



Modélisation dynamique de l'interaction hôte-pathogène appliquée à la primo-infection à VIH

Julia Drylewicz

► To cite this version:

Julia Drylewicz. Modélisation dynamique de l'interaction hôte-pathogène appliquée à la primo-infection à VIH. Sciences du Vivant [q-bio]. Université Victor Segalen - Bordeaux II, 2009. Français.
NNT : . tel-00433430

HAL Id: tel-00433430

<https://theses.hal.science/tel-00433430>

Submitted on 19 Nov 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Victor Segalen Bordeaux 2

Année 2009

Thèse n° 1625

THESE

pour le

**DOCTORAT DE L'UNIVERSITE
VICTOR SEGALEN BORDEAUX 2**

Mention : Sciences, Technologie, Santé

Option : Epidémiologie et Santé Publique

Présentée et soutenue publiquement

Le 30 octobre 2009

Par

Julia DRYLEWICZ

**Modélisation dynamique de l'interaction
hôte-pathogène appliquée à la primo-infection à VIH**

Membres du jury

Madame France MENTRÉ, Professeur, Paris	Rapporteur
Monsieur Hein PUTTER, Assistant Professeur, Leiden	Rapporteur
Monsieur Philippe VANHEMS, Professeur, Lyon	Examinateur
Monsieur Bedr'Eddine AÏNSEBA, Professeur, Bordeaux	Examinateur
Monsieur Rodolphe THIÉBAUT, Chargé de Recherche, Bordeaux	Directeur de thèse
Monsieur Daniel COMMENGES, Directeur de Recherche, Bordeaux	Codirecteur de thèse

Remerciements

A Madame France Mentré

Je vous remercie d'avoir accepté de juger ce travail. Votre expérience en modélisation dynamique et en statistiques m'apporteront beaucoup. Merci d'avoir accepté d'être le rapporteur de ma thèse.

A Monsieur Hein Putter

You honore me to judge this work. I have much admiration for the quality of your work. Your knowledge in biostatistics will bring me a lot. I am very grateful for your presence among my judges.

A Monsieur Philippe Vanhems

Vous me faites un grand honneur en acceptant de participer à mon jury de thèse. Votre grande connaissance sur l'infection à VIH m'apportera assurément un éclairage nouveau sur mon travail. Merci d'avoir accepté d'examiner mes travaux.

A Monsieur Bedr'Eddine Aïnseba

Je vous remercie de l'attention que vous avez bien voulu porter à ma thèse en acceptant de participer à mon jury. Soyez assuré de ma sincère gratitude et de ma profonde considération.

A Monsieur Daniel Commenges

Merci d'avoir partagé avec moi vos connaissances en statistiques. Votre disponibilité, votre souci de transmettre votre savoir et vos critiques constructives m'ont été particulièrement précieux. Veuillez accepter l'expression de ma reconnaissance et de mon profond respect.

A Monsieur Rodolphe Thiébaut

Tu m'as accueillie et fait confiance depuis près de 5 ans. Je te remercie d'avoir su me proposer ce(s) sujet(s) de recherche si passionnant alliant les mathématiques, les statistiques et des applications pratiques. Ta rigueur, ton enthousiasme, ton dynamisme et ta passion pour la recherche m'ont beaucoup apporté et motivé. Merci pour tout ce que tu m'as appris durant ces années et merci d'avoir été disponible et à l'écoute dans mes moments de doutes. Le travail sous ta direction et celle de Daniel a été particulièrement agréable. J'espère continuer à travailler avec vous deux à l'avenir.

Mes remerciements vont également

A la Collaboration CASCADE qui a financé ces travaux et plus particulièrement à Kholoud Porter pour m'avoir fait confiance.

A l'ensemble de l'équipe Biostatistique du Centre de Recherche INSERM U897. Merci de m'avoir accueillie dans cette équipe soudée. Un merci particulier à Guillaume pour les dépannages en tout genre.

A tous ceux que j'ai croisés à l'ISPED qui font de cet endroit une grande famille et un lieu où il fait bon travailler. Un merci tout particulier à Pierre, Majid, Luc, Mélanie, Fanny et tous les autres avec qui j'ai partagé de nombreux fous-rires et de merveilleux moments de détente.

Aux locataires du bureau 45 passés et présents : Cécile Proust-Lima, Réza, Cécile Sommen, Etienne, Linda, Andrey et bien sûr Jérémie pour avoir toujours répondu à mes questions, même les plus simples !

A tous mes "camarades de galère" : Amélie, Laurent-Emmanuel, Aude, Marie, Clément, Xavier, Yann, Mina, Diana, Sylvaine et tous ceux que j'ai croisé à 2D2B. Merci pour ces soirées pendant lesquelles les discussions les plus folles et improbables ont souvent eu lieu !

A mes amis. En particulier Anne, Charlotte, Benjamin, Rémi et le "club des pipelettes" pour votre soutien de tous les jours et votre bonne humeur. Et surtout à Oriane et Audrey pour "ce" qu'on partage depuis tant d'années et pour votre douce folie qui me fait paraître "normale".

A toute ma meute qui est présente et fidèle quoi qu'il arrive. A mes grands-parents qui auraient aimé être là. A mes parents et à mes frères pour leur amour et leur soutien inconditionnel.

Résumé :

La phase de primo-infection à VIH qui correspond aux premières semaines de l'infection est une phase cruciale qui détermine en partie l'évolution clinique de la maladie. La dynamique du virus et des lymphocytes T CD4+ (principale cible) durant cette phase est complexe. D'une part cette phase est asymptomatique dans plus de la moitié des cas et est semblable à un syndrome grippal ; elle passe par conséquent très souvent inaperçue et peu de données sont disponibles. D'autre part la date d'infection est généralement inconnue ce qui complique l'étude de l'évolution au cours du temps des marqueurs viro-immunologiques. Les modèles dynamiques basés sur des systèmes d'équations différentielles permettent de prendre en compte l'interaction complexe et non-linéaire existant entre le VIH et le système immunitaire. De plus, leur aspect mécanistique peut être très utile pour la compréhension de la physiopathologie ou l'effet d'intervention. Cependant l'estimation de ces modèles est complexe et plusieurs méthodes sont disponibles à ce jour. Nous proposons dans cette thèse une méthode d'estimation pour des modèles dynamiques prenant en compte l'incertitude sur la date d'infection. Nous appliquons cette méthode à un jeu de données réelles de 761 séroconverters de la Collaboration CASCADE durant leur première année de suivi. Ces modèles possèdent un grand nombre de paramètres pouvant inclure des effets aléatoires et des variables explicatives. La sélection du meilleur modèle peut nécessiter l'estimation d'un grand nombre de modèles et peut s'avérer très longue. Nous proposons des tests du score pour sélectionner les effets aléatoires et les variables explicatives plus rapidement dans le cadre général des modèles non-linéaires à effets mixtes avec une illustration sur des modèles dynamiques. Enfin, nous proposons de comparer plusieurs modèles biologiques possibles de la primo-infection à VIH pour prédire la dynamique de la charge virale plasmatique et des CD4.

Dynamical modelling of host-pathogen interaction applied to primary HIV infection**Abstract :**

Primary HIV infection which corresponds to the first weeks of infection, is a crucial phase which determines partly the clinical progression. The dynamics of HIV viral load and CD4 count during this phase is complex. First, as it is asymptotic in most of the cases, the primary HIV infection is not observed. Moreover, the date of infection is generally unknown and the evolution of markers with respect to the time of infection is difficult. Dynamical models based on systems of ordinary differential equations (ODE) allow to take into account complex non-linear between HIV and the immune system. Moreover, their mechanistic characteristic can be usefull through the understanding of physiopathology or the effect of an intervention. However, the estimation of this kind of models is complex and several methods are currently available. We propose here a method of estimation for ODE models taking into account the uncertainty of the date of infection. This method is applied to real dataset of 761 seroconverters from the CASCADE Collaboration during their first year of follow-up. Dynamical models have many parameters on which we can include random effects and explanatory variables. The selection of the best model can be time-consuming if we consider the computation's time issues met during estimation. We develop score tests for exploring complex models with an application to HIV dynamics models. Finally, we compare different biological models of primary HIV infection to predict the dynamics of plasma viral load and CD4 count.

MOTS CLÉS : infection par le VIH, primo-infection, modèles dynamiques, données longitudinales, tests du score

KEY WORDS : HIV infection, primary infection, dynamical models, longitudinal data, score tests

LABORATOIRE :

Centre de recherche INSERM U897 Epidémiologie et Biostatistiques
 Université Victor Segalen Bordeaux 2
 146, rue Léo Saignat 33076 BORDEAUX

Productions scientifiques liées à la thèse

Articles liés à la thèse

- *Score tests for exploring complex models : Application to HIV dynamics models.* Drylewicz J., Commenges D. and Thiébaut R. sous presse **Biometrical Journal**
- *Modelling the dynamics of biomarkers during primary HIV infection taking into account the uncertainty of infection date.* Drylewicz J., Guedj J., Commenges D. and Thiébaut R. on behalf of the CASCADE Collaboration. soumis à **Annals of Applied Statistics**

Communications dans un congrès avec comité de lecture

Communications orales

- *Score tests for exploring complex models : Application to HIV dynamics models.* Drylewicz J., Commenges D. and Thiébaut R. for the CASCADE Collaboration.
30th Annual Conference of the International Society for Clinical Biostatistics, Prague (République Tchèque), 23-27 Août 2009
- *Tests du score pour l'exploration de modèles complexes : Application à la modélisation dynamique du VIH.* Drylewicz J., Commenges D. and Thiébaut R.
41ème Journées de Statistique de la SFdS, Bordeaux (France), 25-29 Mai 2009.
- *Modelling the dynamics of biomarkers during primary HIV infection taking into account the uncertainty of infection date.* Drylewicz J., Guedj J., Commenges D. and Thiébaut R. for the CASCADE Collaboration.
XXIVth International Biometric Conference, Dublin (Irlande), 13-18 Juillet 2008

Communications affichées

- *Comparison of three dynamical models during primary HIV infection : Application to the CASCADE dataset.* Drylewicz J., Commenges D. and Thiébaut R. for the CASCADE Collaboration.
16th International HIV Dynamics & Evolution Conference, Oxford (Royaume-Uni), 4-7 Avril 2009
(Poster)
- *Plasma HIV RNA and CD4+ Count Dynamics during Acute Infection in 761 HIV-1-infected Patients : The CASCADE Collaboration.* Drylewicz J., Walker S., Commenges D., Pillay D., Venet A., Masquelier B., Meyer L., Chêne G., Porter K., and Thiébaut R. for the CASCADE Collaboration.
16th Conference on Retroviruses and Opportunistic Infections, Montréal (Canada), 8-11 Février 2009 (Poster). Young Investigator Award.

Table des matières

1 Description de la dynamique des marqueurs du VIH	9
1.1 Infection à VIH	9
1.1.1 Cycle de réplication du virus	9
1.1.2 Histoire naturelle de l'infection	10
1.1.3 Réponses immunes contre le VIH	12
1.1.4 Mesures des marqueurs viro-immunologiques	13
1.2 Etude de la primo-infection à VIH	15
1.2.1 Intérêts de son étude	15
1.2.2 Difficultés liées à son étude	17
2 Modélisation de l'infection à VIH	21
2.1 Modèles descriptifs	21
2.2 Modèles explicatifs	23
2.2.1 Les premiers modèles	23
2.2.2 Modélisation de la primo-infection	25
2.2.3 Conclusion	30
2.3 Méthodes d'estimation	33
2.3.1 Structure générale et notations	33
2.3.2 Approche bayésienne	36
2.3.3 Approche fréquentiste	37
2.3.4 Maximum a posteriori	40
2.3.5 Conclusion	41
2.4 Objectif du travail de thèse	42

3 Estimation d'un modèle ODE en prenant en compte l'incertitude sur la date d'infection	45
4 Tests du score pour l'exploration de modèles complexes	69
5 Dynamique de la primo-infection à VIH	83
5.1 Contexte	83
5.2 Un modèle de base	84
5.2.1 Données	84
5.2.2 Modèle biologique	85
5.2.3 Dynamique de la primo-infection à VIH	87
5.3 Comparaison de modèles biologiques	95
5.3.1 Modèles biologiques	95
5.3.2 Modèles statistiques	98
5.3.3 Dynamique de la primo-infection	98
5.4 Discussion	101
6 Conclusion générale et perspectives	103
7 Matériels complémentaires	109
Bibliographie	113
Annexes	125

Chapitre 1

Description de la dynamique des marqueurs du VIH

1.1 Infection à VIH

Le Virus d'Immunodéficience Humaine (VIH) a été découvert au début des années 80 (Barre-Sinoussi et al., 1983) et a rapidement été considéré comme un enjeu majeur de santé publique. D'après le rapport publié par ONUSIDA (UNAIDS, 2008), le nombre de malades vivant avec le VIH est d'environ 33 millions et le nombre de nouvelles infections en 2007 était d'environ 3 millions.

1.1.1 Cycle de réPLICATION du virus

Le VIH est un rétrovirus appartenant à la famille des lentivirus dont la pathologie est caractérisée par une longue période entre l'infection et les manifestations cliniques. Deux types viraux ont été identifiés : le VIH-1 présent dans le monde entier et le VIH-2 répandu principalement en Afrique de l'Ouest. Les cellules cibles du VIH sont essentiellement les lymphocytes T-CD4+ (CD4), en particulier lorsqu'ils sont activés (Ribeiro et al., 2002; Grossman et al., 2002). Le virus peut également infecter les monocytes/macrophages, les cellules dendritiques et au niveau du système nerveux central, les cellules microgliales.

Pour se multiplier, le virus va utiliser à son propre compte la machinerie de transcription/traduction de la cellule cible. Il doit pour se faire, se lier à une cellule cible, pénétrer dans le cytoplasme puis dans le noyau et intégrer le génome. Le virus se lie à la

membrane cellulaire au niveau d'un récepteur CD4 présent à la surface de la cellule cible grâce à sa glycoprotéine gp120. La fusion des membranes peut alors s'effectuer grâce aux co-récepteurs CCR5 et CXCR4 ; une fois le virus dans le cytoplasme de la cellule hôte, l'ARN viral est transformé en ADN viral par la transcriptase inverse puis dupliqué en ADN double brin. Cet ADN est alors intégré au génome de la cellule hôte grâce à l'intégrase. L'ARN polymérase de l'hôte va alors transcrire l'ADN proviral en ARN messagers qui vont permettre la synthèse de protéines virales (Fig. 1.1).

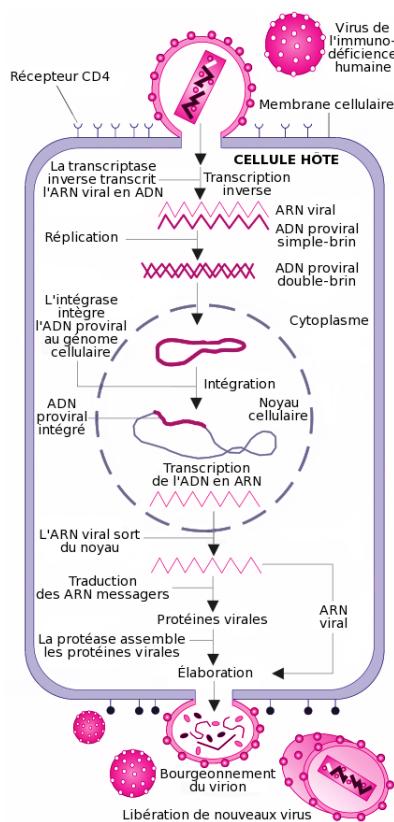


Fig. 1.1 : Cycle de réplication du virus.

1.1.2 Histoire naturelle de l'infection

Comme nous l'avons précisé précédemment, le VIH est un lentivirus et par conséquent une longue période peut se passer avant les manifestations cliniques. L'histoire naturelle de l'infection à VIH peut se diviser en trois phases distinctes (Fig. 1.2). La primo-infection, qui correspond aux premiers mois de l'infection, est caractérisée par un taux de réplication virale important et à une chute brutale du nombre total de CD4 (Clark et al., 1991; Daar

et al., 1991; Piatak et al., 1993). On peut observer un pic de charge virale plasmatique dans les deux à trois semaines qui suivent l'infection (Clark et al., 1991; Daar et al., 1991). Cette phase de primo-infection est accompagnée de symptômes dans moins de 50% des cas et sont similaires à un syndrome grippal (Kahn et Walker, 1998).

Après cette phase, le nombre de CD4 total et la charge virale se stabilisent à des états d'équilibre (Kaufmann et al., 1998; Schacker et al., 1998; Sterling et al., 2001) souvent appelé "setpoint" et le patient entre alors dans la seconde phase de la maladie : la phase dite "latente". Cette phase peut durer plusieurs années et est caractérisée par une charge virale constante et un nombre de CD4 diminuant progressivement. La durée de cette phase est variable et dépend de nombreux facteurs (âge, sexe, consommation de drogues, ...). Il a été montré grâce à des modèles explicatifs (que nous présenterons dans le chapitre suivant), que cette phase dite de "latence clinique" est en fait le siège d'une interaction intense entre le virus et le système immunitaire et que le virus continue de se répliquer activement : Perelson et al. (1996) estiment à 10^{10} le nombre de virions produits quotidiennement.

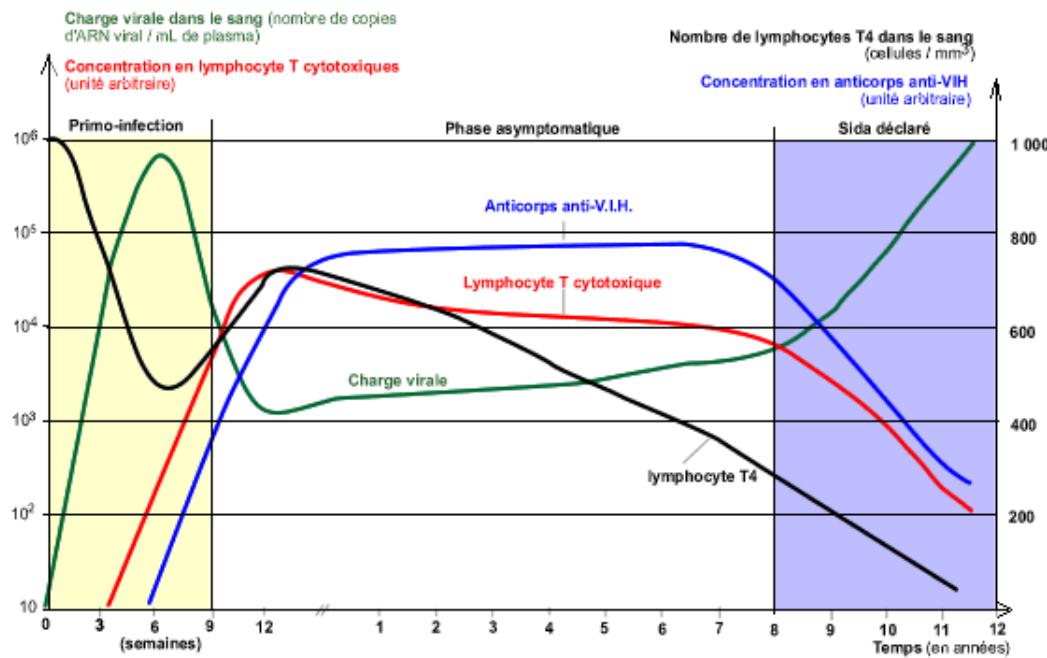


Fig. 1.2 : Evolution de la dynamique des marqueurs viro-immunologiques du VIH en absence de traitement.

La troisième et dernière phase de l'infection à VIH est celle de l'immunodéficience sévère. Durant cette phase, on observe une remontée de la charge virale et un déclin du

nombre de CD4. La survenue d'infections opportunistes majeures définit le stade Syndrome d'Immunodéficience Humaine Acquise (SIDA).

1.1.3 Réponses immunes contre le VIH

Le système immunitaire est à la fois la cible du virus et la ligne de défense. On distingue deux types de réponse immunitaire : la réponse humorale et la réponse cellulaire.

Les réponses humorales sont caractérisées par la production d'anticorps par certains lymphocytes dirigés contre toutes les protéines du virus : les protéines enveloppes (gp120, gp41), les protéines de capsid (p24, p18, nef, ...). Un patient infecté par le VIH développe une concentration élevée d'anticorps anti-VIH généralement entre un et trois mois après l'infection (Moore et al., 1994; Fiebig et al., 2003). Cette réponse est essentiellement ciblée contre les débris viraux (Parren et al., 1997) et n'a par conséquent qu'un faible impact sur la dynamique virale (Burton et al., 2005). Le rôle précis des anticorps dans le contrôle viral reste cependant incertain. En effet, on observe un déclin de la charge virale plasmatique avant l'apparition d'anticorps anti-VIH lors de la primo-infection (Fig. 1.2) et chez certains patients non-progresseurs à long-terme on n'observe pas de réponse anticorps puissante alors que le virus est contrôlé.

La réponse cellulaire est basée sur l'action directe de certains lymphocytes T-CD8+ (dits cytotoxiques) et élimine directement les cellules infectées. Les lymphocytes T-CD4+ auxiliaires spécifiques du VIH, capables de produire des interleukines, IFN γ et TNF α , ont un rôle déterminant dans la réponse acquise car ils amplifient les réponses cytotoxiques (CTL) au VIH. Durant les premières semaines de l'infection, le nombre de CD8 circulant spécifiques au VIH augmente pouvant atteindre jusqu'à 20 fois son niveau normal (Pantaleo et al., 1997). Il a été montré que cette augmentation a lieu très précocement durant l'infection (Koup et al., 1994; Borrow et al., 1994) et coïncide avec le déclin viral (Koup et al., 1994; Musey et al., 1997; Kassutto et al., 2004). L'amplitude de la réponse cytotoxique peut varier en fonction de la force de la réponse due aux anticorps (Appay et al., 2002). Les CTL spécifiques au VIH sont dirigés contre des protéines structurales (gag et env) et régulatrices (tat et rev) du VIH et leurs précurseurs sont présents en quantités importantes dans le sang. Mackewicz et al. (1991) ont également mis en évidence une activité suppressive du VIH médiée par des facteurs dérivées des CD8 lors de

la primo-infection. Ils jouent un rôle majeur dans le contrôle initial du virus avant l'apparition d'anticorps neutralisants efficaces. Les cellules CD8 anti-VIH restent présentes en quantités importantes durant toute la durée de l'infection et sont détectables chez plus de 90% des sujets infectés. Durant la primo-infection, la réponse immunitaire due aux cytokines est essentiellement constituée par l'expression de IFN γ . Graziosi et al. (1993) ont mis en évidence un pic précoce d'expression de IFN γ ainsi qu'une expression plus tardive et plus lente d'IL-10 et TNF α durant la transition entre la phase de primo-infection et la phase chronique. Le pic d'expression de IFN γ coïncide avec l'augmentation des CD8. Les lymphocytes CD4+ auxiliaires permettent grâce à la production de IFN γ lors de la phase précoce de l'infection de contrôler la réplication virale. Les lymphocytes T-CD8 (CD8), et notamment les lymphocytes T cytotoxiques (CTL), sont l'un des principaux mécanismes impliqués dans la lutte virale (Streeck et al., 2009).

1.1.4 Mesures des marqueurs viro-immunologiques

Durant le suivi des patients, deux marqueurs sont usuellement mesurés afin d'évaluer la progression de l'infection : la quantité de virus circulant ainsi que le nombre total de CD4. La mesure des CD4 est généralement obtenue par la technique de cytométrie de flux et s'exprime en nombre de cellules par mm³ (ou μ L). Malgré l'amélioration des techniques de mesures, le nombre de CD4 est connu pour avoir une grande variabilité avec des fluctuations diurnes (Malone et al., 1990). Chez un sujet sain, le nombre de CD4 est estimé en moyenne à 830 cellules/mm³ (écart-type= 288) (Bofill et al., 1992; Maini et al., 1996).

La mesure du nombre de virus circulant, abusivement appelée "charge virale", correspond au nombre de copies d'ARN viral présentes dans le compartiment sanguin (plasma). Elle ne représente pas nécessairement la quantité totale de virus présent dans l'organisme. La quantification est réalisée le plus souvent par PCR et est exprimée en copies d'ARN par mL ou en logarithme de base 10 du nombre de copies, afin de minimiser les variations de faibles amplitudes et d'approcher une distribution normale.

La charge virale et le nombre total de CD4 sont les données les plus fréquemment disponibles dans les études cliniques courantes mais d'autres mesures peuvent être effectuées. Ainsi, certaines cohortes de patients infectés par le VIH mesurent également en

routine d'autres marqueurs tels que le nombre de CD8, le pourcentage de CD4 et de CD8 parmi l'ensemble des lymphocytes T ou l'ADN viral total ou intégré. D'autres marqueurs immunologiques sont quantifiés dans des études plus spécifiques. Par exemple, on peut mesurer les CD8 spécifiques du VIH ou même l'activité spécifique des CTL. Une autre classe de marqueurs intéressante est la mesure d'activation tant des CD8 que des CD4 par l'intermédiaire de marqueurs membranaires tels que CD38 ou HLA-DR (Hunt et al., 2003; Deeks et al., 2004).

1.2 Etude de la primo-infection à VIH

1.2.1 Intérêts de son étude

La phase de primo-infection correspond aux premiers mois de l'infection et détermine en grande partie son évolution clinique (Mellors et al., 1995). Les valeurs pronostiques de la charge virale plasmatique au moment du pic viral et à son état d'équilibre ont été étudiées et l'impact déterminant de la primo-infection sur l'évolution clinique de l'infection a été mis en évidence. En effet, il a été montré que la hauteur du pic de charge virale (entre 10^5 et 10^7 copies/mL) est corrélée au niveau d'équilibre atteint par celle-ci au moment de la phase asymptomatique (Lindback et al., 2000). Ce niveau d'équilibre est lui-même associé à la progression clinique (Mellors et al., 1995) : ainsi, plus le niveau d'équilibre viral d'un patient est élevé, plus sa progression vers le stade SIDA sera rapide. De même, une primo-infection symptomatique est prédictive d'une progression clinique plus rapide qu'une primo-infection asymptomatique (Henrard et al., 1995; Vanhems et al., 1998). Mieux comprendre la constitution de cet équilibre viral permettrait potentiellement d'expliquer pourquoi certains patients sont des progresseurs rapides (stade SIDA déclaré en moins de 7 ans) ou au contraire des non-progresseurs à long terme.

Comme nous l'avons précédemment décrit, la primo-infection à VIH est caractérisée par un taux de réPLICATION virale très important. Dans les semaines qui suivent le pic viral, la charge virale décroît d'un facteur 100 à 1000. La clairance relative de la virémie initiale et le contrôle ultérieur de la réPLICATION virale peuvent être dus à plusieurs composantes de la réponse immunitaire : la réponse immunitaire non-spécifique du virus par les cellules NK, la réponse immunitaire humorale spécifique médiée par des anticorps, la réponse immunitaire cellulaire spécifique médiée par les CTL mais également à la chute du nombre de cellules cibles. En effet, le nombre de CD8 anti-VIH atteint son niveau maximal au moment où la virémie décroît. De plus, une association très forte entre certaines molécules HLA de classe 1, molécules présentant les antigènes viraux aux CD8, et l'absence de progression clinique a été démontrée (Saah et al., 1998; Altfeld et al., 2003).

Améliorer les connaissances sur les mécanismes d'interaction entre le VIH et le système immunitaire durant les premiers mois de l'infection est au cœur de la recherche biomédicale actuelle. Par l'étude de l'effet de traitement précoce mais également dans le cadre

de la recherche vaccinale. L'objectif de mettre sous traitement précocement des individus infectés par le VIH est de rendre la charge virale plasmatique indétectable et ainsi préserver les fonctions des CD4. De nombreuses études se sont donc intéressées à l'effet d'une mise sous traitement antirétroviral durant les premiers mois de l'infection (Fidler et al., 2008; Kilby et al., 2008; Pantazis et al., 2008; Seng et al., 2008; Voirin et al., 2008; ...). Les bénéfices d'une telle intervention sont toutefois limités ; même si les fonctions des CD4 et des CD8 sont mieux préservées, des phénomènes de toxicités et de résistances apparaissent rapidement (Kaufmann et al., 2004). La réponse immunologique à une mise sous traitement précoce est efficace mais la réponse virologique est quant à elle plus discutée. En effet, on observe bien chez certains patients un contrôle du virus mais des rebonds sont souvent observés : ils sont plus précoces chez les patients ayant été traités plus tardivement (Hecht et al., 2006; Steingrover et al., 2008). Se pose donc la question de démarrer le traitement le plus tôt possible après l'infection avant que l'équilibre viral se soit établi. Les rebonds de charge virale observés sont dus à l'existence de cellules infectées latentes qui se réactivent et produisent du virus infectieux (Sedaghat et al., 2007). Ce réservoir de cellules infectées latentes est constitué très tôt durant l'infection, probablement durant la phase de primo-infection (Chun et al., 1998). L'étude de l'effet d'une mise sous traitement précoce rejoint les questions de l'effet de traitement prophylactique pré-exposition (PreP). Connaitre l'efficacité minimale des traitements nécessaires pour rendre la charge virale indétectable dès le début de l'infection est alors essentiel pour mettre au point un traitement préventif efficient. D'autre part, comprendre (et connaître) les caractéristiques de l'infection lui permettant d'échapper au système immunitaire (via la constitution d'un réservoir notamment) sont des éléments pertinents à la réalisation d'un modèle vaccinal. La constitution d'un réservoir de cellules infectées latentes, l'hypervariabilité du VIH et des facteurs génétiques intervenant dans la réponse CTL sont les principaux freins à l'élaboration actuelle d'un vaccin.

Notons enfin que toutes ces études ont été effectuées sur des patients infectés par le VIH-1 et que des différences notables sont observées dans le cas d'une infection par le VIH-2. On sait que le niveau d'équilibre de charge virale est plus faible et celui des CD4 plus élevé (Bock et Markovitz, 2001) chez des patients infectés par le VIH-2. Le déclin des CD4 est plus lent conduisant ainsi à comparer ces patients à des non-progresseurs à

long terme VIH-1 (Rowland-Jones et Whittle, 2007).

Ainsi, mieux comprendre les mécanismes d'interactions qui entrent en jeu durant la phase précoce de l'infection à VIH permettrait de faire des hypothèses physiopathologiques sur les différences observées de progression clinique : progresseurs rapides vs. non-progresseurs à long terme, VIH-1 vs. VIH-2.

