moitié de la valeur du seuil sur les estimations de la moyenne et de l'écart-type de la distribution du logarithme de la charge virale. L'imputation multiple peut également induire des estimations biaisées car les données censurées traitées comme des données manquantes ne sont pas manquantes par hasard (MAR). En effet, celles-ci sont censurées du fait que la valeur est inférieure au seuil de détection. La méthode d'imputation multiple peut être améliorée en tenant compte du mécanisme de censure. Dans ce cas, elle conduit à des estimations proches de celles du maximum de vraisemblance qui sont les moins biaisées. L'hypothèse sous-jacente de ces méthodes est une distribution normale de l'ensemble des observations du logarithme de la charge virale. Lynn [144] et Lyles et al. [145] proposent également ces méthodes pour évaluer la corrélation de la charge virale avec d'autres variables ou bien l'effet de la charge virale sur une variable dichotomique en utilisant une régression logistique.

Plutôt que d'analyser la distribution d'une variable censurée ou son effet en tant que prédicteur, on peut également être intéressé par la régression de la variable censurée sur un certain nombre de variables explicatives. Si on considère la distribution de la variable censurée comme normale, le modèle Tobit et les techniques de régression en analyse de survie peuvent être utilisées [146, 147]. Ceci a été proposé pour l'analyse de la charge virale [148]. En effet, si Y_0 est la valeur initiale non censurée et Y_1 la valeur de la charge virale au cours du suivi potentiellement censurée à gauche, la différence $\Delta = Y_0 - Y_1$ est potentiellement censurée à droite par la valeur $Y_0 - R$ (R étant la valeur du seuil de censure). En pratique, les auteurs analysent la différence de charge virale entre la valeur à la date d'initiation d'un

traitement antirétroviral et un temps fixe ultérieur en utilisant la procédure LIFETEST de SAS® pour la méthode non paramétrique de Kaplan-Meier et la procédure LIFEREG pour l'estimation d'un modèle paramétrique de vie accélérée. Lorsque la proportion de mesures censurées est importante, l'estimation non paramétrique n'est pas toujours disponible [149, 150]. D'autre part, les méthodes paramétriques sont sensibles à l'hypothèse de distribution de la charge virale [149]. Lorsqu'on analyse la réduction de charge virale, on peut être confronté à un problème de censure informative. En effet, si le changement $\Delta = Y_0 - Y_1$ et donc la censure sont dépendants de Y_0, alors il y a censure informative. Pour tester cette hypothèse et la prendre en compte, il est nécessaire d'ajuster le modèle paramétrique sur la valeur initiale Y_0 [148, 149]. Les méthodes paramétriques proposées pour analyser les données censurées sont le plus souvent basée sur une hypothèse de distribution normale de la variable dépendante. Il a été proposé également des mélanges de distribution pour les données observées et les données censurées [143].

Le premier article prenant en compte la censure à droite de données longitudinales gaussiennes proposait un algorithme EM [151]. L'auteur utilisait en effet un modèle linéaire mixte appliqué au logarithme de temps de survie. Pour les données longitudinales gaussiennes censurées à gauche, la plupart des développements méthodologiques ont été réalisés pour des applications à la charge virale du VIH. La première méthode proposée fut une imputation multiple dans le cadre d'un article épidémiologique [56]. Puis, il a été proposé un algorithme MCEM pour prendre en compte une variable dépendante censurée à gauche ou à droite dans le cadre d'un modèle mixte [152]. Par la suite d'autres méthodes d'estimation ont été proposées, toutes basées sur le même modèle [85, 153].

3 Modélisation bivariée de la charge virale et des CD4+

3.1 Motivation

Bien que les modèles pour données longitudinales soient de plus en plus utilisés en médecine, les applications utilisant des modèles multivariés longitudinaux sont encore très rares. Bien que le nombre de marqueurs mesurés chez les patients infectés par le VIH soit très important, les analyses d'étude épidémiologiques sont le plus souvent restreintes à l'explication de l'évolution d'un marqueur à partir d'autres marqueurs (voir section 1.4). Seuls quelques articles méthodologiques ont présenté une modélisation conjointe des lymphocytes T CD4+ et CD8+ [81], des lymphocytes T CD4+ et de la beta-2-microglobuline [82], et de la charge virale avec les CD4+ [83, 105]. Pourtant cette modélisation multivariée se justifie pour plusieurs raisons. Tout d'abord, l'hypothèse qu'un marqueur influence un autre sans que ce dernier n'influence le premier peut être trop restrictive. Par exemple, lors de la primo-infection, l'évolution de la charge virale dépend de la quantité de CD4+ [154] et après l'initiation d'un traitement hautement actif la réponse immunologique (augmentation des CD4+) est influencée par la réponse virologique (décroissance de la charge virale) qui semble survenir plus tôt [105]. L'autre argument à l'utilisation de modèles multivariés est la prise en compte de toute l'information disponible. Le plus souvent, l'étude de l'impact de la charge virale sur l'évolution des CD4+ utilise des mesures de charge virale à des temps donnés (voir par exemple [74]). La modélisation conjointe des deux marqueurs permet de prendre en compte toutes les informations disponibles pour les deux marqueurs qui sont le plus souvent mesurés au même moment.

Jusqu'à présent les méthodes présentées et appliquées dans le VIH étaient implémentées avec des programmes écrits en C/Fortran [105] ou dans le module IML (calcul matriciel) de SAS® [81-83]. Ceci a pour inconvénient une diffusion limitée de ces méthodes ainsi que des temps de calcul souvent supérieur à l'utilisation d'un programme compilé et optimisé.
La disponibilité de procédures d'estimation pour les modèles longitudinaux univariés dans des logiciels de statistique (Procédure MIXED dans SAS® ou fonction lme dans S-PLUS) a contribué, de façon déterminante, au développement de leur utilisation. Certains modèles univariés sont assez simplement généralisables pour une modélisation multivariée en utilisant ces mêmes logiciels. Il nous a donc semblé utile de présenter l'estimation d'un modèle pour données longitudinales bivariées à partir de la procédure SAS® MIXED et de montrer l'intérêt

d'une telle approche pour l'étude des marqueurs du VIH. Cet article a été publié dans la revue *Computer Methods and Programs in Biomedicine.*

3.2 Modélisation bivariée à l'aide de la procédure MIXED

Bivariate linear mixed models using SAS proc MIXED

Rodolphe Thiébaut[a*], Hélène Jacqmin-Gadda[a], Geneviève Chêne[a], Catherine Leport[b], Daniel Commenges[a]

[a] INSERM Unité 330, ISPED, Université Victor Segalen Bordeaux II, 146, rue Léo Saignat 33076, Bordeaux Cedex, France

[b] Hôpital Bichat Claude Bernard, Paris, France

Abstract

Bivariate linear mixed models are useful when analyzing longitudinal data of two associated markers. In this paper, we present a bivariate linear mixed model including random effects or first-order auto-regressive process and independent measurement error for both markers. Codes and tricks to fit these models using SAS Proc MIXED are provided. Limitations of this program are discussed and an example in the field of HIV infection is shown. Despite some limitations, SAS Proc MIXED is a useful tool that may be easily extendable to multivariate response in longitudinal studies.

Keywords: Bivariate random effects model, Bivariate First Order Auto-regressive process, SAS proc MIXED, HIV infection

* Corresponding author. Tel: +33 5 57 57 45 21 Fax: +33 5 56 24 00 81

Email: rodolphe.thiebaut@isped.u-bordeaux2.fr

1. Introduction

Longitudinal data are often collected in epidemiological studies, especially to study the evolution of biomedical markers. Thus, linear mixed models [1], recently available in standard statistical packages [2, 3], are increasingly used to take into account all available information and deal with the intra-subject correlation.

When several markers are measured repeatedly, longitudinal multivariate models could be used, like in econometrics. However, this extension of univariate models is rarely used in biomedicine although it could be useful to study the joint evolution of biomarkers. Into example, in HIV infection, several markers are available to measure the quantity of virus (plasma viral load noted HIV RNA), the status of immune system (CD4+ T lymphocytes which are a specific target of the virus, CD8+ T lymphocytes) or the inflammation process (β2 microglobuline). These markers are associated as the infection measured by HIV RNA induces inflammation and the destruction of immune cells. Several authors have developed methods to fit evolution of CD4 and CD8 cells [4] or CD4 and β2 microglobuline [5]. Amrick Shah et al. [4] used an EM algorithm to fit a bivariate linear random effects model. Sy et al [5] used the Fisher scoring method to fit a bivariate linear random effects model including an integrated Orstein-Uhlenbeck process (IOU). IOU is a stochastic process that includes Brownian motion as special limiting case.

Their programs were implemented using IML module of SAS Software [6]. However, their flexibility is not sufficient to allow a large use by researchers not familiar with IML. Also, the EM algorithm chosen is slow.

In this paper, we propose some tricks to use SAS MIXED procedure in order to fit multivariate linear mixed models to multivariate longitudinal gaussian data. SAS MIXED procedure uses Newton-Raphson algorithm known to be faster than the EM algorithm [7]. In section 2 and 3, we present bivariate linear mixed models and the code used in SAS to fit these models. In section 4, we apply these models to study the joint evolution of HIV RNA and CD4+ T lymphocytes in a cohort of HIV-1 infected patients (APROCO) treated with highly active antiretroviral treatment.

2. Model for bivariate longitudinal gaussian data

We define a general bivariate linear mixed model including a random component, a first order auto-regressive process and an independent error.

Let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$, the response vector for the subject i, with Y_i^k the n_i^k-vector of measurements of the marker k (k=1, 2) with $n_i^1 = n_i^2 = n_i$. If the two markers are independent, we can use the two following models:

$$\begin{cases} Y_i^1 = X_i^1\beta^1 + Z_i^1\gamma_i^1 + W_i^1 + \varepsilon_i^1 \\ Y_i^2 = X_i^2\beta^2 + Z_i^2\gamma_i^2 + W_i^2 + \varepsilon_i^2 \end{cases} \text{ where } \begin{array}{ll} \varepsilon_i^1 \sim N(0, \sigma_{\varepsilon^1}^2 I_{n_i}) & \varepsilon_i^2 \sim N(0, \sigma_{\varepsilon^2}^2 I_{n_i}) \\ \gamma_i^1 \sim N(0, G^1) & \text{and } \gamma_i^2 \sim N(0, G^2) \\ W_i^1 \sim N(0, R_i^1) & W_i^2 \sim N(0, R_i^2) \end{array}$$

where X_i^k is a $n_i \times p^k$ design matrix, β^k is a p^k-vector of fixed effects, Z_i^k is a $n_i \times q^k$ design matrix which is usually a subset of X_i^k, γ_i^k is a q^k-vector of individual random effects with $q^k \le p^k$. W_i^k is a vector of realization of a first order auto-regressive process $w_i^k(t)$ with covariance given by $R_i^k(s,t) = \sigma_{w^k}^2 e^{\lambda^k |t-s|}$ and I_{n_i} is a $n_i \times n_i$ identity matrix.

To take into account correlation between both markers, one could use the following bivariate linear mixed model:

$$Y_i = X_i\beta + Z_i\gamma_i + W_i + \varepsilon_i \text{ with } \begin{cases} \varepsilon_i \sim N(0, \Sigma_i) \\ W_i \sim N(0, R_i) \\ \gamma_i \sim N(0, G) \end{cases}$$

where $X_i = \begin{bmatrix} X_i^1 & 0 \\ 0 & X_i^2 \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, $Z_i = \begin{bmatrix} Z_i^1 & 0 \\ 0 & Z_i^2 \end{bmatrix}$, $\gamma_i = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix}$ and $W_i = \begin{bmatrix} W_i^1 \\ W_i^2 \end{bmatrix}$ is a $2n_i$-vector of realization of a bivariate first order auto-regressive process $w_i(t) = \begin{bmatrix} w_i^1(t) \\ w_i^2(t) \end{bmatrix}$ and $\varepsilon_i = \begin{bmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \end{bmatrix}$ represents independent measurement errors.

The covariance matrix of measurement errors is defined by $\Sigma_i = \Sigma \otimes I_{n_i}$ and $\Sigma = \begin{bmatrix} \sigma_{\varepsilon^1}^2 & 0 \\ 0 & \sigma_{\varepsilon^2}^2 \end{bmatrix}$ (the symbol $\otimes$ represents the Kronecker product). The covariance function of the bivariate auto-regressive process $w_i(t) = \begin{bmatrix} w_i^1(t) \\ w_i^2(t) \end{bmatrix}$ is given by $R_i(s,t) = C \times e^{B|t-s|}$ with $C = \begin{bmatrix} \sigma_{w^1}^2 & \sigma_{w^1w^2} \\ \sigma_{w^1w^2} & \sigma_{w^2}^2 \end{bmatrix}$ is the process covariance matrix at $t = s$ and B is a 2×2 matrix such that (i) the eigenvalues of B have negative real parts, and (ii) C and $D = -(CB + B'C)$ are positive

definite symmetric [5]. The covariance matrix of random effects is the matrix $G = \begin{bmatrix} G^1 & G^{12} \\ G^{12} & G^2 \end{bmatrix}$. With the assumption that γ_i, W_i and ε_i are mutually independent, it is obvious that $\text{var}(Y_i) = V_i = Z_i G_i Z_i^T + R_i + \Sigma_i$.

3. Models using Proc MIXED of SAS software

3.1 Random effects

As described in the documentation [3], multivariate random effects models can be fitted using the statement random and an indicator variable for each marker to define Y_i^k, X_i^k and Z_i^k.

To add an independent error for each response variable in a multivariate random effect model, one must use the repeated statement with the option *GROUP(VAR)* where VAR is a binary variable indicating the response variable concerned (VAR=0 for Y^1 and VAR=1 for Y^2 into example). This option allows estimation of heterogeneous covariance structure, i.e. the variances of the measurement errors are different for each response variable.

An example of SAS code for a bivariate random effect model with random intercept and random slopes is:

```
Proc mixed data=BIV;
class CEN_PAT VAR;
model Y=VAR VAR*T;
random VAR VAR*T /type=UN sub=CEN_PAT;
repeated /type=VC grp=VAR sub=CEN_PAT;
run ;
```

where CEN_PAT is a single identification number of each patient and T is time. In this example, we have $p^1 = p^2 = q^1 = q^2 = 2$ and $X_{ij}^1 = X_{ij}^2 = Z_{ij}^1 = Z_{ij}^2 = \begin{bmatrix} 1 & t_{ij} \end{bmatrix}$ for the measurement j of the subject i. Note that the two markers are independent if $G = \begin{bmatrix} G^1 & 0 \\ 0 & G^2 \end{bmatrix}$.

3.2 First order auto-regressive process

In the repeated statement SAS provides the possibility to fit bivariate models using a Kronecker product notation [8]. For instance, in the bivariate case with 3 repeated measures, the option *type=UN@AR(1)* in the statement *repeated* assumes that the covariance matrix has

the following structure: $\begin{bmatrix} \sigma^2_{w^1} & \sigma_{w^1w^2} \\ \sigma_{w^1w^2} & \sigma^2_{w^2} \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$. Compared with the general bivariate auto-regressive process defined in the previous section, this structure has two important limitations. First, the covariance structure is a first order auto-regressive process for discrete data and assumes the measures are equally spaced for all subjects and for the two markers. In the univariate case, a continuous time AR(1) model, which allows non equally spaced measures, may be fitted using the structure *SP(POW)* but this structure is not available for multivariate models. The second limitation is that the SAS program allows to estimate only one correlation parameter (ρ) for the 'bivariate process' rather than a matrix B. Thus, using this formulation, one assumes that the intra-marker correlation is the same for the two markers, i.e. $Corr\left(w_i^1(s), w_i^1(t)\right) = Corr\left(w_i^2(s), w_i^2(t)\right) = \rho^{|t-s|}$. Moreover, one assumes that inter-marker correlation is proportional to the intra-marker correlation, i.e. $Corr\left(w_i^1(s), w_i^2(t)\right) = \frac{\sigma_{w^1w^2}}{\sigma_{w^1}\sigma_{w^2}}.\rho^{|t-s|}$. Both markers are independent if the covariance matrix has the form $\begin{bmatrix} \sigma^2_{w^1} & 0 \\ 0 & \sigma^2_{w^2} \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$.

To add an independent measurement error for both markers, one must use the option *LOCAL(EXP <effects>)* which produces exponential local effects, *<effects>=VAR* being still the indicator variable of response variable. These local effects have the form $\sigma^2_\varepsilon diag\left[\exp(U\delta)\right]$ where **U** is a full-rank design matrix. PROC MIXED constructs **U** in terms of 1s and –1s for a classification effect and estimates δ.

An example of SAS code to fit a bivariate first-order auto-regressive model is:

```
Proc mixed data=BIV;
class CEN_PAT VAR;
model Y=VAR VAR*T;
repeated VAR /type=UN@AR(1) local=exp(VAR) sub=CEN_PAT;
run ;
```

where T is the time as a continuous variable and VAR is the indicator variable.

The SAS output contains the following covariance parameters estimates: 'VAR UN(x,y)' which correspond to the matrix containing covariance parameter of the auto-regressive

process $\begin{bmatrix} \sigma^2_{w^1} & \sigma_{w^1w^2} \\ \sigma_{w^1w^2} & \sigma^2_{w^2} \end{bmatrix}$, 'EXP VAR' which is the local effect parameter (δ), 'Residual' that we noted r and a parameter called 'AR(1)'. From this output, the parameters of the model could be calculated as: $\sigma^2_{\varepsilon^1} = r \times e^{\delta}$, $\sigma^2_{\varepsilon^2} = r \times e^{-\delta}$ and $\rho = \frac{AR(1)}{r}$.

3.3 Incomplete data

Likelihood based inference used by Proc MIXED is valid whenever the mechanism of missing data is ignorable, that is MAR (Missing at Random), i.e. the availability of the measurement do not depend on the true value of the marker at the same time, and the parameters describing the non-response mechanism are distinct from the model parameters [9]. However, using an auto-regressive process, one must be careful when missing data occur. By default, a dropout mechanism is assumed to be responsible of missing data by MIXED procedure: all missing data are considered to occur after the last observed measurement. To take into account for intermittent missing data (one observation missing between two observed), a class variable must be used in the *repeated* statement indicating the order of observations within a subject. In the following example, the class variable is a copy of the variable time, named 'Tbis' :

```
Proc mixed data=BIV;
class CEN_PAT VAR Tbis;
model Y=VAR VAR*T;
repeated VAR Tbis /type=UN@AR(1) local=exp(VAR) sub=CEN_PAT;
run ;
```

When the measurements of the two markers never occur at the same time because of a design consideration, auto-regressive process can not be used unlike random effects model.

4. Application

4.1 The APROCO Cohort

The APROCO (ANRS-EP11) cohort is a prospective observational cohort ongoing in 47 clinical centres in France. A total of 1,281 HIV-1-infected patients were enrolled from May 1997 to June 1999 at the initiation of their first highly active antiretroviral therapy containing a protease inhibitor. Standardised clinical and biological data including CD4+ cell counts measurements and plasma HIV RNA quantification were collected at baseline (M0), one

month later (M1) and every 4 months (M4-M24) thereafter. In order to ensure sufficient available information, only a sub-sample of patients having both plasma HIV RNA and CD4+ cell counts measurements at M0 and at least two measurements thereafter were included in the analyses. The first measurement after baseline (at one month) was deleted to provide a data set with equally spaced measures. Follow-up data were included until the 24th month; thus patients had a maximum of 7 measures. The study population and evolution of virological response were described elsewhere [10]. Available information at each study time and description of the evolution of both markers were presented in table 1 and figure 1.

4.2 Modeling

To assure normality and homoskedasticity of residuals distribution, variable response was the change in value of marker at time t since the initial visit, i.e. $Y_i^1(t) = \log_{10} HIVRNA(t) - \log_{10} HIVRNA(0)$ and $Y_i^2(t) = CD_4(t) - CD_4(0)$.

Fixed effects included a change of slope intensity at time 4 months as suggested in figure 1. Note that we did not include intercept because $Y_i^1(0) = Y_i^2(0) = 0 \quad \forall i$.

We compared 4 models providing two forms of covariance structure (random effects or auto-regressive process) in two formulations (univariate or bivariate). Univariate and bivariate random effect models were compared using likelihood ratio test as both models were nested. The bivariate model had only four covariance parameters in addition. Comparison of random effects versus auto-regressive process were performed using AIC criteria [11]. A general model including random slopes and a bivariate first order auto-regressive process did not converge as reported in univariate cases by others (see [12] for example).

The model including two random slopes and a measurement error for each marker was:

$$\begin{cases} Y_i^1 = \beta_1^1(t_i \wedge \tau) + \beta_2^1(t_i - \tau)I_{t_i \geq \tau} + \gamma_{1i}^1(t_i \wedge \tau) + \gamma_{2i}^1(t_i - \tau)I_{t_i \geq \tau} + \varepsilon_i^1 \\ Y_i^2 = \beta_1^2(t_i \wedge \tau) + \beta_2^2(t_i - \tau)I_{t_i \geq \tau} + \gamma_{1i}^2(t_i \wedge \tau) + \gamma_{2i}^2(t_i - \tau)I_{t_i \geq \tau} + \varepsilon_i^2 \end{cases}$$

where β_1^k is the first slope before the time $\tau = 4\ months$, β_2^k is the second slope after the time τ and $t_i \wedge \tau$ represents the minimum between t_i and τ.

Moreover, $\begin{pmatrix} \gamma_{1i}^1 \\ \underline{\gamma_{2i}^1} \\ \gamma_{1i}^2 \\ \gamma_{2i}^2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \underline{0} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\gamma_{1i}^1} & \sigma_{\gamma_{1i}^1\gamma_{2i}^1} & \sigma_{\gamma_{1i}^1\gamma_{1i}^2} & \sigma_{\gamma_{1i}^1\gamma_{2i}^2} \\ \sigma_{\gamma_{1i}^1\gamma_{2i}^1} & \sigma^2_{\gamma_{2i}^1} & \sigma_{\gamma_{2i}^1\gamma_{1i}^2} & \sigma_{\gamma_{2i}^1\gamma_{2i}^2} \\ \sigma_{\gamma_{1i}^1\gamma_{1i}^2} & \sigma_{\gamma_{2i}^1\gamma_{1i}^2} & \sigma^2_{\gamma_{1i}^2} & \sigma_{\gamma_{1i}^2\gamma_{2i}^2} \\ \sigma_{\gamma_{1i}^1\gamma_{2i}^2} & \sigma_{\gamma_{2i}^1\gamma_{2i}^2} & \sigma_{\gamma_{1i}^2\gamma_{2i}^2} & \sigma^2_{\gamma_{2i}^2} \end{pmatrix} \right)$

The model including an auto-regressive process and a measurement error was:

$$\begin{cases} Y_i^1 = \beta_1^1 t_i \wedge \tau + \beta_2^1 (t_i - \tau) I_{t_i \geq \tau} + W_i^1 + \varepsilon_i^1 \\ Y_i^2 = \beta_1^2 t_i \wedge \tau + \beta_2^2 (t_i - \tau) I_{t_i \geq \tau} + W_i^2 + \varepsilon_i^2 \end{cases} \text{ where } \begin{bmatrix} W_i^1 \\ W_i^2 \end{bmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, R_i \right)$$

where $R_i = \begin{bmatrix} \sigma^2_{w^1} & \sigma_{w^1 w^2} \\ \sigma_{w^1 w^2} & \sigma^2_{w^2} \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & ... & \rho^7 \\ \rho & 1 & \rho & ... \\ ... & \rho & ... & \rho \\ \rho^7 & ... & \rho & 1 \end{bmatrix}$.

4.3 SAS programming

The initial data set had the following presentation :

CEN_PAT	CD4	RNA	T
1001	166	-3.02635	4
1001	147	-1.96563	8
1001	171	-1.42426	12
1001	355	-1.07208	16
1001	223	-3.38035	20
1001	52	-2.08382	24
1002	-14	-2.84515	4
1002	-123	-2.84515	8

With CEN_PAT being the patient number, CD4 the difference in CD4 cell count since baseline, RNA the difference in HIV RNA since baseline and T the date of measurement in months. The change in slope intensity at 4 months was computed using a data step:

```
Data file; set file;
if T<4 then do ; T1=T;T2=0; end ;
if T ge 4 then do; T1=CP;T2=T-4;
end ;
```

Then, the structure of input data was transformed to allow bivariate modeling. Mainly, it consists in the integration of CD4 and HIV RNA in the same vector (Y here) and an indicator variable (VAR here)

```
Data var0; set file;
VAR=0;Y=RNA;
keep CEN_PAT Y VAR T T1 T2;
```

```
Data var1;set file;
VAR=1;Y=CD4;
keep CEN_PAT Y VAR T T1 T2;

Data biv ;set var0 var1 ;
run ;
```

Thus, a bivariate random effect model was fitted using the code described below.

```
Proc mixed data=BIV CL;
class CEN_PAT VAR;
model Y=VAR*T1 VAR*T2/noint s;
random VAR*T1 VAR*T2/type=UN sub=CEN_PAT G GCORR;
repeated /type=VC grp=VAR sub=CEN_PAT;
run ;
```

The option "CL" requests confidence limits for the covariance parameter estimates. A Satterthwaite approximation is used to construct limits for all parameters that have a default lower boundary constraint of zero. In the statement model, the option "noint" was used to avoid the inclusion of intercepts and "s" to obtain solution for fixed effects.