1.2.2 Difficultés liées à son étude

L'étude de la primo-infection paraît donc être essentielle à une meilleure compréhension de l'infection à VIH dans son ensemble. Malheureusement, peu de données de patients récemment infectés sont disponibles. En effet, l'une des difficultés majeures dans l'étude de la primo-infection à VIH est que la date d'infection est inconnue. La phase de primo-infection étant dans la majorité des cas asymptomatique, elle passe inaperçue et très peu de patients se font diagnostiquer durant cette période, à moins d'avoir eu un comportement à risque. D'après l'InVS (2008), en 2007, parmi les personnes diagnostiquées en France dont le stade clinique est documenté, 12% sont diagnostiqués en primo-infection, 58% durant le stade asymptomatique, 13% en phase symptomatique non-sida et 17% au stade sida.

Ainsi, au vu du peu de données disponibles, des cohortes de patients à risque ont été créées proposant des dépistages réguliers à ses participants, la plus importante étant la cohorte américaine *MACS* (Multicenter AIDS Cohort Study). De nombreuses autres cohortes existent l'*Amsterdam Cohort Study among drug users* ou encore la *Greek haemophilia Study*. Dans le meilleur des cas, nous disposons d'une date de dernière sérologie négative et d'une date de première sérologie positive. Ces deux dates pouvant être espacées de plusieurs mois voire de plusieurs années. La séroconversion (apparition des anticorps anti-VIH) ayant eu lieu entre ces deux dates. Les rares cas où cette fenêtre de temps est courte (moins de 6 mois) correspondent à des patients ayant eu une primo-infection symptomatique et/ou des comportements à risque à répétition. Se limiter à l'étude de ces patients induit un biais de sélection important au vu des conséquences d'une primo-infection symptomatique sur la progression clinique.

En 1997, en Europe, une collaboration de cohortes s'est créée avec pour objectif de mettre en relation des épidémiologistes, des statisticiens, des virologues et des cliniciens. La

collaboration *CASCADE* (pour Concerted Action on SeroConversion to AIDS and Death in Europe) regroupe les données de patients séroconvertis en Europe, au Canada et en Australie et suivent les patients tout au long de leur vie. La date de séroconversion (date d'apparition des anticorps anti-VIH) peut être estimée pour chaque patient soit à partir de la date médiane entre les dates de dernière sérologie négative et de première sérologie positive soit sur la base de preuves biologiques de séroconversion (Western Blot incomplet et charge virale détectable par exemple). Nous avons basé nos travaux sur les données issues de cette collaboration. En novembre 2006, elle incluait 17 240 patients. Cependant, 761 patients seulement ont des données suffisamment répétées durant la première année de suivi pour pouvoir être utilisées dans une étude des dynamiques de la charge virale et des CD4 durant la première année d'infection (correspondant à la primo-infection). Le nombre médian de mesures disponibles par patient est de 4 (écart-interquartile (EIQ)= [3 ; 5]) pour la charge virale et les CD4. Nous disposons donc de l'une des bases de données les plus importantes de séroconvertis. La figure 1.3 représente l'ensemble des données disponibles pour cet échantillon ainsi qu'une prédiction par régression polynomiale locale pour la charge virale plasmatique et les CD4. On peut remarquer que certains patients ont une charge virale détectable au moment de leur dernière sérologie négative. Ainsi, bien qu'ayant une sérologie négative, ils étaient déjà infectés par le VIH. On peut également relever la tendance d'un pic précoce de charge virale comme il est attendu d'observer en primo-infection et démontrant la richesse du jeu de données utilisé.

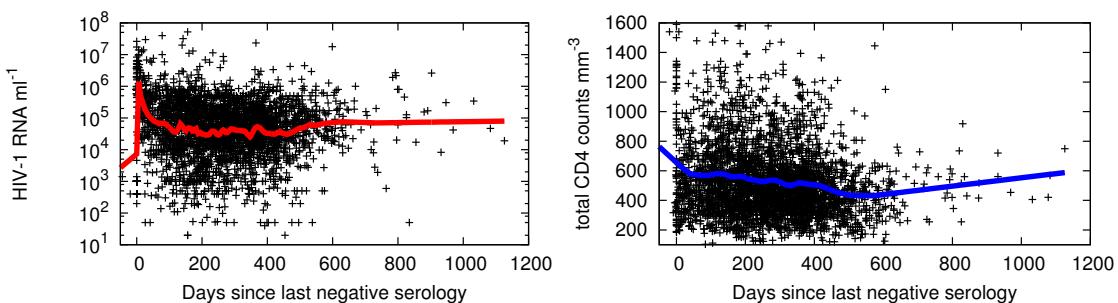


Fig. 1.3 : Données de charge virale et de CD4 disponibles chez 761 séroconvertis issus de la collaboration CASCADE.

L'étude des dynamiques des CD4 et de la charge virale plasmatique durant la primo-

infection à VIH semble donc pertinent afin de mieux comprendre l'évolution de la maladie. Cette période initiale offre l'opportunité d'examiner l'interaction entre le virus et l'hôte immunologiquement compétent au moment où les mécanismes immunologiques qui s'opposent à la réPLICATION et à la dissémination du virus se mettent en place.

Chapitre 2

Modélisation de l'infection à VIH

Plusieurs approches peuvent être envisagées pour modéliser l'évolution des marqueurs viro-immunologiques dans le cadre de l'infection à VIH. La plus courante se base sur des modèles dit “descriptifs” qui vont, comme leur nom l'indique, décrire les variations des marqueurs en fonction du temps. Une seconde approche, qui est développée depuis le début des années 90, se base sur des modèles dit “explicatifs” qui vont, en plus de décrire l'évolution des marqueurs, modéliser leurs interactions de manière mécanistique.

2.1 Modèles descriptifs

Les modèles linéaires mixtes (Laird et Ware, 1982) permettent de modéliser le suivi longitudinal de marqueurs en prenant en compte les variabilités inter-individuelles. Ces modèles peuvent s'écrire, pour un individu i , ayant n_i mesures répétées, sous la forme :

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\alpha}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G}) \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i) \\ \boldsymbol{\alpha}_i \perp \boldsymbol{\epsilon}_i, \forall i \end{cases}$$

où \mathbf{Y}_i est le vecteur des variables à expliquer, $\boldsymbol{\beta}$ le vecteur des coefficients de régression (commun à tous les individus i), $\boldsymbol{\alpha}_i$ est le vecteur des effets aléatoires de l'individu i supposé gaussien d'espérance nulle et de matrice de variance-covariance \mathbf{G} . Les vecteurs $\boldsymbol{\alpha}_i$ représentent les écarts individuels à la tendance globale et $\boldsymbol{\epsilon}_i$ est le vecteur d'erreurs individuelles dues essentiellement aux erreurs de mesures. La matrice \mathbf{X}_i représente l'en-

semble des variables explicatives (en particulier les temps de mesure dans notre contexte) et \mathbf{Z}_i est une sous-matrice de \mathbf{X}_i contenant uniquement les variables explicatives sur lesquelles on place un effet aléatoire (i.e. une variabilité individuelle). Ce type de modélisation permet de décrire l'évolution supposée linéaire d'une population en tenant compte des spécificités de chaque individu et de la corrélation entre les mesures.

Ces modèles ont été utilisés notamment pour décrire l'évolution de la charge virale et des CD4 au cours du temps après une mise sous traitement. Une modélisation linéaire à deux pentes en fonction du temps depuis l'initiation du traitement est souvent utilisée. Ces modèles constituent une première approche dans une tentative d'explication des déterminants de la réponse viro-immunologique à un traitement antirétroviral. Etant donné l'interaction existante entre le VIH et les CD4, il semble pertinent de modéliser leur évolution simultanément (Boscardin et al., 1998) à l'aide de modèles linéaires bivariés. La corrélation entre les deux marqueurs se faisant par l'intermédiaire de la matrice de variance-covariance des effets aléatoires. Thiébaut et al. (2003) montrent la supériorité de cette approche bivariée par rapport à la modélisation séparée des deux marqueurs. Des corrélations négatives ont été mises en évidence entre les pentes de la charge virale et des CD4 (Brown et al., 2001; Thiébaut et al., 2003) : plus le traitement est efficace en terme de chute de la charge virale, plus la remontée des CD4 est importante. Cependant cette corrélation peut évoluer avec le temps (Brown et al., 2001) et les interactions entre les marqueurs ne peuvent être synthétisés par la seule matrice de corrélation.

Bien qu'utilisés très fréquemment, pour leur simplicité d'estimation et d'interprétation, les modèles descriptifs ne prennent pas en compte les interactions biologiques connues existant entre les marqueurs et ignorent ainsi le modèle biologique sous-jacent à l'infection à VIH. Il paraît alors assez naturel de se baser sur une modélisation mécanistique de l'interaction entre le VIH et le système immunitaire.

2.2 Modèles explicatifs

Les modèles explicatifs sont basés sur des connaissances biologiques et permettent de prendre en compte les interactions complexes et non-linéaires existant entre le système immunitaire et le VIH par l'intermédiaire de systèmes d'Equations Différentielles Ordinaires (ODE en anglais). Les paramètres de ces modèles ont une signification biologique, tels que le taux d'infectiosité du virus, le taux de décès de cellules, ...

2.2.1 Les premiers modèles

Les premiers modèles explicatifs se sont intéressés à la dynamique virale après mise sous traitement. Ainsi, dans un numéro de *Nature* de 1995 (Ho et al., 1995; Wei et al., 1995), deux articles utilisent deux modèles très simples de l'interaction entre le VIH et le système immunitaire et estiment ainsi des taux similaires de décès du virus (environ 2 jours^{-1}). Le virus V est produit par des cellules infectées à un taux noté P et meurt à un taux c supposé constant. On a alors $dV/dt = P - cV$. On peut considérer que l'effet d'un traitement mis en place est parfait dans les premiers jours qui suivent son initiation, ce qui signifie qu'aucun virus ne sera produit ($P = 0$). La dynamique virale s'écrit alors $dV/dt = -cV$ (Ho et al., 1995). Durant les premiers jours de la mise sous traitement, on a donc : $V(t) = V_0 \exp(-ct)$, où V_0 est le niveau de charge virale à la mise sous traitement. Si on dispose de données de charge virale dans les premiers jours suivant la mise sous traitement, cette équation permet d'estimer c . Mais cette estimation n'est qu'une valeur minimale de c , car en réalité, le traitement n'est pas parfait et par conséquent $P > 0$. La demi-vie du virus peut être déduite : $t_{1/2} = \ln(2)/c$. Si on suppose que les données ont été recueillies durant la phase asymptomatique de l'infection, alors la charge virale peut-être considérée comme stable et on a $P = cV_0$. On peut donc obtenir une estimation du taux de production P à partir de celle de c . Le même type de raisonnement peut être appliqué pour la dynamique des CD4 afin de déduire leur taux de production P' . Wei et al. (1995) proposent un modèle comparable et donne en plus une estimation de la demi-vie des cellules infectées productrices. Ces deux papiers ont eu un impact considérable car ils ont démontré que le cycle de réplication du virus est extrêmement rapide, bien que le niveau de charge virale soit constant. La phase asymptomatique ne doit donc pas être vue comme

une pause dans l'infection, mais au contraire comme une période de renouvellement intensif du virus et des CD4. Ainsi, Ho et al. (1995) trouvent une corrélation négative entre le niveau des CD4 à la mise sous traitement et leur remontée. Les patients ayant un niveau initial de CD4 plus bas, ont remontée plus importante après la mise sous traitement. La déplétion lymphocytaire devrait donc être interprétée comme une conséquence de la destruction des CD4 (due à la cytopathogénicité du virus) et non comme un déficit de leur production.

Ces estimations ont été affinées par Perelson et al. (1996) en combinant la dynamique des CD4 infectés à celle du virus. Ils présentent un modèle biomathématique de type “proie-prédateur”. Ce modèle est devenu le modèle de référence dans la modélisation de l'interaction VIH-système immunitaire. En l'absence de traitement, ce modèle s'écrit :

$$\begin{cases} \frac{dT}{dt} = \lambda - kVT - dT \\ \frac{dI}{dt} = kVT - \delta I \\ \frac{dV}{dt} = pI - cV \end{cases} \quad (2.1)$$

où T est la concentration de CD4 activés non infectés (les cellules cibles du virus), I est la concentration de CD4 activés infectés et V est la concentration de virus circulant. Les cellules T sont produites à un taux journalier λ , elles s'infectent à un taux k et meurent avec un taux d . Le nombre journalier de nouvelles cellules infectées est le terme kVT qui suit une loi d'action de masse : le nombre de nouvelles cellules infectées I est donc proportionnel au nombre de cellules cibles T et de virions V présents. Chaque cellule infectée I produit p virions par jour. Les cellules I meurent à un taux δ et les virions à un taux c . Les deux principaux mécanismes d'action des traitements antirétroviraux sont des antiprotéases et des inhibiteurs de la transcriptase inverse. Les antiprotéases vont induire la production de virus non-infectieux par les cellules infectées et vont donc jouer sur le paramètre p du modèle (2.1). Il convient alors de distinguer le virus infectieux V_I et le virus non-infectieux V_{NI} dans le modèle. L'efficacité d'une antiprotéase est notée η_{PI} . Les inhibiteurs de la transcriptase inverse vont bloquer (ou du moins ralentir) l'infection de nouvelles cellules ; ils vont modifier le taux d'infectiosité k ; leur efficacité est notée η_{RT} .

Le modèle sous antiprotéase peut s'écrire :

$$\left\{ \begin{array}{l} \frac{dT}{dt} = \lambda - kVT - dT \\ \frac{dI}{dt} = kVT - \delta I \\ \frac{dV_I}{dt} = (1 - \eta_{PI})pI - cV_I \\ \frac{dV_{NI}}{dt} = \eta_{PI}pI - c_V NI \end{array} \right.$$

En faisant comme hypothèses que le nombre de cellules cibles (T) est constant et que l'efficacité de l'antiprotéase est parfaite (i.e. $\eta_{PI}=1$), on peut donner une formule analytique de l'évolution de la charge virale $V = V_I + V_{NI}$ en fonction de δ et de c :

$$V(t) = V_0 \exp(-ct) + \frac{cV_0}{c - \delta} \left\{ \frac{c}{c - \delta} [\exp(-\delta t) - \exp(-ct)] - \delta t \exp(-ct) \right\} \quad (2.2)$$

En réalisant une régression non-linéaire patient par patient, Perelson et al. (1996) estiment ainsi la clairance virale (c) à 3 jours^{-1} (soit une demi-vie d'environ 6 heures) et le taux de décès des cellules infectées (δ) à 1.6 jours^{-1} (soit une demi-vie de 10 heures).

De nombreuses extensions de ces modèles ont été proposées par la suite en intégrant un délai d'action du traitement (Dixit et Perelson, 2004), la pharmacocinétique de la molécule (Huang et al., 2003) ou l'évolution du virus (Ball et al., 2007).

2.2.2 Modélisation de la primo-infection

Comme nous l'avons vu dans le chapitre 1, la phase de primo-infection est cruciale dans l'histoire de l'infection. Comparativement à la mise sous traitement, les dynamiques de la charge virale et des CD4 ont été moins souvent modélisées essentiellement à cause du manque de données disponibles. Néanmoins, plusieurs études de simulations ont été réalisées dans ce contexte : Phillips (1996), Tuckwell et Le Corfec (1998), Wick et al. (2002), Wodarz et Hamer (2007), Burg et al. (2009). Dans l'article de Phillips (1996), il est démontré à l'aide d'un modèle simple (proche de celui proposé par McLean et al. (1991)) que le déclin de l'ARN viral après le pic peut être une conséquence du comportement "proie-prédateur" de l'interaction du VIH avec le système immunitaire et par conséquent indépendant d'une réponse immunitaire spécifique. Il utilise un modèle incluant des cellules CD4 activées non-infectées R , des cellules infectées latentes L , des cellules activées

infectées E et des virions circulant V . Le modèle peut s'écrire sous la forme :

$$\begin{cases} \frac{dR}{dt} = \Gamma\tau - \mu R - \beta RV \\ \frac{dL}{dt} = p\beta RV - \mu L - \alpha L \\ \frac{dE}{dt} = (1-p)\beta RV + \alpha L - \delta E \\ \frac{dV}{dt} = \pi E - \sigma V \end{cases}$$

Les cellules activées non-infectées sont produites à un taux constant $\Gamma\tau$ où Γ est le taux de nouvelles cellules produites et τ la proportion activée, elles meurent à un taux μ indépendant du virus ou sont infectées avec un taux βV . Après l'infection, une proportion p devient latente (L) ; cellules qui peuvent s'activer à un taux α ou mourir à un taux μ indépendamment du virus. Les cellules E sont générées immédiatement après l'infection par les cellules R ou par activation des cellules L et meurent à un taux δ . Enfin, les virions V sont produits à un taux π et meurent à un taux σ .

Ce modèle, bien que simple comparé à la réalité, permet de capturer la dynamique de la charge virale et des CD4 durant la primo-infection (Fig. 2.1). Cependant, il ne permet pas d'expliquer le déclin progressif du nombre de CD4 durant la phase asymptomatique de l'infection.

Certains auteurs ont proposé d'ajouter une composante stochastique dans le système d'équations différentielles pour prendre en compte l'aspect aléatoire des mécanismes biologiques. Une limite des modèles biomathématiques déterministes serait de ne pas intégrer l'aspect intrinsèquement aléatoire des interactions entre le virus et le système immunitaire. Ainsi, Tuckwell et Le Corfec (1998) proposent notamment une amélioration du modèle de primo-infection (2.1) :

$$\begin{cases} \frac{dT}{dt} = \lambda - dT - kVT - \sqrt{kVT}dW \\ \frac{dI}{dt} = kVT - \delta I + \sqrt{kVT}dW \\ \frac{dV}{dt} = pI - cV - \sqrt{kVT}dW \end{cases}$$

où W est un processus Brownien standard, de moyenne 0 et de variance t à la date t : la rencontre virus-CD4 devient alors un processus aléatoire.

D'autres modèles plus ou moins complexes ont été proposés pour modéliser la phase de primo-infection. Ainsi, Wodarz et Hamer (2007) distinguent les cellules spécifiques et non-spécifiques au VIH et incluent également des cellules effectrices (CTL) naïves et mémoires dans leurs modèles. Ou plus récemment, Burg et al. (2009) complexifient le modèle (2.1)

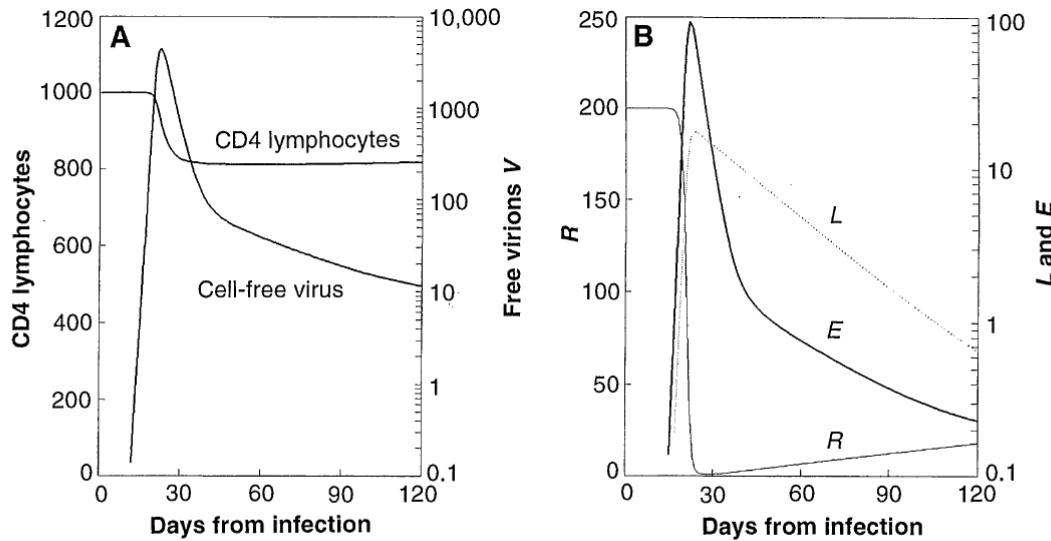


Fig. 1. Temporal changes in the number of (A) total CD4 lymphocytes [$1000(1 - \tau) + R + L + E$] and free virions (V) and (B) activated, uninfected CD4 lymphocytes (R), latently infected cells (L), and actively infected cells (E) in the first 120 days after HIV infection as predicted by the model on the basis of a simulation with iterations of step length 1 hour and 20 min. The total (whole-body) inoculum of HIV was assumed to consist of 100 virions (and no infected cells); thus, there were initially 4×10^{-7} virions in the quantity being considered.

Fig. 2.1 : Phillips, *Science* (1996)

(modèle proche de celui proposé par Ciupe et al. (2006)) en ajoutant un compartiment de cellules effectrices E pour induire une réponse immunitaire suite à l'infection :

$$\begin{cases} \frac{dT}{dt} = s - dT - \beta VT \\ \frac{dI}{dt} = \beta VT - [\alpha d + k_0 E]I \\ \frac{dV}{dt} = \pi I - cV \\ \frac{dE}{dt} = \alpha_E \frac{I}{\theta + I} - d_E E \end{cases}$$

où T représente les cellules cibles du virus, I les cellules infectées et V les virions circulant. Une fonction de saturation des cellules E dépendant du nombre de cellules I est introduite avec une stimulation limite θ . Les cellules effectrices n'étant pas mesurées et le nombre de paramètres du modèle étant grand, ils proposent une simplification du modèle en utilisant une approximation de l'état d'équilibre des cellules E . L'hypothèse faite étant que les cellules effectrices ont une dynamique plus rapide que la phase de primo-infection

(environ 1 mois). Ils s'intéressent alors au modèle suivant :

$$\begin{cases} \frac{dT}{dt} = s - dT - \beta VT \\ \frac{dI}{dt} = \beta VT - [\alpha d + k \frac{I}{I+\theta}] I \\ \frac{dV}{dt} = \pi I - cV \end{cases}$$

L'ajout d'un terme de saturation dans le modèle montre qu'une réponse immunitaire anti-VIH peut être présente dès le pic viral sans l'introduction d'une réponse immunitaire se déclenchant après un certain délai.

La limite majeure de ces modèles est que les paramètres sont fixés : les trajectoires de charge virale et de CD4 sont obtenues par simulation. Les valeurs choisies pour les paramètres sont très rarement justifiées biologiquement et ne sont pas toujours basées sur des résultats antérieurs. Il est alors nécessaire de prendre un certain recul sur les conclusions biologiques issues de ces travaux.

Peu de publications se sont penchées sur l'estimation de modèles dynamiques en primo-infection à VIH. Comme nous l'avons déjà évoqué, peu de données sont disponibles et une autre difficulté se pose quant à l'origine des temps du modèle. En effet, la date d'infection, qui correspond à $t = 0$, est généralement inconnue. Ainsi, la majorité des publications ont choisi de prendre comme origine des temps la date des premiers symptômes et de retrancher à cette date une vingtaine de jours, ce qui fournit une date approximative pour l'infection.

Stafford et al. (2000) estiment les paramètres du modèle (2.1) pour 10 patients en utilisant des données de charge virale uniquement. Cette estimation est effectuée par une méthode des moindres carrés non-linéaires et patient par patient. Ils font l'hypothèse que le nombre de cellules cibles T est constant. Ils proposent également un modèle pour expliquer la dynamique à long terme en faisant dépendre du temps le taux de décès des cellules infectées δ :

$$\delta(V) = \begin{cases} 0 & t < t_1 \\ f(t)V & t \geq t_1 \end{cases}$$

où $f(t) = \frac{\beta}{1+\kappa \exp(-(t-t_1)/\Delta T_1)} - \frac{\beta}{1+\kappa \exp(-(t-t_2)/\Delta T_2)}$. Les trajectoires individuelles ainsi obtenues sont cohérentes avec les données observées (Fig. 2.2). Ils fournissent également une estimation du nombre de reproduction de base R_0 déterminé à partir du système ODE.

R_0 correspond au nombre moyen d'infections secondaires causé par l'introduction d'une cellule infectée dans une population de cellules cibles supposée infinie. Cette quantité s'obtient facilement en multipliant le nombre de virions produits par une cellule infectée durant sa vie de cellule infectée par le nombre de cellules infectées par virion durant sa vie. Elle peut être également obtenue en effectuant le rapport entre k et k_c où k_c est le taux d'infectiosité critique nécessaire pour bifurquer de l'état d'équilibre initial (sans virus) à l'état d'équilibre post-infection. Le nombre de reproduction de base R_0 pour le modèle (2.1) est défini par :

$$R_0 = \frac{\lambda k \pi}{\delta d c}$$

Pour les 10 patients inclus dans leur étude, Stafford et al. (2000) estiment un R_0 médian à 5.7 (min=2.8, max=11.0).

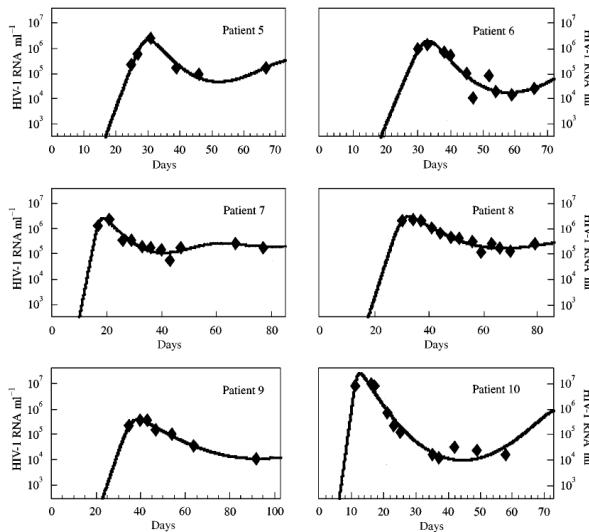


FIG. 2. Theoretical curves using estimated parameters vs. observed viral load data (\blacklozenge) for all patients. Time values of the data points differ from those in Table 1 since all data sets have been translated in time as described in the text.

Fig. 2.2 : Stafford et al., *Journal of Theoretical Biology* (2000)

Little et al. (1999) modélisent la dynamique virale à l'aide d'une simplification du modèle (2.1). Les cellules infectées I sont supposées avoir une distribution exponentielle dépendant du taux de croissance virale noté r_0 : $I(t) = I_0 \exp(r_0 t)$. Ce taux de croissance est alors supposé égal à $\beta T_0 - \delta_I$ où β est le taux d'infectiosité du virus et δ_I le taux de décès des cellules infectées (estimable à partir des données de mise sous traitement). Ils donnent ainsi une estimation de r_0 et de R_0 pour 8 patients ayant eu une primo-infection symptomatique. Cependant, leurs calculs diffèrent de ceux effectués par Stafford et al.

(2000) ; r_0 et R_0 sont déterminés à partir de \hat{r} (le taux de croissance observé de la charge virale), δ_I , T_0 (le nombre de CD4 chez un individu sain) et \hat{T} (le nombre de CD4 observé proche du temps du pic viral), ainsi :

$$r_0 = \frac{T_0}{\hat{T}}(\hat{r} + \delta_I) - \delta_I \quad \text{et} \quad R_0 = \frac{T_0}{\hat{T}}\left(1 + \frac{\hat{r}}{\delta_I}\right)\exp(\hat{r}\tau),$$

où τ correspond au temps nécessaire pour qu'une cellule infectée devienne productrice. Pour le modèle (2.1), Little et al. (1999) estiment un R_0 moyen à 7.1 (min=5.2, max=9.1) pour un modèle sans délai ($\tau=0$) et à 19.3 (min=7.4, max=34.0) pour un délai de 24h (i.e une fois infectée, une cellule met un jour avant de pouvoir produire du virus).

Kaufmann et al. (1998) proposent une classification en “patterns” de la dynamique virale en primo-infection suivant la hauteur du pic et la vitesse de convergence vers le niveau d'équilibre (Fig. 2.3). Ils se basent sur un modèle biexponentiel inspiré de ceux développés en pharmacocinétique (Gibaldi et Perrier, 1982). L'origine des temps correspond à la date de début des symptômes et la charge virale est modélisée sous la forme suivante :

$$V(t) = \frac{A_v k_a}{V_c(k_a - k_{e1})} \left\{ \exp(-k_{e1}t) - \exp(-k_a t) \right\} + \frac{V_R}{V_c k_{e2}} \left\{ 1 - \exp(-k_{e2}t) \right\}$$

où A_v est le nombre de virions relâchés dans le plasma par chaque cellule infectée durant la phase du pic, k_a est le taux de production virale, k_{e1} le taux d'élimination du virus pendant le pic viral, V_R la production journalière de virions durant la seconde phase, k_{e2} le taux d'élimination du virus pendant la seconde phase et V_c le volume sanguin. Ils ne mettent pas en évidence d'association entre l'état d'équilibre viral et le pic ou le déclin après le pic, de même que les taux de déclin des CD4 ne sont pas corrélés avec le pic, le déclin ou l'équilibre viral.