In the same way, a bivariate model with an auto-regressive process and separate measurement errors was fitted using the following code:

```
Proc mixed data=BIV CL;
class CEN_PAT VAR T;
model Y=VAR*T1 VAR*T2 /noint s;
repeated VAR T/type=UN@AR(1) local=exp(VAR) sub=CEN_PAT;
run ;
```

4.4 Results

The bivariate random effects model was significantly better than two separate univariate random effects models (-25194 vs. -25307, likelihood ratio = 226 with 4 degrees of freedom, $p<10^{-4}$, table 2) showing a strong association between the two markers. The bivariate random effect model allows to estimate the correlation matrix between individual slopes for each marker. In this correlation matrix, every element was significantly ($p < 0.05$) different from 1 (table 3). Briefly, the highest correlations were between the slopes of the two markers at the same period: ($\rho\left(\beta_1^{CD4}, \beta_1^{HIVRNA}\right) = -0.41$ before 4 months and $\rho\left(\beta_2^{CD4}, \beta_2^{HIVRNA}\right) = -0.60$ after

4 months). These results were expected because of biological relation between the two markers. Moreover, the second slope of CD4 cell count was highly correlated to the first slope of the same marker $\rho\left(\beta_1^{CD4}, \beta_2^{CD4}\right) = 0.37$.

The bivariate model including a bivariate auto-regressive process was better than the bivariate random effects model despite the restrictive assumption that the two intra-marker correlations are equal (AIC 50386 vs. 50646).

Output obtained with the model including a first order auto-regressive process provide estimations of $\sigma^2_{w^1} = 1.54$, $\sigma^2_{w^2} = 195$ and $\sigma_{w^1 w^2} = -7.00$, significantly different from 0 (Wald test, $p<10^{-4}$). This last result underlines the relationship between the two markers. The parameter $\rho = \frac{3.11}{3.42} = 0.91$ is the correlation between two consecutive measures of CD4 cell count or HIV RNA. Variances of measurement error are calculated as: $\sigma^2_{\varepsilon^1} = 3.42\, e^{3.11} = 77.00$ and $\sigma^2_{\varepsilon^2} = 3.42\, e^{-3.11} = 0.15$.

Thus, the relationship between the two makers was underlined by the correlation between the markers at each period and the improvement of likelihood of the bivariate model compared to two univariate models. Bivariate random effect model offers a direct interpretation of the relationship between the markers without assumption on the dependence of one marker in relation to the other.

5. Conclusion

Bivariate models are useful for longitudinal data in biomedical research and can be computed using standard statistical package like the SAS system. Moreover, the efficiency of the procedure MIXED, which allows quick convergence, should be underlined. However, there are some limitations inherent in the identical intra-marker correlations or in the assumption of constant period between two measurements for the first order auto-regressive covariance structure implemented in the SAS system. Finally, although the number of parameters would dramatically increase, particularly in the case of multivariate random effect model, bivariate models are easily extendable to multivariate models with more than two dependent variables.

References

[1] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics 38 (1982) 963-74.

[2] S-PLUS Programmer's Manual, (Statistical Sciences Inc., Seattle, WA, 1991).

[3] R.C. Littell, G.A. Milliken, W.W. Stroup, R.D. Wolfinger, SAS System for Mixed Models, (SAS Institute, Cary, NC, 1996).

[4] A. Shah, N. Laird, D. Schoenfeld, A random-effects model for multiple characteristics with possibly missing data, JASA 92 (1997) 775-9.

[5] J.P. Sy, J.M. Taylor, W.G. Cumberland, A stochastic model for the analysis of bivariate longitudinal AIDS data, Biometrics 53 (1997) 542-55.

[6] SAS/IML Software : Usage and Reference. Version 6., (SAS Institute Inc., Carry, NC, 1990).

[7] M.J. Lindstrom, D.M. Bates, Newton-Raphson and EM Algorithm for linear mixed-effects models for repeated-measures data, JASA 83 (1988) 1014-22.

[8] A.T. Galecki, General class of covariance structures for two or more repeated factors in longitudinal data analysis, Commun. Statist.-Theory Meth. 23 (1994) 3105-19.

[9] G. Verbeke, G. Molenberghs, Linear mixed models in practice. A SAS-oriented approach, (Springer, New York, 1997).

[10] V. Le Moing, G. Chêne, J.-P. Moatti et al., Predictors of early immunological and virological response in a cohort of HIV-infected patients initiating protease inhibitors: role of adherence to therapy, J Acquir Immune Defic Syndr (2001) (in press).

[11] H. Akaike, A new look at the statistical model identification, IEEE Trans automat contr AC-10 (1974) 716-23.

[12] E. Lesaffre, M. Asefa, G. Verbeke, Assessing the goodness-of-fit of the Laird and Ware model--an example: the Jimma Infant Survival Differential Longitudinal Study, Stat Med 18 (1999) 835-54.

Table 1. Measures of CD4 cell count and HIV RNA during follow-up. APROCO Study (N=988).

Month	Change in CD4 cell count / mm^3 from baseline			Change in log_{10} copies/ml HIV RNA from baseline		
	N	Mean	SD	N	Mean	SD
4	988	97	130	988	-1.95	1.20
8	935	126	147	919	-2.01	1.27
12	901	153	169	894	-2.04	1.34
16	823	176	180	813	-2.03	1.35
20	708	192	190	703	-2.00	1.37
24	534	201	196	530	-1.93	1.37

Table 2. Likelihood of models according to the type of covariance matrix. APROCO Study (N=988).

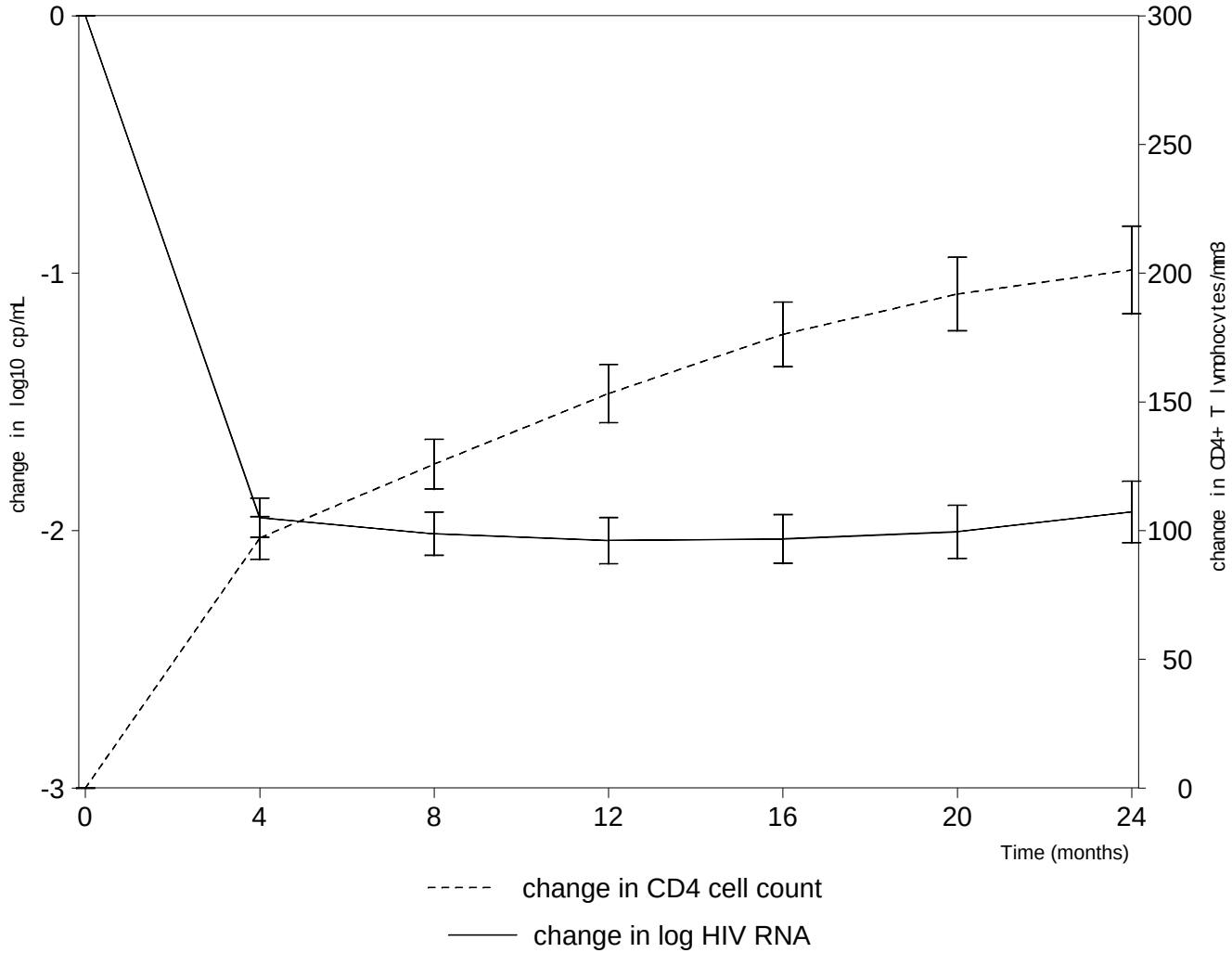
	Log Likelihood	No. of parameters	AIC
Univariate model with two random slopes	-25307	12	50638
Bivariate model with two random slopes	-25194	16	50420
Univariate model with AR(1)	-25313	10	50646
Bivariate model with AR(1)	-25183	10	50386

$AIC = (-2\log likelihood) + 2\,(No.\ of\ \ parameters)$ AR(1) : First order auto-regressive process

Table 3. Estimated correlation matrix of the bivariate model including two random slopes. APROCO Study (N=988).

	First slope of HIV RNA	Second slope of HIV RNA	First slope of CD4+	Second slope of CD4+
First slope of HIV RNA	1			
Second slope of HIV RNA	-0.10	1		
First slope of CD4+	-0.41	0.13	1	
Second slope of CD4+	-0.16	-0.60	0.37	1

Figure 1. Mean change in observed HIV RNA and CD4+ cell count (95% confidence interval) after initiation of an antiretroviral treatment containing a protease inhibitor. APROCO study (N=988).



4 Censure de la charge virale

4.1 Travaux antérieurs

Sont présentées dans la section suivante, d'une part, une méthode de prise en compte des données longitudinales gaussiennes censurées à gauche développée au cours de mon Diplôme d'Etudes Approfondies d'épidémiologie et publiée dans Biostatistics en 2000 [153]. D'autre part, la méthode a été appliquée pour analyser le rôle pronostique de l'évolution de la charge virale après l'initiation d'un traitement antirétroviral hautement actif. On a étudié l'effet des pentes individuelles [155] ou des valeurs estimées au cours du temps [156].

4.1.1 Développement d'une méthode d'estimation pour des données censurées à gauche

Avec Hughes [152], nous fument dans les premiers à publier une méthode de prise en compte de la censure à gauche de la variable dépendante dans le cadre des modèles mixtes [153]. Le principe est de considérer que l'ensemble de la variable réponse $Y_i = \begin{bmatrix} Y_i^o \\ Y_i^c \end{bmatrix}$, comprenant des données observées Y_i^o et des données censurées Y_i^c, suit une loi normale. Ainsi, la contribution à la vraisemblance du vecteur des mesures observées Y_i^o est la densité d'une loi multivariée normale et la contribution du vecteur des mesures censurée Y_i^c est la probabilité conditionnelle à Y_i^o que ces mesures soient inférieures au seuil de détection c_i. Ainsi, la vraisemblance d'un modèle défini avec le vecteur de paramètres θ pour l'ensemble de N sujets peut s'écrire :

$$L(\theta) = \prod_{i=1}^{N} f_{Y_i^o}\left(y_i^o \middle| \theta\right) \Pr\left(Y_i^c < c_i \middle| Y_i^o = y_i^o, \theta\right)$$

Nous avons clairement montré que l'estimation des paramètres fixes et surtout des paramètres de covariance sont moins biaisés avec cette méthode comparativement à l'imputation simple de la valeur du seuil de détectabilité. De plus, pour des structures de covariance un peu évoluée (de type non structuré avec un intercept et une pente aléatoire), notre méthode conduit à des estimations moins biaisées que l'algorithme MCEM proposé par Hughes [152]. Chez des patients traités par bithérapie antirétrovirale dans l'essai thérapeutique ALBI ANRS 070, l'estimation de la réponse virologique initiale moyenne variait de 1 à 2 $\log_{10}$ copies/ml entre la

méthode d'imputation simple et notre méthode. Cette variation était cliniquement très significative.

4.1.2 Applications

Cette méthode a été appliquée pour répondre à des questions épidémiologiques précises notamment sur l'effet de l'évolution de la charge virale plasmatique sur l'évolution clinique. La première de ces applications étudiait l'impact de l'évolution virologique dans l'année suivant la mise en place d'un traitement antirétroviral sur la survenue ultérieure d'événements cliniques ou du décès [155]. Pour cela, un modèle mixte prenant en compte la censure de la charge virale a été utilisé pour obtenir les estimations des pentes individuelles de la charge virale suivant l'initiation du traitement antirétroviral. La pente de la charge virale estimée dans le premier mois était hautement prédictive de la survenue d'événement clinique après un an. De plus, cette réponse initiale dans le premier mois était très corrélée à la réponse virologique au cours de la première année de traitement.

Par la suite, plutôt que d'utiliser les prédictions des pentes individuelles $\left(\hat{\beta}+\hat{\gamma}_i\right)$, les estimations de charge virale $\left(\hat{Y}_i\right)$ ont été directement modélisées en tant que variable dépendante du temps dans un modèle de Cox [156]. Dans notre application, la prise en compte des données de charge virale indétectable a permis de modéliser l'effet de la charge virale sur une échelle continue. Le gain de puissance par rapport à une analyse en classe était significatif. Nous concluons que ce type d'analyse peut être utile notamment lorsque l'analyse en classe de la charge virale conduit à une absence d'effet significatif [156]. Dans cette étude, deux modèles mixtes séparés ont été utilisés pour modéliser respectivement le logarithme en base 10 de la charge virale et la racine carrée des CD4+. Cette dernière transformation a été retenue car elle permettait aux résidus de satisfaire les conditions d'application du modèle et elle conduisait à la vraisemblance maximum comparativement aux autres transformations candidates (voir annexe 1).

4.2 Motivation

Etant donné le gain d'information apporté par la modélisation bivariée et étant donné l'impact de la prise en compte de la censure de la charge virale, cette dernière méthode a été étendue pour estimer les paramètres d'un modèle linéaire mixte bivarié. Ceci a donné lieu à un article

original sous presse dans *Journal of Biopharmaceutical Statistics* et présenté dans la section suivante.

L'application de ce modèle pour la charge virale et les CD4+ dans la cohorte APROCO [74] a donné des résultats très intéressants concernant l'évolution et la corrélation des deux marqueurs. En effet, nous avons objectivé la corrélation continue entre la réponse virologique (décroissance de la charge virale) et la réponse immunologique (croissance des CD4+) tout au long du suivi. De plus, nous avons décrit une décroissance moyenne continue de la charge virale la première année.

Nous avons également étudié la valeur pronostique des estimations obtenues par le modèle bivarié prenant en compte la censure de la charge virale sur la survenue d'un événement clinique dans la Cohorte Aquitaine. Chaque marqueur était pris en compte dans un modèle de Cox en tant que variable dépendante du temps. Le gain d'information obtenu via la modélisation bivariée dans la première étape avait peu d'impact sur les estimations du risque relatif de progression clinique comparativement à l'utilisation des estimations provenant de deux modèles univariés. Nous avons présenté ces résultats lors d'un atelier sur les cohortes observationnelles dans le VIH [157].

4.3 Modélisation bivariée prenant en compte la censure de la charge virale

Title: Bivariate longitudinal model for the analysis of the evolution of HIV RNA and CD4 cell count in HIV infection taking into account left censoring of HIV RNA measures.

Authors: Rodolphe Thiébaut [1], Hélène Jacqmin-Gadda [1], Catherine Leport [2], Christine Katlama [3], Dominique Costagliola [4], Vincent Le Moing [1,2], Philippe Morlat [5], Geneviève Chêne [1] and the APROCO study group.

Affiliations:

(1) INSERM U330, Institut de Santé Publique d'Epidémiologie et de Développement (ISPED), Bordeaux, France

(2) Service des Maladies Infectieuses et Tropicales, Hôpital Bichat-Claude Bernard, Paris, France

(3) Département des Maladies Infectieuses et Tropicales, Hôpital Salpêtrière, Paris, France

(4) INSERM SC4, Centre Coopérateur de données épidémiologiques sur l'immunodéficience humaine, Paris, France

(5) Service de Médecine Interne et de Maladies Infectieuses, Hôpital Saint-André, Bordeaux, France

Financial support: The APROCO study is sponsored and financially supported by the Agence Nationale de Recherches sur le Sida (ANRS, Action Coordonnée n°7). Their financial supports are from associated pharmaceutical companies: Abbot, Boerhinger-Ingelheim, Roche, Bristol Myers Squib, Merck Dohm Chibret, Glaxo-Smithkline. Rodolphe Thiébaut is supported by a grant from Ensemble Contre le SIDA.

Title: Bivariate longitudinal model for the analysis of the evolution of HIV RNA and CD4 cell count in HIV infection taking into account left censoring of HIV RNA measures.

ABSTRACT

We present a bivariate linear mixed model taking into account censored measures of the response variable due to lower quantification limit of the assays. It allows to estimate the correlation between the two response variables and take into account this correlation for the estimation of other model parameters. This model was applied in a large cohort study (APROCO Cohort) to study the evolution under antiretroviral treatment of the two major biomarkers of the progression of Human Immunodeficiency Virus (HIV) infection: plasma HIV RNA and CD4+ T lymphocytes cell count. In a sample of 929 patients who started an highly active antiretroviral therapy, we illustrate the superiority in terms of likelihood of a bivariate model compared to two univariate models and the impact of taking into account the left-censoring of HIV-RNA. Moreover, interpretation of the model parameters allow to confirm the correlation between these two markers throughout the whole follow-up and the continuous decrease of plasma HIV RNA on average. Despite some limitations (distribution assumption, ignorance of missingness process), such model appeared to be very useful to correctly describe the current evolution of important biomarkers in HIV infection.

Key words: bivariate linear mixed model, CD4+ cell count, censoring, HIV infection, HIV RNA

1. INTRODUCTION

Linear mixed models (1) have been increasingly used for the analysis of repeated measures of biomarkers as they allow to take into account all available information. HIV infection is one example where longitudinal analyses are frequent (2). Since the availability of highly active antiretroviral therapy (HAART) for the antiretroviral treatment of Human Immunodeficiency Virus (HIV)-infected patients, opportunistic diseases and mortality have dramatically decreased. Thus, patients case management (3, 4) and clinical trial outcomes (5) are mainly based on biomarkers such as plasma HIV RNA and CD4+ lymphocytes cell count (CD4+). These two biomarkers are highly correlated because CD4+ constitute the target of the HIV whose replication is partly measured by plasma HIV RNA. Many studies reported the association between these markers, often explaining the evolution of one marker according to the other (6-8). The joint modelling of the evolution of the two markers using a bivariate model allows to study the correlation between markers without assumption on the timing of this relation (2, 9, 10).

However, the study of the evolution of plasma HIV RNA is complicated by the lack of sensitivity of the assays used to measure the concentration in plasma. Some measures are undetectable, i.e. below the threshold of the assay and so not quantifiable. The left-censoring of these values must be taken into account to avoid bias yielding an overestimation of the plasma HIV RNA quantity (11-13). We have already proposed a method to estimate a linear mixed model for longitudinal data taking left censoring into account (13). A simulation study and the analysis of a real data set have validated the method and shown that estimates obtained by imputing the threshold are biased.

In this paper, we propose a method to fit a bivariate linear mixed model taking into account censored repeated measures. This is an extension for bivariate data of the previously published work (13). This model provides estimations of the evolution of the markers, of the

correlations between the markers all over the patients' follow-up and allows to take into account undetectable measurements of HIV RNA. We illustrate the usefulness of this kind of model in a large cohort of patients who started a combination of antiretrovirals including a protease inhibitor: The APROCO study (8). In section 2 and 3, we present the bivariate linear mixed model and the method of estimation. The application is presented in section 4. In section 5, we discuss the interest of such model.

2. THE BIVARIATE LINEAR MIXED MODEL

We present the extension of a usual univariate linear mixed model [88] to a bivariate model for two markers. Let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$, the response vector for the subject i=1,…N, with Y_i^k the n_i^k-vector of measurements of the marker k (k=1, 2). The number of measurements may be different for each marker and each subject. A bivariate linear mixed model could be written as follow :

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i \text{ with } \begin{cases} \varepsilon_i \sim N(0,\Sigma_i) \\ \gamma_i \sim N(0,G) \end{cases}$$

and $\beta = \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix}$, $X_i = \begin{bmatrix} X_i^1 & 0 \\ 0 & X_i^2 \end{bmatrix}$, $\gamma_i = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix}$, $Z_i = \begin{bmatrix} Z_i^1 & 0 \\ 0 & Z_i^2 \end{bmatrix}$, $\varepsilon_i = \begin{bmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \end{bmatrix}$.

X_i^k is a $n_i \times p^k$ design matrix of explanatory variables for marker k, β^k is a p^k-vector of fixed effects, Z_i^k is a $n_i \times q^k$ design matrix which is usually a subset of X_i^k, γ_i^k is a q^k-vector of individual random effects with $q^k \leq p^k$. The covariance matrix of measurement errors is a diagonal matrix, denoted by Σ_i and containing the two elements $\sigma^2_{\varepsilon^1}$ and $\sigma^2_{\varepsilon^2}$ representing the measurement error of each marker. The covariance matrix of random effects

is the matrix $G = \begin{bmatrix} G^1 & G^{12} \\ G^{12} & G^2 \end{bmatrix}$. This matrix G is divided in three parts: (i) G^1 the covariance matrix including variances and covariances of random effects for the first marker (ii) G^2 the covariance matrix including variances and covariances of random effects for the second marker and (iii) G^{12} the matrix of covariances between random effects of each marker. This is through this sub-matrix that correlation between the two markers is taken into account. With the assumption that γ_i and ε_i are mutually independent, we obtain $\text{var}(Y_i) = V_i = Z_i G Z_i^T + \Sigma_i$.

3. THE LIKELIHOOD WITH CENSORED MEASURES

For a subject i, let Y_i^o the n_i^O-vector of observed outcomes, Y_i^c the n_i^c-vector of censored outcomes and s_i the n_i^c-vector of measurement thresholds. Thus, the threshold could be different at each measurement. This is useful as the threshold could vary because of the improvement of the assay, into example. After reordering, Y_i, X_i and V_i can be partitioned as: $Y_i = \begin{bmatrix} Y_i^o \\ Y_i^c \end{bmatrix}$, $X_i = \begin{bmatrix} X_i^o \\ X_i^c \end{bmatrix}$ and $V_i = \begin{bmatrix} V_i^o & V_i^{co^t} \\ V_i^{co} & V_i^c \end{bmatrix}$. The likelihood function of the vector θ of the parameters to be estimated is:

$$L(\theta) = \prod_{i=1}^{N} f_{Y_i^o}\left(y_i^o \middle| \theta\right) \Pr\left(Y_i^c < c_i \middle| Y_i^o = y_i^o, \theta\right)$$

$$= \prod_{i=1}^{N} f_{Y_i^o}\left(y_i^o | \theta\right) \int_{H_1} \int_{H_2} \ldots \int_{H_d} f_{Y_i^c | Y_i^o}(u)\, du$$

where $u = \left(u_1, u_2, \ldots, u_{n_i^c}\right)^t$, $H_d =]-\infty, s_{id}]$, $d = 1, \ldots, n_i^c$ and $f_{Y_i^o}$ is the multivariate normal distribution of the vector Y_i^o.

Using the properties of the multivariate normal distribution, we find that the conditional distribution of Y_i^c given Y_i^o is gaussian with expectation $\mu_i^{c/o}$ and covariance $V_i^{c/o}$:

$$\mu_i^{c/o} = X_i^c \beta + V_i^{co} V_i^{o\,-1} \left[Y_i^o - \mu_i^o \right]$$
$$V_i^{c/o} = V_i^c - V_i^{co} V_i^{o\,-1} V_i^{co\,t}$$

For the computation, the likelihood was re-parameterised in terms of the square root for $\sigma^2_{\varepsilon^1}$ and $\sigma^2_{\varepsilon^2}$ and the Cholesky factorisation for the covariance matrix ($G = U^T U$ where U is a upper triangular matrix) to impose positivity constraints. The optimisation was performed using a Marquardt algorithm (14). The multiple integrals were computed using a subregion adaptative algorithm developed by Genz (15). Computational details were presented previously for a univariate mixed model (13).