2.2.3 Conclusion

En conclusion, les modèles dynamiques ont permis de mettre en évidence la dynamique très rapide du virus et des cellules infectées (bien que celles-ci ne soit pas mesurées) en utilisant des données de charge virale après une mise sous traitement antirétroviral. Ils permettent de fournir *in vivo* une estimation de l'efficacité des traitements et peuvent attester de la pertinence d'hypothèses biologiques. Dans le cadre de l'étude de la primo-infection, ils ont notamment permis de fournir une description précise de la dynamique virale durant

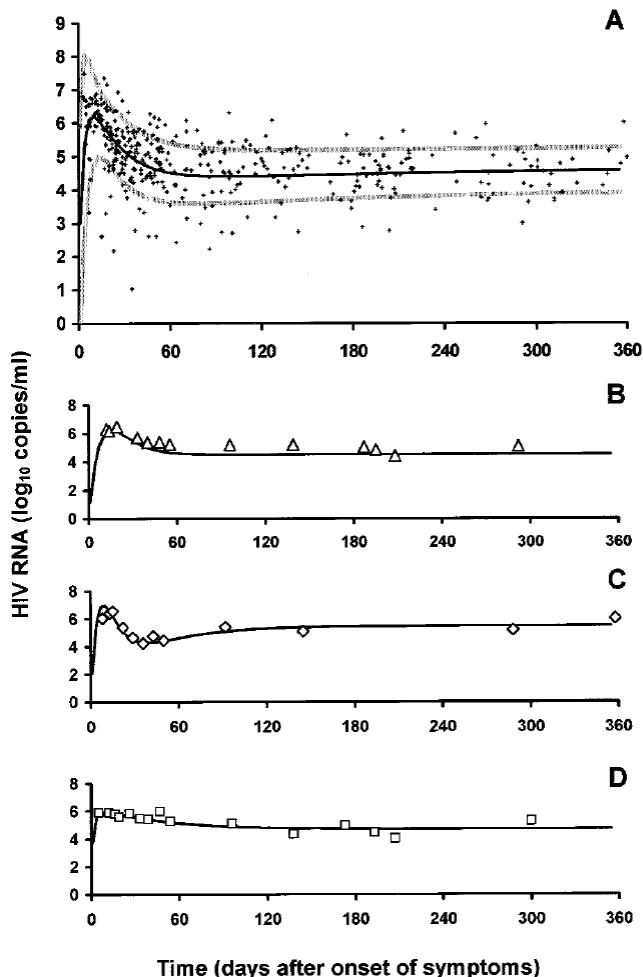


Figure 1. A, HIV-1 load of 41 patients with primary infection (+) and average fitted virus load curve and SD (solid and gray lines, respectively). B–D, Representative examples of viral kinetic patterns 1–3, respectively. Symbols indicate measured HIV RNA; line is fitted curve. Pattern 1 (**B**) was defined by rapid decline of virus load within 90 days to steady state. In pattern 2 (**C**), virus load dropped rapidly to nadir, $>0.25 \log_{10}$ copies/mL lower than final steady state, but increased to steady state after 90 days. Pattern 3 virus load (**D**) approached steady state after 90 days but declined more slowly after peak virus load.

Fig. 2.3 : Kaufmann et al., *The Journal of Infectious Diseases* (1998)

les premières semaines de l'infection en utilisant un modèle simple de l'interaction entre le virus et le système immunitaire. Des estimations du nombre de reproduction de base R_0 ont également été fournies permettant ainsi à Little et al. (1999) de déterminer l'efficacité minimale nécessaire pour éradiquer théoriquement le virus : une extinction virale se traduisant par un $R_0 < 1$.

Les modèles explicatifs, basés sur des connaissances biologiques, imposent, de par leur complexité, de faire des hypothèses souvent fortes (traitement parfait, nombre de cellules cibles constant, production immédiate de virus après l'infection, ...) pour pouvoir estimer les paramètres. Dans la plupart des études publiées, les paramètres étaient estimés patient par patient et présentés en terme de moyenne, médiane ou variance empirique obtenue sur l'ensemble des sujets. Les modèles dynamiques sont donc des outils plus adaptés que les modèles descriptifs pour l'étude de systèmes biologiques complexes mais l'estimation de tels modèles est complexe si aucune simplification du modèle n'est faite.

2.3 Méthodes d'estimation

Nous donnons dans cette section un cadre statistique plus formel à ces modèles ainsi qu'une revue des outils statistiques disponibles pour l'estimation de modèles dynamiques.

2.3.1 Structure générale et notations

Modèle biomathématique pour le système

On considère un échantillon de n patients indépendants. Pour un sujet i ($i = 1, \dots, n$), on note K le nombre de composantes du vecteur d'état \mathbf{X} ($K = 3$ dans le modèle (2.1)). A chaque instant $t \in \mathbb{R}^+$, la valeur de \mathbf{X} est notée $\mathbf{X}^i(t) = (X_1^i(t), \dots, X_K^i(t))$. Le vecteur de paramètres du modèle est noté $\boldsymbol{\xi}^i = (\xi_1^i, \dots, \xi_p^i)$ qui détermine la trajectoire attendue du vecteur d'état :

$$\begin{cases} \frac{dX^i(t)}{dt} = f(X^i(t), \boldsymbol{\xi}^i) \\ X^i(0) = h(\boldsymbol{\xi}^i) \end{cases} \quad (2.3)$$

où f et h sont supposées deux fois différentiables par rapport à $\boldsymbol{\xi}^i$. L'existence d'une condition d'équilibre à t_0 est une hypothèse fréquente, mais pas indispensable : cette hypothèse est souvent contredite par les données réelles.

Dans la mesure où (2.3) n'a pas de solution analytique, certains auteurs proposent de faire certaines hypothèses biologiques comme nous l'avons discuté précédemment. Ces hypothèses permettent de simplifier le modèle de départ et de travailler sur un modèle défini par une solution explicite.

Modèle statistique pour les observations

Les observations ne correspondent pas toujours aux vecteurs d'états du système. Par exemple, on peut distinguer dans le modèle du virus infectieux et du virus non-infectieux alors que la mesure effectuée ne les différencie pas ou encore des cellules cibles des cel-lules infectées mais nous ne mesurons que la somme des deux (les CD4 totaux). De plus, pour des raisons d'adéquation aux hypothèses d'homoscédasticité, on préfère travailler sur des transformations des quantités mesurées ; on travaille généralement en échelle logarithmique pour la charge virale. D'après des travaux basés sur des modèles descriptifs, utiliser une transformation logarithmique ou racine carrée (voire quatrième) des CD4 conduit à

des résidus homoscédastes. De manière plus générale, on observe des fonctions des variables d'états du système $g_m(\cdot)$ de \mathbb{R}^K vers \mathbb{R} . L'ensemble de ces M fonctions est appelé "compartiments observés". On pose y_{ijm} la j -ème observation du m -ème compartiment ($j \leq n_{im}$) observé du sujet i au temps t_{ijm} . Le modèle pour les observations peut s'écrire :

$$\begin{cases} y_{ijm} = g_m(\mathbf{X}(t_{ijm}, \boldsymbol{\xi}^i)) + \epsilon_{ijm} \\ \boldsymbol{\xi}^i \sim \mathcal{N}(\boldsymbol{\phi}', \boldsymbol{\Sigma}') \\ \epsilon_{ijm} = \epsilon_{ijm,mes} + \epsilon_{ijm,ech} \end{cases} \quad (2.4)$$

où ϵ_{ijm} représente les variations intra-sujets non expliquées par le modèle. ϵ_{ijm} peut se décomposer en une erreur d'observation $\epsilon_{ijm,mes}$ et une erreur d'échantillon $\epsilon_{ijm,ech}$. Celles-ci sont supposées gaussiennes et indépendantes conditionnellement à $\boldsymbol{\xi}^i$; en outre $\mathbb{E}(\epsilon_{ijm,mes} | \boldsymbol{\xi}^i) = \mathbb{E}(\epsilon_{ijm,ech} | \boldsymbol{\xi}^i) = 0$.

Nous tenons compte des contraintes existant sur les paramètres, en particulier celle de positivité des paramètres biologiques, en imposant des transformations Ψ aux paramètres. En notant $\tilde{\boldsymbol{\xi}}_l^i = \Psi_l(\xi_l^i)$, $l \leq p$ le vecteur des paramètres individuels reparamétrisés, le système (2.4) devient :

$$\begin{cases} y_{ijm} = g_m(X(t_{ijm}, \tilde{\boldsymbol{\xi}}^i)) + \epsilon_{ijm} \\ \boldsymbol{\xi}^i \sim \mathcal{N}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \\ \epsilon_{ijm} \sim \mathcal{N}(0, \sigma_m^2) \end{cases} \quad (2.5)$$

Les fonctions g_m sont obtenues par résolution numérique du système d'équations différentielles (2.3). Notons que l'essentiel des travaux de la littérature reposent sur une approche simplifiée et g_m a dans ce cas une écriture explicite.

Le modèle statistique pour les observations est une extension du modèle (2.5) :

$$\begin{cases} y_{ijm} = g_m(X(t_{ijm}, \tilde{\boldsymbol{\xi}}^i)) + \epsilon_{ijm} \\ \tilde{\boldsymbol{\xi}}_l^i = \phi_l + \mathbf{z}_l^i \boldsymbol{\beta}_l + \boldsymbol{\omega}_l^i \mathbf{b}^i, \quad l \leq p \end{cases} \quad (2.6)$$

où ϕ_l est l'intercept du l -ème paramètre biologique $\tilde{\boldsymbol{\xi}}_l^i$, \mathbf{z}_l^i et $\boldsymbol{\omega}_l^i$ sont les vecteurs des variables explicatives associées respectivement aux effets fixes et aux effets aléatoires de $\tilde{\boldsymbol{\xi}}_l^i$. Le vecteur $\boldsymbol{\beta}_l$ correspond aux coefficients de régression associés aux effets fixes. On suppose que $\mathbf{b}^i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, où \mathbf{b}^i est le vecteur individuel des effets aléatoires de dimension q et $\mathbf{b}^i | \mathbf{z}^i \perp \epsilon_i | \mathbf{z}^i$, où $\mathbf{z}^i = (z_l^i)_{l=1,\dots,p}$. Soit $\mathbf{A} = (a_{l''l'})_{l'' \leq l' \leq q}$ la matrice triangulaire inférieure de termes diagonaux positifs telle que $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$ (décomposition de Cholesky). On peut écrire

$\mathbf{b}^i = \mathbf{A}\mathbf{u}^i$ avec $\mathbf{u}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$. Notons $\boldsymbol{\phi} = (\phi_l)_{l=1,\dots,p}$, $\boldsymbol{\beta} = (\beta_l)_{l=1,\dots,p}$, $\boldsymbol{\sigma} = (\sigma_m)_{m=1,\dots,M}$ et $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma})$ avec $\dim(\boldsymbol{\theta}) = s$.

Notations

Soit $p(\mathbf{y}_i | \mathbf{z}^i, \mathbf{u}^i; \boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma})$ la vraisemblance de l'individu i sachant les effets aléatoires \mathbf{u}^i et les covariables \mathbf{z}^i . On peut écrire :

$$p(\mathbf{y}_i | \mathbf{z}^i, \mathbf{u}^i; \boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma}) = \prod_{j,m} \frac{1}{\sigma_m \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{ijm} - g_m(X(t_{ijm}, \tilde{\boldsymbol{\xi}}^i))}{\sigma_m} \right)^2 \right]$$

On en déduit la contribution marginale de l'individu i à la vraisemblance observée, en notant Φ la densité de la loi normale standard de dimension q :

$$p(\mathbf{y}_i | \mathbf{z}^i; \boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma}) = \int_{u \in \mathbb{R}^q} \left\{ \prod_{j,m} \frac{1}{\sigma_m \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{ijm} - g_m(X(t_{ijm}, \tilde{\boldsymbol{\xi}}^i))}{\sigma_m} \right)^2 \right] \right\} \Phi(\mathbf{u}) d\mathbf{u} \quad (2.7)$$

Les patients étant supposés indépendants, on peut en déduire la vraisemblance observée sur l'échantillon $p(\mathbf{y} | \mathbf{z}; \boldsymbol{\theta})$ ainsi que la log-vraisemblance $L(\mathbf{y}; \boldsymbol{\theta})$:

$$\begin{aligned} p(\mathbf{y} | \mathbf{z}; \boldsymbol{\theta}) &= \prod_i p(\mathbf{y}_i | \mathbf{z}^i; \boldsymbol{\theta}) \\ L(\mathbf{y}; \boldsymbol{\theta}) &= \sum_i L(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_i \log [p(\mathbf{y}_i | \mathbf{z}^i; \boldsymbol{\theta})] \end{aligned}$$

L'estimateur du maximum de vraisemblance est : $\hat{\boldsymbol{\theta}}_{EMV} = \underset{\boldsymbol{\theta} \in \mathbb{R}^s}{\operatorname{Argmax}} L(\mathbf{y}; \boldsymbol{\theta})$.

Soit $\mathcal{I}_i(\boldsymbol{\theta})$ la Matrice d'Information de Fisher (FIM) individuelle et $\mathcal{I}(\boldsymbol{\theta})$ la FIM sur tout l'échantillon. Pour simplifier les notations, supposons qu'il n'y ait pas de variables explicatives dans le modèle. On a :

$$\mathcal{I}_i(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left(-\frac{\partial^2 L(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) \quad \mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left(-\frac{\partial^2 L(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)$$

Soit \mathbf{H} la Hessienne de la log-vraisemblance $L(\mathbf{y}; \boldsymbol{\theta})$:

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 L(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

et on a $\frac{1}{n} \mathbf{H}(\hat{\boldsymbol{\theta}}_{EMV}) \xrightarrow{p} -\mathcal{I}_i(\boldsymbol{\theta}^*)$ (borne de Rao), si $\boldsymbol{\theta}^*$ est le vecteur des vrais paramètres. Sous réserve de régularité et d'identifiabilité du modèle, l'estimateur du maximum de vraisemblance est asymptotiquement distribué selon une loi normale, d'espérance $\boldsymbol{\theta}^*$ et

de matrice de variance-covariance $\mathcal{I}(\boldsymbol{\theta}^*)^{-1}$: $\sqrt{n} (\hat{\boldsymbol{\theta}}_{EMV} - \boldsymbol{\theta}^*) \xrightarrow[n \rightarrow \infty]{} \mathcal{N}(0, \mathcal{I}_i(\boldsymbol{\theta}^*)^{-1})$. La Hessienne au maximum de vraisemblance fournit donc un estimateur de la matrice de variance-covariance des paramètres estimés. Pris marginalement, ses éléments fournissent une estimation de la variance des paramètres estimés.

2.3.2 Approche bayésienne

Le paradigme bayésien

Le principe des méthodes bayésiennes est de considérer le vecteur paramètre $\boldsymbol{\theta}$ comme une variable aléatoire et non pas inconnu et déterministe. L'incertitude sur $\boldsymbol{\theta}$ est modélisée par une loi de distribution notée $\pi(\boldsymbol{\theta})$ sur Θ (l'espace des paramètres). Cette loi est appelée “loi a priori”. L'inférence se base sur la loi de distribution de $\boldsymbol{\theta}$ conditionnellement aux observations \mathbf{Y} , $\pi(\boldsymbol{\theta}|\mathbf{Y})$ appelée “loi a posteriori”. D'après le théorème de Bayes :

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

où $p(\mathbf{Y}|\boldsymbol{\theta})$ est la vraisemblance du modèle. Le problème est alors d'évaluer la loi a posteriori $\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. La loi a posteriori est alors connue aux constantes de normalisation près. Quand la vraisemblance n'a pas d'expression analytique et les constantes de normalisation des différentes distributions ne pouvant être calculées, l'évaluation de cette loi a posteriori est difficile. Les méthodes de Monte-Carlo par Chaînes de Markov (MCMC) ont été développées pour contourner ce problème.

Méthode bayésienne hiérarchique

Dès 1998, Wu et al. (1998) propose une estimation des paramètres d'un modèle dynamique à l'aide de modèles non-linéaires hiérarchiques. Ils modélisent la dynamique virale après mise sous traitement à l'aide du modèle (2.2). Les modèles hiérarchiques ont l'avantage d'être flexibles et de prendre en compte les données riches et déséquilibrées des patients. Ils permettent également d'inclure des covariables. Ces modèles ont été introduit en pharmacocinétique/pharmacodynamique par Sheiner et al. (1972). Cette méthode a été utilisée pour estimer un modèle non-simplifié de la dynamique de la charge virale par : Putter et al. (2002), Han et Chaloner (2004), Huang et al. (2006), Huang et Wu (2006). Le

modèle bayésien hiérarchique à effets mixtes peut s'écrire d'après les 3 niveaux suivants (Davidian et Giltinan, 1995) :

- *Niveau 1.* Variation intra-sujet :

$$\mathbf{y}_i = g_m(\boldsymbol{\theta}) + \boldsymbol{\epsilon}_i, \text{ avec } [\boldsymbol{\epsilon}_i | \sigma^2, \boldsymbol{\theta}] \sim \mathcal{N}(0, \sigma^2 I_{n_{im}}).$$

- *Niveau 2.* Variation inter-sujet :

$$\boldsymbol{\theta} = \boldsymbol{\phi} + \mathbf{b}^i, \text{ avec } [\mathbf{b}^i | \boldsymbol{\Sigma}] \sim \mathcal{N}(0, \boldsymbol{\Sigma}).$$

- *Niveau 3.* Distribution des hyper-priors :

$\sigma^{-2} \sim Ga(a, b)$, $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Lambda})$ et $\boldsymbol{\Sigma}^{-1} \sim Wi(\boldsymbol{\Omega}, \nu)$, où les distributions Gamma (Ga), normale (\mathcal{N}) et Wishart (Wi) sont mutuellement indépendantes pour simplifier les calculs.

Huang et al. (2006) proposent de combiner l'échantillonnage de Gibbs et l'algorithme de Metropolis-Hastings. Dans leur approche, l'échantillonnage de Gibbs (Geman et Geman, 1984) met à jour σ^{-2} , $\boldsymbol{\phi}$ et $\boldsymbol{\Sigma}^{-1}$ alors que l'algorithme de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) met à jour les $\boldsymbol{\theta}$.

L'approche bayésienne permet d'apporter une certaine souplesse comparée à l'approche fréquentiste car il est possible théoriquement d'estimer tous les paramètres du modèle. En effet, au lieu de fixer certains paramètres nous pouvons les introduire dans le modèle avec des a priori "forts". De plus, le fait d'utiliser les connaissances a priori sur les paramètres est d'autant plus pertinent dans le cadre de modèles déjà basés sur les connaissances biologiques tels que les modèles dynamiques de l'interaction entre le VIH et le système immunitaire. Notons également que les temps de calculs dus à l'utilisation d'approche MCMC peuvent être importants pour des modèles complexes et que les critères de convergence et de sélection de modèles ne sont pas toujours bien définis.

2.3.3 Approche fréquentiste

Nous présentons maintenant deux méthodes d'estimation réalisant une maximisation exacte de la vraisemblance. L'une est fondée sur l'utilisation de l'algorithme Expectation-Maximisation (EM) et l'autre sur l'algorithme de Newton-Raphson (NR).

Méthode basée sur des approximations stochastiques

L'algorithme EM a été proposé par Dempster et al. (1977). Afin d'éviter le calcul des intégrales numériques de (2.7), il propose de poser à l'itération k :

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k) = \mathbb{E} [\log(p(\mathbf{y}, \mathbf{u}|\mathbf{z}; \boldsymbol{\theta})|\mathbf{y}; \boldsymbol{\theta}_k)].$$

On montre que si $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k) \geq Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k)$ alors $L(\mathbf{y}; \boldsymbol{\theta}_k) \geq L(\mathbf{y}; \boldsymbol{\theta})$. Ainsi $L(\mathbf{y}; \cdot)$ augmente avec Q . La suite des paramètres $(\boldsymbol{\theta}_k)$ converge vers un extremum de la vraisemblance. S'appuyant sur ce principe, Dempster et al. (1977) proposent l'algorithme itératif EM. A l'itération k :

- *Etape E* : Calcul de l'espérance conditionnelle de la log-vraisemblance : $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k)$.
- *Etape M* : Maximisation de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k)$, on prend $\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{Argmax}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k)$.

L'algorithme SAEM (pour *Stochastic Approximation EM*) décrit par Kuhn et Lavielle (2005) repose sur une approximation stochastique de l'étape E de l'algorithme EM, passant ainsi outre le problème numérique du calcul intégral. L'étape E est alors divisée en une étape de simulation et une étape d'approximation stochastique.

A l'itération k , on simule une réalisation \mathbf{u}_k de \mathbf{u} d'après une loi conditionnelle $\pi(\mathbf{u}|\mathbf{y}; \boldsymbol{\theta}_k)$ (étape de simulation), puis on réalise une approximation stochastique en mettant à jour Q_k en introduisant une suite décroissante de pas positifs (γ_k) :

$$Q_{k+1}(\boldsymbol{\theta}) = Q_k(\boldsymbol{\theta}) + \gamma_k (\log p(\mathbf{y}, \mathbf{u}_k; \boldsymbol{\theta}_k) - Q_k(\boldsymbol{\theta})).$$

L'étape de maximisation M consiste alors à mettre à jour $\boldsymbol{\theta}_k$ en prenant :

$$\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{Argmax}} Q_{k+1}(\boldsymbol{\theta})$$

Quand la loi conditionnelle $\pi(\mathbf{u}|\mathbf{y}; \boldsymbol{\theta}_k)$ ne peut pas être simulée, Kuhn et Lavielle (2004) ont proposé de combiner leur algorithme avec une méthode de MCMC. L'étape de simulation est alors la seule modifiée. A l'itération k , on simule \mathbf{u}_k par un algorithme de Metropolis-Hasting d'après une probabilité de transition $\Pi_{\boldsymbol{\theta}_k}$ correspondant à une chaîne de Markov de loi stationnaire $\pi(\mathbf{u}|\mathbf{y}; \boldsymbol{\theta}_k)$.

Cette méthode est plus rapide et plus efficace en terme de biais et d'erreur de moindre carré (RMSE) qu'un algorithme EM classique. Cet algorithme a, d'autre part, été complexifié pour prendre en compte la censure à gauche de la charge virale (Samson et al.,

2006) mais également permettre l'estimation de modèles dynamiques stochastiques (Donnet et Samson, 2008).

Méthode basée sur un algorithme de type Newton

L'algorithme de Newton-Raphson est une méthode classique d'estimation par maximum de vraisemblance. C'est un algorithme itératif qui se base sur les fonctions de score (\mathbf{U}) et hessienne (\mathbf{H}) de la vraisemblance d'après le schéma suivant :

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{H}^{-1}(\boldsymbol{\theta}_k)\mathbf{U}(\boldsymbol{\theta}_k).$$

Le plus fréquemment \mathbf{U} et \mathbf{H} sont calculées par différences finies :

$$\begin{aligned}\mathbf{U}(\boldsymbol{\theta}) &\simeq \frac{\text{L}(\mathbf{y}; \boldsymbol{\theta} + d\boldsymbol{\theta}) - \text{L}(\mathbf{y}; \boldsymbol{\theta} - d\boldsymbol{\theta})}{2 * d\boldsymbol{\theta}} \\ \mathbf{H}(\boldsymbol{\theta}) &\simeq \frac{\text{L}(\mathbf{y}; \boldsymbol{\theta} + d\boldsymbol{\theta} + d\boldsymbol{\theta}') - \text{L}(\mathbf{y}; \boldsymbol{\theta} + d\boldsymbol{\theta}) - \text{L}(\mathbf{y}; \boldsymbol{\theta} + d\boldsymbol{\theta}') + \text{L}(\mathbf{y}; \boldsymbol{\theta})}{d\boldsymbol{\theta} d\boldsymbol{\theta}'}\end{aligned}$$

mais des approximations de Monte-Carlo (Tanner, 1996; McCulloch, 1997) et stochastiques (Robbins et Monro, 1951) de l'algorithme NR ont également été proposées.

Une autre méthode d'estimation est celle développée par Guedj et al. (2007) qui se base sur l'algorithme RVS (pour *Robust Variance Scoring*) proposé par Commenges et al. (2006). Cette méthode utilise seulement les dérivées premières de la log-vraisemblance pour l'estimation des paramètres. La vraisemblance $p(\mathbf{y}_i | \mathbf{z}^i; \boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma})$ pour chaque individu i est calculée par quadrature-gaussienne adaptative. Le calcul des scores se réalise en 2 étapes :

- Calcul des scores de la vraisemblance totale conditionnellement aux effets aléatoires : $\mathbf{U}_i(\boldsymbol{\theta} | \mathbf{u}^i)$
- Calcul des scores de la vraisemblance observée par intégration : $\mathbf{U}_i(\boldsymbol{\theta})$

Cette seconde étape utilise la relation de Louis (1982) généralisée par Commenges et Rondeau (2006) :

$$\mathbf{U}_i(\boldsymbol{\theta}) = p(\mathbf{y}_i | \mathbf{z}^i; \boldsymbol{\theta})^{-1} \int_{\mathbb{R}^q} p(\mathbf{y}_i | \mathbf{z}^i, \mathbf{u}^i; \boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma}) \mathbf{U}_i(\boldsymbol{\theta} | \mathbf{u}^i) \Phi(\mathbf{u}) d\mathbf{u}.$$

La Hessienne à l'itération k est approchée par :

$$G(\boldsymbol{\theta}_k) = \sum_{i \leq n} \mathbf{U}_i(\boldsymbol{\theta}_k) \mathbf{U}'_i(\boldsymbol{\theta}_k) + \frac{\nu}{n} U(\boldsymbol{\theta}_k) U'(\boldsymbol{\theta}_k)$$

où ν est un poids, \mathbf{U}_i le vecteur des scores pour l'individu i (\mathbf{U}'_i sa transposée) et $U = \sum_{i \leq n} \mathbf{U}_i$. A convergence, la matrice $\mathbf{G}(\hat{\boldsymbol{\theta}})$ est un estimateur de la matrice d'information $I(\boldsymbol{\theta}^*)$.

Cet algorithme est plus rapide et plus performant que le Newton-Raphson original et Marquardt. De plus, les critères de convergence sont définis clairement en comparaison avec l'approche bayésienne et le SAEM. La prise en compte de données censurées à gauche a également été implémentée. Cependant, sur des modèles complexes, les temps de calcul peuvent être long. L'algorithme doit résoudre le système ODE, calculer une intégrale multiple dont la dimension est égale au nombre d'effets aléatoires inclus et enfin maximiser la vraisemblance. Le nombre d'effets aléatoires est généralement limité à 3 ou 4 pour ces questions numériques.

2.3.4 Maximum a posteriori

Une troisième approche, qui peut être considérée comme intermédiaire entre l'approche bayésienne et l'approche fréquentiste, est celle du maximum a posteriori (MAP). Cette méthode fut abandonnée par les bayésiens au profit des méthodes MCMC mais est toujours utilisée en imagerie et ingénierie biomédicale. Elle permet de combiner les avantages des deux approches car elle prend en compte les a priori sur les paramètres et ne fait pas d'approximation de la vraisemblance.

L'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\theta}}_{EMV}$ est défini par : $\hat{\boldsymbol{\theta}}_{EMV} = \operatorname{Argmax}_{\boldsymbol{\theta} \in \Theta} L(\mathbf{y}; \boldsymbol{\theta})$ et l'estimateur du MAP $\hat{\boldsymbol{\theta}}_{MAP}$ par : $\hat{\boldsymbol{\theta}}_{MAP} = \operatorname{Argmax}_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}|\mathbf{y})$. D'après le théorème de Bayes, on a : $\hat{\boldsymbol{\theta}}_{MAP} = \operatorname{Argmax}_{\boldsymbol{\theta} \in \Theta} \left[L(\mathbf{y}; \boldsymbol{\theta}) + \log(\pi(\boldsymbol{\theta})) \right]$. On peut donc voir le MAP comme une vraisemblance pénalisée au sens classique (Commenges, 2009). Si on suppose que $\boldsymbol{\theta}$ est gaussien de vecteur moyen $\boldsymbol{\mu}$ et matrice de variance-covariance $\boldsymbol{\Omega}$ alors $\hat{\boldsymbol{\theta}}_{MAP}$ maximise la fonction :

$$L(\mathbf{y}; \boldsymbol{\theta}) + \log \left\{ \frac{1}{(2\pi)^{s/2} \det \boldsymbol{\Omega}} \exp \left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right) \right\} \propto L(\mathbf{y}; \boldsymbol{\theta}) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}).$$

Les estimateurs MAP sont asymptotiquement équivalents aux estimateurs du maximum de vraisemblance classiques (Diaconis et Freedman, 1986). Lorsque la taille de l'échantillon tend vers l'infini, l'information contenue dans cet échantillon devient prédominante par rapport à l'information "fixe" apportée par la loi a priori $\pi(\boldsymbol{\theta})$.

Cette méthode permet de prendre en compte les connaissances a priori sur les paramètres et de réaliser une maximisation exacte de la vraisemblance. Elle semble pertinente dans le contexte de l'étude de l'interaction hôte-pathogène étant donné que les paramètres des modèles dynamiques ont une signification biologique. L'algorithme SAEM offre la possibilité d'inclure des a priori sur les paramètres à estimer et le MAP est très facilement implémentable dans la méthode proposée par Guedj et al. (2007).

2.3.5 Conclusion

En conclusion, plusieurs méthodes d'estimation sont disponibles pour les modèles dynamiques. Il faut cependant prendre en considération les difficultés numériques liées à ce type de modélisation. Les calculs sont en effet très lourds avec les méthodes MCMC et le choix entre les différentes méthodes fréquentistes dépend de nombreux critères : biais, précision, efficacité et temps de calculs. Aucune comparaison formelle n'a été jusqu'à présent réalisée dans le cadre des modèles dynamiques.

2.4 Objectif du travail de thèse

Nous avons présenté, dans ce second chapitre, une modélisation basée sur les connaissances biologiques qui permet de prendre en compte l'interaction complexe et non-linéaire qui existe entre le VIH et les cellules du système immunitaire ainsi que les difficultés liées à l'estimation de ce type de modèle. Nous avons exposé dans le chapitre 1, l'intérêt d'étudier la phase de primo-infection à VIH qui permet de mieux comprendre les mécanismes viro-immunologiques précoces de l'infection qui vont régir en partie l'évolution clinique de la maladie. Par ailleurs, nous avons exposé les difficultés liées à l'étude de la primo-infection : manque de données et date d'infection inconnue.

Il nous a semblé pertinent d'utiliser les données de cohortes issues de la collaboration CASCADE qui constitue l'une des plus grandes bases de données de patients séroconverteurs infectés par le VIH. Cette base de données constitue une source d'information importante puisque les patients sont suivis de leur séroconversion jusqu'au décès. Afin d'étudier les dynamiques viro-immunologiques durant la phase de primo-infection, il est nécessaire de développer une méthode d'estimation prenant en compte l'incertitude sur la date d'infection (origine des temps du modèle). Au vu des travaux antérieurs de l'équipe en matière d'estimation de modèles non-linéaires à effets mixtes et plus particulièrement dans le domaine de la modélisation dynamique de l'infection à VIH, nous avons choisi de travailler sur l'algorithme proposé par Guedj et al. (2007).

D'autre part, le nombre de paramètres présents dans les modèles étudiés étant souvent important (6 dans le modèle de Perelson et al. (1996) et 11 dans celui de Guedj et al. (2007)), le choix des effets aléatoires et des variables explicatives à inclure est un problème non-négligeable. En effet, le nombre de modèles à comparer (et donc à estimer) peut devenir très grand et il devient alors impossible de tous les tester. Le développement d'un outil permettant de tester rapidement les différentes possibilités semble alors indispensable pour permettre la sélection de modèles.