4. APPLICATION

4.1 Data: the APROCO study

The APROCO (ANRS-EP11) cohort is a prospective observational cohort ongoing in 47 clinical centers in France. A total of 1,281 HIV-1-infected patients were enrolled from May 1997 to June 1999 at the initiation of their antiretroviral treatment containing a protease inhibitor. Standardised clinical and biological data including CD4+ cell counts measurements and plasma HIV RNA quantification were collected at baseline (M0), and after approximately 1 month (M1) and every 4 months thereafter. The exact time of measurements in days was used in the analyses. In order to ensure sufficient available information, only the sub-sample of patients having both plasma HIV RNA and CD4+ cell counts measurements at M0 and at least two measurements thereafter was included in the analyses (N=929). Analyses focused on the difference between the marker value at a time t and the baseline value at the initiation of HAART ($CD4_t$-$CD4_0$) and ($\log_{10}$ HIV RNA_t- $\log_{10}$ HIV RNA_0). Thus, patients with an undetectable measure of plasma HIV RNA at baseline were excluded to avoid interval

censoring of the difference of plasma HIV RNA value between baseline and other measures during follow-up (N=59). Follow-up was restricted to the 12th month to limit the impact of patients lost to follow up and potential virological rebound. Only 80 (8.6%) patients did not attend the M12 visit. The 929 patients included in the present analysis are accounting for a total of 4513 measures of CD4+ and 4491 measures of plasma HIV RNA ; only 122 (13%) and 142 (15%) patients did not have the maximum of 5 measures until M12 for CD4 cell count and HIV RNA, respectively. CD4+ cell counts were prospectively measured by standardised flow cytometry. All plasma HIV RNA levels were prospectively measured by the assay routinely available in each center with lower limits of detection varying from 1.3 to 2.7 $\log_{10}$ copies of HIV-1 RNA/ml. An overall of 2094 (47%) of HIV RNA measures were left-censored, that is 42% at M1, 66% at M4, 65% at M8 and 62% at M12. Among censored measures, the proportions below the 2.3 $\log_{10}$ copies/ml threshold were 3% at M1, 12% at M4, 31% at M8 and 43% at M12 which could be associated with an increase of the sensitivity of the assays. The proportions of patients with HIV RNA below 2.7 $\log_{10}$ copies/ml were 2% at M0, 56% at M1, 79% at M4, 77% at M8 and 76% at M12. On 731 patients with HIV RNA below 2.7 $\log_{10}$ copies/ml at M4, 15% had a measure above 2.7 $\log_{10}$ copies/ml at M12.

4.2 Model used in the application

Figure 1 displays observed mean change in CD4+ and HIV RNA at the 4 visits replacing censored measures of HIV RNA by half of the quantification limit. Plasma HIV RNA decreased dramatically during the first month (-1.61 $\log_{10}$ copies/ml, 95% confidence interval [CI] -1.67; -1.55) and tended to decrease slowly thereafter (table 1, figure 1). The increase of CD4+ at one month (+ 72 cells/mm^3 at one month, 95% CI +65; +79) was less pronounced than the fall of plasma HIV RNA during the same period (in term of relative variation: +24% vs. -39%, respectively). After one month, CD4+ continued to increase regularly (table 1,

figure 1). This crude description of markers evolution encouraged a piecewise linear formulation of the model using two slopes :

$$\begin{cases} Y_{ij}^1 = \beta_1^1\left(t_{ij} \wedge \tau\right) + \beta_2^1\left(t_{ij} - \tau\right) I_{t_{ij} \geq \tau} + \gamma_{1i}^1\left(t_{ij} \wedge \tau\right) + \gamma_{2i}^1\left(t_{ij} - \tau\right) I_{t_{ij} \geq \tau} + \varepsilon_{ij}^1 \\ Y_{ij}^2 = \beta_1^2\left(t_{ij} \wedge \tau\right) + \beta_2^2\left(t_{ij} - \tau\right) I_{t_{ij} \geq \tau} + \gamma_{1i}^2\left(t_{ij} \wedge \tau\right) + \gamma_{2i}^2\left(t_{ij} - \tau\right) I_{t_{ij} \geq \tau} + \varepsilon_{ij}^2 \end{cases}$$

with

$$\begin{pmatrix} \gamma_{1i}^1 \\ \underline{\gamma_{2i}^1} \\ \gamma_{1i}^2 \\ \gamma_{2i}^2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \underline{0} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\gamma_{1i}^1} & \sigma_{\gamma_{1i}^1 \gamma_{2i}^1} & \sigma_{\gamma_{1i}^1 \gamma_{1i}^2} & \sigma_{\gamma_{1i}^1 \gamma_{2i}^2} \\ \sigma_{\gamma_{1i}^1 \gamma_{2i}^1} & \sigma^2_{\gamma_{2i}^1} & \sigma_{\gamma_{2i}^1 \gamma_{1i}^2} & \sigma_{\gamma_{2i}^1 \gamma_{2i}^2} \\ \sigma_{\gamma_{1i}^1 \gamma_{1i}^2} & \sigma_{\gamma_{2i}^1 \gamma_{1i}^2} & \sigma^2_{\gamma_{1i}^2} & \sigma_{\gamma_{1i}^2 \gamma_{2i}^2} \\ \sigma_{\gamma_{1i}^1 \gamma_{2i}^2} & \sigma_{\gamma_{2i}^1 \gamma_{2i}^2} & \sigma_{\gamma_{1i}^2 \gamma_{2i}^2} & \sigma^2_{\gamma_{2i}^2} \end{pmatrix} \right) \text{ and } \varepsilon_i^k \sim N\left(0, \sigma^2_{\varepsilon^k} I_{n_i^k}\right)$$

Fixed effects, representing the average among the study population, included two slopes for each marker. The first one (β_1^k) represented the short term response (before the time τ) and the second one (β_2^k) the long term response after time τ . The time when the slopes changed was approximated separately for each marker, using a profile of likelihood: the model was fitted for different values of the change point and the time leading to the best likelihood was kept. The best univariate model for the evolution of CD4+ included a change of the intensity of the slope at 40 days. For plasma HIV RNA, this time was at 37 days. For the sake of simplicity, because the likelihood of the models did not change a lot near the first month for the two markers and because these dates highly depends on schedule of measurement (which explains that the two dates were close) a same date of 40 days was chosen for the two markers in the bivariate model. The model did not include intercepts because differences were equal to 0 at baseline. The covariance structure allowed to take into account correlations of repeated measures in a same patient within and between each marker. Random effects consisted in two individual slopes (γ_{1i}^k before τ and γ_{2i}^k after τ) for each marker. Covariances of the individual slopes between and within markers were estimated without a priori hypotheses

(unstructured covariance matrix). Normality and homoskedasticity of residuals were graphically checked for CD4+ (data not shown).

4.3 Results

Results of the bivariate linear mixed model are shown in table 2 and figure 1 (estimated curves). The estimated short term slope (before 40 days) of plasma HIV RNA was -2.03 $\log_{10}$ copies/ml/month (95% CI -2.11; -1.95) and, for the long term slope, -0.58 $\log_{10}$ copies/ml/year (95% CI -0.78;-0.37). For CD4+, these estimations were +66 cells/mm^3/month (95% CI +60; +72) and +78 cells/mm^3/year (95% CI +66; +90), respectively (table 2). The likelihood of the bivariate model was -8899, and thus much greater than the sum of the likelihood of the two univariate models with the same change point of 40 days : -8937 ($p<10^{-4}$ for a chi-squared distribution of 4 degrees of freedom) underlining the association of the parameters between the two markers. Moreover, the estimated standard error of the second slope of HIV RNA (0.10) is smaller than that estimated with a univariate model (0.16). This underlines that information provided by CD4 data in the bivariate model contributes to the estimation of the long term evolution of HIV RNA. However, the second slope was significantly negative in both models. Table 3 shows the estimation of the correlation matrix for the random effects and their standard error obtained by the Delta method. The subject-specific slopes of plasma HIV RNA and CD4+ were negatively correlated in the first period (r = -0.33, standard error [SE] = 0.054) and in the second period (r = -0.37, SE = 0.080). So, all over follow-up, the more plasma HIV RNA decreased, the more CD4+ increased. The other correlation were weak, in particular the correlation within each marker (r = 0.13, SE = 0.074 for HIV RNA and r = -0.047, SE = 0.051 for CD4+).

When left-censoring of plasma HIV RNA was not taken into account and undetectable measures were replaced by the log of half of the quantification limit (i.e. 2.4 $\log_{10}$ copies/ml

for a measure <2.7 $\log_{10}$ copies/ml or 500 copies/ml), the results differed mainly on the intensity of the second slope (table 2). The estimated plasma HIV RNA short term slope was -1.61 $\log_{10}$ copies/ml/month (95% CI -1.66; -1.56) and the estimated long term slope was +0.035 $\log_{10}$ copies/ml/year (95% CI -0.04;+0.11) which was not significantly different from 0, i.e. stable (table 2). The impact of the left-censoring is also illustrated in figure 1 when comparing observed and estimated change in log plasma HIV RNA. Moreover, the estimated correlation between the two slopes of plasma HIV RNA was also modified: r = 0.047, SE = 0.065 for the crude model and r = 0.13, SE 0.074 when left-censoring was handled.

7. DISCUSSION

We have presented a bivariate linear mixed model taking into account censored measures. It was applied in a large cohort study to analyse the evolution of the two major markers of the progression of HIV infection. To our knowledge, a bivariate random effects model of CD4+ and plasma HIV RNA was only reported once by Boscardin et al. (2). Their model was applied in a population of patients included in a clinical trial comparing early versus late single therapy by zidovudine (non HAART regimen). Thus, CD4+ and plasma HIV RNA evolution were modelled using only one slope and without taking into account censored value of HIV RNA. In our study, we have modelled the two phases response of the marker after the initiation of a HAART regimen which allowed to study the correlation between slopes of two different periods (before and after 40 days) without hypothesis on which marker influence the other. Moreover, our analyses take into account censored measures of HIV RNA. Like in other reports of univariate analysis of HIV RNA evolution (11-13), we underlined the importance to take into account left-censoring to avoid a biased estimation of plasma HIV RNA slopes. The discrepancies between observed (imputing half of the threshold for censored values) and estimated evolution of plasma HIV RNA in figure 1 illustrate this issue.

The interpretation of covariance parameters underlines the usefulness of bivariate random effects model. In fact, the evolution of the two markers was divided in two periods (a short term and a long term response) according to previous reports on the dynamics of these markers (16, 17). Thus, we were interested in the correlation between individual slopes within and between the two markers through the two periods. The correlation between plasma HIV RNA and CD4 cell count was confirmed through all the follow-up instead of a determined date (7, 18). This result reinforces the importance of sustaining the virologic response as it is usually recommended (19). However, the estimations must be interpreted carefully in regards to model assumptions. The model used in the application did the hypothesis of a piecewise linear evolution through two slopes and thus did not account for viral rebound compared to more flexible approaches (20). That is why we have restricted the follow up to 12 months after initiation of HAART leading to few rebounds and few withdrawals. The continued estimated decrease of HIV RNA could be partly due to the improvement of the method to quantify plasma viral load. Although the left-censoring was taken into account, the assays with lower limit of quantification brought more information. But, on the other hand, a simple description of the proportion of censored or observed measures of HIV RNA below 2.7 log10 copies/ml showed a decrease of viral load from 1 months to 4 months which contributes to the estimation of the second slope.

We have chosen to take into account the correlation between markers using a bivariate random effects model with an unstructured covariance matrix because we were interested in the interpretation of the correlation between individual slopes. Thus, this model includes many covariance parameters to be estimated. In the model used in the application, the number of parameters was 16 (4 fixed effects and 12 covariance parameters). This increases the

computation time and could conduct to convergence problem. Other approaches have been proposed to take into account correlation between markers with less parameters and more flexibility. A bivariate stochastic process (21) including 4 parameters for the stochastic process or a non-parametric approach (9) were proposed and showed better predictions than two univariate models. However, estimated parameters of these models are difficult to interpret compared to our model. Moreover, univariate models including only a brownian motion (which is a particular case of Integrated Ornstein Uhlenbeck process (22)) for the covariance structure did not fit the data as well as random effects models in our application (Likelihood -5667 vs. -5585 for CD4 cell count and -3466 vs. -3352 for HIV RNA).

A limitation of maximum likelihood approaches (11-13) is that they make stringent gaussian distribution assumptions. Working on the change of the marker value at a time t compared to the baseline value allow to respect homoskedasticity of residuals for CD4 cell count without using a transformation like square root which does not allow direct clinical interpretation (22). The distribution of the HIV RNA residuals and of random effects are difficult to check because of censoring. For residuals, several authors have suggested to compare Kaplan-Meier estimates of the distribution function with those of a normal distribution $N\left(0,\sigma_{\varepsilon^1}^2\right)$ (13, 23) but, in the context of longitudinal data, subject specific residuals computed using empirical Bayes estimates do not have a variance equal to $\sigma_{\varepsilon^1}^2$. For random effects, even if a graphical assessment of the normality of their distribution has been already proposed with a left-censored outcome (12), Empirical Bayes estimates are known to be very dependent on their assumed prior distribution (24).

Another limit is that our model did not take into account informative drop-out. In fact, inferences of linear mixed models are valid if the missingness mechanism is ignorable in the

likelihood (25) which implies that data are missing at random, i.e. MAR (26), that is that missingness mechanism could depend on previously observed measures but does not depend on the current value of the marker. Otherwise, the missingness mechanism could be modelled using selection model (27) or pattern mixture model (28). In the example, the follow-up was restricted to 12 months, limiting the proportion of patients lost to follow-up.

In summary, we have presented a bivariate mixed model taking into account repeated censored measures. Like most of statistical methods to deal with missing data, our approach relies on strong distribution assumption which are difficult to check. However, in respect of these conditions, this model yields better estimations than two univariate models. Moreover, the correlation matrix between random effects could be very informative as illustrated in the application of the present paper.

REFERENCES

1. Laird NM, Ware JH: "Random-effects models for longitudinal data" Biometrics, 38(4), 963-74, 1982.

2. Boscardin WJ, Taylor JM, Law N: "Longitudinal models for AIDS marker data" Stat Methods Med Res, 7(1), 13-27, 1998.

3. Carpenter CCJ, Cooper DA, Fischl MA, Gatell JM, Gazzard BG, Hammer SM, et al.: "Antiretroviral therapy in adults - Updated recommendations of the International AIDS Society-USA Panel" JAMA, 283(3), 381-90, 2000.

4. Delfraissy JF. Prise en charge thérapeutique des personnes infectées par le VIH. Recommandations du groupe d'experts. Paris: Médecine Sciences Flammarion; 1999.

5. Albert JM, Ioannidis JPA, Reichelderfer P, Conway B, Coombs RW, Crane L, et al.: "Statistical issues for HIV surrogate endpoints: Point/counterpoint" Stat Med, 17(21), 2435-62, 1998.

6. Renaud M, Katlama C, Mallet A, Calvez V, Carcelain G, Tubiana R, et al.: "Determinants of paradoxical CD4 cell reconstitution after protease inhibitor-containing antiretroviral regimen" AIDS, 13(6), 669-76, 1999.

7. Staszewski S, Miller V, Sabin C, Schlecht C, Gute P, Stamm S, et al.: "Determinants of sustainable CD4 lymphocyte count increases in response to antiretroviral therapy" AIDS, 13(8), 951-6, 1999.

8. Le Moing V, Thiébaut R, Chêne G, Leport C, Moatti JP, Michelet C, et al.: "Predictors of long-term increase of CD4+ cell count in human immunodeficiency virus-infected patients initiating a protease inhibitor-containing regimen" J Infect Dis, 185, 471-80, 2002.

9. Brown ER, MaWhinney S, Jones RH, Kafadar K, Young B: "Improving the fit of bivariate smoothing splines when estimating longitudinal immunological and virological

markers in HIV patients with individual antiretroviral treatment strategies" Stat Med, 20(16), 2489-504, 2001.

10. Thiébaut R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D: "Bivariate linear mixed models using SAS proc MIXED" Comput Methods Programs Biomed, in press, 2002.

11. Hughes JP: "Mixed effects models with censored data with application to HIV RNA levels" Biometrics, 55(2), 625-9, 1999.

12. Lyles RH, Lyles CM, Taylor DJ: "Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs" Appl Statist, 49(4), 485-97, 2000.

13. Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D: "Analysis of left-censored longitudinal data with application to viral load in HIV infection" Biostatistics, 1(4), 355-68, 2000.

14. Marquardt DW: "An algorithm for least squares estimation of nonlinear parameters" SIAM J., 11(431-41, 1963.

15. Genz A: "Numerical computation of multivariate normal probabilities" J Comput Graph Stat, 1,141-9, 1992.

16. Hammer SM, Squires KE, Hughes MD, Grimes JM, Demeter LM, Currier JS, et al.: "A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team" N Engl J Med, 337(11), 725-33, 1997.

17. Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, et al.: "Decay characteristics of HIV-1-infected compartments during combination therapy" Nature, 387(6629), 188-91, 1997.

18. Kaufmann GR, Bloch M, Zaunders JJ, Smith D, Cooper DA: "Long-term immunological response in HIV-1-infected subjects receiving potent antiretroviral therapy" AIDS, 14(8), 959-69, 2000.

19. Wood E, Yip B, Hogg RS, Sherlock CH, Jahnke N, Harrigan RP, et al.: "Full suppression of viral load is needed to achieve an optimal CD4 cell count response among patients on triple drug antiretroviral therapy" AIDS, 14(13), 1955-60, 2000.

20. Fitzgerald AP, DeGruttola VG, Vaida F: "Modelling HIV viral rebound using non-linear mixed effects models" Stat Med, 21,2093-108, 2002.

21. Sy JP, Taylor JM, Cumberland WG: "A stochastic model for the analysis of bivariate longitudinal AIDS data" Biometrics, 53(2), 542-55, 1997.

22. Taylor JM, Law N: "Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts?" Stat Med, 17(20), 2381-94, 1998.

23. Marschner IC, Betensky RA, DeGruttola V, Hammer SM, Kuritzkes DR: "Clinical trials using HIV-1 RNA-based primary endpoints: Statistical analysis and potential biases" J Acquir Immune Defic Syndr Hum Retrovirol, 20(3), 220-7, 1999.

24. Verbeke G, Molenberghs G. Inference for the random effects. In: Linear mixed model for longitudinal data. New York: Springer; 2000. p. 77-92.

25. Mallinckrodt CH, Clark WS, David SR: "Accounting for dropout bias using mixed-effects models" J Biopharm Stat, 11(1-2), 9-21, 2001.

26. Laird NM: "Missing data in longitudinal studies" Stat Med, 7(1-2), 305-15, 1988.

27. De Gruttola V, Tu XM: "Modelling progression of CD4-lymphocyte count and its relationship to survival time" Biometrics, 50(4), 1003-14, 1994.

28. Siddiqui O, Ali MW: "A comparison of the random-effects pattern mixture model with last- observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts" J Biopharm Stat, 8(4), 545-63, 1998.

Table 1. Measures of CD4 cell count and HIV RNA during follow-up. APROCO Study (N=929).

Month	Change in CD4 cell count / mm^3 from baseline			Change in log_{10} copies/ml HIV RNA from baseline*			
	N	Mean	SD	N	Mean	SD	% censored
1	929	+72	113	929	-1.74	0.88	42
4	929	+101	130	929	-2.05	1.14	66
8	877	+131	147	862	-2.12	1.21	65
12	849	+159	166	842	-2.13	1.30	62

* Undetectable HIV RNA were replaced by the log_{10} of half of the assay threshold.

SD Standard Deviation

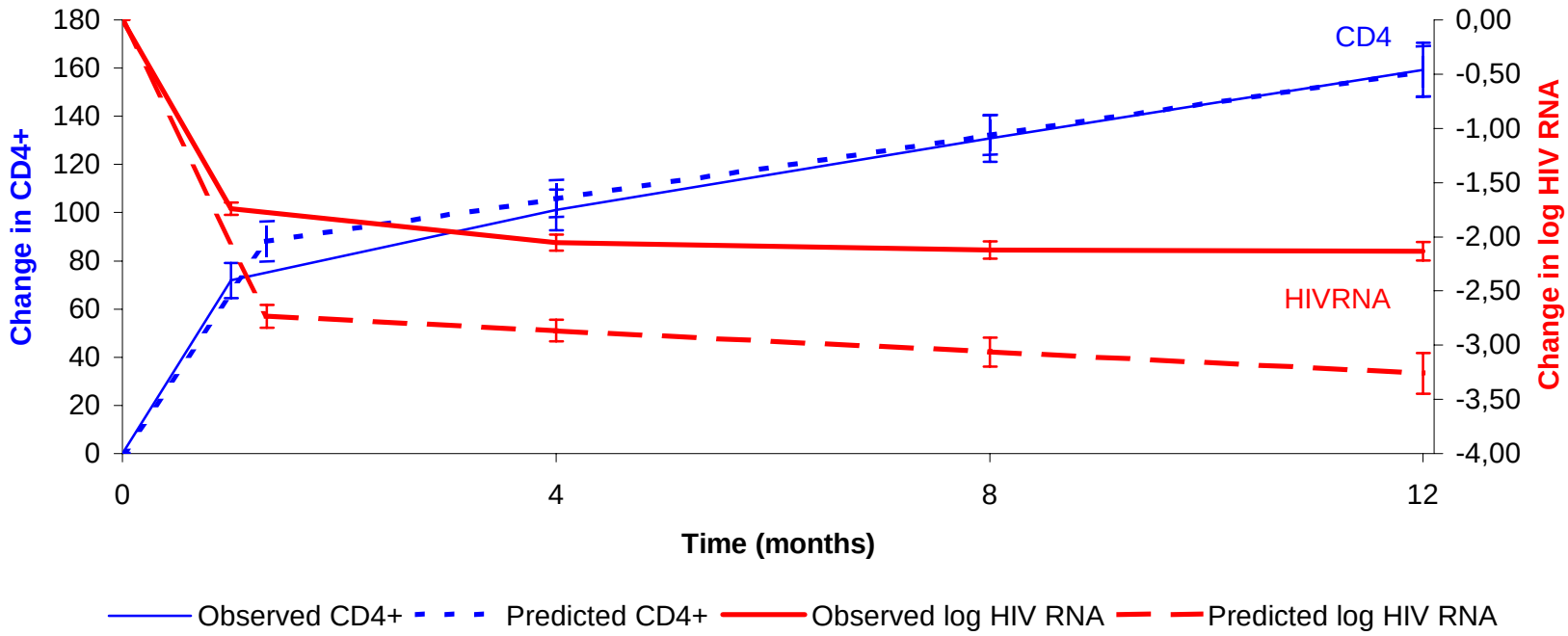
Table 2. Estimation of fixed parameters of linear mixed models for the evolution of CD4 and HIV RNA according to the method used. APROCO Study.

	Analyses taking left-censoring into account		Analyses imputing half of the threshold for undetectable HIV RNA
	Two univariate models	Bivariate model	Bivariate model
Slope 1 of HIV RNA (log_{10} cp/ml/month)	-2.05 (0.05)	-2.03 (0.04)	-1.61 (0.03)
Slope 2 of HIV RNA (log_{10} cp/ml/year)	-0.59 (0.16)	-0.58 (0.10)	+0.035 (0.04)
Slope 1 of CD4+ (cells/mm^3/month)	+66 (3.02)	+66 (3.19)	+66 (3.19)
Slope 2 of CD4+ (cells/mm^3/year)	+78 (5.99)	+78 (6.24)	+78 (5.99)

Table 3. Correlation matrix (standard error) of a bivariate linear mixed model fitting longitudinal data of CD4 cell count and HIV RNA of 929 HIV 1 infected patients during 24 months. APROCO Study.

	First slope of CD4	First slope of HIV RNA	Second slope of CD4	Second slope of HIV RNA
First slope of CD4	1	-0.33 (0.054)	-0.047 (0.051)	0.12 (0.073)
First slope of HIV RNA	-0.33 (0.054)	1	-0.069 (0.067)	0.13 (0.074)
Second slope of CD4	-0.047 (0.051)	-0.069 (0.067)	1	-0.37 (0.080)
Second slope of HIV RNA	0.12 (0.073)	0.13 (0.074)	-0.37 (0.080)	1

Fig. 1. Observed and predicted mean change in log HIV RNA and CD4+ cell count (with 95% confidence interval) after initiation of an antiretroviral treatment containing a protease inhibitor. APROCO study (N=929).



4.4 Utilisation du logiciel SAS® pour la modélisation de données longitudinales censurées

4.4.1 Introduction

Depuis la version 7 de SAS®, une nouvelle procédure est disponible : NLMIXED. Cette procédure a été programmée afin d'estimer les paramètres de modèles non linéaires à effets aléatoires. Les distributions possibles de la variable réponse sont les lois de poisson, binomiale et normale. Il existe aussi une fonction dite générale qui permet de maximiser une fonction de vraisemblance conditionnelle aux effets aléatoires définie par l'utilisateur. C'est cette fonction qu'on propose d'utiliser pour prendre en compte la censure de la charge virale dans le cadre d'un modèle linéaire mixte pour données longitudinales gaussiennes. L'avantage de cette méthode est bien entendu sa disponibilité et sa simplicité pour tout utilisateur du logiciel SAS®. On a comparé cette approche avec la précédente présentée dans la section 4.3 et dans l'article princeps [153] ainsi qu'à la méthode proposée par Lyles [85] utilisant le module IML de SAS®.

4.4.2 Modèle et vraisemblances

Les trois méthodes sont destinées à estimer les paramètres de modèles mixtes en prenant en compte la censure de la variable réponse. Seuls l'écriture de la vraisemblance et/ou la méthode d'estimation des paramètres diffèrent. Le modèle utilisé pour la comparaison est un simple modèle à intercept et pente aléatoire, potentiellement corrélés.

Soit la variable réponse Y_{ij} définie comme le logarithme en base 10 de la charge virale, à la j^{ème} mesure au temps t_{ij} (j=1,…, ni) pour le i^{ème} sujet (i=1,…, N). On distingue pour un sujet i, Y_i^o le vecteur de dimension n_i^O et Y_i^c le vecteur de dimension n_i^c correspondant respectivement aux données observées et aux données censurées.

Le modèle s'écrit $Y_{ij} = \alpha + a_i + (\beta + b_i)\, t_{ij} + e_{ij}$

avec $a_i \sim N(0, \sigma_1^2)$, $b_i \sim N(0, \sigma_2^2)$, $\mathrm{cov}(a_i, b_i) = \sigma_{12}$ et $e_{ij} \sim N(0, \sigma^2)$ l'erreur de mesure indépendante des effets aléatoires.

La vraisemblance selon l'approche présentée dans la section 4.3 et implémentée dans le programme Fortran CENSAD est une formulation marginale du modèle et s'écrit :

$$L(\theta)=\prod_{i=1}^{N} f_{Y_i^o|\theta}\left(Y_i^o|\theta\right)\int_{H_1}\int_{H_2}\ldots\int_{H_{n_i^c}} f_{Y_i^c|Y_i^o,\theta}\left(u_d|Y_i^o,\theta\right)du_1du_2\ldots du_{n_i^c}$$

avec $\theta=\begin{bmatrix}\beta'\\ \alpha'\end{bmatrix}$ le vecteur de paramètres du modèle et $H_d=\left]-\infty, c_{id}\right]$ le domaine d'intégration avec $d=1,2,\ldots,n_i^c$.

Dans l'approche de Lyles [85] et dans NLMIXED, il s'agit d'une formulation hiérarchique de la vraisemblance obtenue en conditionnant sur les effets aléatoires.

$$L(\theta)=\prod_{i=1}^{N}\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{Y_i^o,Y_i^c|a_i,b_i}\left(Y_i^o,Y_i^c|u,w\right) f_{a_i,b_i}(u,w)\,dudw\right]$$

$$=\prod_{i=1}^{N}\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left\{\prod_{j=1}^{n_{io}} f_{Y_{ij}^o|a_i,b_i}\left(Y_{ij}^o|u,w\right)\right\}\left\{\prod_{j=n_{io}+1}^{n_{ic}}\phi_{Y_{ij}^c|a_i,b_i}\left(Y_{ij}^c|u,w\right)\right\}f(u,w)\,dudw\right]$$

Les différences entre les trois méthodes proposées sont résumées dans le tableau suivant.

Tableau 1. Méthodes d'estimation des paramètres d'un modèle à effet mixte prenant en compte la censure à gauche de la variable réponse.

Méthode	Formulation vraisemblance	Algorithme d'optimisation	Méthode d'intégration
CENSAD Fortran [153]	Marginale	Marquardt	Adaptative par sous-région, SADMVN [158]
IML [85]	Hiérarchique	Quasi-Newton ou autres*	Quadrature simple
NLMIXED	Hiérarchique	Quasi-Newton ou autres*	Quadrature adaptative

* Autres algorithmes disponibles : Gradient conjugué, Simplex, Double dogleg, Région de confiance, Newton-Raphson et apparentés

L'approche du programme CENSAD conduit au calcul d'une intégrale multiple de dimension le nombre de mesures censurées par sujet. La fonction à intégrer étant une loi multivariée normale des méthodes numériques efficientes sont disponibles notamment la routine Fortran SADMVN écrite par Genz [158] et basée sur une méthode adaptative par sous-région. Les approches utilisant IML ou NLMIXED conduisent au calcul d'une intégrale multiple aussi large que le nombre d'effets aléatoires. Dans son programme Lyles et al. utilisent une méthode

de quadrature simple. NLMIXED propose plusieurs méthodes d'intégration, par défaut il s'agit de la quadrature adaptative [85]. Quant aux algorithme d'optimisation, CENSAD utilise l'algorithme Marquardt [159], IML et NLMIXED un algorithme de Quasi-Newton (par défaut). Ces algorithmes sont tous deux proches de l'algorithme de Newton-Raphson. Le programme IML de Lyles [85] utilise une routine d'optimisation précompilée (NLPQN); d'autres routines sont disponibles dans le module IML (Gradient conjugué NLPCG, Simplex NLPNMS, Double dogleg NLPDD, Région de confiance NLPTR, Newton-Raphson NLPNRA).

4.4.3 Code SAS®

Le code présenté ci-dessous permet d'estimer les paramètres du modèle présenté en 4.4.2 avec la procédure NLMIXED.

```
proc nlmixed data=test2 QTOL=1E-6 ;
parms sigsq1=0.4 sig12=0.03 sigsq2=0.4 sigsq=0.18
      alpha=3.11 beta=0.37;
bounds sigsq1 > 0, sigsq2 > 0, sigsq > 0;
      pi=2*arsin(1);mu=alpha+beta*time+a_i+b_i*time ;
if observed=1 then ll=(1/(sqrt(2*pi*sigsq)))
            *exp(-(response-mu)**2/(2*sigsq));
if observed=0 then ll=probnorm((response-mu)/sqrt(sigsq));
L=log(ll);
model response ~ general(L);
random a_i b_i ~ normal([0,0],[sigsq1,sig12,sigsq2])
subject=id;
```

La déclaration "parms" permet de définir les paramètres du modèle et de leur donner une valeur initiale. Cette étape est très importante. En cas de problème de convergence, la modification des points de départ est la première chose à faire.

La déclaration "bounds" permet de définir des contraintes sur les paramètres du modèle notamment sur les paramètres de la matrice de variance-covariance.

Les lignes qui apparaissent en gras représente la vraisemblance conditionnelle aux effets aléatoires. On distingue deux situations: les mesures observées ("observed=1") et les mesures censurées ("observed=0"). L'espérance conditionnelle aux effets aléatoires est notée μ. La procédure maximise une log-vraisemblance d'où l'écriture de l'instruction "L=log(ll)". La déclaration "model" permet de définir la variable réponse. La déclaration "random" permet de

définir la distribution des effets aléatoires. L'option "subject=id" indique quelle est l'unité de regroupement des observations.

On peut aisément étendre le modèle univarié à un modèle multivarié en incluant un indicateur pour différencier les deux marqueurs (ici VAR). Dans le code SAS, on retrouve l'espérance conditionnelle aux effets aléatoires des deux marqueurs :

$$\begin{cases} E\left(Y_i^1 \middle| u0v1_i u1v1_i\right) = b0v1 + b1v1*T_i + u0v1_i + u1v1_i*T_i \\ E\left(Y_i^2 \middle| u0v0_i u1v0_i\right) = b0v0 + b1v0*T_i + u0v0_i + u1v0_i*T_i \end{cases}$$

$$\text{avec} \begin{bmatrix} u0v0 \\ u0v1 \\ u1v0 \\ u1v1 \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s1 & & & \\ s2 & s3 & & \\ s4 & s5 & s6 & \\ s7 & s8 & s9 & s10 \end{pmatrix}\right]$$

```
Proc nlmixed data=test ;
parms b0v0=3.4 b1v0=-0.2 b0v1=4.4 b1v1=0.01
s1=29 s2=-0.06 s3=0.2 s4=-11.2 s5=0.3 s6=45
s7=-0.02 s8=-0.097 s9=-0.5 s10=0.05 sce0=94 sce1=10;
bounds s1 >=0,s3>=0,s6>=0,s10>=0,sce0>=0,sce1>=0;
pi=2*arsin(1);
/* VAR1 toujours observée*/
if VAR=1 then do ;
mu=b0v1+b1v1*T+u0v1+u1v1*T ;
L=(1/(sqrt(2*pi*sce1)))*exp(-(Y-mu)**2/(2*sce1));
end;
/* VAR2 potentiellement censurée */
if VAR=0 and OBS=1 then do ;
mu=b0v0+b1v0*T+u0v0+u1v0*T ;
L=(1/(sqrt(2*pi*sce0)))*exp(-(Y-mu)**2/(2*sce0));
end;
if VAR=0 and OBS=0 then do ;
mu=b0v0+b1v0*T+u0v0+u1v0*T ;
L=probnorm((Y-mu)/sqrt(sce0));
end;
ll=log(L);
model Y ~ general(ll);
random u0v0 u0v1 u1v0 u1v1 ~
normal([0,0,0,0],[s1,s2,s3,s4,s5,s6,s7,s8,s9,s10])
subject=id;
```

4.4.4 Résultats

On présente les estimations des paramètres du modèle à effets mixtes en traitant ou non la censure de la charge virale et selon la méthode utilisée. Les données utilisées sont les données simulées par Lyles et al. [85] et disponibles sur le site internet http://www.blackwellpublishers.co.uk/rss/Volumes/Cv49p4.htm. Il s'agit de données simulées avec les paramètres suivants : N=50, ni=5 $\forall$i, α=3, β=0,5, σ_1^2=0,5, σ_2^2=0,1, σ_{12}=-0,1. Le taux de données censurées était de 15%.

Tableau 2. Estimations des paramètres du modèle à effets mixtes en traitant ou non la censure de la charge virale et selon la méthode utilisée. Les données simulées par Lyles et al. [85] : k=50, ni=5, α=3, β=0,5, σ_1^2=0,5, σ_2^2=0,1, σ_{12}=-0,1.

Méthodes	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_{12}$	$\hat{\sigma}_2^2$
Cs non traitée (MIXED)	3.0837 (0.1044)	0.4268 (0.0523)	0.4381 (0.1096)	-0.02579 (0.04173)	0.06554 (0.02851)
Cs traitée (NLMIXED)	2.9401 (0.1283)	0.5046 (0.06203)	0.6554 (0.1681)	-0.1052 (0.06629)	0.08898 (0.03988)
Cs traitée (IML)	2.9401 (0.1283)	0.5046 (0.06203)	0.6554 (0.1681)	-0.1052 (0.06627)	0.08898 (0.03987)
Cs (Censad)	2.9401 (0.1283)	0.5047 (0.06202)	0.6554 (0.1684)	-0.1052 (0.06632)	0.08898 (0.03992)

Les deux résultats principaux sont les suivants :

- ✓ L'absence de prise en compte de la censure de la charge virale conduit à des estimations biaisées notamment de la pente
- ✓ Les estimations prenant en compte la censure de la charge virale sont équivalentes quelle que soit la méthode utilisée.

4.4.5 Discussion

La principale distinction entre le programme Fortran CENSAD d'une part et NLMIXED ou le programme IML de Lyles et al. d'autre part est la dimension de l'intégration multiple. Dans le premier cas, la taille de l'intégrale est égale au nombre de mesures censurées et dans le second

cas, la taille de l'intégrale est égale au nombre d'effets aléatoires. Le choix de l'une au l'autre méthodes dépend donc avant tout du type de données (proportion de données censurées) et du modèle. Pour cette intégration, le programme Fortran utilise une routine écrite par Genz [160]. Celui-ci rapporte une bonne efficacité de sa méthode pour des dimensions inférieures à dix. Le programme de Lyles [85] sous IML utilise une méthode de quadrature simple pour le calcul des intégrales multiples. Or cette méthode est clairement moins bonne que la quadrature adaptative [161] utilisée par NLMIXED par défaut. Dans notre expérience, nous avons inclus jusqu'à quatre effets aléatoires avec donc dix paramètres de covariance dans NLMIXED. De plus, cette procédure présente l'avantage de pouvoir contraindre certains paramètres de covariance à 0 ce qui permet de tester des hypothèses sur un seul paramètre de covariance avec un test du rapport de vraisemblance et de limiter le nombre de paramètres de covariance si nécessaire.

Il est très facile d'estimer les paramètres d'un modèle bivarié à l'aide de cette procédure en incluant simplement un indicateur pour différencier les deux marqueurs. Cependant, le nombre de paramètres de covariance sera une limite. Parmi les trois programmes comparés, seul CENSAD permet l'ajout d'une erreur autocorrélée via un processus auto-régressif d'ordre 1 ou un mouvement brownien. Quant au programme IML de Lyles et al., il a été écrit pour estimer un modèle à intercept et pente aléatoire. Pour estimer les paramètres de tout autre modèle, le programme doit être modifié.

5 Sortie d'étude informative

5.1 Etat de la question

Les articles méthodologiques concernant la modélisation conjointe de données longitudinales et d'un temps de survie ou de sortie d'étude avec une application portant sur le VIH sont relativement nombreux [33, 84, 85, 129, 162]. On peut également se référer à deux revues de la littérature [83, 163]. Cependant, seuls Lyles et al. [85] ont proposé une méthode pour prendre en compte deux des difficultés méthodologiques rencontrées dans la modélisation de données longitudinales du VIH. En effet, ils proposent un modèle mixte a intercept et pente aléatoire prenant en compte la censure de la charge virale et la sortie d'étude informative. Ainsi, jusqu'à présent, aucune méthode n'a été proposée pour prendre en compte les trois difficultés méthodologiques citées dans la section 1.5. Nous avons donc proposé un modèle bivarié à effets mixtes prenant en compte l'incomplétude des marqueurs liée aux données manquantes informatives ou à la censure d'un ou des deux marqueurs. La méthode d'estimation des paramètres du modèle est basée sur une méthode de maximisation de la vraisemblance utilisant un algorithme de Marquardt. Elle a été programmée en Fortran 90. Elle a été appliquée sur les données d'une collaboration internationale regroupant une vingtaine de cohorte de patients dont la date de séroconversion est connue ou correctement estimée (CASCADE Concerted Action on SeroConversion to AIDS and Death in Europe). L'objectif épidémiologique était d'étudier la réponse viro-immunologique chez les patients infectés par le VIH-1 et traités pour la première fois par un traitement antirétroviral hautement actif en fonction de différentes variables d'intérêt dont la toxicomanie intraveineuse ou le délai entre la séroconversion et l'initiation du traitement. Le modèle présenté a donc été développé pour répondre à cette question ce qui explique l'utilisation de ces données pour l'illustration de la méthode.

Le suivi des patients a été censuré lors de la modification du traitement initial du patient (définissant une sortie d'étude). En effet, la modification d'un traitement antirétroviral survient assez fréquemment mais à des temps très différents en fonction des individus. De plus, ce changement de traitement entraîne souvent une modification de l'évolution des marqueurs, ce qui a amené certains auteurs a utilisé des méthodes alternatives à la modélisation linéaire paramétrique [164, 165]. Cependant, on peut s'attendre à la génération des données manquantes informative dans la mesure ou une grande partie de ces modifications de traitement sont liées à une mauvaise efficacité du traitement évaluée par le clinicien d'après

l'évolution des marqueurs viro-immunologique mais aussi sur d'autre informations non présentes dans le modèle. Ceci justifie donc l'utilisation d'une modélisation conjointe des marqueurs et du délai jusqu'à la modification du traitement.

L'application de la méthode sur les données de CASCADE montre un impact important de chacun des problèmes méthodologiques à savoir la corrélation des deux marqueurs, la censure de la charge virale et la présence de données manquantes informatives. Dans l'article présentant cette méthode (section suivante 5.2), on souligne sa principale limite à savoir le caractère complètement paramétrique du modèle.

De plus, il faut noter que l'application de la méthode a été réalisée en modélisant la différence de la valeur des marqueurs en fonction de leur valeur initiale à l'initiation du traitement antirétroviral afin d'assurer la normalité et l'homoscédasticité des résidus et de réduire le nombre de paramètres de covariance. Cette approche permettait ainsi de ne pas inclure d'intercept. Dans le modèle bivarié, onze paramètres étaient économisés. De plus, ceci évitait de modéliser des intercepts aléatoires dont la distribution n'est peut être pas noramle. Toutefois, ceci a engendré deux difficultés : un problème de régression vers la moyenne et une hypothèse d'indépendance et de normalité sur la différence des erreurs de mesure. Le premier problème peut être en partie régler par l'ajout de la valeur initiale en tant que variable explicative fixe dans le modèle [166]. Le second problème qui concerne l'erreur de mesure peut être pris en compte en modifiant la structure de covariance. En effet, si on veut prendre en compte la corrélation entre l'erreur de mesure sur la valeur initiale et l'erreur de mesure sur la valeur ultérieure, on peut définir $\Sigma_i = (1_{n_i} + I_{n_i})\sigma_\varepsilon^2$ au lieu de $\Sigma_i = I_{n_i}\sigma_\varepsilon^2$ [93].

5.2 Modélisation bivariée prenant en compte la censure de la charge virale et la sortie d'étude informative

Title: 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.

Rodolphe Thiébaut[1,*,†], Hélène Jacqmin-Gadda[1], Abdel Babiker[2], Daniel Commenges[1] and the CASCADE Collaboration[‡]

1 INSERM U 330, Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France

2 Medical Research Council Clinical Trials Unit, University College London Medical School, 222 Euston Road, London NW1 2DA, UK

‡ see appendix

* Correspondence to: Rodolphe Thiébaut, ISPED - INSERM U330, Case 11, Université de Bordeaux 2, 146 Rue Leo Saignat 33076 Bordeaux Cedex. France

†E-mail: rodolphe.thiebaut@isped.u-bordeaux2.fr

Grants

CASCADE is funded through a grant from the European Union [QLK2-2000-01431] and has received additional funding from GlaxoSmithKline. R. Thiébaut is supported by a fellowship from the Charity Ensemble Contre le SIDA (Sidaction).

SUMMARY

Several methodological issues occur in the context of the longitudinal study of HIV markers evolution. Three of them are of particular importance: i) correlation between CD4+ T lymphocytes (CD4+) and plasma HIV RNA, CD4+ being the target of the virus partly measured by plasma HIV RNA; ii) left-censoring of HIV RNA due to a lower quantification limit; iii) and potential informative dropout. We propose a full parametric approach to estimate parameters of a joint model including a bivariate linear mixed model for the two markers and a lognormal survival model for the time to dropout. We apply the model to data from patients starting antiretroviral treatment in the CASCADE collaboration where all of the three issues needed to be addressed.

KEYWORDS: Bivariate mixed model, repeated measurements, left-censoring, informative dropout, HIV infection

1. INTRODUCTION

Since 1996, two major improvements have been made in the management of patients infected by the human immunodeficiency virus 1 (HIV-1): (i) the availability of highly active antiretroviral therapy (HAART) leading to a substantial reduction in mortality [2] and disease progression [167] and (ii) the availability of assays for the quantification of virus load in blood (plasma HIV RNA) used in addition to CD4+ T lymphocytes cell count (CD4+) which is the major target of HIV and a key agent for the immune system. They have both separately demonstrated their prognostic value [35] on disease progression (opportunistic diseases and/or death). Because of the rarity of clinical events, they are used as outcomes in many clinical trials and in routine monitoring of patients in clinical practice. So, the response of these two markers after the initiation of HAART is of interest in the evaluation of efficacy of antiretroviral treatment and the factors associated with this efficacy. In our example, we were interested in the differential response to HAART according to factors such as the likely mode of HIV transmission. Some previous epidemiological studies reported a worse response in patients infected through intravenous drug use [36, 51] which could be due to poor adherence to treatment. If this kind of information is confirmed, these patients would need closer monitoring than others.

Random effects models [88] are increasingly used to deal with repeated measures in epidemiological studies. There are several methodological issues in the analysis of data from longitudinal studies of the evolution of HIV markers. Three of them are of particular importance: i) correlation between CD4+ and HIV RNA; ii) left-censoring of HIV RNA and iii) informative dropout. Firstly, because CD4+ and HIV RNA markers are intrinsically correlated, a bivariate random effects model is not only necessary for the estimation of the correlation between the two markers, but has also been shown to provide a better fit to data than two separate univariate models [83, 168]. Secondly, due to a lower quantification limit of

the assays used to quantify HIV RNA, measures may be undetectable, i.e. left-censored. This problem is more pertinent when analysing the evolution of HIV RNA after the initiation of a HAART regimen because a lot of patients experience a fall of HIV RNA below assays quantification limit. Crude approaches like imputation of viral load half the limit of the assay threshold leads to biased estimations of model parameters and their standard deviations [152, 153]. Thirdly, incomplete follow up, leading to censored longitudinal data, because of death or withdrawal from the study is likely to be informative, i.e. associated with the marker trajectory [33, 84, 85]. If this dropout process can be classified as missing at random that is depending only on the observed values of the major markers of HIV disease progression and if this process is ignorable, then a simple likelihood could be adapted [169]. If not, estimations of the mixed model would be biased. For example, when studying natural history of CD4+, methods that ignore informative dropouts lead to overoptimistic statements about the marker trends, because subjects with worse CD4+ evolution tend to have shorter follow-up times and hence are weighted less in the estimation of the group rate of the average marker decline [170]. Various approaches have been proposed to obtain unbiased inference of the longitudinal profiles in the context of informative censoring due to dropout. These were mainly based on the joint modelling of the longitudinal and the dropout processes where random effects represent the underlying link between the two processes [33, 84, 85, 128, 130]. In our example, we chose to censor follow up at the time when the patient's antiretroviral treatment was significantly modified or stopped, because the pattern of marker evolution will then be too complicated and the use of linear models would not be appropriate [171]. We assume that the modification of the treatment depends on the marker response (whether observed or not) and potentially on some other factors not necessarily measured in the study (e.g. patient's adherence to treatment) but associated with the marker response. Modelling the dropout process by using a random-effects-dependent selection model [117] favour a

relationship between treatment modification and trend in markers rather than with absolute value of the marker. Moreover, in our example, data were from open cohorts with follow-up based on routine clinical practice in different sites and not on fixed predetermined visit times for all patients. In such an unbalanced design, outcome-dependent selection models are not suitable because the lack of a discrete set of measurement times for the outcome [116].

Although the three issues have been separately addressed, they have never been considered simultaneously. To be able to make accurate inferences, we combined a bivariate mixed model for the markers with a lognormal survival model of time-to-dropout taking into account left-censoring of HIV RNA using a full parametric approach. This model was applied to a large multi-cohort study including HIV-1 infected patients with a known or well estimated date of contamination (the Concerted Action on SeroConversion to AIDS or Death in Europe CASCADE collaboration) described elsewhere [2]. Our main aim was to investigate the influence of HIV mode of transmission on viral and immunological response to HAART.

2. METHODS

2.1 Model

Let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$, the response vector for the subject i=1,…N, with Y_i^k the n_i^k-vector of measurements of the marker k (k=1, 2). The number of measurements may be different for each marker and each subject. A bivariate linear mixed model may be written as follows:

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i \text{ with } \begin{cases} \varepsilon_i \sim N(0, \Sigma_i) \\ \gamma_i \sim N(0, G) \end{cases}$$

and $\beta = \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix}$, $X_i = \begin{bmatrix} X_i^1 & 0 \\ 0 & X_i^2 \end{bmatrix}$, $\gamma_i = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix}$, $Z_i = \begin{bmatrix} Z_i^1 & 0 \\ 0 & Z_i^2 \end{bmatrix}$, $\varepsilon_i = \begin{bmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \end{bmatrix}$.

X_i^k is a $n_i \times p^k$ design matrix of explanatory variables for marker k, β^k is a p^k-vector of fixed effects, Z_i^k is a $n_i \times q^k$ design matrix which is usually a subset of X_i^k, γ_i^k is a q^k-vector of individual random effects with $q^k \leq p^k$ (for k=1, 2). The covariance matrix of measurement errors is a diagonal matrix, denoted by Σ_i and containing the two elements $\sigma^2_{\varepsilon^1}$ and $\sigma^2_{\varepsilon^2}$ representing the measurement error of each marker. The covariance matrix of random effects, G, can be partitioned in four sub-matrices: (i) G^1 the covariance matrix including variances and covariances of random effects for the first marker (ii) G^2 the covariance matrix including variances and covariances of random effects for the second marker and (iii) $G^{12} = G^{21}$ the matrix of covariances between random effects of each marker. This is through this sub-matrix that correlation between the two markers is taken into account.

The dropout process can be defined by two types of event:

- informative events, i.e. patients who left before the scheduled end of the study because of clinical progression, discontinuation of treatment or any potential informative reason that is associated with the latent evolution of the marker,

- patients followed until the end of the study or loss to follow up for non-informative reasons.