Plusieurs modèles sont fréquemment utilisés pour décrire la phase primo-infection à VIH : le modèle de base présenté par Perelson et al. (1996), un modèle incluant des cellules quiescentes et des cellules activées, ou encore un modèle incluant une réponse immunitaire. En outre, il semble pertinent de les comparer en terme d'estimation de paramètres (taux de décès des cellules cibles, des cellules infectées, nombre de reproduction de base

R_0), de dynamique (temps du pic viral, nombre de cellules infectées à l'équilibre, ...) mais surtout en terme statistique et d'adéquation aux données.

La présentation de nos travaux est la suivante : dans le chapitre 3, nous proposons une méthode d'estimation prenant en compte l'incertitude sur la date d'infection pour des modèles dynamiques. Cette méthode est illustrée par l'estimation des paramètres de deux modèles dynamiques pour 761 patients issus de la collaboration CASCADE. Dans le chapitre 4, nous développons des tests du score pour l'exploration de modèles complexes avec une application à la modélisation dynamique du VIH et dans le chapitre 5, nous réalisons une comparaison de différents modèles dynamiques ayant été estimés à partir des données de 761 patients durant la première année de suivi. Une conclusion générale prend place dans le chapitre 6.

Chapitre 3

Estimation d'un modèle ODE en prenant en compte l'incertitude sur la date d'infection

Les chapitres 1 et 2 ont montré l'apport des modèles dynamiques dans l'étude de l'interaction entre le système immunitaire et le VIH ainsi que l'intérêt d'étudier la phase de primo-infection à VIH. L'une des difficultés majeures dans l'étude de l'infection à VIH est que la date d'infection est inconnue. Or cette date correspond à l'origine des temps pour un modèle dynamique de primo-infection. Cependant, pour certains patients, nous disposons d'une date de dernière sérologie négative et d'une date de première sérologie positive. Ces deux dates nous renseignent sur la date de séroconversion (date d'apparition des anticorps anti-VIH) et par conséquent sur la date d'infection.

Nous proposons donc dans cette partie une méthode d'estimation pour un modèle ODE prenant en compte l'incertitude sur l'origine des temps dans le cadre de l'infection à VIH basée sur celle décrite par Guedj et al. (2007). En supposant que la date d'infection pour un sujet i (τ_i) a une densité connue f_{τ_i} de support $[b_{inf}^i, b_{sup}^i]$. La vraisemblance observée pour le sujet i est donnée par :

$$\int_{u \in \mathbb{R}^q} \int_{b_{inf}^i}^{b_{sup}^i} p(\mathbf{y}_i | \tau^i, \mathbf{z}^i, \mathbf{u}^i; \boldsymbol{\phi}, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\sigma}) f_{\tau_i}(s) \Phi(\mathbf{u}) ds d\mathbf{u}$$

La vraisemblance observée totale s'obtient donc en intégrant sur les effets aléatoires et sur la fenêtre de dates d'infection possible. La densité de la date d'infection f_{τ_i} a été

définie à partir des travaux de Fiebig et al. (2003) et prend en compte le délai nécessaire à la production d'anticorps par un individu infecté mais également celui nécessaire pour avoir un niveau détectable d'anticorps. En effet, un patient peut avoir un test sérologique négative mais être déjà infecté par le VIH et peut même avoir une charge virale détectable.

Nous avons appliqué cette méthode à un jeu de données réelles de séroconvertisseurs issus de la collaboration CASCADE. Nous avons proposé deux modèles de primo-infection. Le modèle de Perelson et al. (1996) :

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dT^*}{dt} = \gamma VT - \mu_{T^*} T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{cases}$$

Les compartiments sont supposé être à l'équilibre avant l'infection, ce qui peut se traduire par : $T_{0-} = \lambda/\mu_T$, $T_{0-}^* = 0$ et $V_{0-} = 0$. Après l'introduction de virus dans le système ($V_0 = 10^{-6}$ virions/mm³), l'état d'équilibre initial est perturbé et le système converge vers un nouvel état d'équilibre si R_0 ($R_0 = \frac{\gamma\lambda\pi}{\mu_T\mu_{T^*}\mu_V}$) est supérieur à 1 :

$$\begin{cases} \bar{T} = \frac{\mu_{T^*}\mu_V}{\gamma\pi} \\ \bar{T}^* = \frac{\lambda\gamma\pi - \mu_T\mu_{T^*}\mu_V}{\gamma\pi\mu_{T^*}} \\ \bar{V} = \frac{\lambda\gamma\pi - \mu_T\mu_{T^*}\mu_V}{\gamma\mu_{T^*}\mu_V} \end{cases}$$

Ce modèle fait l'hypothèse qu'une cellule, une fois infectée, devient immédiatement capable de produire des virions, alors qu'il a été montré *in vitro* qu'un délai d'environ un jour était nécessaire pour qu'une cellule infectée devienne productrice (Kiernan et al., 1990; Barbosa et al., 1994). Pour prendre en compte ce délai, nous proposons un second modèle en introduisant un compartiment supplémentaire. Ce modèle distingue les cellules infectées non-productrices I des cellules infectées productrices P comme précédemment proposé par Nowak et al. (1997) :

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dI}{dt} = \gamma VT - \mu_I I - \delta I \\ \frac{dP}{dt} = \delta I - \mu_I P \\ \frac{dV}{dt} = \pi P - \mu_V V \end{cases}$$

Le taux pour qu'une cellule infectée devienne productrice est noté δ . L'équilibre avant infection pour ce modèle peut s'écrire : $T_{0-} = \lambda/\mu_T$, $I_{0-} = 0$, $P_{0-} = 0$ et $V_{0-} = 0$.

Le système se stabilise à un nouvel équilibre après introduction de virus si R_0 ($R_0 = \frac{\lambda\gamma\pi\alpha}{\mu_T\mu_I\mu_V(\mu_I+\alpha)}$) est supérieur à 1 :

$$\left\{ \begin{array}{l} \bar{T} = \frac{\mu_I\mu_V(\mu_I+\alpha)}{\gamma\pi\alpha} \\ \bar{I} = \frac{\lambda}{\mu_I+\alpha} - \frac{\mu_T\mu_V\mu_I}{\gamma\pi\alpha} \\ \bar{P} = \frac{\alpha}{\mu_I} \left(\frac{\lambda}{\mu_I+\alpha} - \frac{\mu_T\mu_V\mu_I}{\gamma\pi\alpha} \right) \\ \bar{V} = \frac{\pi\alpha}{\mu_V\mu_I} \left(\frac{\lambda}{\mu_I+\alpha} - \frac{\mu_T\mu_V\mu_I}{\gamma\pi\alpha} \right) \end{array} \right.$$

Ce travail fait l'objet d'un article soumis dans *Annals of Applied Statistics*.

Submitted to the Annals of Applied Statistics

MODELING THE DYNAMICS OF BIOMARKERS DURING PRIMARY HIV INFECTION TAKING INTO ACCOUNT THE UNCERTAINTY OF INFECTION DATE

BY J. DRYLEWICZ ^{1,2 *}, J. GUEDJ ³, D. COMMENGES ^{1,2} AND R.
THIÉBAUT ^{1,2,†} ON BEHALF OF THE CASCADE COLLABORATION

(1) *INSERM, U897 Epidemiology and Biostatistics, (2) Bordeaux 2 Victor Segalen University, (3) Bar-Ilan University*

During primary HIV infection, the kinetics of plasma virus concentrations and CD4+ cell count is very complex. Parametric and non-parametric models have been suggested for fitting repeated measurements of these markers. Alternatively, mechanistic approaches based on ordinary differential equations have also been proposed. These latter models are constructed according to biological knowledge and take into account the complex non-linear interactions between virus and cells. However, estimating the parameters of these models is difficult. A main difficulty in the context of primary HIV infection is that the date of infection is generally unknown. For some patients, the date of last negative HIV test is available in addition to the date of first positive HIV test (seroconverters). In this paper, we propose a likelihood-based method for estimating the parameters of dynamical models using a population approach and taking into account the uncertainty of the baseline time. We applied this method to a sample of 761 HIV-infected patients from the Concerted Action on Seroconversion to AIDS and Death in Europe (CASCADE).

1. Introduction.

Primary Human Immunodeficiency Virus (HIV) infection is a crucial period during HIV infection history where there is a viral burst due to the spread of the virus through target cells, mainly CD4+ T lymphocytes (CD4). The dynamics of markers at that time is believed to partly determine the evolution of the infection in the future ([Mellors et al., 1995](#)). For instance, the peak of viral load has been shown to be predictive of the viral setpoint, i.e. the plasma HIV RNA level at which patients often stay for several years ([Lindback et al., 2000](#)) unless they are treated. This viral setpoint is associated with clinical progression ([Mellors et al., 1995](#)).

*J. Drylewicz received a PhD grant from CASCADE Concerted Action on Seroconversion to AIDS and Death in Europe (European Commission Framework Program 6)

†Corresponding author: rodolphe.thiebaut@isped.u-bordeaux2.fr

Keywords and phrases: Dynamical model, non-linear models, ordinary differential equations, HIV dynamics, primary infection

Parametric and non parametric descriptive models have been suggested for fitting repeated measurements of CD4 and HIV RNA (or viral load) (Dubin *et al.*, 1994; Desquibet *et al.*, 2004; Pantazis *et al.*, 2005; Hecht *et al.*, 2006; Geskus *et al.*, 2007). A mechanistic approach based on ordinary differential equations (ODE) has also been suggested (Phillips, 1996; De Boer and Perelson, 1998). These mathematical models present several advantages. First, they are based on biological knowledge. Therefore, the parameters may have direct biological meaning and the relationship between markers is modeled through biological mechanisms rather than parametric correlation structures. Second, this type of dynamical model is able to capture complex interaction between markers. For instance, these models can predict the decrease of viral load following the peak as a consequence of the limitation of target cells (Phillips, 1996). The first models used in this context gave important insights on the dynamics of the infection and how to control it (Nowak *et al.*, 1997b; Little *et al.*, 1999). Numerous attempts to improve models have been published (De Boer and Perelson, 1998; Wick, 1999; Stafford *et al.*, 2000). Most often, the parameters of these models are not estimated and values which appear “reasonable” are chosen. Indeed, estimating the parameters in such models is difficult. To improve identifiability, random effects models (or population approach) can be used. However, the combination of non-linear ODE systems and random effects leads to difficult numerical problems (Putter *et al.*, 2002; Filter *et al.*, 2005; Huang *et al.*, 2006; Samson *et al.*, 2006; Guedj *et al.*, 2007a; Cao *et al.*, 2008).

The dynamics of the biomarkers during primary HIV infection is quite complex. In the few studies, where ODE models have been used in the context of primary infection, the parameters were estimated with the viral load data only (Kaufmann *et al.*, 1998; Little *et al.*, 1999; Stafford *et al.*, 2000; Lindback *et al.*, 2000; Ciupe *et al.*, 2006) or the individual fit of the CD4 counts data was very bad (Murray *et al.*, 1998). Moreover, the parameters were not estimated with random effects models.

In the primary HIV infection context, the system is in a “trivial” equilibrium state without virus and it is disrupted by the introduction of a small quantity of viruses (the “infection event”). The date of infection must thus be known (in contrast to studies of the response to a treatment initiation far from the infection). This date is most often unknown; bringing an additional difficulty. At best, the only available information is the date of the last seronegative test and first seropositive test or first detectable HIV RNA. In some cohort studies (such as the Multicenter AIDS Cohort study (Muñoz *et al.*, 1992)) the last seronegative date is prospectively identified during follow-up. In most cases, patients are seropositive at entry into the

cohort and the last seronegative date is retrospectively recorded. Usually, the date of seroconversion is imputed at the midpoint between last negative and first positive HIV serology test (Fidler *et al.*, 2007). Several methods have been suggested for estimating this date according to the marker values (Berman, 1990; DeGruttola *et al.*, 1991; Dubin *et al.*, 1994; Geskus, 2000). Geskus *et al.* (2007) performed a conditional mean imputation to estimate the date of infection that was used as the baseline for linear mixed models.

In the present paper, we propose a method for estimating the parameters of dynamic models using a population approach taking into account the uncertainty of the infection date. We estimate the parameters of a dynamical model using HIV RNA and CD4 count data in a large collaboration including cohorts of patients with a documented date of negative and positive HIV serologies (HIV seroconverters).

There are several direct applications of such an approach. First, we can estimate fundamental parameters such as virion clearance, the infected cell death rate (Perelson *et al.*, 1997), virus infectivity (Wilson *et al.*, 2007) or the basic reproductive number R_0 (Little *et al.*, 1999). Second, the huge variability in the peak level of HIV RNA and of the viral setpoint (Kaufmann *et al.*, 1998; Little *et al.*, 1999) may be explained by the variability of the parameters in the population that carry a direct biological interpretation (cell death rates, virus infectivity, virus production,...). Third, the effect of patient characteristics or interventions could be checked through their biological mechanism and by using a model able to fit complex trajectories. For instance, one could look at the effect of an antiretroviral regimen combining reverse transcriptase inhibitors (blocking the infection of new cells) and protease inhibitors (generating a higher proportion of non-infectious virion) (Wu and Ding, 1999) and to define individually the optimal time of treatment initiation. It would also be interesting to look at the effect of the transmission of mutated viruses (Fidler *et al.*, 2006) or host genetics characteristics (Fellay *et al.*, 2007) on the early dynamics. Furthermore, the probability distribution of the date of infection could be useful at the population level for defining the incidence of HIV infection or the rate of disease progression and to estimate the induction time (delay between the date of infection and AIDS stage).

The paper is organized as follows. In section 2, we describe a method based on likelihood maximization for the estimation of parameters of a general population ODE model with an unknown origin of time. In section 3, we describe two HIV dynamics models and we present a simulation study. In section 4, we show the results for a sample of 761 HIV-infected patients from the Concerted Action on Seroconversion to AIDS and Death in Europe

(CASCADE). A conclusion is given in section 5.

2. Statistical model and inference.

2.1. General statistical model for systems and observations.

First, we describe the model for a known date of infection. We consider an ODE model; for subject i with $i = 1, \dots, n$, we can write:

$$\begin{cases} \frac{dX^i(t)}{dt} = f(X^i(t), \xi^i) \\ X^i(0) = h(\xi^i) \end{cases}$$

where $X^i(t) = (X_1^i(t), \dots, X_K^i(t))$ is the vector of the K components and ξ^i is a vector of p individual natural (or biological) parameters which appear in the ODE system. We assume that f and h are twice differentiable with respect to ξ^i .

Second, a model is introduced for the ξ^i to allow inter-individual variability:

$$\begin{cases} \tilde{\xi}_l^i = \Psi_l(\xi_l^i) \\ \tilde{\xi}_l^i = \phi_l + z_l^i \beta_l + \omega_l^i u^i, \quad l \leq p \end{cases}$$

where ϕ_l is the intercept, z_l^i and ω_l^i are the vectors of explanatory variables associated respectively to the fixed and to the random effects of the l th biological parameter. The β_l 's are vectors of regression coefficients associated to the fixed effects. We assume $u^i \sim \mathcal{N}(0, \Sigma)$, where u^i is the individual vector of random effects of dimension q and Σ its variance matrix which we assume to be diagonal for sake of simplicity.

Third, we construct a model for the observations. Generally, we do not directly observe all the components of X^i , but rather $M \leq K$ functions of X^i ; we call these functions “observable components”. We observe Y_{ijm} , the j th measurement of the m th observable component for subject i at date d_{ijm} . If we know the date of infection τ_i , we can compute the time since infection $t_{ijm} = d_{ijm} - \tau_i$ and we assume that:

$$Y_{ijm} = g_m(X(t_{ijm}, \tilde{\xi}^i)) + \epsilon_{ijm} \quad j = 1, \dots, n_{im} \quad m = 1, \dots, M$$

where the ϵ_{ijm} are independent Gaussian measurement errors with zero mean and variances σ_m^2 . The $g_m(\cdot)$, $m = 1, \dots, M$, are twice differentiable functions of \mathbb{R}^K to \mathbb{R} .

2.2. The issue of unknown date of infection.

We assume that the date of infection τ_i has a known density f_{τ_i} with a support $[b_{inf}^i, b_{sup}^i]$. We denote by d_-^i (resp. d_+^i) the date of last seronegative test (resp. first seropositive test) for patient i . For very long intervals it would be relevant to take the density proportional to HIV incidence. Since we work with intervals of moderate length (3 years maximum), the incidence can be considered approximately constant. Therefore, we took uniform densities between d_-^i and $b_{sup}^i = \min(d_+^i, d_{i11})$, where d_{i11} is the date of first HIV RNA detectable measurement.

However, another problem is that patients may have been infected before the last seronegative date. Fiebig *et al.* (2003) determined that the HIV test becomes positive on average 89 days (95% Confidence Interval=47-130) after infection. The window of possible infection is thus $[b_{inf}^i, b_{sup}^i]$, where $b_{inf}^i = (d_-^i - 130)$. However, it is less likely that a patient has been infected 130 days before d_-^i than 5 days before d_-^i for instance. Therefore, we assumed linearly increasing densities between b_{inf}^i and d_-^i starting with $f_{\tau_i}(b_{inf}^i) = 0$. Thus, the density f_{τ_i} is defined by :

$$f_{\tau_i}(s) = \begin{cases} 0 & \text{if } s \leq b_{inf}^i \\ \frac{f_{\tau_i}(d_-^i)(s-b_{inf}^i)}{d_-^i - b_{inf}^i} & \text{if } b_{inf}^i \leq s \leq d_-^i \\ f_{\tau_i}(d_-^i) & \text{if } s \geq d_-^i \end{cases}$$

where $f_{\tau_i}(d_-^i) = \frac{2}{2b_{sup}^i - d_-^i - 3b_{inf}^i}$.

2.3. Log-likelihood.

The model is complicated by the detection limits ζ_{ij} of assays leading to left-censored observations of HIV RNA for the j th measurement for subject i . We define HIV RNA as the first observable compartment ($m = 1$) and we observe Y_{ij1} or $\{Y_{ij1} < \zeta_{ij}\}$. The left-censored observations are taken into account as in Thiébaut *et al.* (2006) and Guedj *et al.* (2007a). Noting $\delta_{ij} = \mathbf{1}_{\{Y_{ij1} > \zeta_{ij}\}}$, if we assume the date of infection τ_i known, the full likelihood given the random effects is given by:

$$\begin{aligned} \mathcal{L}_{\mathcal{F}_i|u^i, \tau_i} = & \prod_{j=1}^{n_{i1}} \left\{ \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{ij1} - g_1(X(t_{ij1}, \tilde{\xi}^i))}{\sigma_1} \right)^2 \right] \right\}^{\delta_{ij}} \left\{ \Phi \left(\frac{\zeta_{ij} - g_1(X(t_{ij1}, \tilde{\xi}^i))}{\sigma_1} \right) \right\}^{1-\delta_{ij}} \\ & \prod_{\substack{m=1, M \\ j=1, n_{im}}} \left\{ \frac{1}{\sigma_m \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{Y_{ijm} - g_m(X(t_{ijm}, \tilde{\xi}^i))}{\sigma_m} \right)^2 \right] \right\} \end{aligned}$$

where Φ is the cumulative distribution function of the standard univariate normal distribution. The individual likelihood given the date of infection is:

$$\mathcal{L}_{\mathcal{O}_i|\tau_i} = \int_{\mathbb{R}^q} \mathcal{L}_{\mathcal{F}_i|u^i, \tau_i}(u) \phi(u) du$$

where ϕ is the multivariate normal density of $\mathcal{N}(0, I_q)$. To determine the observed individual likelihood ($\mathcal{L}_{\mathcal{O}_i}$), we integrate $\mathcal{L}_{\mathcal{O}_i|\tau_i}$ on $[b_{inf}^i; b_{sup}^i]$:

$$\mathcal{L}_{\mathcal{O}_i} = \int_{b_{inf}^i}^{b_{sup}^i} \mathcal{L}_{\mathcal{O}_i|\tau_i}(s) f_{\tau_i}(s) ds = \int_{b_{inf}^i}^{b_{sup}^i} \int_{\mathbb{R}^q} \mathcal{L}_{\mathcal{F}_i|u^i, \tau_i}(u, s) \phi(u) f_{\tau_i}(s) du ds$$

We note $L_{\mathcal{F}_i|u^i, \tau_i} = \log \mathcal{L}_{\mathcal{F}_i|u^i, \tau_i}$ and $L_{\mathcal{O}_i} = \log \mathcal{L}_{\mathcal{O}_i|\tau_i}$. The global observed log-likelihood is $L_{\mathcal{O}} = \sum_{i \leq n} L_{\mathcal{O}_i}$. The integral on time is calculated by recursive adaptive Gauss-Legendre quadrature.

2.4. Likelihood maximization.

A Newton-like method was used for maximizing the likelihood. This method uses the first derivatives of the log-likelihood (the scores) (Commenges *et al.*, 2006). The observed scores can be obtained by applying twice Louis' formula (Louis, 1982):

$$U_{\mathcal{O}_i} = \frac{\partial L_{\mathcal{O}_i}}{\partial \theta} = (\mathcal{L}_{\mathcal{O}_i})^{-1} \int_{b_{inf}^i}^{b_{sup}^i} \mathcal{L}_{\mathcal{O}_i|\tau_i}(s) U_{\mathcal{O}_i|\tau_i}(s) f_{\tau_i}(s) ds$$

where:

$$U_{\mathcal{O}_i|\tau_i}(\cdot) = (\mathcal{L}_{\mathcal{O}_i|\tau_i}(\cdot))^{-1} \int_{\mathbb{R}^q} \mathcal{L}_{\mathcal{F}_i|u^i, \tau_i}(u, \cdot) U_{\mathcal{F}_i|u^i, \tau_i}(u) \phi(u) du$$

The scores $U_{\mathcal{F}_i|u^i, \tau_i}$ are determined by differentiating $L_{\mathcal{F}_i|u^i, \tau_i}$ by θ , where $\theta = ((\phi_l)_{l=1,p}, (\beta_l)_{l=1,p}, (u_l)_{l=1,q}, \sigma = (\sigma_l)_{l \leq M})$.

2.5. A posteriori estimations and distribution of the date of infection.

Individual parameters $\tilde{\xi}^i$ were estimated by the empirical Bayes estimators $\hat{\tilde{\xi}}^i$, where $\hat{\tilde{\xi}}_l^i = \hat{\phi}_l + z_l^i \hat{\beta}_l + w_l^i \hat{u}^i$ and \hat{u}^i is the posterior mode of the density of u^i given the data. Then, individual trajectories were predicted by computing $\hat{X}^i = X(t, \hat{\tilde{\xi}}^i)$.

An estimator of the *a posteriori* distribution of the date of infection, $f_{\tau_i|Y_i}$, can then be computed at the estimated individual parameters $\hat{\tilde{\xi}}^i$. We denote by Y_i the vector of measurements for subject i and by $f_{Y_i|\tau_i}$, the probability

density function of Y_i given the date of infection. Applying Bayes' formula, we have:

$$f_{\tau_i|Y_i}(\cdot|Y_i) = \frac{f_{Y_i|\tau_i}(Y_i|\cdot)f_{\tau_i}(\cdot)}{\int_{b_{inf}^i}^{b_{sup}^i} f_{Y_i|\tau_i}(Y_i|s)f_{\tau_i}(s)ds}$$

3. Models for primary HIV infection.

3.1. Models for the biological system.

We considered two ODE models for HIV infection: the “basic model” and the “productive cells model”. The “basic model” has three compartments: T (non-infected CD4), T^* (productively infected CD4) and V (free virion)(Nowak *et al.*, 1997b; Stafford *et al.*, 2000; Perelson, 2002). This model can be written as:

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dT^*}{dt} = \gamma VT - \mu_{T^*} T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{cases}$$

The uninfected CD4 cells enter the blood circulation at rate λ , and die at the rate μ_T . They can be infected by the virus at the rate γ . The infected CD4 die at the rate μ_{T^*} and produce virions at the rate π . Virions die at the rate μ_V and can also infect other CD4 cells (Figure 1 (a)). Model parameters are summarized in Table 1.

TABLE 1
Parameters of the dynamic models.

Parameter	Meaning (per day)
λ	Rate of T cell production per μL
μ_{T^*}	Death rate of T^* cells
π	Number of virions per T^* cells
μ_T	Death rate of T cells
γ	Infection rate of T cells per virion per μL
μ_V	Clearance of free virions
α	Rate to become productive cells

We assume that the model is at equilibrium before infection, so at $t = 0_-$, we have:

$$\begin{cases} T_{0_-} = \frac{\lambda}{\mu_T} \\ T_{0_-}^* = 0 \\ V_{0_-} = 0 \end{cases}$$

The initial inoculum of virions is fixed at 10^{-6} virions/ mm^3 (Ciupe *et al.*, 2006; Stafford *et al.*, 2000) which is equivalent to 5 virions for 5 liters of

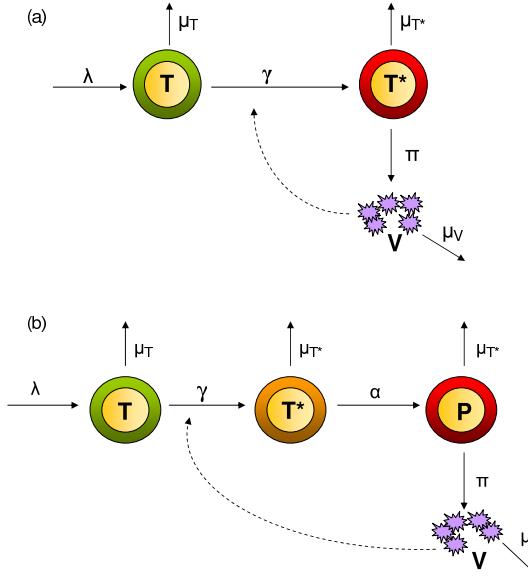


FIG 1. Graphical representation of “the basic model” (a) and “the productive cells model” (b).

blood. The introduction of virions in the system at $t = 0$, disrupts the initial stability and the system stabilizes to a new equilibrium if the reproductive number R_0 ($R_0 = \frac{\lambda\gamma\pi}{\mu_T\mu_{T^*}\mu_V}$) is higher than 1:

$$\begin{cases} \bar{T} = \frac{\mu_{T^*}\mu_V}{\gamma\pi} \\ \bar{T}^* = \frac{\lambda\gamma\pi - \mu_T\mu_{T^*}\mu_V}{\gamma\pi\mu_{T^*}} \\ \bar{V} = \frac{\lambda\gamma\pi - \mu_T\mu_{T^*}\mu_V}{\gamma\mu_{T^*}\mu_V} \end{cases}$$

The second model (the “productive cells model”) differentiates non-productive infected cells (T^*) and productive infected cells (P) (Figure 1 (b)) as previously suggested in Nowak *et al.* (1997a). This second model includes one more parameter α (the rate to become productive infected cells). The model can be written as:

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dT^*}{dt} = \gamma VT - \mu_{T^*} T^* - \alpha T^* \\ \frac{dP}{dt} = \alpha T^* - \mu_{T^*} P \\ \frac{dV}{dt} = \pi P - \mu_V V \end{cases}$$

The equilibrium before infection is:

$$\begin{cases} T_{0-} = \frac{\lambda}{\mu_T} \\ T_{0-}^* = 0 \\ P_{0-} = 0 \\ V_{0-} = 0 \end{cases}$$

After the introduction of virions, the system stabilizes to a new equilibrium if R_0 ($R_0 = \frac{\lambda\gamma\pi\alpha}{\mu_T\mu_{T^*}\mu_V(\mu_{T^*}+\alpha)}$) is higher than 1:

$$\begin{cases} \bar{T} = \frac{\mu_{T^*}\mu_V(\mu_{T^*}+\alpha)}{\gamma\pi\alpha} \\ \bar{T}^* = \frac{\lambda}{\mu_{T^*}+\alpha} - \frac{\mu_T\mu_V\mu_{T^*}}{\gamma\pi\alpha} \\ \bar{P} = \frac{\alpha}{\mu_{T^*}} \left(\frac{\lambda}{\mu_{T^*}+\alpha} - \frac{\mu_T\mu_V\mu_{T^*}}{\gamma\pi\alpha} \right) \\ \bar{V} = \frac{\pi\alpha}{\mu_V\mu_{T^*}} \left(\frac{\lambda}{\mu_{T^*}+\alpha} - \frac{\mu_T\mu_V\mu_{T^*}}{\gamma\pi\alpha} \right) \end{cases}$$

3.2. Statistical models.

We can construct statistical models as in section 2.1. We took $\Psi_l(.) = \log(.)$ for all l to ensure positivity of parameters. The individual parameters are $\tilde{\xi}^i = (\tilde{\lambda}^i, \tilde{\mu}_{T^*}^i, \tilde{\pi}^i, \tilde{\mu}_T^i)$ with $\tilde{\lambda}^i = \log \lambda^i$, $\tilde{\mu}_{T^*}^i = \log \mu_{T^*}^i$, $\tilde{\pi}^i = \log \pi^i$ and $\tilde{\mu}_T^i = \log \mu_T^i$. Because of identifiability issues, the values of the other parameters were taken according to the literature: $\mu_V = 20.0 \text{ day}^{-1}$ (Ramratnam *et al.*, 1999) and $\gamma = 0.00027 \text{ virions}^{-1} \text{ day}^{-1} \mu L^{-1}$ (Wilson *et al.*, 2007). The parameter α is fixed at 1 day^{-1} because the time lag between virus entry and virus production is about 1 day (Kiernan *et al.*, 1990; Barbosa *et al.*, 1994).

The observed components were the base-10 logarithm of HIV RNA load (number of virions per μL) and the fourth root of total CD4 count (number of cells per μL). For the “basic model”, $g_1(X) = \log_{10}(V)$ and $g_2(X) = (T+T^*)^{0.25}$ where $X = (T, T^*, V)$. For the “productive cells model”, $g_1(X) = \log_{10}(V)$ and $g_2(X) = (T+T^*+P)^{0.25}$ where $X = (T, T^*, P, V)$. These transformations are commonly used for achieving normality and homoscedasticity of measurement error distributions (Thiébaut *et al.*, 2003). We have:

$$\begin{aligned} Y_{ij1} &= g_1(X(t_{ij1}, \tilde{\xi}^i)) + \epsilon_{ij1} \quad j = 1, \dots, n_{i1} \\ Y_{ij2} &= g_2(X(t_{ij2}, \tilde{\xi}^i)) + \epsilon_{ij2} \quad j = 1, \dots, n_{i2} \end{aligned}$$

where ϵ_{ij1} and ϵ_{ij2} are independent Gaussian with zero mean and variances σ_{VL}^2 and σ_{CD4}^2 respectively.