Using a survival model, any informative dropout is considered as observed dropout $(\delta_i = 1)$ and any other event leads to right-censoring $(\delta_i = 0)$ because the patient is known to be followed at least until this date. In other words, the dropout process that generates the informative right-censoring of longitudinal profiles, is itself right-censored, but non informatively, by the scheduled end of the study.

In our approach, we used a lognormal survival model to take into account the informative dropout. Following Schluchter et al. notations [127], let C_i be the natural logarithm of the time from baseline to the scheduled end of follow-up or to a dropout considered as non-informative. T_i^o is the natural logarithm of the time from baseline to an event that would define dropout as informative. So, we observe the pair $\left(T_i = \min\left[T_i^o, C_i\right], \delta_i\right)$ for each patient.

We assume that the natural logarithm T_i^o is correlated to the random effects γ_i through the q-vector of covariances **B** such that

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

2.2 Likelihood

The likelihood of the joint model fitting longitudinal markers and time to dropout can be written:

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

with

$$f_{Y_i|\gamma_i}\left(Y_i \,|\gamma_i\right) = f_{Y_i|\gamma_i}\left(Y_i^1, Y_i^2|\gamma_i\right) = \left\{\prod_{j=1}^{n_{i1}} f_{Y_{ij}^1|\gamma_i}\left(Y_{ij}^1|\gamma_i\right)\right\}\left\{\prod_{j=1}^{n_{i2}} f_{Y_{ij}^2|\gamma_i}\left(Y_{ij}^2|\gamma_i\right)\right\}.$$

$f_{T_i^o|\gamma_i}(.|.)$ and $F_{T_i^o|\gamma_i}(.|.)$ are the conditional probability density function and cumulative distribution function of T_i^o given γ_i, respectively. Using properties of the multivariate normal distribution, $f_{T_i^o|\gamma_i}(.|.)$ is Gaussian with expectation $\mu_{T_o} + \mathbf{B}^T\mathbf{G_i^{-1}}\boldsymbol{\gamma}_\mathbf{i}$ and covariance $\sigma^2_{T_o} - \mathbf{B}^T\mathbf{G_i^{-1}}\mathbf{B}$.

If there was no dropout , all dropout times would be censored. Then, the likelihood (1) would reduce to a simpler one accounting for the contribution of Y alone only if $F_{T_i^o|\gamma_i}(T_i^o|\gamma_i) = 0$. This last condition signifies that no dropout was observed because dropout was not possible.

When one or more markers present left-censored values, the likelihood needs to be modified. For a subject i, let Y_i^o the n_i^O-vector of observed outcomes, Y_i^c the n_i^c-vector of censored outcomes and s_i the n_i^c-vector of measurement thresholds. To take into account observed and censored values and using the independence between the observations given the random effects, the contribution to the likelihood of the vector $Y_i|\gamma_i$ can be written:

$$f_{Y_i|\gamma_i}(Y_i|\gamma_i) = f_{Y_i^o|\gamma_i}(Y_i^o|\gamma_i)\Pr(Y_i^c < s_i|\gamma_i) = \left\{\prod_{j=1}^{n_{io}} f_{Y_{ij}^o|\gamma_i}(Y_{ij}^o|\gamma_i)\right\}\left\{\prod_{j=1}^{n_{ic}} F_{Y_{ij}^c|\gamma_i}(s_{ij}|\gamma_i)\right\} \quad (2)$$

with $f_{Y_{ij}^c|\gamma_i}(.|.)$ and $F_{Y_{ij}^c|\gamma_i}(.|.)$ are the conditional probability density function and cumulative distribution function of Y_i^c given γ_i, respectively.

Finally, the likelihood of a linear mixed model for longitudinal bivariate gaussian data accounting for left-censoring and informative dropout is obtained by combining (1) and (2).

2.3 Algorithm

Model parameters were estimated by direct maximisation of the likelihood (1) using a Marquardt algorithm [159]. Multiple integration on random effects was performed by the Monte Carlo method using 2000 simulations per subject. In order to impose a positivity

constraint to variance parameters, the likelihood was reparameterised in terms of the square root of variance of measurement errors ($\sigma^2_{\varepsilon^1}$ and $\sigma^2_{\varepsilon^2}$) and the Cholesky decomposition for the covariance matrix of random effects and time to dropout $\begin{pmatrix} \mathbf{G} & \mathbf{B} \\ \mathbf{B}^T & \sigma^2_{T_o} \end{pmatrix} = \mathbf{U}^T\mathbf{U}$ (where U is an upper triangular matrix). The estimate of the variance-covariance matrix of the fixed effects was minus the inverse of the matrix of the second derivatives of the likelihood. The derivatives were computed numerically by finite difference (central difference for the first derivatives and forward difference for the second ones).

3 APPLICATION

3.1 Objectives

As mentioned in introduction, we were interested in the analysis of the marker response after the initiation of antiretroviral treatment according to patients characteristics. One of those characteristics is the HIV transmission category and in particular intravenous drug users (IDU). Because the residuals of longitudinal models of CD4+ on natural scale do not appear to satisfy the normality and homoskedasticity assumptions, some transformations like logarithm or square root are often used. However, this makes interpretation of the parameters estimates difficult. In our example, we worked with change from baseline for which the model assumptions are more tenable (see figure 1). Moreover, working with change in CD4+ and change in HIV RNA follows some guidelines which define a good response to treatment in term of a HIV RNA decline of at least 1 $\log_{10}$ copies/mL and CD4+ increase of 50 cells/mm^3 by four to eight weeks [42]. So, in the following application, the main objective was to study the change in markers according to HIV transmission group.

3.2 The data

In the CASCADE project, 994 patients were treated by a HAART regimen after 1996 and had no previous experience of antiretroviral treatment. HAART treatment was defined as three or more antiretroviral drugs, containing at least one of these major drugs: a protease inhibitor, a non-nucleoside reverse transcriptase inhibitor or abacavir. In the present study, we selected 494 patients who must have a measurement of CD4+ and HIV RNA at the time of treatment prescription and at least one measurement thereafter to be able to calculate at least one difference between marker value during follow-up and marker value at baseline. HIV RNA value at baseline must be detectable, that is not censored, to avoid a potential interval censoring of the change of marker value. The evolution of the two markers is shown in figures 2a and 2b. Patient's follow-up is based on clinical practice that is every 3 to 6 months. Follow-up is censored at the time of a modification of at least one major drug because this modification is likely to modify significantly the evolution of the markers. The probability of being still treated by the first regimen at 12 and 24 months were 53% (95% confidence interval [CI]: 48; 57) and 27% (CI: 22; 31), respectively (see figure 3). The median follow-up until the date of analysis or modification of the first line regimen was 10 months (Interquartile range [IR]: 4; 18). Median number of available measurements on each marker before censoring was 4 (IR: 3; 7). A total of 1143 (57%) measures of HIV RNA were left censored. The detection limit varied according to the assay used to quantify HIV RNA: 37% were ultra-sensitive that is with a threshold below or equal to 1.7 $\log_{10}$ copies/mL (Roche 1.5, Chiron b-DNA 3.0 and local assays). The other assays used (Roche 1.0, Chiron b-DNA 2.0, NASBA and other local assay) tend to quantify HIV RNA with a higher value compared to ultra-sensitive ones so all analyses were adjusted for the assay type by an indicator for ultra-sensitivity (denoted by 'US' in model equation below). There were 69 (14%) patients infected through intravenous drug use.

3.3 Model

Our outcome was the change in the markers $\tilde{Y}_{ij}^{k} = Y_{ij}^{k} - Y_{io}^{k}$ (for k=1, 2). In addition to homoscedasticity of residuals and ease of interpretation, this also has the advantage of reducing the number of covariance parameters because no intercept was included in the model. The evolution of the two markers after the initiation of HAART is classically bi-phasic with a change point around two weeks for HIV RNA [172]. Thus, we used a piecewise linear model with one slope representing the short term response and a second slope representing the long term response. The time of change of slope (τ) was first estimated using separate univariate models accounting for left-censoring for HIV RNA. The estimated change time for HIV RNA was earlier than for CD4+: $\hat{\tau}_1 = 1.3$ months (CI 1.2-1.4) and $\hat{\tau}_2 = 1.6$ months (CI 1.4-1.8) (see also figure 2). However, these estimations are highly depending on measures schedule; in fact, HIV RNA change point is known to be in the first two weeks after treatment initiation [172]. So, for the sake of simplicity, we decided to fix a common $\tau = 1.5$ months for the two markers for all subjects. This choice could lead to underestimate variance parameters but the estimation of an individual change point for each marker increases numerical difficulties. We studied the effect of IDU on the marker evolution through interactions with slopes. IDU in the model below is a binary variable indicating patients infected through intravenous drug use.

The model was:

$$\begin{cases} \tilde{Y}_{ij}^{1} = \beta_1^1 t_{1ij} + \beta_2^1 t_{2ij} + \beta_3^1 t_{1ij} \times IDU + \beta_4^1 t_{2ij} \times IDU + \beta_5^1 \times US + \gamma_{1i}^1 t_{1ij} + \gamma_{2i}^1 t_{2ij} + \varepsilon_{ij}^1 \\ \tilde{Y}_{ij'}^{2} = \beta_1^2 t_{1ij'} + \beta_2^2 t_{2ij'} + \beta_3^2 t_{1ij'} \times IDU + \beta_4^2 t_{2ij'} \times IDU + \gamma_{1i}^2 t_{1ij'} + \gamma_{2i}^2 t_{2ij'} + \varepsilon_{ij'}^2 \\ T_i = \mu_t + e_i \end{cases}$$

with $t_{1ij} = t_{ij} \wedge \tau$ and $t_{2ij} = (t_{ij} - \tau) I_{t_{ij} \geq \tau}$, $\varepsilon_i^k \sim N\left(0, \sigma_{\varepsilon^k}^2 I_{n_i^k}\right)$ and

$$\begin{pmatrix} \gamma_{1i}^1 \\ \underline{\gamma_{2i}^1} \\ \gamma_{1i}^2 \\ \underline{\gamma_{2i}^2} \\ T_i^o \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \mu_{T^o} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\gamma_{1i}^1} & \sigma_{\gamma_{1i}^1\gamma_{2i}^1} & \sigma_{\gamma_{1i}^1\gamma_{1i}^2} & \sigma_{\gamma_{1i}^1\gamma_{2i}^2} & \sigma_{\gamma_{1i}^1 T^o} \\ \sigma_{\gamma_{1i}^1\gamma_{2i}^1} & \sigma^2_{\gamma_{2i}^1} & \sigma_{\gamma_{2i}^1\gamma_{1i}^2} & \sigma_{\gamma_{2i}^1\gamma_{2i}^2} & \sigma_{\gamma_{2i}^1 T^o} \\ \sigma_{\gamma_{1i}^1\gamma_{1i}^2} & \sigma_{\gamma_{2i}^1\gamma_{1i}^2} & \sigma^2_{\gamma_{1i}^2} & \sigma_{\gamma_{1i}^2\gamma_{2i}^2} & \sigma_{\gamma_{1i}^2 T^o} \\ \sigma_{\gamma_{1i}^1\gamma_{2i}^2} & \sigma_{\gamma_{2i}^1\gamma_{2i}^2} & \sigma_{\gamma_{1i}^2\gamma_{2i}^2} & \sigma^2_{\gamma_{2i}^2} & \sigma_{\gamma_{2i}^2 T^o} \\ \sigma_{\gamma_{1i}^1 T^o} & \sigma_{\gamma_{2i}^1 T^o} & \sigma_{\gamma_{1i}^2 T^o} & \sigma_{\gamma_{2i}^2 T^o} & \sigma^2_{T^o} \end{pmatrix} \right)$$

3.4 Results

Estimations of CD4+ evolution and HIV RNA are presented in table 1 and 2, respectively. The crude univariate model of change in CD4+ estimated a significantly increasing first slope and second slope (86 cells per month before 1.5 months and 69 cells per year after). This second slope was significantly lower in IDU patients, tending to decline (-71 cells per year). In this univariate model, the estimation of the second slope in non IDU patients became non significantly different from 0 (switching from 69 cells to 6 cells per year) when dropout was taken into account. In fact, there was a positive correlation between the duration of first antiretroviral treatment and the CD4+ slopes (r = 0.42 and r = 0.87 for the first and the second slope, respectively). In other words, patients with a poor immune response (a decline of CD4+ or a weak increase) tend to have their treatment modified earlier than the others. Thus, their CD4 response is weighted less in the estimation of the average slopes leading to an overoptimistic estimation of response when time to treatment modification was not handled. In the bivariate model, which takes into account for the correlation between individual slopes of the two markers and informative dropout, the estimated second slope of CD4+ was significantly smaller than the estimate obtained from the crude univariate model (+25 cells per year). The evolution of CD4+ estimated under this model is illustrated in figure 3a.

The crude univariate model of change in $\log_{10}$ HIV RNA imputing half of the threshold for left-censored values estimated a significant decline (-1.47 $\log_{10}$ copies/mL/month during the first 1.5 months). There was no significant modification in long-term slope and no significant

effect of IDU. However, when left-censoring was taken into account, the second slope indicated a significant decline in non IDU patients (-0.16 log_{10} copies/mL/year) and significantly higher in IDU patients (+0.47 log_{10} copies/mL/year). This effect was expected as a simple imputation for left-censored values has previously been described to leads to underestimation of the true value of HIV RNA. When dropout was taken into account in addition to left-censoring, there was no evidence of long term decline in HIV RNA or of IDU effect. In fact, there was a negative correlation between the duration of first antiretroviral treatment and the HIV RNA slopes (r = -0.22 and r = -0.38 for the first and the second slope, respectively). In other words, patients with a poor viral response (an increase of HIV RNA or a weak decrease) tend to have their treatment modified earlier than the others. So, adjusting for the time to treatment modification corrects for the overestimation of viral response. In the bivariate model accounting for left-censoring, informative dropout and correlation between markers, the increase of HIV RNA in IDU patients remained significant (+0.58 log_{10} copies/mL/year) but not the decrease in non IDU. So, the impact of handling left-censored values of HIV RNA was obvious (figure 3b). However, the effects of modelling informative dropout and bivariate modelling on fixed effects estimation were less important for HIV RNA parameters than for CD4+ parameters.

The bivariate model was clearly better than two univariate models for fitting the change in HIV RNA and CD4+ (likelihood -6534 for 27 parameters vs. -7202 for 25 parameters). This full model allowed also the estimation of the correlation matrix between individual slopes of each marker and time to dropout. The strongest correlations were between the slopes of HIV RNA and the slopes of CD4+: r = -0.48, std = 0.20 for the first slopes and r = -0.73, std = 0.18 for the second slopes. Thus, throughout the follow-up, better viral response (declining HIV RNA) was associated with better immunological response (increasing CD4+). Moreover, the correlations of time to dropout with random slopes were significant only with the CD4+

slopes (r = 0.28, std = 0.11 for the first slope and r = 0.38, std = 0.097 for the second slope). This indicates that when both markers were taken into account, only individual slopes of CD4+ were correlated with the duration of the first treatment. As expected, the better the immune response, the longer was the duration of the first regimen.

3.5 Model assumptions

The distribution of absolute CD4 counts is known to be highly skewed. In the above model, we used change in maker value from baseline level as outcome variable. This has resulted in residuals that did not appear to deviate from the assumptions of normality and homoscedasticity (see figure 1). However, one disadvantage of the change from baseline as outcome variable is that baseline marker levels can affect the response because of the problem of regression to the mean. So, we studied the effect of IDU on slopes with a univariate model of square root of CD4+ including one intercept and two slopes to check if the model on the absolute value led to the same conclusions. The model was:

$$Y_{ij} = \beta_0 + \beta_1 t_{1ij} + \beta_2 t_{2ij} + \beta_3 t_{1ij} \times IDU + \beta_4 t_{2ij} \times IDU + \gamma_{0i} + \gamma_{1i} t_{1ij} + \gamma_{2i} t_{2ij} + \varepsilon_{ij} \qquad (3)$$

where Y_{ij} was the square root of CD4+ (leading to better fit in terms of normality and homoscedasticity of residuals).

The estimation of fixed intercept was 19.4 $\sqrt{cells}$ /mm^3 (CI: 18.9; 20.0), with an increase of 1.9 $\sqrt{cells}$ /mm^3/month (CI: 1.6; 2.2) during the first 1.5 months, followed by a weaker but significant increase 0.83 $\sqrt{cells}$ /mm^3/year (CI: 0.27; 1.4). Like in model of change in CD4+, the effect of IDU was not significant on first slope $\hat{\beta}_3 = 0.18$ $\sqrt{cells}$ /mm^3/year (CI: -0.55; 0.91), but it was on the second slope $\hat{\beta}_4 = -2.48$ $\sqrt{cells}$ /mm^3/year (CI: -3.6; -1.4).

One of the main assumptions using the proposed method is the lognormality of the time to dropout (defined by the modification of the first antiretroviral treatment). The estimation of

the survival function for a lognormal distribution of the time to dropout and the estimation using the non parametric Kaplan-Meier estimator are represented in figure 2. The estimated lognormal survival function was always included within the 95% confidence interval of non parametric estimations.

We also did a sensitivity analysis using a univariate mixed model of square root of CD4+ and a Cox proportional hazard model rather than a lognormal survival model as proposed by Henderson et al [131]. According to their notation, we assume the joint model:

$$\begin{cases} Y_{ij} = \beta_0 + \beta_1 t_{1ij} + \beta_2 t_{2ij} + \beta_3 t_{1ij} \times IDU + \beta_4 t_{2ij} \times IDU + W_{1i}(t_{ij}) + \varepsilon_{ij} \\ \lambda_i(t) = \alpha_0(t) \exp(W_{2i}(t_{ij})) \end{cases}$$

with,

$W_{1i}(t_{ij}) = U_{0i} + U_{1i} t_{1ij} + U_{2i} t_{2ij}$ and $W_{2i}(t_{ij}) = \gamma_0 U_{0i} + \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3 W_{1i}(t_{ij})$.

Thus, compared to our model, the proportional hazard model has less assumption because the baseline risk $\alpha_0(t)$ is not specified and because the dependence of the marker on the event time is a function of random effects ($\gamma_0 U_{0i} + \gamma_1 U_{1i} + \gamma_2 U_{2i}$) and also of the current value of the latent process $W_{1i}(t_{ij})$. Estimations of the parameters of the longitudinal model were $\hat{\beta}_0 = 19.0$, $\hat{\beta}_1 = 1.8$, $\hat{\beta}_2 = 0.87$, $\hat{\beta}_3 = 0.22$, $\hat{\beta}_4 = -2.35$ with Henderson et al. model. All these estimations were close to those obtained from model (3) and were included in their confidence intervals.

4. DISCUSSION

We presented a joint model including a bivariate linear mixed model and a lognormal survival model applied to two major markers of HIV infection. This approach deals with two kinds of incomplete outcome. The marker could be missing because of an informative dropout or could be left-censored because of a lack of sensitivity of the assay. In our application, the event defining the dropout was a modification of the first HAART regimen. Using our joint model, we found a significant correlation between the immune response measured by individual random slopes of CD4+ and the time to treatment modification. Moreover, we found a poorer viral and immunological responses to treatment in IDU patients. This may be due to IDU patients having poorer adherence to treatment or to other confounding factors not included in the model.

We demonstrated that the bivariate modelling increased the model likelihood and changed significantly the estimations of fixed effects compared to univariate models even after accounting for left-censoring and informative dropout. Moreover, the bivariate model provides estimates of the correlation between the two markers through the whole follow-up. The importance of accounting for left-censoring of HIV-RNA rather than using simple imputation has also been previously demonstrated [85, 152, 153]. The effect of informative dropout on mixed model estimates has also been shown in the context of HIV infection [33, 84, 85]. In the present application, there was a clear correlation between CD4+ random slopes and time to treatment modification which shows that the probability of treatment modification is dependent on the past, current and future values of CD4+ which measure with error the immune response to treatment [117].

The main drawback in this kind of joint model lies in the parametric assumptions. In particular, the survival model was assumed to be a lognormal model for ease of computation. This kind of assumptions needs sensitivity analyses [173]. We did one in our application comparing our approach with a less stringent one [131]. Estimates of fixed effects (which were of primary interest) were very similar. Because the extension of Henderson et al. model to the bivariate model accounting for left-censoring of HIV RNA is difficult, we did a sensitivity analysis on the univariate evolution of CD4+. Given the results of this sensitivity analysis and the empirical comparison with Kaplan-Meier curves, we considered the lognormal assumption of the survival model appropriate for our application. Random effects were assumed to have a multivariate normal distribution. A recent paper [174] showed in the context of joint modelling that the impact of random effects distribution on fixed effects estimations is limited particularly for within subject effects such as slopes. Another limitation of our approach is the number of random effects allowed in the model. In fact, increasing the number of random effects will increase the dimension of multiple integrals used to calculate the likelihood and also the number of covariance parameters as we assumed an unstructured covariance matrix leading to numerical difficulties.

A fundamental extension of such joint model would be the use of a stochastic process like the bivariate formulation of the IOU process [82] and a semi-parametric survival model to relax model assumptions.

Acknowledgements

We thank Didier Renard for giving us an adaptation of Henderson et al. program. We thank Geneviève Chêne for her useful comment on this paper.

APPENDIX

Steering Committee: Valerie Beral, Roel Coutinho, Janet Darbyshire (Project Leader), Julia Del Amo, Noël Gill (Chairman), Christine Lee, Laurence Meyer, Giovanni Rezza,.

Co-ordinating Centre: Kholoud Porter (Scientific Co-ordinator), Abdel Babiker, A Sarah Walker, Janet Darbyshire, Freya Tyrer.

Collaborators: Aquitaine cohort, France: Francois Dabis, Rodolphe Thiébaut, Geneviève Chêne, Sylvie Lawson-Ayayi; SEROCO cohort, France: Laurence Meyer, Faroudy Boufassa; German cohort, Germany: Osamah Hamouda, Klaus Fischer; Italian Seroconversion Study, Italy: Patrizio Pezzotti, Giovanni Rezza; Greek Haemophilia cohort, Greece: Giota Touloumi, Angelos Hatzakis, Anastasia Karafoulidou, Olga Katsarou; Edinburgh Hospital cohort, United Kingdom: Ray Brettle; Madrid cohort, Spain: Julia Del Amo, Jorge del Romero; Amsterdam Cohort Studies among homosexual men and drug users, the Netherlands: Liselotte van Asten, Birgit van Benthem, Maria Prins, Roel Coutinho; Copenhagen cohort, Denmark: Ole Kirk, Court Pedersen; Valencia IDU cohort, Spain: Ildefonso Hernández Aguado, Santiago Pérez-Hoyos; Oslo and Ulleval Hospital cohorts, Norway: Anne Eskild, Johan N Bruun, Mette Sannes; Royal Free haemophilia cohort, United Kingdom: Caroline Sabin, Christine Lee; UK Register of HIV Seroconverters, United Kingdom: Anne M Johnson, Andrew N Phillips, Abdel Babiker, Janet H Darbyshire, Noël Gill, Kholoud Porter; Swiss HIV cohort, Switzerland: Patrick Francioli, Philippe Vanhems, Matthias Egger, Martin Rickenbach; Sydney AIDS Prospective Study, Australia: David Cooper, John Kaldor, Lesley Ashton; Sydney Primary HIV Infection cohort, Australia: David Cooper, John Kaldor, Lesley Ashton, Jeanette Vizzard; Badalona IDU cohort, Spain: Robeto Muga Bustamente; Lyon Primary Infection cohort, France: Philippe Vanhems; MRC Biostatistics Unit, Cambridge, United Kingdom: Nicholas E Day, Daniela De Angelis.

REFERENCES

1. Survival after introduction of HAART in people with known duration of HIV-1 infection. The CASCADE Collaboration. Concerted Action on SeroConversion to AIDS and Death in Europe. Lancet 2000; 355:1158-1159.

2. Egger M, Hirschel B, Francioli P, Sudre P, Wirz M, Flepp M, Rickenbach M, Malinverni R, Vernazza P, Battegay M. Impact of new antiretroviral combination therapies in HIV infected patients in Switzerland: prospective multicentre study. Swiss HIV Cohort Study. British Medical Journal 1997; 315:1194-1199.

3. Mellors JW, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP, Rinaldo CR, Jr. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. Annals of Internal Medicine 1997; 126:946-954.

4. Munoz A, Carey V, Saah AJ, Phair JP, Kingsley LA, Fahey JL, Ginzburg HM, Polk BF. Predictors of decline in CD4 lymphocytes in a cohort of homosexual men infected with human immunodeficiency virus. Journal of Acquired Immune Deficiency Syndromes 1988; 1:396-404.

5. Egger M, May M, Chene G, Phillips AN, Ledergerber B, Dabis F, Costagliola D, Monforte AD, deWolf F, Reiss P, Lundgren JD, Justice AC, Staszewski S, Leport C, Hogg RS, Sabin CA, Gill MJ, Salzberger B, Sterne JAC. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. Lancet 2002; 360:119-129.

6. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982; 38:963-974.

7. Boscardin WJ, Taylor JM, Law N. Longitudinal models for AIDS marker data. Statistical Methods in Medical Research 1998; 7:13-27.

8. Thiébaut R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D. Bivariate linear mixed models using SAS proc MIXED. Computer Methods and Programs in Biomedicine 2002; 69:249-256.

9. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. Biometrics 1999; 55:625-629.

10. Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics 2000; 1:355-368.

11. De Gruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics 1994; 50:1003-1014.

12. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. Statistics in Medicine 1999; 18:1215-1233.

13. Lyles RH, Lyles CM, Taylor DJ. 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 (Applied Statistics) 2000; 49:485-497.

14. Laird NM. Missing data in longitudinal studies. Statistics in Medicine 1988; 7:305-315.

15. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. Epidemiology 2002; 13:347-355.

16. Pawitan Y, Self S. Modelling disease marker processes in AIDS. Journal of the American Statistical Association 1993; 88:719-726.

17. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics 1997; 53:330-339.

18. Fitzgerald AP, DeGruttola VG, Vaida F. Modelling HIV viral rebound using non-linear mixed effects models. Statistics in Medicine 2002; 21:2093-2108.

19. Little R. Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 1995; 90:1112-1121.

20. Hogan J, Laird N. Model-based approaches to analyzing incomplete longitudinal and failure time data. Statistics in Medicine 1997; 16:259-272.

21. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. Statistics in Medicine 1992; 11:1861-1870.

22. Marquardt DW. An algorithm for least squares estimation of nonlinear parameters. SIAM Journal 1963; 11:431-441.

23. Yeni PG, Hammer SM, Carpenter CCJ, Cooper DA, Fischl MA, Gatell JM, Gazzard BG, Hirsch MS, Jacobsen DM, Katzenstein DA, Montaner JSG, Richman DD, Saag MS, Schechter M, Schooley RT, Thompson MA, Vella S, Volberding PA. Antiretroviral treatment for adult HIV infection in 2002 - Updated recommendations of the international AIDS Society-USA panel. Journal of the American Medical Association 2002; 288:222-235.

24. Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho DD. Decay characteristics of HIV-1-infected compartments during combination therapy. Nature 1997; 387:188-191.

25. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics 2000; 1:465-480.

26. Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. Biometrics 2001; 57:7-14.

27. Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. Biometrics 2002:accepted for publication.

28. Sy JP, Taylor JM, Cumberland WG. A stochastic model for the analysis of bivariate longitudinal AIDS data. Biometrics 1997; 53:542-555.

Table 1. Estimations of fixed effects parameters (first slope before 1.5 month, second slope thereafter) of change in CD4+ count / 100, imputing half of the threshold for left-censored HIV RNA measures ("Crude"), handling for left-censoring of HIV RNA ("Left-censoring"), and handling for informative drop-out ("Drop-out"). CASCADE collaboration (N = 494).

Model	In others than IDU $(\hat{\beta} \quad \hat{\sigma})$				Correction for IDU $(\hat{\beta} \quad \hat{\sigma})$			
	First slope per month		Second slope per year		On first slope per month		On second slope per year	
UNIVARIATE								
Crude	**0.86**	**0.068**	**0.69**	**0.096**	0.075	0.19	**-1.40**	**0.31**
Drop out	**0.80**	**0.070**	0.062	0.15	0.28	0.18	**-1.14**	**0.30**
BIVARIATE								
Left-censoring & drop out	**0.88**	**0.070**	**0.25**	**0.12**	0.31	0.18	**-1.36**	**0.28**

*** Parameters significantly different from 0 in bold**

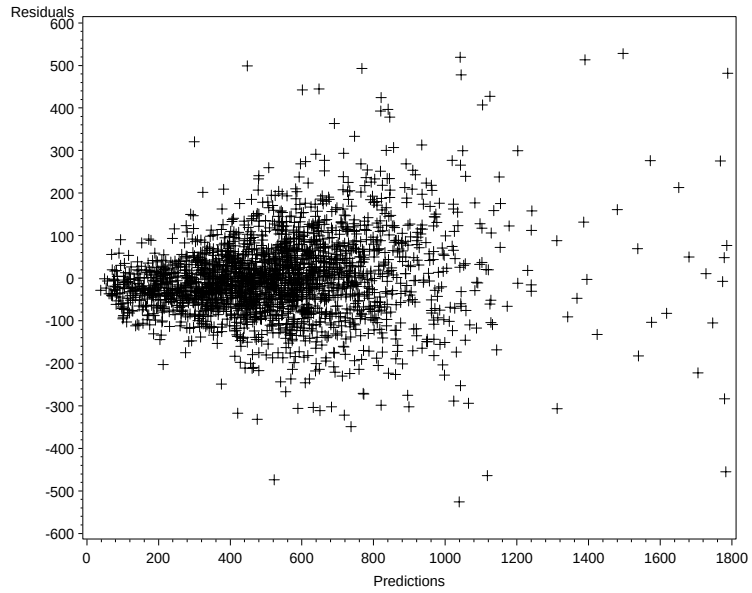
Table 2. Estimations of fixed effects parameters (first slope before 1.5 month, second slope thereafter) of change in $\log_{10}$ HIV RNA, imputing half of the threshold for left-censored HIV RNA measures ("Crude"), handling for left-censoring of HIV RNA ("Censoring"), and handling for informative drop-out ("Drop-out"). CASCADE collaboration (N = 494).

	In others than IDU $(\hat{\beta} \quad \hat{\sigma})$				Correction for IDU $(\hat{\beta} \quad \hat{\sigma})$			
Model	First slope per month		Second slope per year		On first slope per month		On second slope per year	
UNIVARIATE								
Crude	**-1.47**	**0.043**	0.021	0.061	0.21	0.11	0.25	0.16
Left-censoring	**-1.89**	**0.062**	**-0.16**	**0.12**	0.23	0.16	**0.63**	**0.30**
Left-censoring & drop out	**-1.95**	**0.064**	-0.036	0.17	0.20	0.16	0.60	0.32
BIVARIATE								
Left-censoring & drop out	**-1.89**	**0.057**	-0.15	0.10	0.14	0.14	**0.73**	**0.26**

*** Parameters significantly different from 0 in bold**

Figure 1. Residuals plot from a univariate linear mixed model of CD4+ without handling informative dropout: (a) natural scale and (b) change from baseline. CASCADE collaboration (N=494).

(a)



(b)

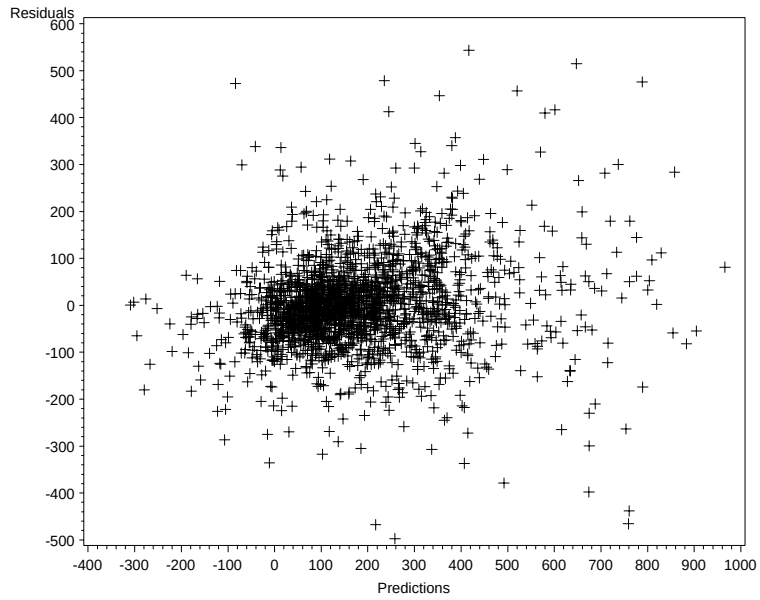


Figure 2a. Observed change in CD4+ after the initiation of a HAART regimen and predicted values using the loess method. CASCADE collaboration (N=494).

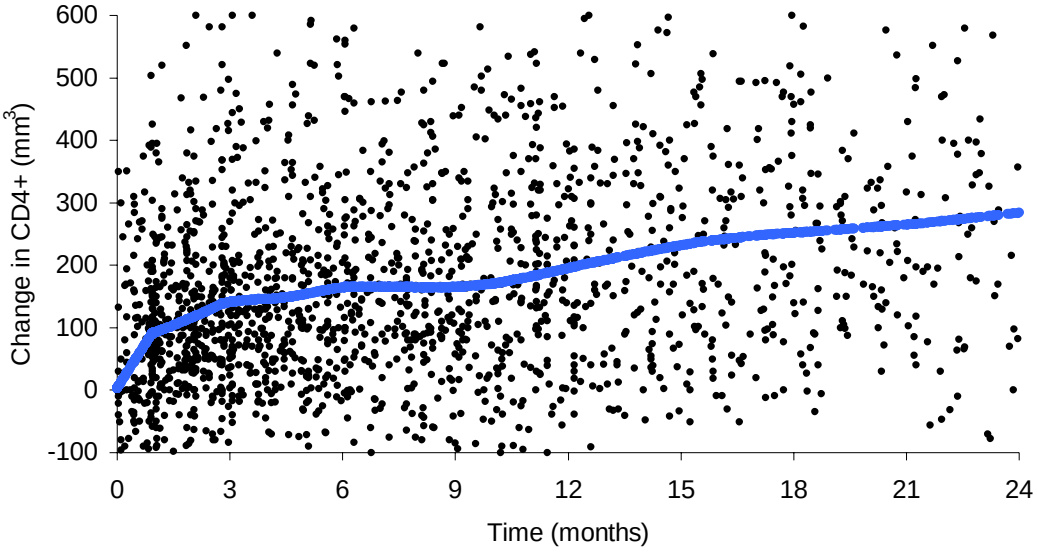


Figure 2b. Observed change in $\log_{10}$ HIV RNA after the initiation of a HAART regimen and predicted values using the loess method (left-censored values were replaced by half of the threshold). CASCADE collaboration (N=494).

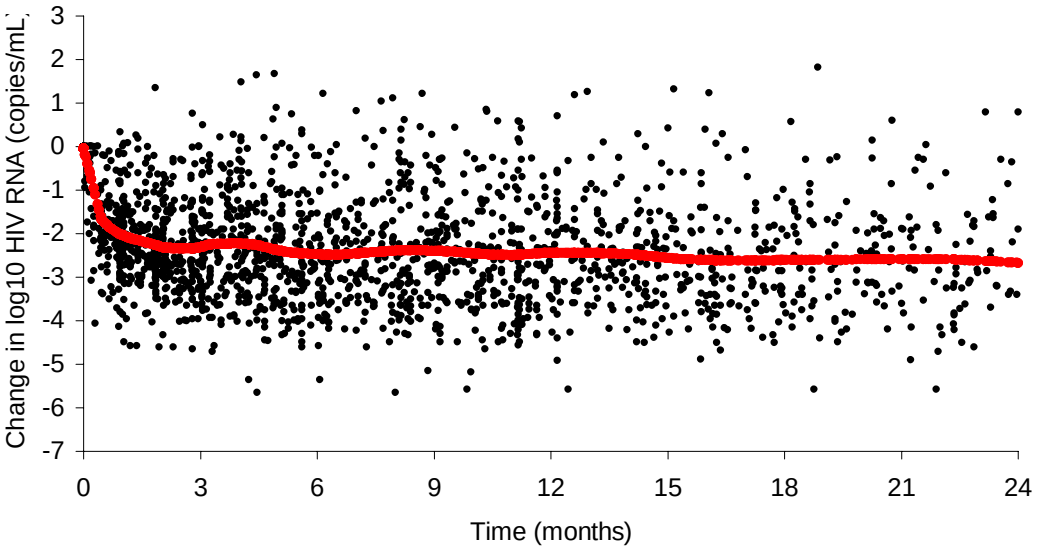
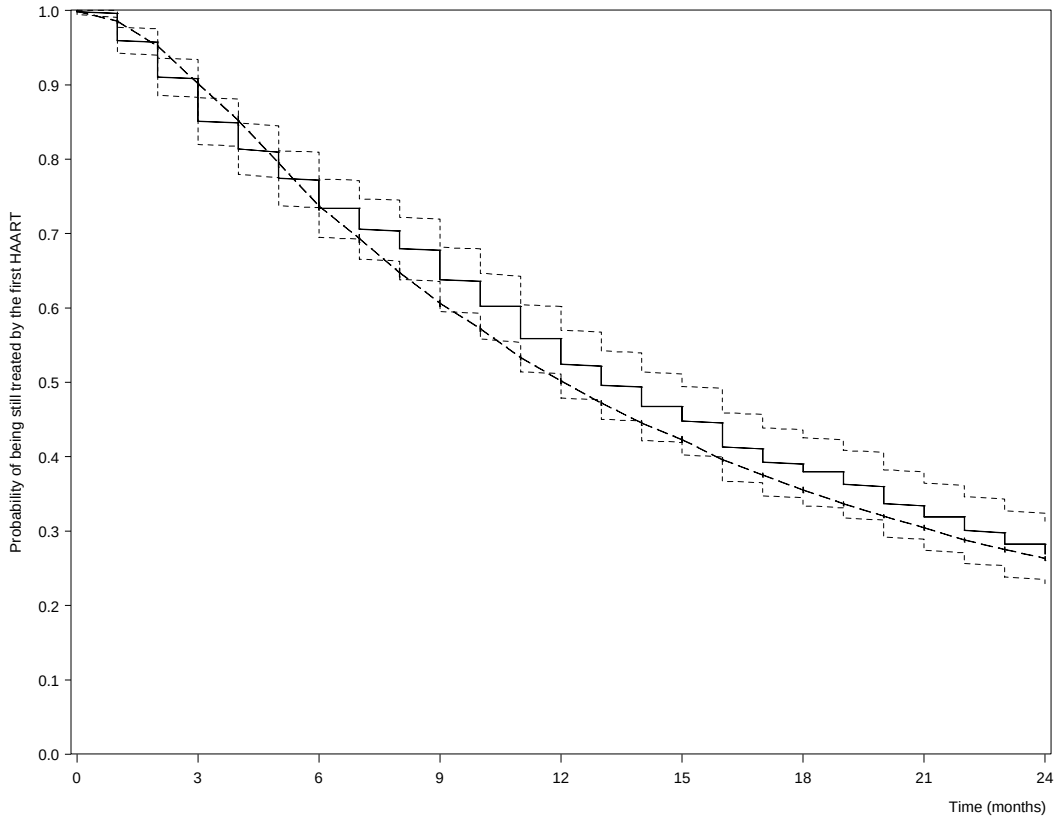


Figure 3. Probability of being still treated by the first HAART regimen: Kaplan-Meier estimate (plain line) and log-normal survival distribution function (dashed line). CASCADE collaboration (N=494).



N at risk	494	433	361	299	228	184	144	107	81
Probability (%)	100	89	76	66	53	45	38	32	27

Figure 4a. Predicted change in CD4+ after the initiation of HAART for non IDU patients according to the model used: crude model (univariate model which did not account for informative drop-out), univariate model (accounting for informative drop-out), full model (accounting for informative drop-out and correlation between HIV RNA and CD4+). CASCADE collaboration (N=494).

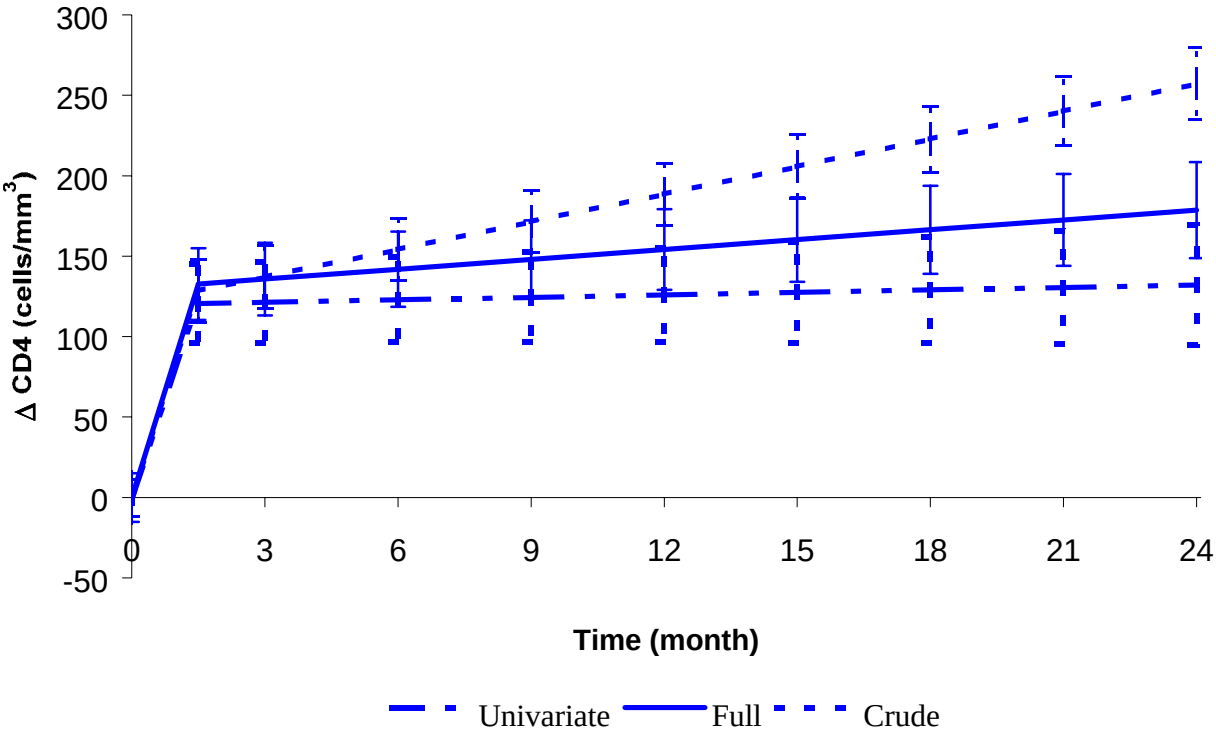
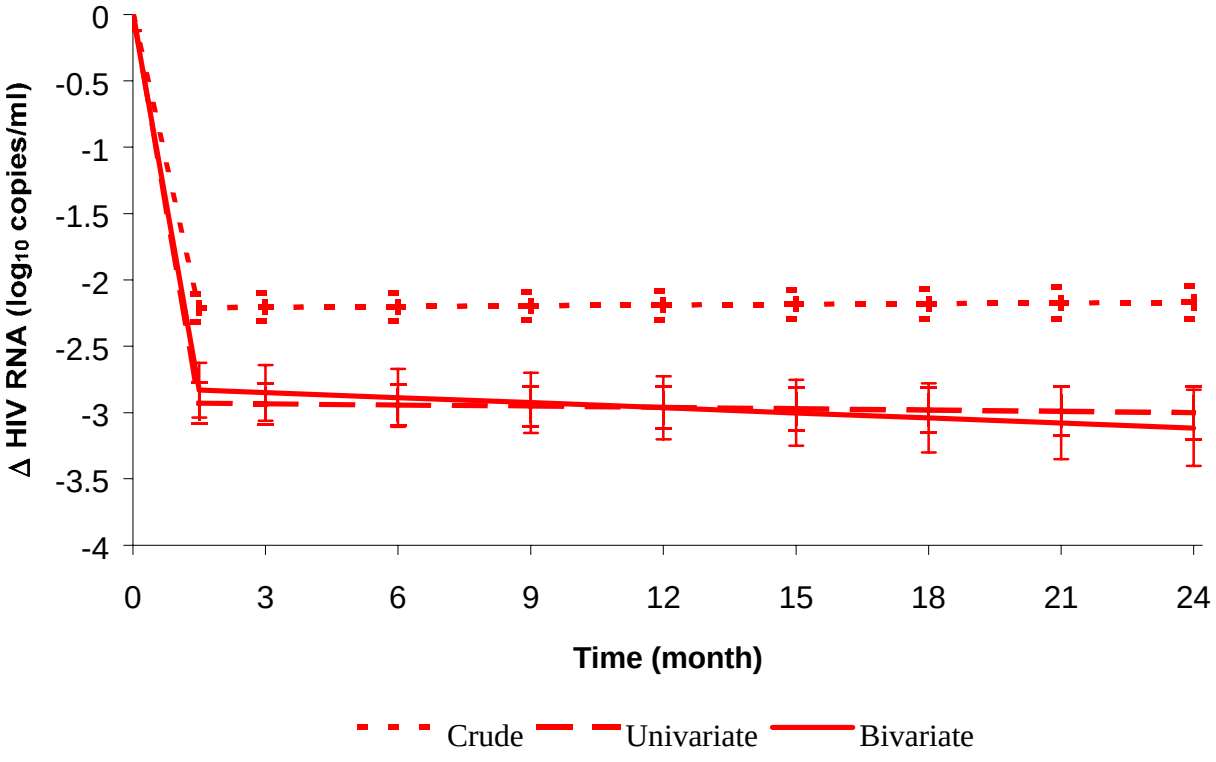


Figure 4b. Predicted change in HIV RNA after the initiation of HAART for non IDU patients according to the model used: crude model (univariate model imputing half of the threshold for left-censored values which did not account for informative drop-out), univariate model (accounting for informative drop-out and left-censoring), bivariate model (accounting for informative drop-out, left-censoring and correlation between HIV RNA and CD4+). CASCADE collaboration (N=494).



6 Discussion

En épidémiologie, les données manquantes sont très fréquentes du fait notamment de l'utilisation d'études observationnelles ou des contraintes liées à l'analyse telles que la modification des traitements. La modélisation conjointe des données longitudinales et d'un temps de sortie d'étude devrait donc être de plus en plus appliquée. De plus, l'étude de l'effet de l'évolution de marqueur (de substitution ou non) sur la survenue d'événements cliniques est aussi une application de ces modèles conjoints.

Une difficulté notable est la définition du processus de sortie d'étude. Dans une première application de la méthode présentée en section 5.2, le processus de données manquantes était défini selon que le patient était encore suivi à la date de point ou non [175]. La sortie d'étude était définie comme prématurée lorsqu'un patient n'était pas décédé et qu'il n'avait pas été revu dans l'année qui précédait la date de point. Cela concernait 14% des patients inclus. La prise en compte de cette sortie prématurée avait peu d'effet sur les autres paramètres du modèle. Ceci provient probablement du fait que les causes de sortie d'étude prématurée sont hétérogènes depuis la disponibilité des traitements antirétroviraux hautement actifs. En effet, certains patients sont perdus de vue du fait d'un mauvais état de santé ou, au contraire, certains abandonnent leur suivi hospitalier du fait d'un bon état de santé. Une des solutions serait donc de prendre en compte les différentes causes de sortie d'étude [133]. Le problème majeur est que l'information des causes de sortie d'étude doit être recueillies.

Des extensions de cette modélisation conjointe ont été présentées dans notre travail afin de prendre en compte plusieurs marqueurs potentiellement censurés à gauche. D'autres développements ont été proposés notamment des modèles plus souples pour les effets aléatoires du modèle mixte [174] ou pour le modèle de survie [131]. Coupler ces différentes approches permettrait une plus grande souplesse du modèle pour des applications différentes et surtout la réalisation aisée d'études de sensibilité.
Cependant, un des problèmes majeurs de ces approches est lié aux difficultés numériques engendrées par le nombre de paramètres et les calculs complexes d'intégrales multiples ou de dérivées. Ces difficultés sont en partie évitées par l'approche bayésienne qui par exemple permet de s'affranchir de l'intégration sur les effets aléatoires. De plus, en utilisant des a priori vagues, cette approche diffère peu de l'approche fréquentiste. En revanche, on peut critiquer la lenteur de convergence et surtout la définition de la convergence en comparaison avec

l'approche fréquentiste. Parmi les approches fréquentistes utilisées dans ce contexte, on distingue l'algorithme EM et les algorithmes de type Newton-Raphson. Lindström et Bates décrivaient ce dernier algorithme comme étant le plus rapide [95]. Cependant, l'algorithme EM semble particulièrement adapté pour ces problématiques liées à des données incomplètes. De plus, des évolutions récentes de l'algorithme EM (PX-EM, par exemple) semblent permettre une convergence plus rapide de l'algorithme.