3.3. Simulation study.

We simulated a sample of 100 subjects during primary HIV infection using the “basic model” for simplicity. The parameter values were defined according to published studies or close to the estimates of our application (see next section). Dates for the HIV serology test (negative and positive, respectively) were simulated according to the prior distribution defined in section 2.1 and with a window of 50 days between last negative and first positive date.

The times of measurements t_{ijm} after infection were days 1, 8, 15, 113 and 197 for each subject i and observed component m . We included independent random effects for the first two parameters (λ and μ_{T^*}). Therefore, the vector of parameters to estimate was $(\xi^i, \sigma_{\tilde{\lambda}}, \sigma_{\tilde{\mu}_{T^*}}, \sigma_{VL}, \sigma_{CD4})$. Results are presented in Table 2. The 95% confidence intervals included the true values of parameters.

TABLE 2
Values of parameters and 95% confidence intervals of estimated parameters in logarithmic scale for 100 simulated patients.

Parameter	True value	Estimated value	95% Confidence Interval
λ	3.524	3.672	[3.004; 4.340]
$\tilde{\mu}_{T^*}$	-0.979	-1.006	[-1.136; -0.876]
$\tilde{\pi}$	5.558	5.544	[5.516; 5.572]
$\tilde{\mu}_T$	-3.824	-3.698	[-4.331; -3.065]
$\sigma_{\tilde{\lambda}}$	0.365	0.382	[0.362; 0.402]
$\sigma_{\tilde{\mu}_{T^*}}$	0.392	0.302	[0.129; 0.475]
σ_{VL}	1.018	1.028	[0.945; 1.111]
σ_{CD4}	0.247	0.234	[0.227; 0.267]

$\sigma_{\tilde{\lambda}}$ and $\sigma_{\tilde{\mu}_{T^*}}$ are the standard deviations of random effect for $\tilde{\lambda}$ and $\tilde{\mu}_{T^*}$
 σ_{VL} and σ_{CD4} are the standard deviations of the observation error of $\log_{10}(V)$ and $(T + T^*)^{0.25}$.

We tried different starting values by increasing values from 10% to 50% and we obtained the same convergence point up to small differences on the third significant digit. The convergence criterion was achieved after less than 50 iterations. We also tried varying the width of the window between the last negative and first positive HIV serology test (100 and 300 days) and succeeded in estimating parameters but with more iterations (sometimes after up to 80 iterations).

4. Application to the CASCADE dataset.

4.1. The study sample.

The study sample came from the Concerted Action on Seroconversion to AIDS and Death in Europe (CASCADE) study. CASCADE is an international collaboration pooling data from seroconverters in Europe, Canada and Australia. The CASCADE study has been described in detail elsewhere ([CASCADE, 2003](#)). Data were pooled in the 2006 update from 20 participating cohorts. We selected HIV seroconverters if their HIV interval test (delay between the date of last seronegative test and the date of first seropositive test) was less than 3 years, if they did not receive any antiretroviral treatment during the first year following the date of seropositive test and if they had more than 3 measurements of CD4 or viral load during this first year of follow-up with the first detectable viral load measurement during the first three months after the date of seropositive test. A total of 761 patients met these criteria.

4.2. Models used.

We used the two models defined in section 3.1. The vector of natural parameters to be estimated for subject i was: $\xi^i = (\lambda^i, \mu_{T^*}^i, \pi^i, \mu_T^i)$ for the two models. The other parameters were assumed to be known as described in section 3.2. The windows for possible dates of infection were fixed as defined in section 2.2.

Random effects were selected according to a forward selection procedure. Starting with a model without random effect, we introduced random effects successively on each parameter and selected the one leading to the best likelihood improvement. Then, we added a new random effect and continued until the new model was not rejected by a likelihood test ratio.

4.3. Results.

Selected patients had a median of four measurements for CD4 and for HIV RNA (InterQuartile Range [IQR]=[3; 5]). Most of the patients were infected by sex between men (71%). Follow-up was censored after 1 year beyond seropositive HIV test, resulting in a median follow-up after the first seropositive test of 195 days (IQR=[119; 260]). The median of the delay between the dates of last seronegative and first seropositive test was 170 days (IQR=[91; 273]).

The fit of the “productive cells model” was better than the “basic model” (AIC= 14300 vs. 15010). The improvement brought by this model can be considered as large ([Commenges et al., 2008](#)). Estimates of the parameters of the two models are presented in Table 3.

For the “productive cells model”, on the natural scale, the mean half-life ($\log(2)/\text{Death rate}$) of infected and uninfected cells was 0.40 and 21 days,

TABLE 3
Estimated parameters and standard-errors (SE) on logarithmic scale for the “basic” and the “productive cells” models; for meaning of parameters see Table 1.

Parameter	Basic model		Productive cells model	
	Estimate	SE	Estimate	SE
λ	3.49	0.032	3.33	0.034
$\mu_{\tilde{T}^*}$	0.62	0.046	0.54	0.045
$\tilde{\pi}$	5.27	0.041	6.06	0.036
$\mu_{\tilde{T}}$	-3.26	0.019	-3.40	0.034
$\sigma_{\tilde{\lambda}}$	0.16	0.008	0.41	0.057
$\sigma_{\tilde{\mu}_{T^*}}$	0.42	0.043	0.37	0.052
σ_{VL}	0.82	0.017	0.75	0.017
σ_{CD4}	0.29	0.012	0.27	0.012
AIC	15010		14300	

$\sigma_{\tilde{\lambda}}$ and $\sigma_{\tilde{\mu}_{T^*}}$ are the standard deviations of random effect for $\tilde{\lambda}$ and $\tilde{\mu}_{T^*}$.
 σ_{VL} and σ_{CD4} are the standard deviations of the observation error of $\log_{10}(V)$ and $(T + T^*)^{0.25}$ ($(T + T^* + P)$ for the “productive cells model”, respectively).

respectively. At the time of the viral peak (median viral load of 5.68 \log_{10} copies/mL, IQR= [5.44; 5.96]), the estimated median number of infected cells was 64 cells/ μ L (IQR= [40; 113]). At the same time, the median number of productive cells/ μ L was 24 (IQR= [14; 46]) among 557 CD4 cells/ μ L (IQR= [460; 658]). The median of the estimated CD4 setpoints was 500 cells/ μ L (IQR= [396; 627]) which is close to the median of observed setpoints (based on at least one measurement after 365 days available in 430 patients): 470 cells/ μ L (IQR= [371; 615]). At the same time, the median of the estimated number of infected cells was 8 (IQR= [6; 12]) including a median of 3 productive cells/ μ L (IQR= [2; 5]). The median of viral setpoints was also close to that observed in 403 patients: 4.77 \log_{10} copies/mL (IQR= [4.59; 4.98]) vs. 4.59 (IQR= [4.04; 5.00]).

We found very different *a posteriori* individual distributions of the date of infection depending on the width of the window and the quantity of available information for each patient. Figure 2 shows the *a posteriori* distributions displayed on the window of infection (see section 2.2), where 0 represents the last seronegative date for six patients chosen for exhibiting different shapes of the *a posteriori* distribution. For patients with data clearly before the setpoint, the model was able to restrict the possible dates of infection. For instance, patient 132 had an interval of possible dates of infection of 273 days and our estimation restricted this interval to 90 days. However, for other patients and especially for some patients with a wide interval between last negative and first positive HIV serology test, the *a posteriori* distribution had local maxima. Analyzes of marker dynamics according to the local maxima

suggested quite different possibilities according to the available information (Figure 3).

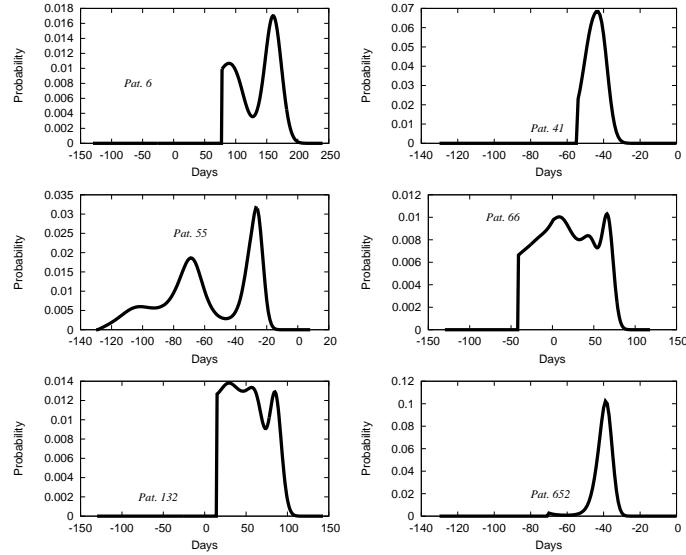


FIG 2. *A posteriori individual distributions of the date of infection for patient 6, 41, 55, 66, 132 and 652 from the “productive cells model”, where day 0 is the date of last negative HIV test.*

For each patient, we took the date with the highest probability and plotted predicted individual trajectories for each marker from this date. The fits of the model were satisfactory. The estimated trajectories are in agreement with those reported in the literature (Little *et al.*, 1999; Lindback *et al.*, 2000; Stafford *et al.*, 2000). Individual predicted fits and observed values are shown in Figure 4 for six patients.

5. Discussion.

In this paper, we present a method for estimating parameters of random effects models based on differential equations when the origin of time is unknown, taking also into account unbalanced data and left-censored measures. This method was applied to a cohort of HIV patients during primary infection using repeated measurements of two markers: plasma HIV RNA and CD4+ T cell counts. The model gave reasonable fits although the kinetics of the markers was very complex due to non-linear interaction between the virus and the target cells. Thanks to the population approach, we were able

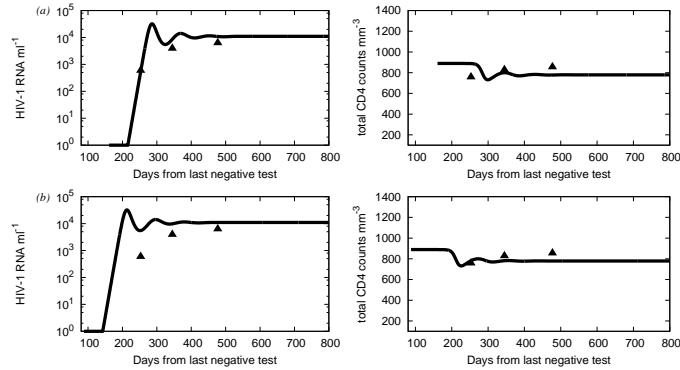


FIG 3. Predicted fits from the two highest probable dates of infection ((a) the global maximum and (b) the local maximum) and observed values of HIV RNA level and total CD4 count for patient 6 from the “productive cells model”.

to describe the dynamics of markers during primary infection using data from several hundreds of patients.

Predictions were obviously driven by the structure of the ODE system that is based on biological knowledge. Compared to a descriptive model, this mechanistic approach brings additional information. Typically, solutions of ODE systems have oscillations: the oscillations are damped as time passes so that the trajectory gets closer to the steady state value. Observations are partly consistent with the oscillating behavior. The peak of viral load, the reality of which is generally admitted, is nothing but the first oscillation. It can be noted that the oscillations are generally more damped for more complex systems (Burg *et al.*, 2009). In conclusion, the oscillatory trajectories produced by our model are for a part the expression of a real phenomenon.

Predicted viral and CD4 setpoints (defined as markers values at equilibrium) were quite similar to those that could be observed. Interestingly, the estimated value of infected cells (that are not observed) was in agreement with previous studies reporting a low concentration of productive infected cells (Chun *et al.*, 1997).

Because the approach was based on a mechanistic model, the parameters have some biological meanings. For instance, the mean half-life of infected cells was estimated at 0.40 days. Previous estimations of this parameter were mainly performed in patients treated by antiretroviral therapy using analytical solution of the same dynamical model, but assuming perfect treatment efficacy, that is no new cell infection after treatment onset. Estimation of infected cells half-life has been refined when more potent an-

tiretoviral therapy became available. The estimate was revised from 1.1 (Perelson *et al.*, 1996) to 0.7 day (Faulkner *et al.*, 2003), that is closer to ours. Similarly, it is interesting to note that our estimate of virus production (428 virions/cells/day) is of the same order as what have been reported (Dimitrov *et al.*, 1993; Levy *et al.*, 1996; Haase, 1999). However, the biological relevance of the present model could still be questioned with regards to other processes that have been ignored. For instance, the cytotoxic T-cell response plays a role in the control of the infection and the decrease of the viral load by killing the infected cells (Stafford *et al.*, 2000), although the efficiency of this immunological response is debated (Asquith and McLean, 2007). Moreover, our estimate of target CD4 cells half-life is a mixture of the half-life of quiescent cells and activated cells that have very different half-lives (McLean and Michie, 1995; Vrisekoop *et al.*, 2008).

The uncertainty on the date of infection was taken into account to improve the accuracy of the estimates as compared to performing simple imputations. In addition, we were able to derive the individual posterior distribution of the date of infection according to the model. When enough information is available the date of infection can be estimated with good accuracy (Figure 2). The sensitivity of the estimates to the assumptions about the priors (section 2.2) was roughly proportional to the amount of data available as illustrated by the posterior infection time distribution of patient 41 compared to patient 66 (Figure 2). We also performed a sensitivity analysis which demonstrated a certain robustness of the method. The lower bound of the infection time distribution could be better defined by using additional information such as antibody dynamics. Unfortunately, this information was not available in the present dataset.

The main limitation for the use of dynamical models is the parameter identifiability that led us to fix the value of γ (infectivity) and μ_V (viral clearance). The clearance was fixed to 20 day^{-1} according to recent studies with highly repeated measures of HIV RNA after initiation of antiretroviral treatment (Perelson *et al.*, 1997; Ramratnam *et al.*, 1999). The chosen value of μ_V may influence the estimation of other parameters such as π , as the viral setpoint is essentially determined by the ratio $\frac{\pi}{\mu_V}$ (Nowak *et al.*, 1997a). Practical identifiability may be improved by measuring more compartments such as infected cells or by increasing the number of repeated measurements (Guedj *et al.*, 2007b). This is an important point to consider in future studies as the issue of identifiability precludes the comparison with more complex models.

Finally, this method can be applied in other areas where either the model is simpler or the amount of measured information greater, so that identifiability

is less an issue.

Acknowledgements.

The authors would like to thank S. Walker for suggestions and criticisms concerning the manuscript.

References.

- Asquith, B. and McLean, A. (2007). In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(15), 6365–6370.
- Barbosa, P., Charneau, P., Dumey, N., and Clavel, F. (1994). Kinetic analysis of HIV-1 early replicative steps in a coculture system. *AIDS Research and Human Retroviruses*, **10**(1), 53–59.
- Berman, S. (1990). A stochastic model for the distribution of HIV latency time based on T4 counts. *Biometrika*, **77**(4), 733–741.
- Burg, D., Rong, L., Neumann, A., and Dahari, H. (2009). Mathematical modeling of viral kinetics under immune control during primary hiv-1 infection. *Journal of Theoretical Biology*. doi:10.1016/j.jtbi.2009.04.010.
- Cao, J., Fussmann, G., and Ramsay, J. (2008). Estimating a Predator-Prey Dynamical Model with the Parameter Cascades Method. *Biometrics*, **64**(3), 959–967.
- CASCADE, C. (2003). Determinants of survival following HIV-1 seroconversion after the introduction of HAART: results from CASCADE. *The Lancet*, **362**(9392), 1267–1274.
- Chun, T., Carruth, L., Shen, X., Taylor, H., et al. (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, **387**(6629), 183–188.
- Ciupe, M., Bivort, B., Bortz, D., and Nelson, P. (2006). Estimating kinetic parameters from HIV primary infection data through the eyes of three different mathematical models. *Mathematical Biosciences*, **200**(1), 1–27.
- Commenges, D., Jacqmin-Gadda, H., Proust, C., and Guedj, J. (2006). A Newton-like algorithm for likelihood maximization: the robust variance scoring algorithm. *arXiv:math.ST/0610402*.
- Commenges, D., Sayyareh, A., Letenneur, L., Guedj, J., and Bar-Hen, A. (2008). Estimating a difference of Kullback–Leibler risks using a normalized difference of AIC. *Annals of Applied Statistics*, **2**(3), 1123–1142.
- De Boer, R. and Perelson, A. (1998). Target cell limited and immune control models of HIV infection: a comparison. *Journal of Theoretical Biology*, **190**(3), 201–214.
- DeGruttola, V., Lange, N., and Dafni, U. (1991). Modeling the Progression of HIV Infection. *Journal of the American Statistical Association*, **86**(415), 569–577.
- Desquillet, L., Goujard, C., Rouzioux, C., Sinet, M., Chaix, M., Sereni, D., Boufassa, F., Delfraissy, J., Meyer, L., PRIMO, and Groups, S. S. (2004). Does Transient HAART During Primary HIV-1 Infection Lower the Virological Set-Point? *AIDS*, **18**(18), 2361–2369.
- Dimitrov, D., Willey, R., Sato, H., Chang, L., Blumenthal, R., and Martin, M. (1993). Quantitation of human immunodeficiency virus type 1 infection kinetics. *Journal of Virology*, **67**(4), 2182–2190.
- Dubin, N., Berman, S., Marmor, M., Tindall, B., Des Jarlais, D., and Kim, M. (1994). Estimation of time since infection using longitudinal disease-marker data. *Statistics in Medicine*, **13**(3), 231–244.

- Faulkner, N., Lane, B., Bock, P., and Markowitz, D. (2003). Protein phosphatase 2A enhances activation of human immunodeficiency virus type 1 by phorbol myristate acetate. *Journal of Virology*, **77**(3), 2276–2281.
- Fellay, J., Shianna, K., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, **317**(5840), 944–947.
- Fidler, S., Fraser, C., Fox, J., Tamm, N., Griffin, J., and Weber, J. (2006). Comparative potency of three antiretroviral therapy regimes in primary HIV infection. *AIDS*, **20**(2), 247–252.
- Fidler, S., Fox, J., Touloumi, G., Pantazis, N., Porter, K., Babiker, A., Weber, J., et al. (2007). Slower CD4 Cell Decline Following Cessation of a 3 Month Course of HAART in Primary HIV Infection: Findings From an Observational Cohort. *AIDS*, **21**(10), 1283–1291.
- Fiebig, E., Wright, D., Rawal, B., Garrett, P., Schumacher, R., Peddada, L., Heldebrant, C., Smith, R., Conrad, A., Kleinman, S., et al. (2003). Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS*, **17**(13), 1871–1879.
- Filter, R., Xia, X., and Gray, C. (2005). Dynamic HIV/AIDS parameter estimation with application to a vaccine readiness study in Southern Africa. *IEEE Transactions on Biomedical Engineering*, **52**(5), 784–791.
- Geskus, R. (2000). On the inclusion of prevalent cases in HIV/AIDS natural history studies through a marker-based estimate of time since seroconversion. *Statistics in Medicine*, **19**(13), 1753–1769.
- Geskus, R., Prins, M., Hubert, J., Miedema, F., Berkhout, B., Rouzioux, C., Delfraissy, J., and Meyer, L. (2007). The HIV RNA setpoint theory revisited. *Retrovirology*, **4**(1), 65–74.
- Guedj, J., Thiébaut, R., and Commenges, D. (2007a). Maximum Likelihood Estimation in Dynamical Models of HIV. *Biometrics*, **63**, 1198–1206.
- Guedj, J., Thiébaut, R., and Commenges, D. (2007b). Practical Identifiability of HIV Dynamics Models. *Bulletin of Mathematical Biology*, **69**(8), 2493–2513.
- Haase, A. (1999). Population biology of HIV-1 infection: Viral and CD4+ T Cell Demographics and Dynamics in Lymphatic Tissues. *Annual Reviews in Immunology*, **17**(1), 625–656.
- Hecht, F., Wang, L., Collier, A., Little, S., Markowitz, M., Margolick, J., Kilby, J., Daar, E., Conway, B., Holte, S., et al. (2006). A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute HIV infection. *Journal of Infectious Diseases*, **194**(6), 725–733.
- Huang, Y., Liu, D., and Wu, H. (2006). Hierarchical Bayesian Methods for Estimation of Parameters in a Longitudinal HIV Dynamic System. *Biometrics*, **62**(2), 413–423.
- Kaufmann, G., Cunningham, P., Kelleher, A., Zaunders, J., Carr, A., Vizzard, J., Law, M., Cooper, D., et al. (1998). Patterns of viral dynamics during primary human immunodeficiency virus type 1 infection. *Journal of Infectious Diseases*, **178**, 1812–1815.
- Kiernan, R., Marshall, J., Bowers, R., Doherty, R., and McPhee, D. (1990). Kinetics of HIV-1 replication and intracellular accumulation of particles in HTLV-I-transformed cells. *AIDS research and human retroviruses*, **6**(6), 743–752.
- Levy, J., Ramachandran, B., Barker, E., Guthrie, J., Elbeik, T., and Coffin, J. (1996). Plasma Viral Load, CD4+ Cell Counts, and HIV-1 Production by Cells. *Science*, **271**(5249), 670–671.
- Lindblad, S., Karlsson, A., Mittler, J., Blaxhult, A., Carlsson, M., Briheim, G., Sonnerborg, A., and Gaines, H. (2000). Viral dynamics in primary HIV-1 infection. Karolinska

- Institutet Primary HIV Infection Study Group. *AIDS*, **14**(15), 2283–2291.
- Little, S., McLean, A., Spina, C., Richman, D., and Havlir, D. (1999). Viral Dynamics of Acute HIV-1 Infection. *Journal of Experimental Medicine*, **190**(6), 841–850.
- Louis, T. (1982). Finding the observed Information matrix when using the *EM* algorithm. *Journal of the Royal Statistical Society. Series B*, **44**, 226–233.
- McLean, A. and Michie, C. (1995). *in vivo* estimates of division and death rates of human t lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 3707–3711.
- Mellors, J., Kingsley, L., Rinaldo, C., Todd, J., Hoo, B., Kokka, R., and Gupta, P. (1995). Quantitation of HIV-1 RNA in Plasma Predicts Outcome after Seroconversion. *Annals of Internal Medicine*, **122**(8), 573–579.
- Munoz, A., Carey, V., Taylor, J., Chmiel, J., Kingsley, L., Van Raden, M., and Hoover, D. (1992). Estimation of time since exposure for a prevalent cohort. *Statistics in Medicine*, **11**(7), 939–952.
- Murray, J., Kaufmann, G., Kelleher, A., and Cooper, D. (1998). A model of primary HIV-1 infection. *Mathematical Biosciences*, **154**(2), 57–85.
- Nowak, M., Bonhoeffer, S., Shaw, G., and May, R. (1997a). Anti-viral Drug Treatment: Dynamics of Resistance in Free Virus and Infected Cell Populations. *Journal of Theoretical Biology*, **184**(2), 203–217.
- Nowak, M., Lloyd, A., Vasquez, G., Wiltrot, T., Wahl, L., Bischofberger, N., Williams, J., Kinter, A., Fauci, A., Hirsch, V., et al. (1997b). Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection. *Journal of Virology*, **71**(10), 7518–7525.
- Pantazis, N., Touloumi, G., Walker, A., and Babiker, A. (2005). Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Journal of the Royal Statistical Society Series C(Applied Statistics)*, **54**(2), 405–423.
- Perelson, A. (2002). Modelling viral and immune system dynamics. *Nature Reviews Immunology*, **2**(1), 28–36.
- Perelson, A., Neumann, A., Markowitz, M., Leonard, J., and Ho, D. (1996). Viral dynamics in human immunodeficiency virus type 1 infection. *Science*, **271**, 1582–1586.
- Perelson, A., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M., and Ho, D. (1997). Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature*, **387**, 188–191.
- Phillips, A. (1996). Reduction of HIV concentration during acute infection: independence from a specific immune response. *Science*, **271**(5248), 497–499.
- Putter, H., Heisterkamp, S., Lange, J., and de Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in Medicine*, **21**(15), 2199–2214.
- Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J., Markowitz, M., Moore, J., Perelson, A., and Ho, D. (1999). Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *The Lancet*, **354**, 1782–1786.
- Samson, A., Lavielle, M., and Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Computational Statistics and Data Analysis*, **51**(3), 1562–1574.
- Stafford, M., Cao, Y., Ho, D., Corey, L., and Perelson, A. (2000). Modeling Plasma Virus Concentration and CD4+ T Cell Kinetics during Primary HIV Infection. *Journal of Theoretical Biology*, **203**, 285–301.
- Thiébaut, R., Jacqmin-Gadda, H., Leport, C., Katlama, C., D., C., Le Moing, V., Morlat,

- P., Chene, G., and the APROCO Study Group (2003). 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. *Journal of Biopharmaceutical Statistics*, **13**, 271–282.
- Thiébaut, R., Guedj, J., Jacqmin-Gadda, H., Chêne, G., Trimoulet, P., Neau, D., and Commenges, D. (2006). Estimation of dynamical model parameters taking into account undetectable marker values. *BMC Medical Research Methodology*, **6**(1), 1–10.
- Vrisekoop, N., den Braber, I., de Boer, A., Ruiter, A., Ackermans, M., van der Crabben, S., Schrijver, E., Spierenburg, G., Sauerwein, H., Hazenberg, M., et al. (2008). Sparse production but preferential incorporation of recently produced naïve T cells in the human peripheral pool. *Proceedings of the National Academy of Sciences*, **105**(16), 6115–6120.
- Wick, D. (1999). The Disappearing CD4+ T Cells in HIV Infection: a Case of Over-stimulation? *Journal of Theoretical Biology*, **197**(4), 507–516.
- Wilson, D., Mattapallil, J., Lay, M., Zhang, L., Roederer, M., and Davenport, M. (2007). Estimating the Infectivity of CCR5-Tropic Simian Immunodeficiency Virus SIVmac251 in the Gut? *Journal of Virology*, **81**(15), 8025–8029.
- Wu, H. and Ding, A. (1999). Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics*, **55**, 410–418.

Appendix: CASCADE Collaboration.

Steering Committee: Julia Del Amo (Chair), Laurence Meyer (Vice Chair), Heiner Bucher, Geneviève Chêne, Deenan Pillay, Maria Prins, Magda Rosinska, Caroline Sabin, Giota Touloumi
Co-ordinating Centre: Kholoud Porter (Project Leader), Sara Lodi, Sarah Walker, Abdel Babiker, Janet Darbyshire
Clinical Advisory Board: Heiner Bucher, Andrea de Luca, Martin Fisher, Roberto Muga
Collaborators: Australia Sydney AIDS Prospective Study and Sydney Primary HIV Infection cohort (John Kaldor, Tony Kelleher, Tim Ramacciotti, Linda Gelgor, David Cooper, Don Smith); Canada South Alberta clinic (John Gill); Denmark Copenhagen HIV Seroreverter Cohort (Louise Bruun Jorgensen, Claus Nielsen, Court Pedersen); Estonia Tartu Ulikool (Irja Lutsar); France Aquitaine cohort (Geneviève Chêne, Francois Dabis, Rodolphe Thiébaut, Bernard Masquelier), French Hospital Database (Dominique Costagliola, Marguerite Guiguet), Lyon Primary Infection cohort (Philippe Vanhems), SEROCO cohort (Laurence Meyer, Faroudy Boufassa); Germany German cohort (Osamah Hamouda, Claudia Kucherer); Greece Greek Haemophilia cohort (Giota Touloumi, Nikos Pantazis, Angelos Hatzakis, Dimitrios Paraskevis, Anastasia Karafoulidou); Italy Italian Seroconversion Study (Giovanni Rezza, Maria Dorrucci, Benedetta Longo, Claudia Balotta); Netherlands Amsterdam Cohort Studies among homosexual men and drug users (Maria Prins, Liselotte van Asten, Akke van der Bij, Ronald Geskus, Roel Coutinho); Norway Oslo and Ullevål Hospital cohorts (Mette Sannes, Oddbjorn Brubakk, Anne Eskild, Johan N Bruun); Poland National Institute of Hygiene (Magdalena Rosinska); Portugal Universidade Nova de Lisboa (Ricardo Camacho); Russia Pasteur Institute (Tatyana Smolskaya); Spain Badalona IDU hospital cohort (Roberto Muga), Barcelona IDU Cohort (Patricia Garcia de Olalla), Madrid cohort (Julia Del Amo, Jorge del Romero), Valencia IDU cohort (Santiago Pérez-Hoyos, Ildefonso Hernandez Aguado); Switzerland Swiss HIV cohort (Heiner Bucher, Martin Rickenbach, Patrick Franciolli); Ukraine Perinatal Prevention of AIDS Initiative (Ruslan Malyuta); United Kingdom Edinburgh Hospital cohort (Ray Brettle), Health Protection Agency (Valerie Delpech, Sam Lattimore, Gary Murphy, John Parry, Noel Gill), Royal Free haemophilia cohort (Caroline Sabin, Christine Lee), UK Register of HIV Seroconverters (Kholoud Porter, Anne Johnson, Andrew Phillips, Abdel Babiker, Janet Darbyshire, Valerie Delpech), University College London (Deenan Pillay), University of Oxford (Harold Jaffe).