La modélisation employée dans cette thèse est descriptive. Il s'agissait de résumer au mieux l'évolution des marqueurs et d'analyser l'effet sur cette évolution de variables explicatives mesurées. Il existe une autre approche plus explicative utilisée également pour la modélisation de la charge virale et des CD4+ dans le cadre de l'infection par le VIH : les équations différentielles. Cette approche peut être considérée comme explicative puisque le modèle est construit à partir des hypothèses physiopathologiques d'interaction entre ces deux marqueurs ou plus précisément d'interaction entre le virus et les cellules cibles [176]. On peut ainsi évaluer l'effet d'une molécule en prenant en compte le lieu d'action exacte de la molécule (l'enzyme de transcription inverse pour les inhibiteurs de cette enzyme, par exemple) [177]. Ce type de modélisation a donc été récemment proposé comme alternative aux critères de jugement classiques dans les essais thérapeutiques sur le VIH [46]. Cependant, de nombreux aspects sont encore à développer pour ce type de modèle. Tout d'abord, ils ont le plus souvent été utilisés à des fins de simulation plutôt que d'estimation sur des données réelles. L'approche par effets aléatoires pour estimer les paramètres de ces modèles n'a été proposée que récemment [178]. De plus, l'ajout de composantes stochastiques reflétant à la fois le caractère aléatoire de la relation entre le virus et les cellules et l'incertitude du modèle par rapport à la vérité biologique est récent et peu développé [154, 179]. En pratique, l'utilisation de cette approche pour l'analyse de données réelles nécessite de trouver les solutions de systèmes d'équations différentielles stochastiques non linéaires et également d'estimer les paramètres de ces solutions. Il s'agit donc d'une perspective de recherche nécessitant une approche multidisciplinaire.

7 Conclusion

L'infection par le VIH est un bon exemple de pathologie où les questions clinico-épidémiologiques sont de plus en plus complexes nécessitant des outils biostatistiques toujours plus sophistiqués. Heureusement, cette complexité est souvent liée à une amélioration de la prise en charge des patients infectés par le VIH. En effet, la disponibilité de la charge virale plasmatique a permis de mieux comprendre la physiopathologie du VIH et de mieux suivre les patients ; Les traitements antirétroviraux ont rendu l'évolution des marqueurs viro-immunologiques non linéaire parce qu'ils sont efficaces. Il est toutefois difficile de promouvoir l'utilisation de méthodes biostatistiques complexes quand bien même elles sont utiles. L'utilité de ces méthodes est souvent la première question qui suit leur présentation. Dans l'exemple de la censure de la charge virale, les techniques naïves d'analyse statistique sous-estiment de plus d'un log la décroissance de la charge virale après l'initiation d'un traitement antirétroviral. Cette sous-estimation est cliniquement hautement significative. De plus, ne pas analyser la charge virale en continu parce qu'elle est censurée peut engendrer une perte de puissance faisant faussement conclure à une absence d'effet pronostique. Montrer que ces méthodes sont utiles ne suffit pas à leur promotion : il faut qu'elles soient accessibles au plus grand nombre. Il est donc nécessaire que les chercheurs rendent disponibles leur méthode dans des logiciels classiques à l'instar du modèle de Diggle et Kenward programmé sous S-PLUS (module OSWALD) [114].

Un autre point concerne les limites de ces méthodes : aucune méthode biostatistique ne remplacera un recueil de données adaptées. Dans l'exemple des données manquantes informatives, il est crucial de recueillir les raisons de sortie d'étude ou de modification de traitement plutôt que de tester des hypothèses invérifiables de dépendance à des données absentes. Dans le cas contraire, quelle que soit la complexité du modèle, celui-ci repose sur des hypothèses nécessaires à la modélisation de données manquantes. Ces limites doivent être gardées à l'esprit en particulier lors de l'interprétation des paramètres du modèle dans les applications.

Références bibliographiques

1. Centers for Disease Control. Pneumocystis Pneumonia - Los Angeles. *Morbidity and Mortality Weekly Report* 1981; **30**:250-252.

2. Survival after introduction of HAART in people with known duration of HIV-1 infection. The CASCADE Collaboration. Concerted Action on SeroConversion to AIDS and Death in Europe. *Lancet* 2000; **355**:1158-1159.

3. Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, Berger EA. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science* 1996; **272**:1955-1958.

4. Levy JA. Infection by human immunodeficiency virus--CD4 is not enough. *The New England Journal of Medicine* 1996; **335**:1528-1530.

5. Pantaleo G, Graziosi C, Fauci AS. New concepts in the immunopathogenesis of human immunodeficiency virus infection. *The New England Journal of Medicine* 1993; **328**:327-335.

6. Pantaleo G, Graziosi C, Demarest JF, Butini L, Montroni M, Fox CH, Orenstein JM, Kotler DP, Fauci AS. HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. *Nature* 1993; **362**:355-358.

7. Fauci AS. Multifactorial nature of human immunodeficiency virus disease: implications for therapy. *Science* 1993; **262**:1011-1018.

8. Centers for Disease Control. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Morbidity and Mortality Weekly Report* 1992; **41**:1-19.

9. Autran B, Carcelain G, Li TS, Blanc C, Mathez D, Tubiana R, Katlama C, Debre P, Leibowitch J. Positive effects of combined antiretroviral therapy on CD4+ T cell homeostasis and function in advanced HIV disease. *Science* 1997; **277**:112-116.

10. Chun TW, Fauci AS. Latent reservoirs of HIV: obstacles to the eradication of virus. *Proceedings of the National Academy of Sciences of the United States of America* 1999; **96**:10958-10961.

11. Chun TW, Davey RT, Jr., Engel D, Lane HC, Fauci AS. Re-emergence of HIV after stopping therapy. *Nature* 1999; **401**:874-875.

12. Malone JL, Simms TE, Gray GC, Wagner KF, Burge JR, Burke DS. Sources of variability in repeated T-helper lymphocyte counts from human immunodeficiency virus type 1-infected patients: total lymphocyte count fluctuations and diurnal cycle are important. *Journal of Acquired Immune Deficiency Syndromes* 1990; **3**:144-151.

13. Raboud JM, Montaner JS, Conway B, Haley L, Sherlock C, MV OS, Schechter MT. Variation in plasma RNA levels, CD4 cell counts, and p24 antigen levels in clinically stable men with human immunodeficiency virus infection. *The Journal of Infectious Diseases* 1996; **174**:191-194.

14. Anastassopoulou CG, Touloumi G, Katsoulidou A, Hatzitheodorou H, Pappa M, Paraskevis D, Lazanas M, Gargalianos P, Hatzakis A. Comparative evaluation of the QUANTIPLEX HIV-1 RNA 2.0 and 3.0 (bDNA) assays and the AMPLICOR HIV-1 MONITOR v1.5 test for the quantitation of human immunodeficiency virus type 1 RNA in plasma. *Journal of Virological Methods* 2001; **91**:67-74.

15. Elbeik T, Charlebois E, Nassos P, Kahn J, Hecht FM, Yajko D, Ng V, Hadley K. Quantitative and cost comparison of ultrasensitive human immunodeficiency virus type 1 RNA viral load assays: Bayer bDNA quantiplex versions 3.0 and 2.0 and Roche PCR Amplicor monitor version 1.5. *Journal of Clinical Microbiology* 2000; **38**:1113-1120.

16. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in Medicine* 1989; **8**:415-425.

17. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8**:431-440.

18. Polk BF, Fox R, Brookmeyer R, Kanchanaraksa S, Kaslow R, Visscher B, Rinaldo C, Phair J. Predictors of the acquired immunodeficiency syndrome developing in a cohort of seropositive homosexual men. *The New England Journal of Medicine* 1987; **316**:61-66.

19. Taylor JM, Fahey JL, Detels R, Giorgi JV. CD4 percentage, CD4 number, and CD4:CD8 ratio in HIV infection: which to choose and how to use. *Journal of Acquired Immune Deficiency Syndromes* 1989; **2**:114-124.

20. Fahey JL, Taylor JM, Detels R, Hofmann B, Melmed R, Nishanian P, Giorgi JV. The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. *The New England Journal of Medicine* 1990; **322**:166-172.

21. Phillips AN, Lee CA, Elford J, Webster A, Janossy G, Griffiths PD, Kernoff PB. p24 antigenaemia, CD4 lymphocyte counts and the development of AIDS. *AIDS* 1991; **5**:1217-1222.

22. Ledergerber B, Egger M, Erard V, Weber R, Hirschel B, Furrer H, Battegay M, Vernazza P, Bernasconi E, Opravil M, Kaufmann D, Sudre P, Francioli P, Telenti A. AIDS-related opportunistic illnesses occurring after initiation of potent antiretroviral therapy: the Swiss HIV Cohort Study. *Journal of the American Medical Association* 1999; **282**:2220-2226.

23. Chêne G, Binquet C, Moreau JF, Neau D, Pellegrin I, Malvy D, Ceccaldi J, Lacoste D, Dabis F. Changes in CD4+ cell count and the risk of opportunistic infection or death after highly active antiretroviral treatment. Groupe d'Epidemiologie Clinique du SIDA en Aquitaine. *AIDS* 1998; **12**:2313-2320.

24. Choi S, Lagakos SW, Schooley RT, Volberding PA. CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Annals of Internal Medicine* 1993; **118**:674-680.

25. De Gruttola V, Wulfsohn M, Fischl MA, Tsiatis A. Modeling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *Journal of Acquired Immune Deficiency Syndromes* 1993; **6**:359-365.

26. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. *Statistics in Medicine* 1993; **12**:835-842.

27. Fleming TR. Surrogate markers in AIDS and cancer trials. *Statistics in Medicine* 1994; **13**:1423-1435.

28. Albert JM, Ioannidis JPA, Reichelderfer P, Conway B, Coombs RW, Crane L, Demasi R, Dixon DO, Flandre P, Hughes MD, Kalish LA, Larntz K, Lin DY, Marschner IC, Munoz A, Murray J, Neaton J, Pettinelli C, Rida W, Taylor JMG, Welles SL. Statistical issues for HIV surrogate endpoints: Point/counterpoint. *Statistics in Medicine* 1998; **17**:2435-2462.

29. Hughes MD, Daniels MJ, Fischl MA, Kim S, Schooley RT. CD4 cell count as a surrogate endpoint in HIV clinical trials: a meta-analysis of studies of the AIDS Clinical Trials Group. *AIDS* 1998; **12**:1823-1832.

30. Raboud J, Reid N, Coates RA, Farewell VT. Estimating risks of progressing to AIDS when covariates are measured with error. *Journal of the Royal Statistical Society: Series A* 1993; **156**:393-406.

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

32. Hu P, Tsiatis AA, Davidian M. Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* 1998; **54**:1407-1419.

33. De Gruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**:1003-1014.

34. Mellors JW, Rinaldo CR, Jr., Gupta P, White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 1996; **272**:1167-1170.

35. Mellors JW, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP, Rinaldo CR, Jr. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine* 1997; **126**:946-954.

36. Egger M, May M, Chene G, Phillips AN, Ledergerber B, Dabis F, Costagliola D, Monforte AD, deWolf F, Reiss P, Lundgren JD, Justice AC, Staszewski S, Leport C, Hogg RS, Sabin CA, Gill MJ, Salzberger B, Sterne JAC. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *Lancet* 2002; **360**:119-129.

37. Ghani AC, deWolf F, Ferguson NM, Donnelly CA, Coutinho R, Miedema F, Goudsmit J, Anderson RM. Surrogate markers for disease progression in treated HIV infection. *Journal of Acquired Immune Deficiency Syndromes* 2001; **28**:226-231.

38. O'Brien WA, Hartigan PM, Martin D, Esinhart J, Hill A, Benoit S, Rubin M, Simberkoff MS, Hamilton JD. Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. Veterans Affairs Cooperative Study Group on AIDS. *The New England Journal of Medicine* 1996; **334**:426-431.

39. Aboulker JP, Babiker AG, Flandre P, Gazzard B, Loveday C, Nunn AJ, Goudsmit J, Huraux JM, vanderNoorda J, Weiss R, Boucher C, Schurrman R, BrunVezinet F, Descamps D, Jeffries D, Tedder R, Weber J, Darbyshire JH, Reiss P, Weverling G. An evaluation of HIV RNA and CD4 cell count as surrogates for clinical outcome. *AIDS* 1999; **13**:565-573.

40. Human immunodeficiency virus type 1 RNA level and CD4 count as prognostic markers and surrogate end points: a meta-analysis. HIV Surrogate Marker Collaborative Group. *AIDS Research and Human Retroviruses* 2000; **16**:1123-1133.

41. Gilbert PB, DeGruttola V, Hammer SM, Kuritzkes DR. Virologic and regimen termination surrogate end points in AIDS clinical trials. *Journal of the American Medical Association* 2001; **285**:777-784.

42. Yeni PG, Hammer SM, Carpenter CCJ, Cooper DA, Fischl MA, Gatell JM, Gazzard BG, Hirsch MS, Jacobsen DM, Katzenstein DA, Montaner JSG, Richman DD, Saag MS, Schechter M, Schooley RT, Thompson MA, Vella S, Volberding PA. Antiretroviral treatment for adult HIV infection in 2002 - Updated recommendations of the international AIDS Society-USA panel. *Journal of the American Medical Association* 2002; **288**:222-235.

43. Guidelines for using antiretroviral agents among HIV-infected adults and adolescents. Recommendations of the Panel on Clinical Practices for Treatment of HIV. *Recommendations and reports : Morbidity and mortality weekly report* 2002; **51**:1-55.

44. Kaplan JE, Masur H, Holmes KK. Guidelines for preventing opportunistic infections among HIV-infected persons--2002. Recommendations of the U.S. Public Health Service and the Infectious Diseases Society of America. *Recommendations and reports : Morbidity and mortality weekly report* 2002; **51**:1-52.

45. Delfraissy JF. *Prise en charge des personnes infectées par le VIH. Recommandations du groupe d'experts.* Paris: Médecine-Sciences Flammarion, 2002:384 p.

46. De Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials: Summary of a National Institutes of Health Workshop. *Controlled Clinical Trials* 2001; **22**:485-502.

47. Molina JM, Chêne G, Ferchal F, Journot V, Pellegrin I, Sombardier MN, Rancinan C, Cotte L, Madelaine I, Debord T, Decazes JM. The ALBI trial: A randomized controlled trial comparing stavudine plus didanosine with zidovudine plus lamivudine and a regimen alternating both combinations in previously untreated patients infected with human immunodeficiency virus. *The Journal of Infectious Diseases* 1999; **180**:351-358.

48. Journot V, Chêne G, Joly P, Savès M, Jacqmin-Gadda H, Molina JM, Salamon R. Viral load as a primary outcome in human immunodeficiency virus trials: A review of statistical analysis methods. *Controlled Clinical Trials* 2001; **22**:639-658.

49. Phillips AN, Grabar S, Tassie JM, Costagliola D, Lundgren JD, Egger M. Use of observational databases to evaluate the effectiveness of antiretroviral therapy for HIV infection: comparison of cohort studies with randomized trials. *AIDS* 1999; **13**:2075-2082.

50. Dunn D, Babiker A, Hooker M, Darbyshire J. The dangers of inferring treatment effects from observational data: a case study in HIV infection. *Controlled Clinical Trials* 2002; **23**:106-110.

51. Munoz A, Carey V, Saah AJ, Phair JP, Kingsley LA, Fahey JL, Ginzburg HM, Polk BF. Predictors of decline in CD4 lymphocytes in a cohort of homosexual men infected with human immunodeficiency virus. *Journal of Acquired Immune Deficiency Syndromes* 1988; **1**:396-404.

52. De Gruttola V, Lange N, Dafni U. Modeling the progression of HIV infection. *Journal of the American Statistical Association* 1991; **86**:569-577.

53. Galai N, Vlahov D, Margolick JB, Chen K, Graham NM, Munoz A. Changes in markers of disease progression in HIV-1 seroconverters: a comparison between cohorts of injecting drug users and homosexual men. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 1995; **8**:66-74.

54. Prins M, Robertson JR, Brettle RP, Aguado IH, Broers B, Boufassa F, Goldberg DJ, Zangerle R, Coutinho RA, vandenHoek A. Do gender differences in CD4 cell counts matter? *AIDS* 1999; **13**:2361-2364.

55. Henrard DR, Phillips JF, Muenz LR, Blattner WA, Wiesner D, Eyster ME, Goedert JJ. Natural history of HIV-1 cell-free viremia. *Journal of the American Medical Association* 1995; **274**:554-558.

56. Paxton WB, Coombs RW, McElrath MJ, Keefer MC, Hughes J, Sinangil F, Chernoff D, Demeter L, Williams B, Corey L. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with > or = 400 CD4 lymphocytes: implications for applying measurements to individual patients. National Institute of Allergy and Infectious Diseases AIDS Vaccine Evaluation Group. *The Journal of Infectious Diseases* 1997; **175**:247-254.

57. Keet IP, Janssen M, Veugelers PJ, Miedema F, Klein MR, Goudsmit J, Coutinho RA, de Wolf F. Longitudinal analysis of CD4 T cell counts, T cell reactivity, and human immunodeficiency virus type 1 RNA levels in persons remaining AIDS-free despite CD4 cell counts <200 for >5 years. *The Journal of Infectious Diseases* 1997; **176**:665-671.

58. Lyles CM, Dorrucci M, Vlahov D, Pezzotti P, Angarano G, Sinicco A, Alberici F, Alcorn TM, Vella S, Rezza G. Longitudinal human immunodeficiency virus type 1 load in the Italian seroconversion study: Correlates and temporal trends of virus load. *The Journal of Infectious Diseases* 1999; **180**:1018-1024.

59. Sterling TR, Lyles CM, Vlahov D, Astemborski J, Margolick JB, Quinn TC. Sex differences in longitudinal human immunodeficiency virus type 1 RNA levels among seroconverters. *The Journal of Infectious Diseases* 1999; **180**:666-672.

60. Gandhi M, Bacchetti P, Miotti P, Quinn TC, Veronese F, Greenblatt RM. Does patient sex affect human immunodeficiency virus levels? *Clinical Infectious Diseases* 2002; **35**:313-322.

61. Napravnik S, Poole C, Thomas JC, Eron Jr JJ. Gender Difference in HIV RNA Levels: A Meta-Analysis of Published Studies. *Journal of Acquired Immune Deficiency Syndromes* 2002; **31**:11-19.

62. Farzadegan H, Hoover DR, Astemborski J, Lyles CM, Margolick JB, Markham RB, Quinn TC, Vlahov D. Sex differences in HIV-1 viral load and progression to AIDS. *Lancet* 1998; **352**:1510-1514.

63. Staszewski S, Miller V, Sabin C, Carlebach A, Berger AM, Weidmann E, Helm EB, Hill A, Phillips A. Virological response to protease inhibitor therapy in an HIV clinic cohort. *AIDS* 1999; **13**:367-373.

64. Ghani AC, Henley WE, Donnelly CA, Mayer S, Anderson RM. Comparison of the effectiveness of non-nucleoside reverse transcriptase inhibitor-containing and protease inhibitor-containing regimens using observational databases. *AIDS* 2001; **15**:1133-1142.

65. Deeks SG, Hecht FM, Swanson M, Elbeik T, Loftus R, Cohen PT, Grant RM. HIV RNA and CD4 cell count response to protease inhibitor therapy in an urban AIDS clinic: Response to both initial and salvage therapy. *AIDS* 1999; **13**:F35-F43.

66. Ledergerber B, Egger M, Opravil M, Telenti A, Hirschel B, Battegay M, Vernazza P, Sudre P, Flepp M, Furrer H, Francioli P, Weber R. Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. *Lancet* 1999; **353**:863-868.

67. Grabar S, LeMoing V, Goujard C, Leport C, Kazatchkine MD, Costagliola D, Weiss L. Clinical outcome of patients with HIV-1 infection according to immunologic and virologic response after 6 months of highly active antiretroviral therapy. *Annals of Internal Medicine* 2000; **133**:401-410.

68. Le Moing V, Chêne G, Carrieri MP, Besnier JM, Masquelier B, Salamon R, Bazin C, Moatti JP, Raffi F, Leport C. Clinical, biologic, and behavioral predictors of early immunologic and virologic response in HIV-infected patients initiating protease inhibitors. *Journal of Acquired Immune Deficiency Syndromes* 2001; **27**:372-376.

69. Carrieri P, Cailleton V, Le Moing V, Spire B, Dellamonica P, Bouvet E, Raffi F, Journot V, Moatti JP. The dynamic of adherence to highly active antiretroviral therapy: results from the French National APROCO cohort. *Journal of Acquired Immune Deficiency Syndromes* 2001; **28**:232-239.

70. Phillips AN, Staszewski S, Weber R, Kirk O, Francioli P, Miller V, Vernazza P, Lundgren JD, Ledergerber B. HIV viral load response to antiretroviral therapy according to the baseline CD4 cell count and viral load. *Journal of the American Medical Association* 2001; **286**:2560-2567.

71. Renaud M, Katlama C, Mallet A, Calvez V, Carcelain G, Tubiana R, Jouan M, Caumes E, Agut H, Bricaire F, Debre P, Autran B. Determinants of paradoxical CD4 cell reconstitution after protease inhibitor-containing antiretroviral regimen. *AIDS* 1999; **13**:669-676.

72. Staszewski S, Miller V, Sabin C, Schlecht C, Gute P, Stamm S, Leder T, Berger A, Weidemann E, Hill A, Phillips A. Determinants of sustainable CD4 lymphocyte count increases in response to antiretroviral therapy. *AIDS* 1999; **13**:951-956.

73. Deeks SG, Barbour JD, Martin JN, Swanson MS, Grant RM. Sustained CD4+ T cell response after virologic failure of protease inhibitor-based regimens in patients with human immunodeficiency virus infection. *The Journal of Infectious Diseases* 2000; **181**:946-953.

74. Le Moing V, Thiébaut R, Chêne G, Leport C, Moatti JP, Michelet C, Fleury H, Herson S, Raffi F. Predictors of long-term increase of CD4+ cell count in human immunodeficiency virus-infected patients initiating a protease inhibitor-containing regimen. *The Journal of Infectious Diseases* 2002; **185**:471-480.

75. Moore AL, Mocroft A, Madge S, Devereux H, Wilson D, Phillips AN, Johnson M. Gender differences in virologic response to treatment in an HIV- positive population: A cohort study. *Journal of Acquired Immune Deficiency Syndromes* 2001; **26**:159-163.

76. Zaccarelli M, Barracchini A, De Longis P, Perno CF, Soldani F, Liuzzi G, Serraino D, Ippolito G, Antinori A. Factors related to virologic failure among HIV-positive injecting drug users treated with combination antiretroviral therapy including two nucleoside reverse transcriptase inhibitors and nevirapine. *AIDS Patient Care and STDS* 2002; **16**:67-73.

77. Pezzotti P, Pappagallo M, Phillips AN, Boros S, Valdarchi C, Sinicco A, Zaccarelli M, Rezza G. Response to highly active antiretroviral therapy according to duration of HIV infection. *Journal of Acquired Immune Deficiency Syndromes* 2001; **26**:473-479.

78. Moyle GJ, Gazzard BG, Cooper DA, Gatell J. Antiretroviral therapy for HIV infection. A knowledge-based approach to drug selection and use. *Drugs* 1998; **55**:383-404.

79. Connors M, Kovacs JA, Krevat S, Gea-Banacloche JC, Sneller MC, Flanigan M, Metcalf JA, Walker RE, Falloon J, Baseler M, Feuerstein I, Masur H, Lane HC. HIV infection induces changes in CD4+ T-cell phenotype and depletions within the CD4+ T-cell repertoire that are not immediately restored by antiviral or immune-based therapies. *Nature Medicine* 1997; **3**:533-540.

80. Gorochov G, Neumann AU, Kereveur A, Parizot C, Li T, Katlama C, Karmochkine M, Raguin G, Autran B, Debre P. Perturbation of CD4+ and CD8+ T-cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. *Nature Medicine* 1998; **4**:215-221.

81. Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association* 1997; **92**:775-779.

82. Sy JP, Taylor JM, Cumberland WG. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics* 1997; **53**:542-555.

83. Boscardin WJ, Taylor JM, Law N. Longitudinal models for AIDS marker data. *Statistical Methods in Medical Research* 1998; **7**:13-27.

84. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in Medicine* 1999; **18**:1215-1233.

85. Lyles RH, Lyles CM, Taylor DJ. 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* 2000; **49**:485-497.

86. Chavance M. [Modeling correlated data in epidemiology: mixed or marginal model?]. *Revue d'Epidemiologie et de Sante Publique* 1999; **47**:535-544.

87. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine* 1998; **17**:1261-1291.

88. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963-974.

89. Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics* 1988; **44**:959-971.

90. Jones RH, Boadi-Boateng F. Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* 1991; **47**:161-175.

91. Chi E, Reinsel G. Models for longitudinal datawith random effects and AR(1) errors. *Journal of the American Statistical Association* 1989; **84**:452-459.

92. Lesaffre E, Asefa M, Verbeke G. Assessing the goodness-of-fit of the Laird and Ware model--an example: the Jimma Infant Survival Differential Longitudinal Study. *Statistics in Medicine* 1999; **18**:835-854.

93. Taylor JM, Cumberland WG, Sy JP. A stochastic model for the analysis of longitudinal AIDS data. *Journal of the American Statistical Association* 1994; **89**:727-736.

94. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. New York: Springer, 2000:568 p.

95. Lindstrom MJ, Bates DM. Newton-Raphson and EM Algorithm for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 1988; **83**:1014-1022.

96. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models*. Cary, NC: SAS Institute, 1996:656 p.