J. DRYLEWICZ (E-MAIL: julia.drylewicz@isped.u-bordeaux2.fr)
D. COMMENGES (E-MAIL: daniel.commenges@isped.u-bordeaux2.fr)
R. THIÉBAUT (E-MAIL: rodolphe.thiebaut@isped.u-bordeaux2.fr)
INSERM, U897 EPIDEMIOLOGY AND BIOSTATISTICS
BORDEAUX 2 VICTOR SEGALEN UNIVERSITY
146, RUE LEO SAIGNAT, F-33076 BORDEAUX, FRANCE

J. GUEDJ (E-MAIL: guedj_jeremie@yahoo.fr)
THE FACULTY OF LIFE SCIENCES
BAR-ILAN UNIVERSITY
VIRAL DYNAMICS MODELING LABORATORY
RAMAT-GAN 52900, ISRAEL

MODELING PRIMARY HIV INFECTION

21

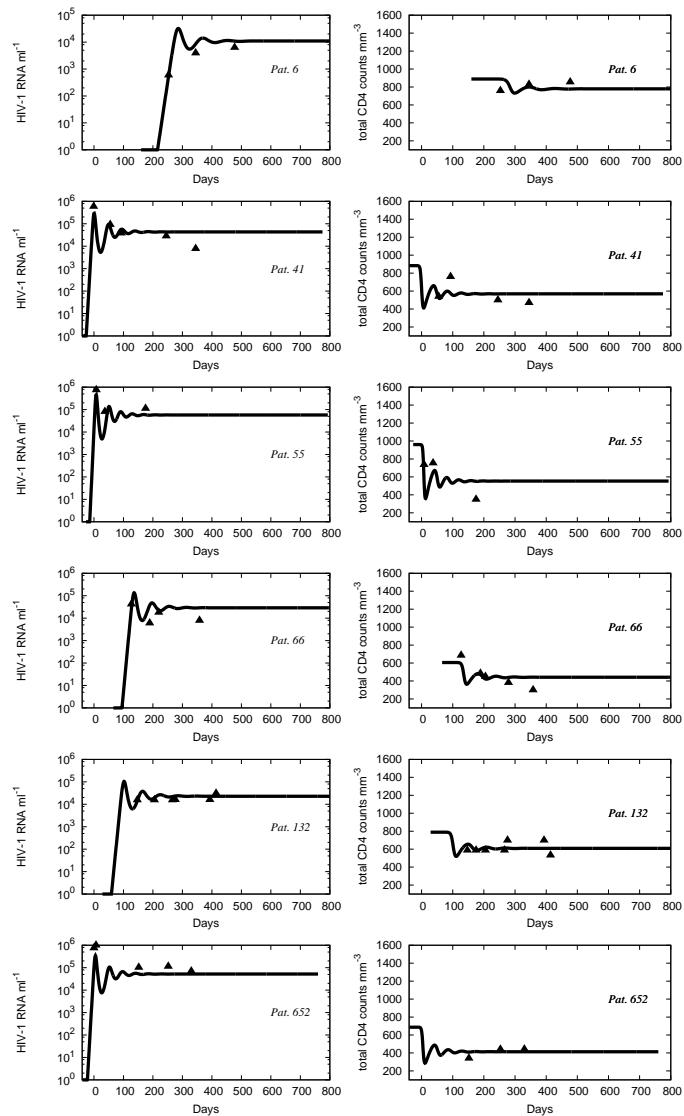


FIG 4. Individual predicted fits from the most probable date of infection and observed values of HIV RNA level and total CD4 count for patient 6, 41, 55, 66, 132 and 652 from the “productive cells model”, where day 0 is the date of last negative HIV test.

Chapitre 4

Tests du score pour l'exploration de modèles complexes

Comme nous l'avons évoqué dans le chapitre 2, l'estimation des modèles ODE est complexe et souvent longue. De plus, ces modèles possèdent de nombreux paramètres biologiques incluant des effets fixes, des effets aléatoires et des variables explicatives. La sélection de ces deux derniers types de variables conduit à l'estimation d'un grand nombre de modèles. Les tests classiquement utilisés sont le test du rapport de vraisemblance et le test de Wald. Ils nécessitent cependant d'estimer tous les modèles pour pouvoir les comparer et sélectionner le meilleur. Dans le cadre de la modélisation dynamique, l'estimation peut être longue et il paraît alors impossible d'estimer tous les modèles. Le développement d'un outil pour tester rapidement la présence d'effets aléatoires et de variables explicatives dans nos modèles sans avoir à tous les estimer semble donc pertinent.

Nous proposons dans cette partie d'utiliser les tests du score pour sélectionner les variables à inclure dans le modèle. Ces tests ont deux avantages majeurs : seul le modèle sous l'hypothèse nulle doit être estimé et la statistique de test prend une forme simple : linéaire des résidus pour les tests d'effet de variables explicatives et quadratique pour les tests d'homogénéité. Nous avons ainsi développé deux tests du score : un pour les variables explicatives et un test d'homogénéité.

Ce travail fait l'objet d'un article accepté par *Biometrical Journal*.

Score Tests for Exploring Complex Models: Application to HIV Dynamics Models

Julia Drylewicz^{1,2}, Daniel Commenges^{*1,2}, and Rodolphe Thiébaut^{1,2}

¹ INSERM, U897 Epidemiology and Biostatistics Research Center, Bordeaux, F-33076 France

² University of Bordeaux 2, Bordeaux, F-33076 France

Summary

In biostatistics, more and more complex models are being developed. This is particularly the case in system biology. Fitting complex models can be very time-consuming, since many models often have to be explored. Among the possibilities are the introduction of explanatory variables and the determination of random effects. The particularity of this use of the score test is that the null hypothesis is not itself very simple; typically, some random effects may be present under the null hypothesis. Moreover, the information matrix cannot be computed, but only an approximation based on the score. This paper examines this situation with the specific example of HIV dynamics models. We examine the score test statistics for testing the effect of explanatory variables and the variance of random effect in this complex situation. We study type I errors and the statistical powers of this score test statistics and we apply the score test approach to a real dataset of HIV-infected patients.

Key words: Homogeneity, Score test, Random effects, Asymptotic distribution, HIV dynamical model

1 Introduction

The three main statistics used for constructing statistical tests are the likelihood ratio, the maximum likelihood estimator and the score. Score tests have the advantage of simplicity for two reasons: (i) only the model under the null hypothesis needs to be fitted; (ii) the score statistic is often simple in form. Point (i) is especially relevant if the alternative involves a model which is numerically more challenging than the one involved under the null hypothesis. Because of point (ii) the score test approach has yielded many conventional tests such as the χ^2 , linear rank tests and the log-rank test and is also a good basis for constructing permutation tests (Commenges, 2003). Indeed, the score statistic is typically a linear form of residuals for testing regression coefficients, and a quadratic form for homogeneity testing. Point (i) is especially relevant for homogeneity testing because the null hypothesis of homogeneity is much simpler than the alternative. Since the breakthrough made by Liang (1987), homogeneity testing has flourished and been applied to generalized linear models (Hamerle, 1990; Jacqmin-Gadda and Commenges, 1995; Lin, 1997; Zhu and Zhang, 2006) and to the proportional hazards model (Commenges and Andersen, 1995); see (Commenges and Jacqmin-Gadda, 1997; Goeman *et al.*, 2006) for general studies. The score test has been used for testing the fit of a model (Le Cessie and Van Houwelingen, 1995), in multiple testing situations (Liquet and Commenges, 2005; Cook *et al.*, 2007; Commenges and Liquet, 2008), in genetic linkage (Commenges, 1994; Dudoit and Speed, 1999; Putter *et al.*, 2002; Lebrec and van Houwelingen, 2007) and in transmission/disequilibrium tests as well as in association studies (Clayton, 1999; Goeman *et al.*, 2004) and familial aggregation studies (Houwing-Duistermaat *et al.*, 1995).

In biostatistics, more and more complex models are being developed. This is particularly the case in system biology. Often, biological mechanisms are represented by a system of differential equations. The parameters of this system may vary from one subject to another and this can be taken into account by random effects. This kind of model involves three layers of numerical complexity for a direct maximum

* Corresponding author: e-mail: daniel.commenges@isped.u-bordeaux2.fr, Phone: +33 557 571 182 Fax: +33 556 240 081

likelihood approach: differential equations must be numerically solved; multiple integrals must be computed; and the likelihood must be maximized. This is feasible by a combination of algorithms (Guedj *et al.*, 2007a) computing the first derivative of the log-likelihood by differentiated differential equations and using a Newton-like algorithm (Commenges *et al.*, 2006) which approximates the Hessian of the log-likelihood using the scores. In spite of these tricks, maximization of the likelihood remains a time-consuming task, while many models often have to be explored. Among the possibilities are the introduction of explanatory variables and the determination of random effects. Here the score test approach may be useful, allowing these possibilities to be explored without having to fit more complex models. The particularity of this use of the score test is that the null hypothesis is not itself very simple; this is mainly due to the presence of random effects in the null hypothesis. Moreover, the information matrix cannot be computed, but only an approximation based on the score.

The aim of this paper is to study score tests for both explanatory variables and random effects as a tool for exploring complex models. In section 2, we examine score test statistics for testing the effect of explanatory variables and the random effect in this complex situation. In section 3, we study type I errors and the statistical powers of these score test statistics in a simulated model of HIV dynamics. In section 4, we apply the score tests to a real dataset of HIV-infected patients. A conclusion is given in section 5.

2 Score Tests for Complex Models

2.1 The non-linear mixed model

For subject i with $i = 1, \dots, n$, we consider a model which specifies the distribution of the observed vector $\mathbf{Y}_i = (Y_{ij}, j = 1, \dots, n_i)$:

$$Y_{ij} = g(t_{ij}, \tilde{\xi}^i) + \epsilon_{ij}, \quad j = 1, \dots, n_i, \quad (1)$$

where Y_{ij} is the j th measurement for subject i at the time t_{ij} , the ϵ_{ij} are independent Gaussian measurement errors with zero mean and variances σ^2 . The function $g(\cdot)$ is a twice differentiable (generally non-linear) function. The individual parameters $\tilde{\xi}^i = (\tilde{\xi}_l^i, l = 1, \dots, p)$ are modeled as a linear form: $\tilde{\xi}_l^i = \phi_l + \omega_l u_l^i + \mathbf{z}_l^i \beta_l$, where ϕ_l is the intercept, ω_l is the standard deviation of the random effect and \mathbf{z}_l^i is a vector of explanatory variables of the l -th parameter. The β_l 's are vectors of regression coefficients. We assume $\mathbf{u}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, where $\mathbf{u}^i = (u_l^i, l = 1, \dots, p)$ is the individual vector of random effects. We denote by $\theta = (\phi_l, \omega_l, \beta_l, l = 1, \dots, p)$ the set of parameters of the model, itself denoted $(P^\theta)_{\theta \in \Theta}$.

We denote by $\mathcal{L}_{\mathcal{Y}_i|u^i}^\theta(Y_i|\mathbf{u})$, the likelihood of observations for subject i given that the random effects \mathbf{u}^i take the value \mathbf{u} . Given \mathbf{u}^i , the Y_{ij} are independent, so $\mathcal{L}_{\mathcal{Y}_i|u^i}^\theta(Y_i|\mathbf{u}) = \prod_j \mathcal{L}_{\mathcal{Y}_{ij}|u^i}^\theta(Y_{ij}|\mathbf{u})$ where:

$$\mathcal{L}_{\mathcal{Y}_{ij}|u^i}^\theta(Y_{ij}|\mathbf{u}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(Y_{ij} - g(t_{ij}, \tilde{\xi}^i))^2}{2\sigma^2} \right\}.$$

The observed log-likelihood for subject i is:

$$L_i^\theta = \log \int_{\mathbb{R}^p} \mathcal{L}_{\mathcal{Y}_i|u^i}^\theta(Y_i|\mathbf{u}) \phi(\mathbf{u}) d\mathbf{u},$$

where ϕ is the multivariate normal density of $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. We denote $L = L_1^\theta + \dots + L_n^\theta$, the global log-likelihood.

2.2 Score test for explanatory variables

We want to test a possible effect of an explanatory variable \mathbf{z}_l^i on $\tilde{\xi}_l^i$. The null hypothesis H_0 is “ $\beta_l = 0$ ” which defines a submodel $(P^\theta)_{\theta \in \Theta_0}$, where Θ_0 is the subset of Θ such that $\beta_l = 0$. The class of

alternative hypotheses is the model $(P^\theta)_{\theta \in \Theta}$, where β_l is different from 0. The score for β_l , $U_{\beta_l}^*(\theta)$, is $\partial L / \partial \beta_l|_\theta$. If H_0 is true, the true probability P^* belongs to $(P^\theta)_{\theta \in \Theta_0}$; that is, there is a $\theta^* \in \Theta_0$ such that $P^{\theta^*} = P^*$. Because $U_{\beta_l}^* = U_{\beta_l}^1 + \dots + U_{\beta_l}^n$, a direct application of the central limit theorem shows that $U_{\beta_l}^*(\theta^*) / \sqrt{\text{Var } U_{\beta_l}^*(\theta^*)}$ has a standard normal distribution. Denote by $\hat{\theta}_0$ the maximum likelihood estimator of θ^* in $(P^\theta)_{\theta \in \Theta_0}$. If H_0 is true, $\hat{\theta}_0$ is consistent for θ^* , and by the continuous mapping theorem we have the same asymptotic distribution for the score test statistic:

$$S = \frac{U_{\beta_l}^*(\hat{\theta}_0)}{\sqrt{\widehat{\text{Var}} U_{\beta_l}^*(\hat{\theta}_0)}}, \quad (2)$$

where $\widehat{\text{Var}} U_{\beta_l}^*(\hat{\theta}_0)$ is a consistent estimator of $\text{Var } U_{\beta_l}^*(\hat{\theta}_0)$.

The scores $U_{\beta_l}^i(\hat{\theta}_0)$ are easily obtained using scores computed during the optimization procedure. We have $U_{\beta_l}^i(\hat{\theta}_0) = \mathbf{z}_l^i U_{\phi_l}^i(\hat{\theta}_0)$; since a Newton-like algorithm computes the gradient of the log-likelihood at each iteration, $U_{\phi_l}^i(\hat{\theta}_0)$, $i = 1, \dots, n$ are by-products of these computations at the last iteration.

The variance of the score $U_{\beta_l}^*(\hat{\theta}_0)$ under H_0 is asymptotically (Cox and Hinkley, 1974):

$$I_{\beta_l, \beta_l} - I_{\beta_l, \theta_{\setminus \beta_l}} (I_{\theta_{\setminus \beta_l}, \theta_{\setminus \beta_l}})^{-1} (I_{\beta_l, \theta_{\setminus \beta_l}})^T, \quad (3)$$

where θ is partitioned as $\theta = (\beta_l, \theta_{\setminus \beta_l})$ and $I_{\beta_l, \beta_l} = \sum E(\frac{\partial L^\theta}{\partial \beta_l})^2$, $I_{\theta_{\setminus \beta_l}, \theta_{\setminus \beta_l}} = \sum E(\frac{\partial L^\theta}{\partial \theta_{\setminus \beta_l}} (\frac{\partial L^\theta}{\partial \theta_{\setminus \beta_l}})^T)$, $I_{\beta_l, \theta_{\setminus \beta_l}} = \sum E(\frac{\partial L^\theta}{\partial \beta_l} (\frac{\partial L^\theta}{\partial \theta_{\setminus \beta_l}})^T)$, and all is calculated at $\hat{\theta}_0$.

Conventionally, these blocks of the information matrix are estimated by the corresponding blocks of the Hessian of the log-likelihood. For complex models, the Hessian cannot always be computed. For instance, Guedj *et al.* (2007a) use a quasi-Newton algorithm which avoids computing the Hessian. In that case an empirical estimator of the variance of the score can be used. Using the law of large numbers, we have:

$$n^{-1} \sum_{i=1}^n U_{\beta_l}^{i2}(\theta^*) \xrightarrow{p} \text{Var } U_{\beta_l}^i(\theta^*).$$

If H_0 is true, $\hat{\theta}_0$ converges in probability toward θ^* . Applying the continuous mapping theorem we have:

$$n^{-1} \sum_{i=1}^n U_{\beta_l}^{i2}(\hat{\theta}_0) \xrightarrow{p} \text{Var } U_{\beta_l}^i(\theta^*).$$

It follows that if we take $\widehat{\text{Var}} U_{\beta_l}^*(\hat{\theta}_0) = U_{\beta_l}^{12}(\hat{\theta}_0) + \dots + U_{\beta_l}^{n2}(\hat{\theta}_0)$ in (2), S has an asymptotic standard normal distribution under H_0 . The test, however, will be slightly conservative because $\text{Var } U_{\beta_l}^*(\hat{\theta}_0)$ is slightly smaller than $\text{Var } U_{\beta_l}^*(\theta^*)$. This is reflected by the correction term in (3) (that is neglected in this approach).

A global score test for explanatory variables can also be defined. Here the null hypothesis is $H_0 : \beta_l = 0, l = 1, \dots, p$ and the test statistic is $S_G = \mathbf{U}^{\bullet T} [\widehat{\text{Var}} \mathbf{U}^{\bullet}]^{-1} \mathbf{U}^{\bullet}$, where \mathbf{U}^{\bullet} is the vector of scores $U_{\beta_l}^*(\hat{\theta}_0)$, for $l = 1, \dots, p$ and $\widehat{\text{Var}} \mathbf{U}^{\bullet}$ is the empirical variance of \mathbf{U}^{\bullet} as above. This statistic has an asymptotic normal χ_p^2 distribution under H_0 .

2.3 Score test of homogeneity

Let us now look at the problem of testing whether there is a random effect u_l^i on $\tilde{\xi}_l^i$. The null hypothesis H_0 is “ $\omega_l = 0$ ”. This defines a sub-model $(P^\theta)_{\theta \in \Theta'_0}$. The class of alternative hypotheses is the model

$(P^\theta)_{\theta \in \Theta}$, where ω_l is different from 0. For this we can use the general form of the score test of homogeneity developed in section 2 of Commenges and Jacqmin-Gadda (1997). We denote by $\mathcal{L}_{\mathcal{Y}|u_l^i}^\theta$ the global likelihood conditional on the u_l^i and by $\mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta$ the likelihood conditional on the u_l^i for the j th measurement of subject i .

The general robust score statistic for H_0 is $T_{WPC} = \mathbf{U}^T \mathbf{W}^* \mathbf{U} - \text{tr}(\hat{\mathbf{I}}_\varepsilon \mathbf{W}^*)$, where \mathbf{U} is the vector of dimension $N = n_1 + \dots + n_n$ of $U_l^{ij} = \partial \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta / \partial \varepsilon_{ijl} |_{\hat{\theta}_0}$ with $\varepsilon_{ijl} = \omega_l u_l^i$, $\hat{\mathbf{I}}_\varepsilon$ is the matrix whose entries are $-\partial^2 \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta / \partial \varepsilon_{ijl} \partial \varepsilon_{ij'l} |_{\hat{\theta}_0}$ and \mathbf{W}^* is the correlation matrix of u_l^i minus the identity matrix. As recommended in Commenges and Jacqmin-Gadda (1997), we retain the weighted pairwise correlation (WPC) term which is more robust to overdispersion than the crude score statistic. The matrix \mathbf{W}^* is a square matrix of dimension N which can be denoted by $\text{diag}(\mathbf{1}^*, \mathbf{1}^*, \dots, \mathbf{1}^*)$, where each $\mathbf{1}^*$ is a matrix of dimension n_i with element ij equal to $1 - \delta_{ij}$ (δ_{ij} , the Kronecker symbol).

The scores U_l^{ij} are obtained by using the fact that $g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)$ depends on ε_{ijl} through $\tilde{\boldsymbol{\xi}}_l^i = \phi_l + \varepsilon_{ijl}$. We have:

$$\frac{\partial \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta}{\partial \varepsilon_{ijl}} = \frac{1}{\sigma^2} \left(Y_{ij} - g(t_{ij}, \tilde{\boldsymbol{\xi}}^i) \right) \frac{\partial g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)}{\partial \varepsilon_{ijl}}$$

and

$$\frac{\partial \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta}{\partial \phi_l} = \frac{1}{\sigma^2} \left(Y_{ij} - g(t_{ij}, \tilde{\boldsymbol{\xi}}^i) \right) \frac{\partial g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)}{\partial \phi_l}$$

Applying the chain rule of differentiation, we can see that:

$$\frac{\partial g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)}{\partial \varepsilon_{ijl}} = \frac{\partial g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)}{\partial \tilde{\boldsymbol{\xi}}_l^i} \frac{\partial \tilde{\boldsymbol{\xi}}_l^i}{\partial \varepsilon_{ijl}} = \frac{\partial g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)}{\partial \tilde{\boldsymbol{\xi}}_l^i} = \frac{\partial g(t_{ij}, \tilde{\boldsymbol{\xi}}^i)}{\partial \phi_l}.$$

So, we have: $\partial \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta / \partial \varepsilon_{ijl} |_{\hat{\theta}_0} = \partial \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta / \partial \phi_l |_{\hat{\theta}_0}$.

In the conventional setting there is no random effect under H_0 . Since the Y_{ij} are independent according to the random effects, the conditional log-likelihood is the sum: $\log \mathcal{L}_{\mathcal{Y}|u_l^i}^\theta = \sum_{i,j} \log \mathcal{L}_{\mathcal{Y}_{ij}|u_l^i}^\theta$. It follows that the matrix $\hat{\mathbf{I}}_\varepsilon$ is diagonal, and $\text{tr}(\hat{\mathbf{I}}_\varepsilon \mathbf{W}^*) = \mathbf{0}$. This simplification, however, does not occur in the complex situation at hand, and this raises a challenging issue since $\hat{\mathbf{I}}_\varepsilon$ involves the second derivative of a complex likelihood, a computation that we wish to avoid. It seems that the informative term is still $\mathbf{U}^T \mathbf{W}^* \mathbf{U}$ but in the complex situation we do not have $E[\mathbf{U}^T \mathbf{W}^* \mathbf{U}] = \mathbf{0}$ under H_0 : the second term is needed to ensure $E[T_{WPC}] = 0$. We propose to keep $\mathbf{U}^T \mathbf{W}^* \mathbf{U}$ for the test statistic but we have to find a way of computing its expectation under H_0 . Specifically our statistic is:

$$T = \mathbf{U}^T \mathbf{W}^* \mathbf{U} = \sum_{i=1}^n T_i, \text{ where } T_i = \sum_{\substack{j, j' \leq n_i \\ j' \neq j}} U_l^{ij} U_l^{ij'}$$

Clearly that the expectation of T is equal to zero if U_l^{ij} and $U_l^{ij'}$ are independent under H_0 (the conventional case), and generally different from zero if they are not (the complex case). Therefore, we propose the test statistic defined by:

$$S_H = \frac{T - \widehat{E}_{H_0}[T]}{\sqrt{\widehat{\text{Var}}_{H_0} T}},$$

where $\widehat{E}_{H_0}[T]$ and $\widehat{\text{Var}}_{H_0} T$ are estimators of the expectation and variance of T under H_0 . It has an asymptotic $\mathcal{N}(0, 1)$ distribution under H_0 . This allows the construction of a one-side test, because under the alternative, $E_{H_1}(T) > E_{H_0}(T)$. Therefore, H_0 is rejected for large positive values of S_H . To obtain $\widehat{E}_{H_0}[T]$ and $\widehat{\text{Var}}_{H_0} T$, we use a parametric bootstrap. That is, we simulate K datasets of n patients under P^{θ_0} ; for each replication we compute the statistic $T_{(k)}$, and $\widehat{E}_{H_0}[T]$ and $\widehat{\text{Var}}_{H_0} T$ are the empirical mean and variance of these values.

For non-linear mixed effects models, we need to compute a multiple integral of dimension q (the number of random effects under H_0) to compute the likelihood and hence each $T_{(k)}$. Although the computation of S_H is not very simple, it is much quicker than fitting a model under the alternative hypothesis, which is necessary for likelihood ratio or Wald tests.

2.4 Ascendent strategy

The score test can be applied to guide an ascendent strategy in the construction of a complex model. Starting with a model with few random effects and no explanatory variables, we can perform score tests for introducing explanatory variables and additional random effects. There may be competition between explanatory variables and random effects. Indeed, random effects are often interpreted as reflecting the effect of omitted explanatory variables. Therefore, if we find an explanatory variable and a random effect for the same parameter, we include the explanatory variable first. Then, we can test the presence of a random effect on this parameter to check for unexplained variability.

3 Simulation Study

In this section, we study the performance of the score tests for explanatory variables and random effects. We take the example of a conventional HIV dynamics model.

3.1 HIV dynamics model

We consider an Ordinary Differential Equations (ODE) model for the interaction between HIV and the immune system. The model has been mainly used after treatment initiation (Perelson *et al.*, 1996; Nowak *et al.*, 1997; Bonhoeffer *et al.*, 1997), but also in primary infection as in Stafford *et al.* (2000). For subject i , we consider three compartments: T^i (target cells: CD4+ T lymphocytes), I^i (infected cells) and V^i (free virions). The model can be written as a system of ordinary differential equations:

$$\frac{dT^i}{dt} = \lambda^i - \gamma V^i T^i - \mu_T^i T^i, \quad \frac{dI^i}{dt} = \gamma V^i T^i - \mu_I^i I^i, \quad \frac{dV^i}{dt} = \pi^i I^i - \mu_V V^i. \quad (4)$$

The target cells enter the blood circulation at rate λ^i (partly due to thymic production), they can be infected by the virus at the rate γ and die at the rate μ_T^i . The infected CD4 die at the rate μ_I^i and produce virions at the rate π^i . Virions die at the rate μ_V and can also infect other target cells.

Because of identifiability issues (Guedj *et al.*, 2007b), γ and μ_V are taken to be known and common for all subjects. We can construct a statistical model as in section 2.1, where the individual parameters are $\tilde{\xi}^i = (\tilde{\lambda}^i, \tilde{\mu}_I^i, \tilde{\pi}^i, \tilde{\mu}_T^i)$ with $\tilde{\lambda}^i = \log \lambda^i$, $\tilde{\mu}_I^i = \log \mu_I^i$, $\tilde{\pi}^i = \log \pi^i$ and $\tilde{\mu}_T^i = \log \mu_T^i$. The log-transform ensures the positivity of the estimates of the biological parameters which appear in equation (4). Generally, we do not directly observe all the components of the system, but rather $M < 3$ functions of $\mathbf{X}^i = (T^i, I^i, V^i)$. We observe Y_{ijm} , the j th measurement of the m th observable component for subject i at time t_{ijm} . The function g defined in equation (1) is then $g = (g_1, \dots, g_M)$ so we have:

$$Y_{ijm} = g_m(\mathbf{X}^i(t_{ijm}, \tilde{\xi}^i)) + \epsilon_{ijm}, \quad j = 1, \dots, n_{im} \quad m = 1, \dots, M,$$

Table 1 Parameters of the HIV dynamics model: meaning and values used for the simulation

Parameter	Meaning (per day)	Value	Reference
λ	Rate of T cell production per μL	32.7	Drylewicz <i>et al.</i> (2009)
μ_I	Death rate of I cells	1.86	Drylewicz <i>et al.</i> (2009)
π	Number of virions per I cells	194.0	Drylewicz <i>et al.</i> (2009)
μ_T	Death rate of T cells	0.04	Drylewicz <i>et al.</i> (2009)
γ	Infection rate of T cells per virion	0.00027	Phillips (1996); Davenport <i>et al.</i> (2006)
μ_V	Clearance of free virions	20.0	Ramratnam <i>et al.</i> (1999)
$\sigma_{\tilde{\lambda}}$	Standard deviation of random effect on $\tilde{\lambda}$	0.30	
$\sigma_{\tilde{\mu}_I}$	Standard deviation of random effect on $\tilde{\mu}_I$	0.20	

where the ϵ_{ijm} are independent Gaussian measurement errors with zero means and variances σ_m^2 . Specifically, the observed components were the base-10 logarithm of HIV RNA load ($g_1(\mathbf{X}) = \log_{10}(V)$) and the fourth root of total CD4 count ($g_2(\mathbf{X}) = (T + I)^{0.25}$). These transformations are commonly used for achieving normality and homoscedasticity of measurement error distributions (Thiébaut *et al.*, 2003).

3.2 Simulations

Data were simulated from model (4) using values of parameters from Drylewicz *et al.* (2009) (see table 1). We assume that the model is at equilibrium before infection, so at $t = 0$, we have (omitting the superscript "i" for notational simplification):

$$T_0 = \frac{\lambda}{\mu_T}, I_0 = 0, V_0 = 0.$$

The initial inoculum of virions is fixed at 10^{-6} virions/mm³ (Ciupe *et al.*, 2006; Stafford *et al.*, 2000) which is equivalent to 5 virions for 5 liters of blood. The introduction of virions in the system disrupts the initial stability and the system stabilizes to a new equilibrium if the basic reproductive number R_0 is higher than 1 (De Leenheer and Smith, 2003):

$$\bar{T} = \frac{\mu_I \mu_V}{\gamma \pi}, \bar{I} = \frac{\lambda \gamma \pi - \mu_T \mu_I \mu_V}{\gamma \pi \mu_I}, \bar{V} = \frac{\lambda \gamma \pi - \mu_T \mu_I \mu_V}{\gamma \mu_I \mu_V}.$$

The times of measurement t_{ijm} were: 0, 5, 10, 15, 20, 25, 30, 40, 50 and 60 for each subject i and observed component m . We simulated data for 100 patients. We included random effects for two parameters: λ and μ_I (see table 1).

We computed the type I errors and the statistical powers of score tests according to the regression coefficient of the explanatory variables (binary or continuous) and the standard deviation of the random effect. For binary explanatory variables, we used a Bernoulli distribution with $p=0.20$ which is similar to the distribution of the variable *gender* in the application. For continuous explanatory variables, we used a normal distribution with expectation 35 and standard deviation 9.5 (similar to the distribution of the variable *age* in the application). The expectation and the variance under H_0 for the test on random effects were calculated using 100 datasets simulated under H_0 . Results for 100 datasets are presented in figures 1 and 2.

The type I errors for the score tests were around 0.04 and 0.05, for explanatory variables and random effects, respectively. The speed of convergence toward 1 (when the effect to be detected increases) is different according to the parameters. The score test is more powerful to detect an explanatory variable on λ and μ_T than on μ_I and π . Similarly, score tests of homogeneity are more powerful to detect a random

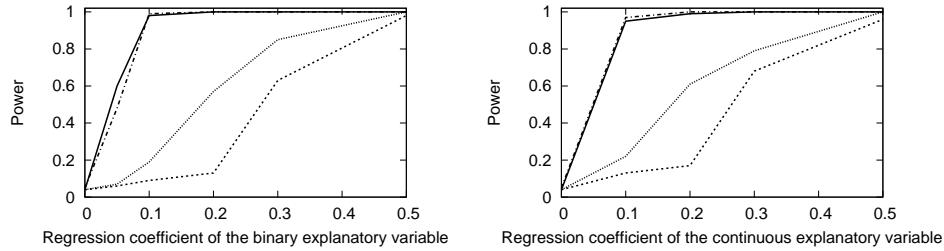


Fig. 1 Power of the score test for binary and continuous explanatory variables for λ (solid line), μ_I (dashed line), π (dotted line) and μ_T (dotted-dashed line)

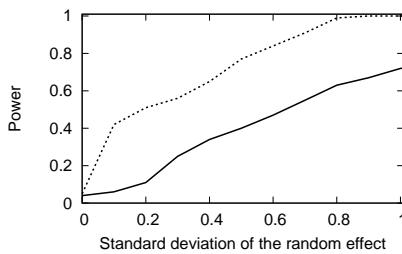


Fig. 2 Power of the homogeneity score test for π (solid line) and μ_T (dashed line)

effect on μ_T than on π . The initial equilibrium of the ODE is defined by the ratio λ/μ_T , which may explain why a small variation of one of these parameters is more easily detected than for other parameters. Moreover, the total CD4 counts ($T + I$) is observed and this gives more information on T than on I because $T \gg I$ during all the dynamics. We made simulations assuming that we observed T , I and V . In this case, the powers for π and μ_I increased as fast as the powers for λ and μ_T (not shown).