97. Smith AFM. A general Bayesian linear model. *Journal of the Royal Statistical Society: Series B* 1973; **35**:67-75.

98. Henderson CR. *Applications of Linear Models in Animal Breeding*. Guelph, Canada: University of Guelph Press, 1984:423 p.

99. Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 1996; **91**:217-221.

100. Verbeke G, Molenberghs G. Case studies. *Linear mixed models for longitudinal data*. New York: Springer; 2000:405-411.

101. Carey VJ, Rosner BA. Analysis of longitudinally observed irregularly timed multivariate outcomes: regression with focus on cross-component correlation. *Statistics in Medicine* 2001; **20**:21-31.

102. Gray SM, Brookmeyer R. Estimating a treatment effect from multidimensional longitudinal data. *Biometrics* 1998; **54**:976-988.

103. Schluchter MD. Estimating correlation between alternative measures of disease progression in a longitudinal study. Modification of Diet in Renal Disease Study. *Statistics in Medicine* 1990; **9**:1175-1188.

104. Jones RH. Multivariate models. *Longitudinal data with serial correlation: a state-space approach*. London: Chapman & Hall; 1993:156-185. Monographs on statistics and applied probability 47;

105. Brown ER, MaWhinney S, Jones RH, Kafadar K, Young B. Improving the fit of bivariate smoothing splines when estimating longitudinal immunological and virological markers in HIV patients with individual antiretroviral treatment strategies. *Statistics in Medicine* 2001; **20**:2489-2504.

106. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* 2000; **56**:1047-1054.

107. Cumberland WG, Rohde CA. A multivariate model for growth of populations. *Theoretical Population Biology* 1977; **11**:127-139.

108. Taylor JM, Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* 1998; **17**:2381-2394.

109. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**.

110. Heyting A, Tolboom JT, Essers JG. Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in Medicine* 1992; **11**:2043-2061.

111. Shih WJ, Quan H. Testing for treatment differences with dropouts present in clinical trials--a composite approach. *Statistics in Medicine* 1997; **16**:1225-1239.

112. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases* 1967; **20**:637-648.

113. Little R, Rubin D. *Statistical analysis with missing data*. New York: John Wiley & Sons, 1987:278 p.

114. Diggle P, Kenward M. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C* 1994; **43**:49-93.

115. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; **44**:175-188.

116. Hogan J, Laird N. Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* 1997; **16**:259-272.

117. Little R. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:1112-1121.

118. Rubin D. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473-489.

119. Ridout MS. Testing for random dropouts in repeated measurement data. *Biometrics* 1991; **47**:1617-1619; discussion 1619-1621.

120. Park T, Lee SY. A test of missing completely at random for longitudinal data with missing observations. *Statistics in Medicine* 1997; **16**:1859-1871.

121. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13-22.

122. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine* 1998; **17**:2723-2732.

123. Troxel AB, Harrington DP, Lipsitz SR. Analysis of longitudinal data with non-ignorable non-monotone missing value. *Journal of the Royal Statistical Society: Series C* 1998; **47**:425-438.

124. Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: Theory and application. *Annals of Statistics* 1981; **9**:861-869.

125. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1989; **45**:939-955.

126. Mori M, Woodworth GG, Woolson RF. Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring. *Statistics in Medicine* 1992; **11**:621-631.

127. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 1992; **11**:1861-1870.

128. Pawitan Y, Self S. Modelling disease marker processes in AIDS. *Journal of the American Statistical Association* 1993; **88**:719-726.

129. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 1996; **15**:1663-1685.

130. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53**:330-339.

131. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**:465-480.

132. Wang Y, Taylor JM. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* 2001; **96**:895-905.

133. Xu J, Zeger SL. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C* 2001; **50**:375-387.

134. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982; **69**:331-342.

135. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**:27-37.

136. Bycott P, Taylor J. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* 1998; **17**:2061-2077.

137. LaValley MP, DeGruttola V. Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine* 1996; **15**:2289-2305.

138. Self S, Pawitan Y. Modeling a marker of disease progression and onset of disease. In: Jewell N, Dietz K, Farewell V, eds. *AIDS epidemiology: methodological issues*. Boston: Birkhaüser; 1992:231-255.

139. Molenberghs G, Williams PL, Lipsitz SR. Prediction of survival and opportunistic infections in HIV-infected patients: a comparison of imputation methods of incomplete CD4 counts. *Statistics in Medicine* 2002; **21**:1387-1408.

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

141. Huang W, Zeger SL, Anthony JC, Garrett E. Latent variable model for joint analysis of multiple repeated measures and bivariate event times. *Journal of the American Statistical Association* 2001; **96**:906-914.

142. Singh A, Nocerino J. Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometrics and Intelligent Laboratory Systems* 2002; **60**:69-86.

143. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995; **51**:1570-1578.

144. Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine* 2001; **20**:33-45.

145. Lyles RH, Fan D, Chuachoowong R. Correlation coefficient estimation involving a left censored laboratory assay variable. *Statistics in Medicine* 2001; **20**:2921-2933.

146. Kaplan E, Meier P. Non parametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457-481.

147. Amemiya T. Tobit models: a survey. *Journal of Econometrics* 1984; **24**:3-61.

148. Marschner IC, Betensky RA, DeGruttola V, Hammer SM, Kuritzkes DR. Clinical trials using HIV-1 RNA-based primary endpoints: Statistical analysis and potential biases. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 1999; **20**:220-227.

149. Hughes MD. Analysis and design issues for studies using censored biomarker measurements with an example of viral load measurements in HIV clinical trials. *Statistics in Medicine* 2000; **19**:3171-3191.

150. Flandre P, Durier C, Descamps D, Launay O, Joly V. On the use of magnitude of reduction in HIV-1 RNA in clinical trials: Statistical analysis and potential biases. *Journal of Acquired Immune Deficiency Syndromes* 2002; **30**:59-64.

151. Pettitt AN. Censored observations, repeated measures and mixed effects models: an approach using the EM algorithm and normal errors. *Biometrika* 1986; **73**:635-643.

152. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. *Biometrics* 1999; **55**:625-629.

153. Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* 2000; **1**:355-368.

154. Tuckwell HC, Le Corfec E. A stochastic model for early HIV-1 population dynamics. *Journal of Theoretical Biology* 1998; **195**:451-463.

155. Thiébaut R, Morlat P, Jacqmin-Gadda H, Neau D, Mercie P, Dabis F, Chêne G. Clinical progression of HIV-1 infection according to the viral response during the first year of antiretroviral treatment. *AIDS* 2000; **14**:971-978.

156. Thiébaut R, Chêne G, Jacqmin-Gadda H, Morlat P, Mercié P, Dupon M, Neau D, Ramaroson H, Dabis F, Salamon R. 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* 2002:in press.

157. Thiébaut R, Jacqmin-Gadda H, Ramaroson H, Dabis F, Morlat P, Mercié P, Dupon M, Neau D, Chêne G. CD4+ cell count and HIV RNA as time dependent prognostic factors in HIV-1 infected patients newly treated with antiretroviral combinations. Aquitaine Cohort, 2000. 5th International Workshop on HIV Observational Databases. Monte Carlo, 2001.

158. Genz A. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* 1992; **1**:141-149.

159. Marquardt DW. An algorithm for least squares estimation of nonlinear parameters. *SIAM Journal* 1963; **11**:431-441.

160. Genz A. Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics* 1993; **25**:400-413.

161. Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society: Series C* 2001; **50**:325-335.

162. Hogan JW, Laird NM. Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research* 1998; **7**:28-48.

163. Raab GM, Parpia T. Random effects models for HIV marker data: practical approaches with currently available software. *Statistical Methods in Medical Research* 2001; **10**:101-116.

164. Huang W, DeGruttola V, Fischl M, Hammer S, Richman D, Havlir D, Gulick R, Squires K, Mellors J. Patterns of plasma human immunodeficiency virus type 1 RNA response to antiretroviral therapy. *The Journal of Infectious Diseases* 2001; **183**:1455-1465.

165. Verotta D, Schaedeli F. Non-linear dynamics models characterizing long-term virological data from AIDS clinical trials. *Mathematical Biosciences* 2002; **176**:163-183.

166. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *British Medical Journal* 2001; **323**:1123-1124.

167. Egger M, Hirschel B, Francioli P, Sudre P, Wirz M, Flepp M, Rickenbach M, Malinverni R, Vernazza P, Battegay M. Impact of new antiretroviral combination therapies in HIV infected patients in Switzerland: prospective multicentre study. Swiss HIV Cohort Study. *British Medical Journal* 1997; **315**:1194-1199.

168. Thiébaut R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D. Bivariate linear mixed models using SAS proc MIXED. *Computer Methods and Programs in Biomedicine* 2002; **69**:249-256.

169. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988; **7**:305-315.

170. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. *Epidemiology* 2002; **13**:347-355.

171. Fitzgerald AP, DeGruttola VG, Vaida F. Modelling HIV viral rebound using non-linear mixed effects models. *Statistics in Medicine* 2002; **21**:2093-2108.

172. Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho DD. Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* 1997; **387**:188-191.

173. Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics* 2001; **57**:7-14.

174. Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* 2002:accepted for publication.

175. Thiébaut R, Jacqmin-Gadda H, Chêne G, Commenges D. Bivariate longitudinal study of CD4+ cell count and HIV RNA taking into account informative drop-out and left-censoring of HIV RNA values. 17th International Workshop on Statistical Modelling. Chania, Crete, 2002:623-628.

176. Ding AA, Wu HL. A comparison study of models and fitting procedures for biphasic viral dynamics in HIV-1 infected patients treated with antiviral therapies. *Biometrics* 2000; **56**:293-300.

177. Ding AA, Wu HL. Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics. *Mathematical Biosciences* 1999; **160**:63-82.

178. Wu HL, Ding AA, DeGruttola V. Estimation of HIV dynamic parameters. *Statistics in Medicine* 1998; **17**:2463-2485.

179. Tan WY, Wu H. Stochastic modeling of the dynamics of CD4+ T-cell infection by HIV and some Monte Carlo studies. *Mathematical Biosciences* 1998; **147**:173-205.

180. Vittinghoff E, Malani HM, Jewell NP. Estimating patterns of CD4 lymphocyte decline using data from a prevalent cohort of HIV infected individuals. *Statistics in Medicine* 1994; **13**:1101-1118.

181. McNeil AJ, Gore SM. Statistical analysis of zidovudine (AZT) effect on CD4 cell counts in HIV disease. *Statistics in Medicine* 1996; **15**:75-92.

182. Judge GG, Hill RC, Griffiths WE, Lütkepohl H, Lee T-C. *Introduction to the theory and practice of econometrics*. New York: 1988:1024 p.

Listes des publications

Articles originaux

Thiébaut R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D. Bivariate linear mixed models using SAS proc MIXED. Computer Methods and Programs Biomedicine 2002;69:249-56.

Thiébaut R, Jacqmin-Gadda H, Leport C, Katlama C, Costagliola D, Le Moing V, Morlat P, Chêne G and the APROCO study group. Joint longitudinal modeling of virologic and immunologic response to highly active antiretroviral therapy in HIV-infected patients. APROCO Cohort 1997-2000. Journal of Biopharmaceutical Statistics (en revision)

Thiébaut R, Jacqmin-Gadda H, Babiker A, Commenges D and the CASCADE Collaboration. Bivariate longitudinal study of CD4+ cell count and HIV RNA handling informative drop-out and left-censoring of HIV RNA values
Statistics in Medicine (soumis)

Communications dans un congrès avec comité de lecture

17th International Workshop on Statistical Modelling, Chania – 07/2002
Thiébaut R, Jacqmin-Gadda H, Chêne G, Commenges D.
Bivariate longitudinal study of CD4+ cell count and HIV RNA taking into account informative drop-out and left-censoring of HIV RNA values. *[communication orale O32]*

6th International Workshop on HIV Observational Databases, Sintra – 03/2002
Thiébaut R, Jacqmin-Gadda H, Chêne G, Commenges D.
Bivariate longitudinal study of CD4+ cell count and HIV RNA in patients with known date of seroconversion taking into account informative drop-out and left-censoring of HIV RNA values *[communication affichée session 1]*

Journées "Modèles Mixtes et Biométrie", Paris - 01/2002
Thiébaut R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D.
Modèles linéaires mixtes bivariés à l'aide de SAS proc MIXED. *[communication orale]*

First CASCADE workshop on statistical modelling of longitudinal markers in HIV infection, Bordeaux - 11/2001
Thiébaut R, Jacqmin-Gadda H, Chêne G, Commenges D.
Bivariate longitudinal study of CD4+ cell count and HIV RNA in patients with known date of seroconversion taking into account informative drop-out and left-censoring of HIV RNA values. *[communication orale]*

22nd Annual Conference of The International Society for Clinical Biostatistics, ISCB, Stockholm – 08/2001
Thiébaut R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D
Analysis of the evolution of HIV RNA and CD4 cell count in HIV infection taking into account left-censoring of HIV RNA measures. [*communication orale O35*]

5th International Workshop on HIV Observational Databases, Monte Carlo – 04/2001
Thiébaut R, Jacqmin-Gadda H, Ramaroson H, Dabis F, Morlat P, Mercié P, Dupon M, Neau D, Chêne G and the Groupe d'Epidémiologie Clinique du SIDA en Aquitaine (GECSA)
CD4+ cell count and HIV RNA as time dependent prognostic factors in HIV-1 infected patients newly treated with antiretroviral combinations. Aquitaine Cohort, 2000. [*communication orale session 2*]

Premier congrès de Biométrie et Epidémiologie (Société Française de biométrie et Association des Epidémiologistes de Langue Française), Vannes – 10/1999
Thiébaut R, Journot V, Jacqmin-Gadda H, Ferchal F, Chêne G.
Analyse de la réduction de la charge virale plasmatique du VIH-1 tenant compte des données mesurées indétectables dans l'essai thérapeutique ALBI-ANRS 070. *[communication orale]*

Annexes

Annexe 1 : Liste des abréviations

ADN : Acide DésoxyriboNucléïque

APROCO : AntiPROtéase Cohorte, Cohorte française de patients traités par un traitement antirétroviral contenant au moins une antiprotéase

ARN : Acide RiboNucléïque

AR(1) : Processus autorégressif d'ordre 1

CASCADE : Concerted Action on SeroConversion to AIDS and Death in Europe, collaboration multi-cohorte regroupant des patients à date de séroconversion connue

CD4+ : Lymphocytes T CD4+ totaux circulants

CRD : Completely Random Drop-out, sortie d'étude complètement aléatoire (voir MCAR)

EM : Algorithme EM pour "Expectation Maximisation"

GEE : Generalized Estimating Equations, équations d'estimation généralisées

HAART : Highly Active Antiretroviral Therapy, traitement antirétroviral hautement actif composé de plusieurs molécules (habituellement trois, i.e. trithérapies)

ID : Informative Drop-out, sortie d'étude informative (voir MNAR)

IOU : processus d'Ornstein-Uhlenbeck intégré (Integrated Ornstein-Uhlenbeck)

MAR : Missing At Random, données manquantes aléatoirement

MCAR : Missing Completely At Random, données manquantes complètement aléatoirement

MCEM : "Monte Carlo Expectation Maximisation" algorithme

MCMC : Markov Chain Monte Carlo, algorithme utilisé dans les approches bayésiennes

ML : Maximum Likelihood, maximum de vraisemblance

MNAR : Missing Not At Random, données manquantes non aléatoires

OU : processus d'Ornstein-Uhlenbeck ou processus autorégressif d'ordre 1 en temps continu

PMLE : Pseudo Maximum Likelihood Estimator

RD : Random Drop-out

SIDA : Syndrome d'ImmunoDéficience Acquise (AIDS pour Acquired ImmunoDeficiency Syndrome)

VIH : Virus de l'Immunodéficience Humaine (HIV pour Human Immunodeficiency Virus)

Annexe 2 : Transformation des CD4+

Sur l'échelle naturelle, la numération des lymphocytes T CD4+ conduit a des résidus souvent hétéroscédastes avec une variance augmentant avec la valeur de la prédiction (figure 3). Il est donc souvent nécessaire de travailler sur une transformation des CD4+. Les transformations les plus fréquentes sont le logarithme népérien [135] et les racines [52, 108, 180, 181]. Il a également été proposé de travailler sur les différences par rapport à la valeur initiale [93]. Ce dernier point engendre des modifications plus complexes qu'une simple transformation et il est détaillé en annexe 4.

Lorsque plusieurs transformations sont éligibles suite à l'étude des résidus, une difficulté réside dans le choix de la transformation à garder pour la suite des analyses. Une façon relativement simple est de comparer les vraisemblances corrigées par le Jacobien de la transformation [182].

Soit X une variable aléatoire continue de densité connue $f_X(.)$ et $Y = g(X)$ avec g une fonction bijective monotone de X. Il existe donc une fonction inverse de $g(.)$ notée $g^{-1}(.)$. On peut alors calculer la densité de $Y = g(X)$ notée $f_Y(.)$:

$$f_Y(y) = f_X\left[g^{-1}(y)\right] \cdot \left|\frac{\partial g^{-1}(y)}{\partial y}\right|$$

Cette expression signifie que pour trouver la distribution de Y, il faut résoudre $y = g(x)$ pour x et substituer cette solution pour x dans $f_X(x)$ et ensuite multiplier par la valeur absolue de la dérivée $\frac{\partial g^{-1}(y)}{\partial y}$. Dans ce contexte, cette dérivée est le Jacobien de la transformation et $\left|\frac{\partial g^{-1}(y)}{\partial y}\right|$ est la valeur absolue du Jacobien. Cette théorie s'applique également au cas multivarié où $X = X_1,...X_n$ et $Y = Y_1,...Y_n$ sont des vecteurs de variables aléatoires. Le Jacobien de la transformation est :

$$J = \det\begin{bmatrix} \frac{\partial g^{-1}(y_1)}{\partial y_1} & \cdots & \frac{\partial g^{-1}(y_1)}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g^{-1}(y_n)}{\partial y_1} & \cdots & \frac{\partial g^{-1}(y_n)}{\partial y_n} \end{bmatrix}$$ où 'det' correspond au déterminant de la matrice.

En terme de vraisemblance, on a :

$$L(\theta) = \prod_{i=1}^{N} \left[f_{X_i} \left[g^{-1}(y_i) \right] \cdot |J_i| \right]$$

Avec les transformations simples de type racine carrée ou logarithme, J est diagonale. D'où la log-vraisemblance :

$$l(\theta) = \sum_{i=1}^{N} \log \left[f_{X_i} \left[g^{-1}(y_i) \right] \right] + \sum_{i=1}^{N} \log \left| \det \begin{bmatrix} \frac{\partial g^{-1}(y_{i1})}{\partial y_{i1}} & & \\ & \ddots & \\ & & \frac{\partial g^{-1}(y_{in})}{\partial y_{in}} \end{bmatrix} \right|$$

Par exemple, si on travaille sur la racine carrée des CD4+, la densité connue $f_X(.)$ correspond à la vraisemblance du modèle pour la variable transformée. On cherche à connaître cette vraisemblance pour l'échelle naturelle $Y = g(X) = X^2$. Autrement dit, $X = g^{-1}(Y) = Y^{1/2}$. Il faudra corriger la log-vraisemblance du modèle estimé avec la variable transformée par le Jacobien défini à l'aide de la fonction dérivée $\frac{\partial g^{-1}(y)}{\partial y} = \frac{1}{2} \cdot y^{-\frac{1}{2}}$.

Dans le cadre de l'étude sur les données de la multicohorte CASCADE, on a étudié différentes transformations possibles car les résidus étaient hétéroscédastes sur l'échelle naturelle (figure 3). Les log-vraisemblances corrigées étaient les suivantes :

Echelle	Log-vraisemblance corrigée
Naturelle	-75004
Logarithme	-79818
Racine carrée	-73585
Racine cubique	-73617
Racine quatrième	-73763

On a donc retenu la transformation racine carrée qui conduisait à la meilleure vraisemblance et à des résidus homoscédastes (figure 4).

Figure 3. Résidus d'un modèle mixte univarié pour les CD4+ sur l'échelle naturelle. CASCADE (916 sujets, 1 mesure/6.8 mois en médiane).

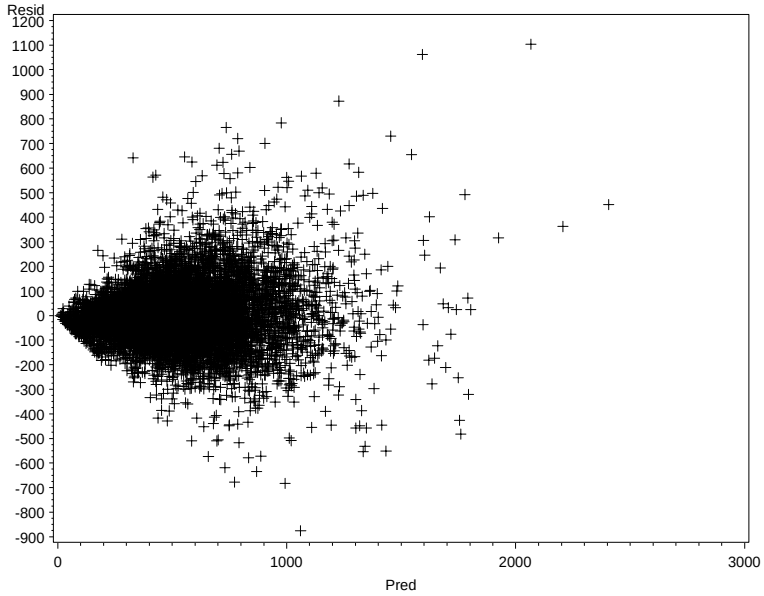
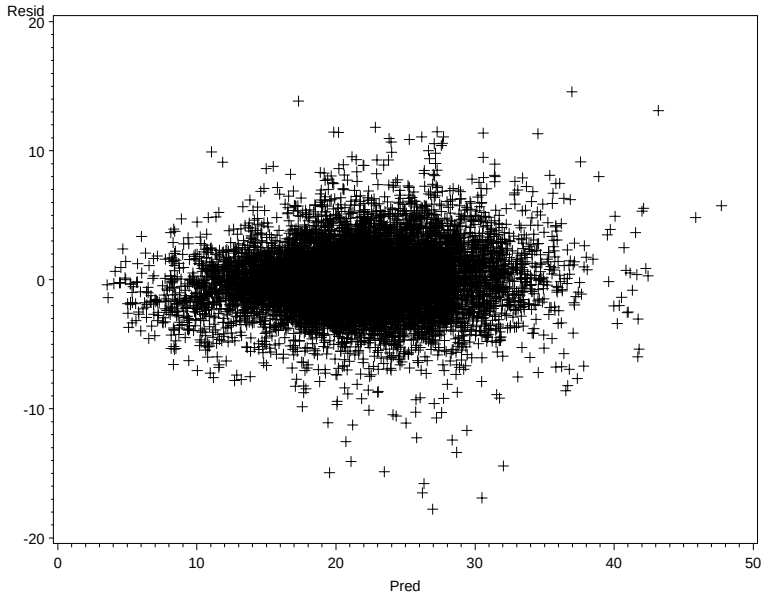


Figure 4. Résidus d'un modèle mixte univarié pour la racine carrée des CD4+. CASCADE (916 sujets, 1 mesure/6.8 mois en médiane).



Une fois obtenues les estimations du modèle mixte pour la variable dépendante transformée, l'interprétation ou l'utilisation des prédictions nécessite une transformation inverse. Par exemple, dans l'analyse de la valeur prédictive de l'évolution des CD4+ [156], la meilleure transformation pour le modèle mixte était également la racine carrée.

Pour faciliter l'interprétation de l'évolution des prédictions de CD4+ ou leur effet au sein d'un modèle de Cox, les prédictions issues du modèle mixte pour la racine carrée ont été retransformées sur l'échelle naturelle. L'espérance conditionnelle des CD4+ a été calculée à l'aide de la formule suivante car la transformation par racine carrée n'est pas linéaire (voir notations dans la section 2.1.1) :

$$E(CD4_{it} / \gamma_i) = \left\{E(\sqrt{CD4}_{it} / \gamma_i)\right\}^2 + Var(\sqrt{CD4}_{it} / \gamma_i) = \left(X_{it}\beta + Z_{it}\gamma_i\right)^2 + \sigma_\varepsilon^2$$

Cette espérance conditionnelle des CD4+ a pu être utilisée à la place des prédictions individuelles de CD4+ dans un modèle de Cox en tant que variable dépendante du temps, par exemple.

Pour la représentation sur des figures, on désire plutôt calculer l'espérance marginale des CD4+. Dans ce cas, on utilise la formule suivante :

$E\left(CD4_{it}\right) = \left\{E(\sqrt{CD4}_{it})\right\}^2 + Var(\sqrt{CD4}_{it}) = \left(X_{it}\beta\right)^2 + Var(Y_{it})$ où $Var(\sqrt{CD4}_{it})$ est le t^{ème} élément sur la diagonale de la matrice $Z_i^1 G Z_i^1$ plus $\sigma_{\varepsilon^1}^2$. Les matrices de variables explicatives sont définies selon les figures qu'on désire représenter.