We also studied the robustness of the score tests to misspecification of the random effects in terms of type I error. We simulated 100 datasets with three random effects (on λ , μ_I and π with standard deviations of 0.30, 0.20 and 0.30 respectively). We included only the first two under the null hypothesis; thus, H_0 was incorrectly specified. We tested a possible effect of a binary explanatory variable on each parameter and a possible random effect on the fourth parameter (μ_T). The results for 100 simulations are presented in table 2. The type I errors were around 0.05 when the model was misspecified. Therefore, it seems that the score tests are robust to misspecification of random effects.

To study the competition between explanatory variables and random effect which may arise in the ascendent strategy demonstrated in section 2.4, we simulated a dataset of 100 patients with a binary explanatory variable on μ_T ($\tilde{\mu}_T^i = \phi_{\mu_T} + z_{\mu_T}^i \beta_{\mu_T}$, with $\beta_{\mu_T} = 0.3$) with the same distribution as previously described. The score test of homogeneity for $H_0 : \omega_{\mu_T} = 0$, not including the explanatory variable, was significant ($S_H = 2.17$, $p=0.015$) and the score test for the explanatory variable on this parameter ($H_0 : \beta_{\mu_T} = 0$) was also significant ($S = 4.25$, $p=2.10^{-5}$). We included the explanatory variable in the model and the score test of homogeneity was no longer significant ($S_H = 0.72$, $p=0.24$).

The computation time for simulating 100 datasets under H_0 for the score test on random effects increases linearly with the number of patients and the number of measurements times. Indeed, we need to simulate the same number of patients and the same times of measurement as in the testing dataset. We studied the power of the score test statistic by simulating only 1 and 10 datasets under H_0 . For a standard-deviation of a random effect of 0.1, the difference in power was on average 15% and 2% (for 1 and 10

Table 2 Type 1 errors of score tests for explanatory variables and random effects for well-specified and misspecified models

	Parameter	Well-specified	misspecified
Explanatory Variables	λ	0.04	0.05
	μ_I	0.04	0.05
	π	0.04	0.04
	μ_T	0.04	0.05
Random effects	π	0.04	0.03
	μ_T	0.05	0.04

datasets respectively); for a standard deviation of 0.3 it was 7% and 1% respectively. Thus, simulating 10 datasets seems to be sufficient for a faster exploration.

In spite of this computation time, these score tests are much quicker than performing a likelihood ratio test. Indeed, estimating another model can be very time-consuming. The time for fitting an ODE model ranges from hours to days according to the model and the dataset. In contrast, computing the score test for explanatory variables takes seconds and the score test of homogeneity just a few minutes.

4 Application

We applied these tests to a real dataset from the Concerted Action on Seroconversion to AIDS and Death in Europe (CASCADE) study. CASCADE is an international collaboration pooling data from sero-converters in Europe, Canada and Australia. The CASCADE study has been described in detail elsewhere (CASCADE, 2003) in the context of primary HIV infection. This period is the first phase of HIV infection and the main difficulty in this context is that the date of infection is unknown. This date is the origin of time for our model. However, the patients included in this collaboration have a date of last negative HIV test and a date of first positive HIV test. These dates give a time interval for the date of infection. We adapted the algorithm of estimation to take into account the uncertainty of the date of infection (Drylewicz *et al.*, 2009). Moreover, the left-censoring of the HIV RNA load due to the detection limit is taken into account in the computation of the likelihood (Thiébaut *et al.*, 2006; Guedj *et al.*, 2007a).

We selected patients if the delay between their last negative and first positive HIV test was less than 3 years, if they did not receive any antiretroviral treatment during the first year following the date of seropositive test and if they had more than 3 measurements of CD4 or viral load during the first year of follow-up with the first detectable viral load measurement during the first three months after the date of seropositive test. A total of 761 patients met these criteria.

We fitted the model described in section 3.1 with random effects on λ and μ_I and no explanatory variable as in Drylewicz *et al.* (2009). Then, we tested two explanatory variables: *gender* and *age* at inclusion. In the sample of 761 patients, the median age at inclusion was 34 years (InterQuartile Range [IQR]=21-40) and 84% were men (N=640). We simultaneously tested each variable on the four estimated parameters using the global test previously described (this test follows a Chi-square distribution with 4 degrees of freedom). For the variable *gender*, the test was not significant: $S_G = 1.06$, $p=0.90$. When *age* was taken as a continuous variable, the global test was not significant: $S_G = 0.84$, $p=0.93$. However, if we considered *age* as a binary variable coded 1 for older patients (≥ 50 years) and 0 for younger (50 years old is the 95-th percentile), the global test was significant ($S_G = 13.03$, $p=0.01$). In fact, the rationale of testing a threshold effect for older age is that aging tends to compromise the immune response (Dorshkind *et al.*, 2009) and the ability of the thymus to produce new CD4+ T cells (Douek *et al.*, 1998). We tested individually the effects of *age* on all the parameters and found a significant effect for μ_I ($S = -1.99$, $p=0.04$). This means that subjects over 50 have a lower death rate of infected T cells, which may reflect a less efficient immune response as compared to younger patients.

We also tested the presence of random effect on π and μ_T . To determine the expectation and the variance under the null hypothesis, we simulated 10 datasets of 761 patients under H_0 with the same measurement design as the real dataset, and we assumed that the date of infection was the mode of the *a posteriori* distribution computed during the optimization procedure. The tests for additional random effects were not significant at the 0.05 level: $S_H = 1.42$ ($p=0.08$) and $S_H = 1.55$ ($p=0.06$), for π and μ_T respectively.

5 Conclusion

We have developed score tests in non-standard situations that arise in complex modeling. Indeed, choosing random effects and explanatory variables in non-linear models including several parameters can be difficult and time-consuming. Score tests can be used in an ascendent strategy for exploring a family of complex models.

With regard to efficiency, score tests have optimal properties (Cox and Hinkley, 1974). Therefore, the only issue to be discussed is the effect of using the statistic $\mathbf{U}^T \mathbf{W}^* \mathbf{U}$ in the complex situation rather than the complete score statistic which is $\mathbf{U}^T \mathbf{W}^* \mathbf{U} - \text{tr}(\hat{\mathbf{I}}_\varepsilon \mathbf{W}^*)$. We must disregard the term $\text{tr}(\hat{\mathbf{I}}_\varepsilon \mathbf{W}^*)$ in complex situations where we cannot reliably compute the second derivative of the log-likelihood. It is difficult to prove that this term does not carry much information but three arguments can be put forward: (i) it involves only the second derivative of the log-likelihood; (ii) it does not appear in the simple situation; (iii) simulations studies show the tests have good power. For μ_T , for instance, the power is above 0.80 when the standard-deviation is 0.50. This standard-deviation bearing on the logarithm of the biological parameters can be interpreted as the coefficient of variation. As noted by a reviewer, robustness could be an issue since we compute the expectation of $\mathbf{U}^T \mathbf{W}^* \mathbf{U}$ by simulation under the null hypothesis. However, most tests are based on the distribution of the statistic under the null hypothesis. Second, we use the WPC version of the score statistics, which is more robust to overdispersion than the crude score statistics (Commenges and Jacqmin-Gadda, 1997). Third, a simulation study shows good robustness to misspecification of other random effects.

These tests can be used in particular in dynamics models where the number of parameters and the estimation computation time is high. The number of models to be explored can be huge and these score tests can help in selecting the variables with the largest effect, while estimating all possible models is impossible. The illustration from the CASCADE study is promising because it already yields important results in a very complex situation (due to the unknown date of infection). The ascendent strategy should prove a powerful tool for more thoroughly exploring HIV dynamics models. The proposed methods could also be useful for exploring other complex models

Acknowledgements J.Drylewicz received a PhD grant from CASCADE Concerted Action on Seroconversion to AIDS and Death in Europe (European Commission Framework Program 6). CASCADE has been funded through grants from the European Union BMH4-CT97-2550, QLK2-2000-01431, QLRT-2001-01708 and LSHP-CT-2006-018949.

Conflict of Interests Statement
All authors have declared no conflict of interest.

References

- Bonhoeffer, S., May, R., Shaw, G., and Nowak, M. (1997). Virus dynamics and drug therapy. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(13), 6971–6976.
- CASCADE, C. (2003). Determinants of survival following HIV-1 seroconversion after the introduction of HAART: results from CASCADE. *The Lancet*, **362**(9392), 1267–1274.

- Ciupe, M., Bivort, B., Bortz, D., and Nelson, P. (2006). Estimating kinetic parameters from HIV primary infection data through the eyes of three different mathematical models. *Mathematical Biosciences*, **200**(1), 1–27.
- Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *The American Journal of Human Genetics*, **65**(4), 1170–1177.
- Commenges, D. (1994). Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic. *Genetic epidemiology*, **11**(2), 189–200.
- Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, **15**(2), 171–185.
- Commenges, D. and Andersen, P. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis*, **1**(2), 145–156.
- Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society: Series B (Methodological)*, **59**(1), 157–171.
- Commenges, D. and Liquet, B. (2008). Asymptotic distribution of score statistics for spatial cluster detection with censored data. *Biometrics*, **64**(4), 1287–1289.
- Commenges, D., Jacqmin-Gadda, H., Proust, C., and Guedj, J. (2006). A Newton-like algorithm for likelihood maximization: the robust variance scoring algorithm. *arXiv:math.ST/0610402*.
- Cook, A., Gold, D., and Li, Y. (2007). Spatial cluster detection for censored outcome data. *Biometrics*, **63**(2), 540–549.
- Cox, D. and Hinkley, D. (1974). *Asymptotic theory*, chapter 9. Chapman & Hall, London.
- Davenport, M., Zhang, L., Shiver, J., Casmiro, D., Ribeiro, R., and Perelson, A. (2006). Influence of peak viral load on the extent of CD4+ T-Cell depletion in Simian HIV Infection. *Journal of Acquired Immune Deficiency Syndromes*, **41**(3), 259.
- De Leenheer, P. and Smith, H. (2003). Virus dynamics: a global analysis. *SIAM Journal on Applied Mathematics*, **63**(4), 1313–1327.
- Dorshkind, K., Montecino-Rodriguez, E., and Signer, R. (2009). The ageing immune system: is it ever too old to become young again? *Nature Review Immunology*, **9**(41), 57–62.
- Douek, D., McFarland, R., Keiser, P., Gage, E., Massey, J., Haynes, B., Polis, M., Haase, A., Feinberg, M., Sullivan, J., et al. (1998). Changes in thymic function with age and during the treatment of HIV infection. *Nature*, **396**(6712), 690–5.
- Drylewicz, J., Guedj, J., Commenges, D., Thiébaut, R., and on behalf of the CASCADE Collaboration (2009). Modelling the dynamics of biomarkers during primary HIV infection taking into account the uncertainty of infection date. *submitted*.
- Dudoit, S. and Speed, T. (1999). A score test for linkage using identity by descent data from sibships. *Annals of Statistics*, **27**, 943–986.
- Goeman, J., van de Geer, S., de Kort, F., and van Houwelingen, H. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**(1), 93–99.
- Goeman, J., van de Geer, S., and van Houwelingen, H. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), 477–493.
- Guedj, J., Thiébaut, R., and Commenges, D. (2007a). Maximum likelihood estimation in dynamical models of HIV. *Biometrics*, **63**, 1198–1206.
- Guedj, J., Thiébaut, R., and Commenges, D. (2007b). Practical identifiability of HIV dynamics models. *Bulletin of Mathematical Biology*, **69**(8), 2493–2513.
- Hamerle, A. (1990). On a simple test for neglected heteroscedasticity in panel studies. *Biometrics*, **46**, 193–199.

- Houwing-Duistermaat, J., Derkx, B., Rosendaal, F., and Van Houwelingen, H. (1995). Testing familial aggregation. *Biometrics*, **51**, 1292–1301.
- Jacqmin-Gadda, H. and Commenges, D. (1995). Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association*, **90**(432), 1237–1246.
- Le Cessie, S. and Van Houwelingen, H. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, **51**(2), 600–614.
- Lebrec, J. and van Houwelingen, H. (2007). Score test for linkage in generalized linear models. *Hum Hered*, **64**(1), 5–15.
- Liang, K. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika*, **74**(2), 259–264.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**(2), 309–326.
- Liquet, B. and Commenges, D. (2005). Computation of the p-value of the maximum of score tests in the generalized linear model; application to multiple coding. *Statistics and Probability Letters*, **71**(1), 33–38.
- Nowak, M., Lloyd, A., Vasquez, G., Wiltrot, T., Wahl, L., Bischofberger, N., Williams, J., Kinter, A., Fauci, A., Hirsch, V., et al. (1997). Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection. *Journal of Virology*, **71**(10), 7518–7525.
- Perelson, A., Neumann, A., Markowitz, M., Leonard, J., and Ho, D. (1996). Viral dynamics in human immunodeficiency virus type 1 infection. *Science*, **271**, 1582–1586.
- Phillips, A. (1996). Reduction of HIV concentration during acute infection: independence from a specific immune response. *Science*, **271**(5248), 497–499.
- Putter, H., Sandkuyl, L., van Houwelingen, J., et al. (2002). Score test for detecting linkage to quantitative traits. *Genetic Epidemiology*, **22**(4), 345–355.
- Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J., Markowitz, M., Moore, J., Perelson, A., and Ho, D. (1999). Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *The Lancet*, **354**, 1782–1786.
- Stafford, M., Cao, Y., Ho, D., Corey, L., and Perelson, A. (2000). Modeling plasma virus concentration and CD4+ T cell kinetics during primary HIV infection. *Journal of Theoretical Biology*, **203**, 285–301.
- Thiébaut, R., Jacqmin-Gadda, H., Leport, C., Katlama, C., D., C., Le Moing, V., Morlat, P., Chene, G., and the APRICO Study Group (2003). 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. *Journal of Biopharmaceutical Statistics*, **13**, 271–282.
- Thiébaut, R., Guedj, J., Jacqmin-Gadda, H., Chêne, G., Trimoulet, P., Neau, D., and Commenges, D. (2006). Estimation of dynamical model parameters taking into account undetectable marker values. *BMC Medical Research Methodology*, **6**(1), 1–10.
- Zhu, H. and Zhang, H. (2006). Generalized score test of homogeneity for mixed effects models. *Annals of Statistics*, **34**(3), 1545.

CASCADE Collaboration

Steering Committee: Julia Del Amo (Chair), Laurence Meyer (Vice Chair), Heiner Bucher, Geneviève Chêne, Deenan Pillay, Maria Prins, Magda Rosinska, Caroline Sabin, Giota Touloumi
Co-ordinating Centre: Kholoud Porter (Project Leader), Sara Lodi, Sarah Banbury, Kate Coughlin, Sarah Walker, Abdellah Babiker, Janet Darbyshire
Clinical Advisory Board: Heiner Bucher, Andrea de Luca, Martin Fisher, Roberto Muga
Collaborators: Australia Sydney AIDS Prospective Study and Sydney Primary HIV Infection cohort (John Kaldor,

Tony Kelleher, Tim Ramacciotti, Linda Gelgor, David Cooper, Don Smith); Canada South Alberta clinic (John Gill); Denmark Copenhagen HIV Seroconverter Cohort (Louise Bruun Jorgensen, Claus Nielsen, Court Pedersen); Estonia Tartu Ulikool (Irja Lutsar); France Aquitaine cohort (Geneviève Chêne, Francois Dabis, Rodolphe Thiébaut, Bernard Masquelier), French Hospital Database (Dominique Costagliola, Marguerite Guiguet), Lyon Primary Infection cohort (Philippe Vanhems), SEROCO cohort (Laurence Meyer, Faroudy Boufassa); Germany German cohort (Osamah Hamouda, Claudia Kucherer); Greece Greek Haemophilia cohort (Giota Touloumi, Nikos Pantazis, Angelos Hatzakis, Dimitrios Paraskevis, Anastasia Karafoulidou); Italy Italian Seroconversion Study (Giovanni Rezza, Maria Dorrucci, Benedetta Longo, Claudia Balotta); Netherlands Amsterdam Cohort Studies among homosexual men and drug users (Maria Prins, Janne van der Helm, Anneke Krol, Hanneke Schuitemaker); Norway Oslo and Ulleval Hospital cohorts (Mette Sannes, Oddbjorn Brubakk, Anne Eskild, Johan N Bruun); Poland National Institute of Hygiene (Magdalena Rosinska); Portugal Universidade Nova de Lisboa (Ricardo Camacho); Russia Pasteur Institute (Tatyana Smolskaya); Spain Badalona IDU hospital cohort (Roberto Muga, Jordi Tor), Barcelona IDU Cohort (Patricia Garcia de Olalla, Joan Cayla), Madrid cohort (Julia Del Amo, Jorge del Romero), Valencia IDU cohort (Santiago Pérez-Hoyos, Ildefonso Hernandez Aguado); Switzerland Swiss HIV Cohort Study (Heiner C. Bucher, Martin Rickenbach, Patrick Francioli); Ukraine Perinatal Prevention of AIDS Initiative (Ruslan Malyuta); United Kingdom Edinburgh Hospital cohort (Ray Brettle), Health Protection Agency (Valerie Delpech, Sam Lattimore, Gary Murphy, John Parry, Noel Gill), Royal Free haemophilia cohort (Caroline Sabin, Christine Lee), UK Register of HIV Seroconverters (Kholoud Porter, Anne Johnson, Andrew Phillips, Abdel Babiker, Janet Darbyshire, Valerie Delpech), University College London (Deenan Pillay), University of Oxford (Harold Jaffe).

Chapitre 5

Dynamique de la primo-infection à VIH

5.1 Contexte

Comme nous l'avons expliqué dans le chapitre 1, la primo-infection est une période cruciale de l'infection à VIH. La primo-infection est caractérisée par un taux de réPLICATION virale très important associé à une chute du nombre de lymphocytes T-CD4+ dans le sang. Quelques études se sont intéressées à l'étude de la dynamique virale durant cette phase précoce de l'infection (Kaufmann et al., 1998; Little et al., 1999; Stafford et al., 2000 ;...). Elles se basaient sur un petit nombre de patients, hautement sélectionnés mais ayant des données très riches. Ainsi, la date du pic de virémie a pu être estimée ainsi que sa valeur (Kaufmann et al., 1998). D'autres auteurs ont calculé le nombre de reproduction de base R_0 qui reflète le nombre de nouvelles cellules infectées par une seule cellule infectée. Cette quantité peut permettre d'estimer l'efficacité nécessaire à une intervention de prévention de l'infection (Little et al., 1999).

Dans cette partie, nous proposons de décrire la dynamique viro-immunologique de la primo-infection à VIH et d'estimer les paramètres biologiques de 761 patients provenant de cohortes de séroconvertisseurs de la collaboration CASCADE. Nous estimons tout d'abord le modèle proposé par Perelson et al. (1996) et étudions l'effet de variables explicatives sur les différents paramètres biologiques. Nous proposons ensuite quatre modèles dynamiques de l'interaction du VIH et du système immunitaire. Ces modèles font des

hypothèses biologiques différentes de cette interaction complexe. Nous les comparons en terme d'estimation de paramètres, de dynamique mais également en terme statistiques au modèle de Perelson et al. (1996).

5.2 Un modèle de base

Une partie du travail de cette section a été présentée par poster (n° 512) à la 16ème Conference on Retroviruses and Opportunistic Infection à Montréal (Canada, 8-11 Février 2009).

5.2.1 Données

L'échantillon d'étude est issu de la collaboration CASCADE (Concerted Action on Seroconversion to AIDS and Death in Europe). CASCADE est une collaboration internationale qui regroupe les données de séroconvertisseurs en Europe, au Canada et en Australie (CASCADE, 2003). Les objectifs de cette étude sont d'aborder des questions sur l'histoire naturelle de l'infection à VIH-1 qui ne peuvent pas être correctement étudiées par une seule étude de cohorte. Les données utilisées sont issues de la base 2006 des vingt cohortes participantes. Nous avons sélectionné les séroconvertisseurs ayant un délai entre leur dernière sérologie négative (d_-) et leur première sérologie positive (d_+) inférieur à 3 ans, n'ayant pas suivi de traitement antirétroviral et ayant au moins 3 mesures de CD4 ou de charge virale durant la première année de suivi (avec la première mesure de charge virale détectable dans les trois mois suivant la date de sérologie positive). Nous avons inclus 761 patients.

Les caractéristiques des patients sont présentés dans le tableau 5.1. Les patients sélectionnés ont en médiane 4 mesures de CD4 et de charge virale (Ecart InterQuartile (EIQ)= [3 ; 5]). La majorité des patients a été infectée par voie homosexuelle (71%). Le suivi a été censuré 1 an après d_+ conduisant à un suivi médian de 195 jours après d_+ (EIQ= [119 ; 260]). Le délai médian entre d_- et d_+ était de 170 jours (EIQ=[91 ; 273]). La figure 5.1 représente les données disponibles des 761 patients inclus à partir de la date de dernière sérologie négative ainsi qu'une prédiction par régression polynomiale locale pour la charge virale et le nombre de CD4. On peut remarquer que certains patients ont

une charge virale détectable au moment de leur dernière sérologie négative. Ainsi, bien qu’ayant une sérologie négative, ils étaient déjà infectés par le VIH. On peut également relever la tendance d’un pic précoce de charge virale comme il est attendu d’observer en primo-infection et démontrant la richesse du jeu de données utilisé.

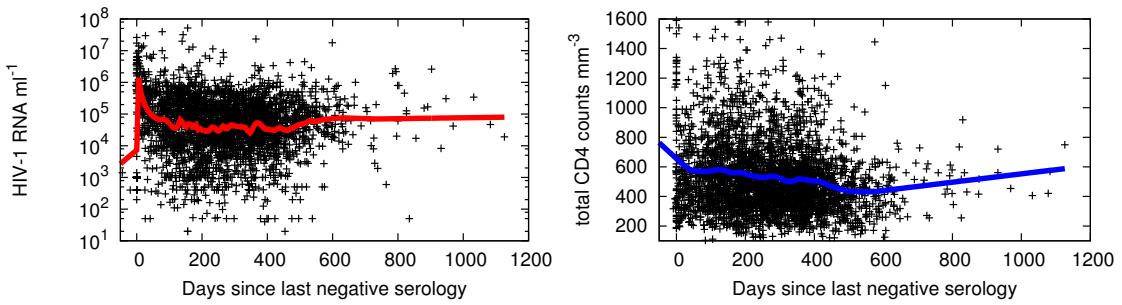


Fig. 5.1 : Données disponibles de charge virale plasmatique (gauche) et du nombre de CD4 (droite) pour les 761 patients inclus de la Collaboration CASCADE.

5.2.2 Modèle biologique

Nous proposons d’utiliser un “modèle de base” de l’interaction entre le VIH et le système immunitaire. L’un des modèles les plus utilisés dans la littérature est le modèle proposé par Perelson et al. (1996). Ce modèle comporte trois compartiments (Fig. 5.5a) : des cellules CD4 cibles T , des cellules infectées productrices T^* et des virions circulant V . Il possède 6 paramètres biologiques et a souvent été utilisé pour modéliser la réponse viro-immunologique après une mise sous traitement mais également dans le contexte de la primo-infection à VIH (Stafford et al., 2000).

Tous les paramètres ont été reparamétrisés par une transformation logarithmique pour s’assurer de leur positivité. Pour un sujet i , le vecteur de paramètres est : $\tilde{\boldsymbol{\xi}}^i = (\tilde{\lambda}^i, \tilde{\mu}_{T^*}^i, \tilde{\pi}^i, \tilde{\mu}_T^i)$. Pour des problèmes d’identifiabilité, nous avons fixé les autres paramètres à des valeurs issues de la littérature : $\mu_V = 20.0 \text{ jour}^{-1}$ (Ramratnam et al., 1999), $\mu_Q = 0.00014 \text{ jour}^{-1}$ (Mclean et Michie, 1995), $\rho = 0.017 \text{ jour}^{-1}$ (Ribeiro et al., 2002) et $\gamma = 0.00027 \text{ virions}^{-1} \text{ jour}^{-1} \mu\text{L}^{-1}$ (Wilson et al., 2007). Les compartiments observés sont le logarithme en base 10 de la charge virale (nombre de virions par μL) et la racine quatrième du nombre total de CD4 (nombre de cellules par mm^3) : $CD4_{total} = (T + T^*)^{0.25}$. Ces transformations sont couramment utilisées pour s’assurer de l’homoscédasticité des

Tab. 5.1 : Caractéristiques des 761 patients de la Collaboration CASCADE inclus dans cette étude.

Caractéristiques		N (%)
Total		761 (100)
Sexe		
	Homme	639 (84)
	Femme	122 (16)
Groupe d'exposition		
	Homosexuel	540 (71)
	Hétérosexuel	267 (22)
	Toxicomane	23(3)
	Inconnu	20(2.6)
	Hétérosexuel + Toxicomane	4 (0.5)
	Autre	4 (0.5)
	Homosexuel + Toxicomane	3 (0.4)
Coinfection VHC		
	Non	548 (71)
	Oui	91 (12)
	Inconnu	122 (16)
Séroconversion symptomatique		
	Non	86 (12)
	Oui	186 (24)
	Inconnu	489 (64)
		Médiane (EIQ*)
Age à la première sérologie positive		34 (21;40)
Délai entre dernière sérologie négative et première sérologie positive (en jours)		170 (91 ; 273)
Nombre de mesures de CD4		4 (2;5)
Nombre de mesure de charge virale plasmatique		4 (2;5)

* Ecart InterQuartile

erreurs de mesures (Thiébaut et al., 2003)

Les paramètres ont été estimés par maximum de vraisemblance prenant en compte l'incertitude sur la date d'infection (Drylewicz et al., 2009) et la censure à gauche de la charge virale plasmatique (Thiébaut et al., 2006). Pour chaque patient i , nous disposons d'une date de dernière sérologie négative d_-^i et d'une date de première sérologie positive d_+^i . Comme dans Drylewicz et al. (2009), nous supposons que la date d'infection d'un patient (τ^i) se situe dans l'intervalle $[binf^i; bsup^i]$, où $binf^i = d_-^i - 130$ et $bsup^i$ est la date la plus récente entre d_+^i et la date de première charge virale détectable. Les effets aléatoires ont été sélectionnés en utilisant le test du score (Drylewicz et al., 2009) présenté dans le chapitre 4.

5.2.3 Dynamique de la primo-infection à VIH

Principaux paramètres dynamiques

Le modèle final incluait des effets aléatoires pour les deux premiers paramètres (λ et μ_{T^*}). En échelle naturelle, la demi-vie moyenne ($\ln(2)/\text{Taux de décès}$) des cellules cibles et infectées était 17 et 0.37 jours respectivement. Au moment du pic viral (valeur médiane $5.97 \log_{10} \text{ copies/mL}$, EIQ= [5.59 ; 6.18]), le nombre médian estimé de cellules infectées était de 97 cellules/ μL (EIQ= [41 ; 159]) parmi 580 CD4 cellules/ μL (EIQ= [502 ; 645]) et la charge virale plasmatique médiane estimée était de $5.97 \log_{10} \text{copies/mL}$ (EIQ= [5.59 ; 6.18]).

Tab. 5.2 : Prédictions au moment du pic viral (en jours après l'infection) pour les cellules cibles (en cellules/ mm^3), cellules infectées (en cellules/ mm^3) et virions circulant (en \log_{10} copies/mL).

	Min	Q1	Médiane	Q3	Max
Cellules cibles (T)	55	352	470	590	1405
Cellules infectées (T^*)	0.01	41	97	159	377
CD4 total ($T + T^*$)	358	502	580	645	1405
Virions circulant (V)	2.09	5.59	5.97	6.18	6.56
Date du pic viral	8	13	16	23	500

La médiane du nombre de CD4 estimé à l'équilibre était de 470 cellules/ μL (EIQ= [378 ; 590]) ce qui était similaire au niveau observé médian (basé sur la dernière mesure disponible après 365 jours de suivi chez 430 patients) : 470 (EIQ= [371 ; 615]). Au même moment, la médiane du nombre de cellules infectées estimé était de 7 (EIQ= [4 ; 12]). Les prédictions médianes de l'équilibre viral étaient aussi proche de celles observées chez 403 patients : $4.86 \log_{10} \text{ copies/mL}$ (EIQ= [4.57 ; 5.05]) vs. 4.59 (EIQ= [4.04 ; 5.00]). Les tableaux 5.2 et 5.3 présentent la distribution des CD4 totaux et de la charge virale plasmatique pour les 761 patients inclus au moment du pic viral et du setpoint.

De plus, nous avons étudié la corrélation entre les paramètres dynamiques tel que la charge virale à l'équilibre ou la date du pic viral en fonction des paramètres estimés. Ainsi, nous avons trouvé que 62% des variations du setpoint viral étaient expliquées par les variations du taux de décès des cellules infectées (μ_{T^*}). Ce paramètre expliquait également 98% du setpoint des CD4 totaux et 52% du niveau de la charge virale au moment du pic.

Tab. 5.3 : Valeurs prédictes et observées des CD4 totaux (en cellules/mm³) et de la charge virale plasmatique (en log₁₀copies/mL) au moment de l'équilibre.

	Min	Q1	Médiane	Q3	Max
Nombre prédict de cellules cibles (T)	108	367	463	587	1413
Nombre prédict de cellules infectées(T^*)	0.01	3.82	7.46	11.52	44.74
Nombre de CD4 total prédict ($T + T^*$)	152	378	470	590	1415
Nombre de CD4 observé	120	371	470	615	1211
Nombre de virions prédict (V)	2.09	4.57	4.86	5.05	5.63
Charge virale plasmatique observée	1.69	4.04	4.59	5.00	6.16

Nous avons trouvé une corrélation positive entre le niveau de charge virale plasmatique au moment du pic viral et le setpoint viral ($r=0.98$) : plus un patient à un pic de charge virale élevé plus son setpoint viral sera élevé. Ce résultat est similaire à ceux publiés précédemment (Kaufmann et al., 1999; Lindback et al., 2000).

Pour chaque patient, nous avons calculé une distribution *a posteriori* de la date d'infection en fonction de la fenêtre d'infection possible (Fig. 5.2). Les distributions observées étaient très variées et dépendent de la taille de la fenêtre mais également du nombre de mesures disponibles pour chaque patient. Plus un patient a une fenêtre courte et des mesures répétées proches du pic viral plus sa distribution *a posteriori* sera précise. Nous avons également représenté les trajectoires prédictes de charge virale plasmatique et du nombre de CD4 à partir de la date la plus probable d'infection pour chaque patient. Les trajectoires obtenues sont comparables à celles de la littérature (Kaufmann et al., 1998 ; Lindback et al., 2000 ; Stafford et al., 2000) malgré des oscillations marquées. Des trajectoires individuelles et les valeurs observées sont présentées dans la figure 5.3.

Prédiction du temps de progression clinique

Nous avons fait une étude supplémentaire pour voir si les valeurs individuelles prédictes d'équilibre étaient associées au temps de progression clinique (apparition d'une infection opportuniste, décès ou mise sous traitement antirétroviral). Nous avons utilisé un modèle de Cox pour décrire la progression clinique en fonction de la dynamique de primo-infection : paramètres biologiques (λ et μ_{T^*}), setpoint prédict et observé.

Parmi les 761 patients inclus, 719 ont commencé un traitement antirétroviral, 2 ont été classés stade SIDA avant la mise sous traitement et 1 est décédé après un délai médian

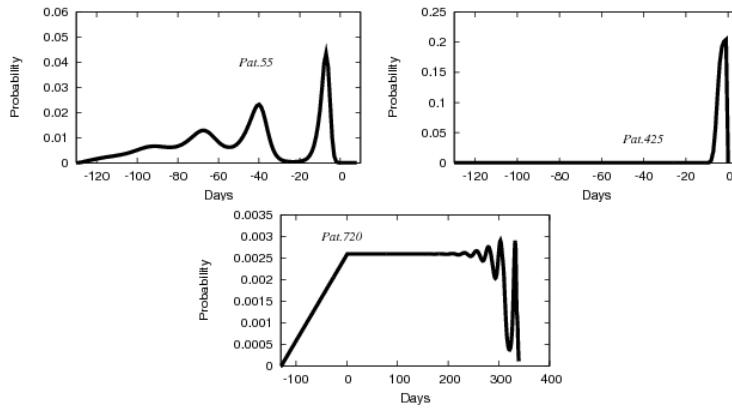


Fig. 5.2 : Distributions *a posteriori* de la date d'infection pour trois patients, où le jour 0 est la date de dernière sérologie négative.

Tab. 5.4 : Evènements apparaissant durant le suivi des 761 patients

Caractéristiques		N
Total		761
Progression clinique	SIDA	102
	Décès avant SIDA	6
	Total	108
Mise sous traitement (ART)	Yes	719
	No	42
Evènements	ART + SIDA + Décès	54
	ART + SIDA + Vivant	46
	ART + Décès + pas SIDA	5
	ART + pas SIDA + Vivant	614
	pas ART + SIDA + Décès	2
	pas ART + Décès + pas SIDA	1
	pas ART + pas SIDA + Vivant	39
		Médiane (EIQ*)
Délai (en mois) entre	ART et SIDA	12 (2-26)
	ART et Décès	32 (19-53)
	SIDA et Décès	20 (7-32)
Délai (en années) entre	Sérologie négative et ART	1.9 (1.2-3.5)
	Sérologie négative et SIDA	7.9 (5.9-8.7)
	Sérologie négative et Décès	5.7 (3.9-7.5)
	Sérologie négative et 1er évènement	1.8 (1.2-3.2)

*Ecart InterQuartile

entre la dernière sérologie négative et le premier événement de 1.8 ans (EIQ= [1.2 ; 3.3]). Le tableau 5.4 présente les évènements survenus dans l'échantillon d'étude. Nous avons trouvé un risque de progression clinique plus élevé pour les patients ayant un plus haut niveau

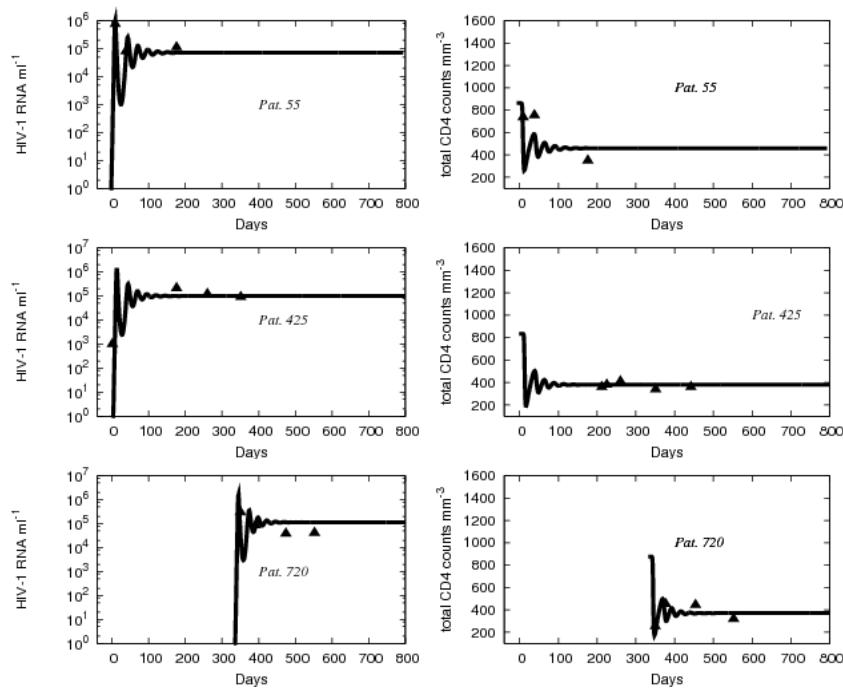


Fig. 5.3 : Trajectoires individuelles prédictes à partir de la date d'infection la plus probable et valeurs observées de charge virale plasmatique et du nombre de CD4 pour trois patients, où le jour 0 est la date de dernière sérologie négative.

de setpoint viral (Hazard Ratio [HR]= 2.33 pour 1 \log_{10} copies/mL, $p<.001$) et un niveau plus bas de CD4 à l'équilibre (HR= 0.82 pour 100 cells/ μL , $p<.001$). De plus, les setpoints estimés (HR= 0.76 pour 100 CD4 de plus; HR=2.78 pour 1 \log_{10} de plus) et le taux de décès des cellules infectées individuel (HR=0.88 pour 0.1 unité de plus, $p<.0001$) étaient de meilleurs prédicteurs de la progression clinique que les setpoints observés (définis par la dernière mesure observée durant la première année de suivi) : AIC=7996 pour le taux de décès des cellules infectées vs. 8220 pour le setpoint viral observé.

Estimation du R_0

L'estimation des paramètres de modèles dynamiques permettent également de fournir une estimation du nombre de reproduction de base R_0 qui correspond au nombre moyen d'infections secondaires causé par l'introduction d'une cellule infectée dans une population de cellules cibles supposée infinie. Le nombre de nouvelles infections dues à une seule cellule infectée est obtenu en multipliant le nombre de virions produit durant la vie d'une cellule infectée par le nombre de cellules infectées par chaque virion durant sa vie (Bonhoeffer

et al., 1997).

Pour le “modèle de base”, le R_0 médian estimé était de 1.45 (EIQ= [1.43 ; 2.30]). Ces valeurs sont plus faibles que celles précédemment publiées (Little et al., 1999; Stafford et al., 2000). Les modèles utilisés dans ces deux études étaient similaires au nôtre mais les hypothèses biologiques faites et leur échantillon d’étude étaient un peu différents. Ils n’incluaient que des patients ayant eu une primo-infection symptomatique et ayant des données très répétées de charge virale plasmatique. La date d’infection était estimée à partir de la date de début des symptômes ou d’une date d’exposition au VIH. Ils ont utilisé le “modèle de base” pour étudier la dynamique de la charge virale plasmatique uniquement.

Stafford et al. (2000) utilisent la même méthode que la nôtre pour estimer R_0 mais ils supposent que les cellules cibles T sont uniquement des cellules activées ; le nombre de cellules activées est faible dans le sang. De plus, ils supposent que ce nombre est constant au cours de l’infection ($T_0 = 10$ cellules/mm³) pour simplifier le modèle. Cette hypothèse a un impact majeur sur l’estimation des paramètres et sur celle du R_0 . Ainsi, le R_0 médian pour 10 patients était de 5.7 (min=2.8, max=11.0).

Little et al. (1999) modélisent la dynamique virale à l’aide d’une simplification du “modèle de base”. Les cellules infectées sont supposées avoir une distribution exponentielle dépendant du taux de croissance virale noté r_0 : $I(t) = I_0 \exp(r_0 t)$. Ce taux de croissance est alors supposé égal à $\beta T_0 - \delta_I$ où β est le taux d’infectiosité du virus, δ_I le taux de décès des cellules infectées et T_0 le nombre de CD4 avant infection. Le paramètre δ_I est estimé à partir des données de mise sous traitement et r_0 et R_0 sont déterminés à partir de \hat{r} (le taux de croissance observé de la charge virale), δ_I , T_0 et \hat{T} (le nombre de CD4 observé proche du temps du pic viral) par les formules suivantes :

$$r_0 = \frac{T_0}{\hat{T}}(\hat{r} + \delta_I) - \delta_I \quad \text{et} \quad R_0 = \frac{T_0}{\hat{T}}\left(1 + \frac{\hat{r}}{\delta_I}\right),$$

La valeur de T_0 a été fixée à différentes valeurs : 500, 1000 et 1500 cellules/mm³. Ils estimaient un R_0 moyen à 7.1 (min=5.2, max=9.1).

La figure 5.4 représente la distribution de R_0 dans la population obtenue à partir de nos estimations ; les valeurs estimées par les deux études présentées peuvent être obtenues à partir de nos estimations. Utilisant les données de Little et al. (1999), nous avons calculé

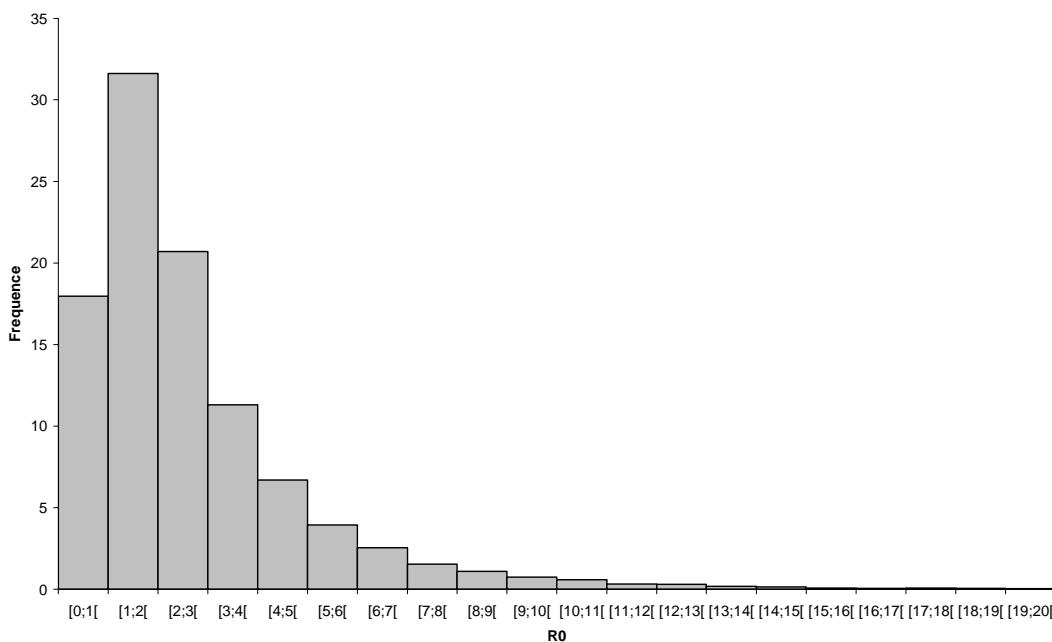


Fig. 5.4 : Distribution estimée de R_0 obtenue à partir de nos estimations.

pour leurs patients, les valeurs des effets aléatoires individuels d'après les estimations de notre modèle en utilisant le mode *a posteriori* de la densité des effets aléatoires sachant les données (Guedj et al., 2007; Drylewicz et al., 2009). Nous avons défini par chaque patient un fenêtre d'infection possible. Nous avons supposé que la date de première sérologie positive était la date de première mesure de charge virale détectable et la date de dernière sérologie négative a été fixée à 30 jours avant la date de première sérologie positive. Le tableau 5.5 présente les estimations avec les deux méthodes.

Tab. 5.5 : Estimation de R_0 pour les 4 patients issus de Little et al., 1999.

Patient	Notre méthode	Little et al. (1999)
059	2.1	5.2
004	2.1	9.1
045	2.0	5.9
056	1.8	8.1
Moyenne	2.0	7.1

Les estimations du R_0 obtenues pour ces quatre patients en utilisant notre méthode sont très similaires à celles obtenues sur notre échantillon de 761 patients. Les différences observées avec les résultats de Little et al. (1999) ne semblent pas conséquentes.

dues à la sélection des patients mais bien à la méthode de calcul du R_0 .

Estimer R_0 lors de la primo-infection peut permettre de déterminer l'efficacité minimale nécessaire pour une intervention de prévention de l'infection. Une telle intervention peut être un vaccin, une circoncision masculine (Auvert et al., 2005) ou une prophylaxie pre/post-exposition (Grant et al., 2005). En effet, l'infection se maintient dans l'organisme si et seulement si R_0 est supérieur à 1. Il est donc possible d'estimer l'efficacité nécessaire pour prévenir une infection à partir des estimations obtenues sur CASCADE. L'infectiosité du virus (γ) doit diminuer de $(1 - 1/R_0)\%$, soit 46% (EIQ= [30 ;57]) d'après nos estimations de paramètres.

Effet de variables explicatives

Nous avons utilisé les tests du score présentés dans le chapitre 4 (Drylewicz et al., 2009) pour tester l'effet possible de variables explicatives sur chaque paramètres. Nous avons testé un modèle incluant l'une des variables explicatives suivantes versus un modèle sans variable explicatif (modèle estimé sous l'hypothèse nulle). Nous nous sommes intéressés aux variables : *sex*, *séroconversion symptomatique*, *groupe d'exposition* (toxicomane vs. autres) et *âge à la première serologie positive*. Nous n'avons trouvé aucun effet des variables *sex*, *séroconversion symptomatique*, *groupe d'exposition* et *âge* pris en variable continue. Le test était significatif quand nous avons considéré *âge* comme une variable binaire codée 1 pour >50 ans et 0 pour les plus jeunes (50 ans étant le 95ième percentile) : p=0.04 pour le taux de décès des cellules infectées (μ_{T^*}). Nous avons ensuite estimé le modèle incluant cette variable explicative et trouvé un taux de décès des cellules infectées plus élevé pour les patients les plus jeunes : 1.79 vs. 0.80 jours⁻¹. Nous avons également trouvé un effet significatif de la variable *coinfection VHC* pour le paramètre λ (taux de production des cellules cibles) : p=0.003. Ce résultat suggère que les patients coinfestés par le VHC ont un taux de production de cellules cibles plus bas que les autres. Nous n'avons pas inclus cette variable dans notre modèle car nous n'avons aucune information sur la date d'infection par le VHC et nous ne pouvons pas savoir si les patients étaient déjà infectés au moment de leur primo-infection à VIH.

De plus, une tendance en fonction de la période d'infection sur la dynamique a été observée (Tab. 5.6). Le nombre de reproduction de base R_0 augmente, suggérant un chan-

gement dans la pathogénicité du virus ou un recrutement de patients plus récemment infectés. De même, nous avons observé un setpoint pour les CD4 plus bas dans les périodes les plus récentes (2001-2006) et un setpoint viral plus élevé. Ces résultats sont en accord avec ceux d'autres études (Vanhems et al., 1999; Dorrucci et al., 2007; Crum-Cianflone et al., 2009).

Tab. 5.6 : Charge virale et nombre de CD4 médian (EIQ) observé et prédit au moment de l'équilibre et estimation du R_0 suivant la période d'infection.

	1985-1995 (N=131)	1996-2000 (N=316)	2001-2006 (N=314)	1985-2006 (N=761)
R_0	1.6 (1.4 ; 1.8)	1.7 (1.5 ; 2.2)	2.0 (1.7 ; 2.7)	1.8 (1.5 ; 2.3)
Viral setpoint observé	4.5 (4.1 ; 4.9)	4.5 (3.9 ; 4.9)	4.7 (4.1 ; 5.1)	4.6 (5.0 ; 5.0)
Viral setpoint prédit	4.6 (4.4 ; 4.8)	4.7 (7.5 ; 4.9)	4.9 (4.7 ; 5.1)	4.8 (4.6 ; 5.0)
Setpoint observé des CD4	585 (430 ; 779)	493 (390 ; 614)	424 (314 ; 516)	471 (371 ; 615)
Setpoint prédict des CD4	629 (499 ; 798)	517 (414 ; 430)	440 (325 ; 551)	500 (398 ; 627)

Limites du modèle de base

Cette section illustre ce qui est faisable et le type de questions auxquelles on peut répondre avec une telle approche. Cependant, le “modèle de base” donnait des résultats inattendus : une date de pic viral assez précoce (16 jours après l'infection), une estimation du R_0 faible et des trajectoires très oscillantes. D'autre part, d'après la distribution estimée de R_0 , 17% des patients présentent un $R_0 < 1$. Ces patients pourraient représenter ceux que l'on appelle des “elite controllers” (patients avec une charge virale chutant à des valeurs indétectable sans traitement antirétroviral). Cependant, le nombre de ce type de patients est supposé être très faible, environ 0.36% dans la French Hospital Database HIV (Grabar et al., 2009). Enfin, la simplicité de ce modèle peut être discutée aux regards de plusieurs mécanismes biologiques importants soulignés dans la littérature et détaillés dans la section suivante.

5.3 Comparaison de modèles biologiques

5.3.1 Modèles biologiques

Le “modèle de base” fait l’hypothèse qu’une cellule, une fois infectée, devient immédiatement capable de produire des virions, alors qu’il a été montré *in vitro* qu’un délai d’environ un jour était nécessaire pour qu’une cellule infectée devienne productrice (Kiernan et al., 1990; Barbosa et al., 1994). Ce délai peut avoir un impact important sur l’estimation de R_0 (Lloyd, 2001). Pour prendre en compte ce délai, nous introduisons un compartiment supplémentaire au “modèle de base” en distinguant les cellules infectées non-productrices I des cellules infectées productrices P (Fig. 5.5b) comme précédemment proposé par Nowak et al. (1997).

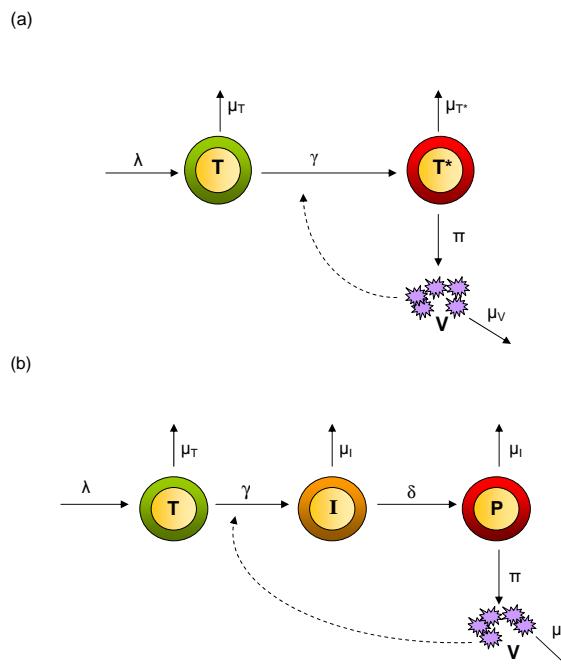


Fig. 5.5 : “Modèle de base” (a) et “modèle cellules productrices” (b). Les cellules cibles T sont produites avec un taux λ , sont infectées à un taux γ et meurent à un taux μ_T . Les virions V sont produits à un taux π par les cellules infectées productrices T^* (ou P pour le modèle (b)) et meurent à un taux μ_V . Les cellules infectées deviennent productrices à un taux δ et meurent à un taux μ_{T^*} (μ_I pour le modèle (b)).

Ces deux modèles font l’hypothèse que toutes les cellules CD4+ sont des cellules cibles pour le virus. Hors, il a été montré que le VIH infecte préférentiellement les cellules CD4 qui prolifèrent ; cellules dites “activées” (Grossman et al., 2002). Nous proposons donc un troisième modèle (“modèle cellules activées”, Fig. 5.6a) qui distingue des cellules

quiescentes (Q) des cellules activées (A). Nous supposons qu'il existe une activation de base α (présente avant l'infection) et une activation dépendante de la concentration en virus α' .

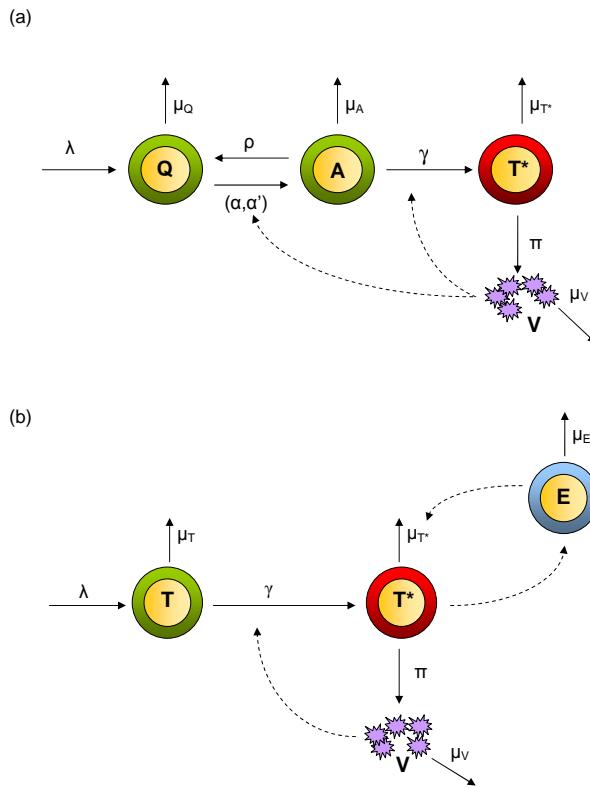


Fig. 5.6 : (a) “Modèle cellules activées” : les cellules quiescentes Q sont produites à un taux λ , sont activées avec un taux de base α et dépendant de la charge virale α' et meurent à un taux μ_Q . Les cellules activées A retournent à l'état quiescent avec un taux ρ , sont infectées avec un taux γ et meurent à un taux μ_A . Les cellules infectées T^* produisent π nouveaux virions par jour et meurent à un taux μ_{T^*} . Les virions V meurent à un taux μ_V . (b) “Modèle cellules effectrices” : Les cellules cibles T sont produites avec un taux λ , sont infectées à un taux γ et meurent à un taux μ_T . Les virions V sont produits à un taux π par les cellules infectées productrices T^* et meurent à un taux μ_V . Les cellules effectrices E sont produites proportionnellement au nombre de cellules infectées, modifient leur taux de décès et meurent à un taux μ_E .

Dans les trois premiers modèles, nous ne prenons pas en compte la réponse immunitaire spécifique au VIH. Le déclin de la charge virale est alors due à la réduction du nombre de cellules cibles (Phillips, 1996). Cependant, les lymphocytes T cytotoxiques (CTL), responsables de la destruction des cellules infectées peuvent jouer un rôle important dans cette dynamique (Koup et al., 1994; Musey et al., 1997). Plusieurs extensions du “modèle de base” ont ainsi été suggérées (Wodarz et al., 1998; Stafford et al., 2000; Wodarz et

Nowak, 2000; Burg et al., 2009). Cette réponse peut être modélisée de différentes manières : en faisant dépendre du temps le taux de décès des cellules infectées ou en introduisant un nouveau compartiment de cellules effectrices. Comme aucune mesure de cellules effectrices n'était disponible, nous avons choisi d'utiliser le modèle proposé par Burg et al. (2009) (Fig. 5.6b). Utilisant l'hypothèse de quasi-équilibre de cellules effectrices E , nous avons un terme supplémentaire pour le compartiment des cellules infectées. Celles-ci meurent à un taux μ_{T^*} et à un taux du à la réponse spécifique au VIH : $\kappa T^*/(T^* + \theta)$; le terme θ étant un terme de saturation. Toutes ces hypothèses sont complémentaires et il semble alors assez naturel de les combiner dans un “modèle complet” faisant ainsi la distinction entre les cellules quiescentes Q , les cellules activées A (cibles du virus), les cellules infectées non-productrices I , les cellules infectées productrices P et incluant une réponse spécifique au VIH (Fig. 5.7).

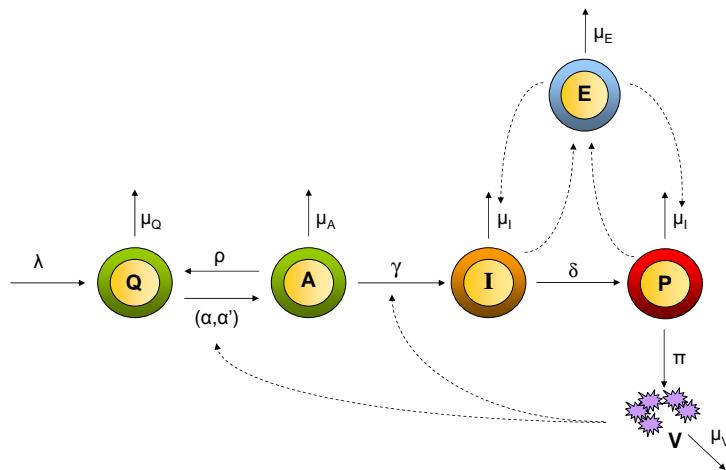


Fig. 5.7 : “Modèle complet” : les cellules quiescentes Q sont produites à un taux λ , sont activées avec un taux de base α et dépendant de la charge virale α' et meurent à un taux μ_Q . Les cellules activées A retournent à l'état quiescent avec un taux ρ , sont infectées avec un taux γ et meurent à un taux μ_A . Les cellules infectées I deviennent productrices à un taux δ et produisent alors π nouveaux virions par jour et meurent à un taux μ_I . Les virions V meurent à un taux μ_V . Les cellules effectrices E sont produites proportionnellement au nombre de cellules infectées (I et P) et modifient leur taux de décès. Les cellules E meurent à un taux μ_E .

5.3.2 Modèles statistiques

Tous les paramètres ont été reparamétrisés par une transformation logarithmique pour s’assurer de leur positivité. Pour un individu i , les vecteurs de paramètres sont $\boldsymbol{\xi}^i = (\tilde{\lambda}^i, \tilde{\mu}_I^i, \tilde{\pi}^i, \tilde{\mu}_T^i)$ pour le “modèle cellules productrices”, $\tilde{\boldsymbol{\xi}}^i = (\tilde{\mu}_{T^*}^i, \tilde{\alpha}^i, \tilde{\lambda}^i, \tilde{\alpha}', \tilde{\pi}^i, \tilde{\mu}_T^i)$ pour le “modèle cellules activées”, $\tilde{\boldsymbol{\xi}}^i = (\tilde{\lambda}^i, \tilde{\mu}_{T^*}^i, \tilde{\pi}^i, \tilde{\mu}_A^i, \tilde{\kappa}^i)$ pour le “modèle cellules effectrices” et $\tilde{\boldsymbol{\xi}}^i = (\tilde{\mu}_I^i, \tilde{\alpha}^i, \tilde{\lambda}^i, \tilde{\alpha}', \tilde{\pi}^i, \tilde{\mu}_A^i, \tilde{\kappa}^i)$ pour le “modèle complet”.

Pour des problèmes d’identifiabilité, nous avons fixé les autres paramètres à des valeurs issues de la littérature : $\mu_V = 20.0 \text{ jour}^{-1}$ (Ramatnam et al., 1999), $\mu_Q = 0.00014 \text{ jour}^{-1}$ (McLean et Michie, 1995), $\rho = 0.017 \text{ jour}^{-1}$ (Ribeiro et al., 2002) et $\gamma = 0.00027 \text{ virions}^{-1} \text{ jour}^{-1} \mu\text{L}^{-1}$ (Wilson et al., 2007) pour les modèles “cellules productrices” et “cellules effectrices” et $\gamma = 0.05 \text{ virions}^{-1} \text{ jour}^{-1} \mu\text{L}^{-1}$ pour le “modèle cellules activées” et le “modèle complet” (informations supplémentaires en section 7). Le paramètre δ est fixé à 1 jour^{-1} , le temps entre l’infection d’une cellule et sa capacité à produire des virions étant estimé à environ 1 jour (Kiernan et al., 1990; Barbosa et al., 1994); le paramètre θ est fixé à 2 (Burg et al., 2009). L’inoculum initial de virions est fixé à $10^{-6} \text{ virions/mm}^3$ (Ciupe et al., 2006; Stafford et al., 2000) ce qui correspond à 5 virions pour 5 litres de sang.

Les compartiments observés sont le logarithme en base 10 de la charge virale (nombre de virions par μL) et la racine quatrième du nombre total de CD4 (nombre de cellules par mm^3) : pour le “modèle cellules effectrices” $CD4_{total} = (T + T^*)^{0.25}$, pour le “modèle cellules productrices” $CD4_{total} = (T + I + P)^{0.25}$, pour le “modèle cellules activées” $CD4_{total} = (Q + A + T^*)^{0.25}$ et pour le “modèle complet” $CD4_{total} = (Q + A + I + P)^{0.25}$. Les paramètres ont été estimés avec la même méthode que celle décrite dans la section précédente.

5.3.3 Dynamique de la primo-infection

Le modèle final incluait des effets aléatoires pour les deux premiers paramètres du “modèle cellules productrices” (λ et μ_I) et du “modèle cellules effectrices” (λ et μ_{T^*}) et pour les trois premiers paramètres pour le “modèle cellules activées” (λ , μ_{T^*} et α) et le “modèle complet” (λ , μ_I et α).

Le meilleur modèle était le “modèle complet” (plus petit AIC). Le tableau 5.7 résume les dynamiques obtenues suivant le modèle. La demi-vie moyenne des cellules cibles ($\ln(2)/\text{Taux de décès}$) était similaire pour le “modèle de base” et le “modèle cellules pro-

ductrices” (18 et 21 jours). Elle était plus faible pour le “modèle cellules activées” et le “modèle complet” (13 et 4 jours) et plus élevée pour le “modèle cellules effectrices” (36 jours). La demi-vie des cellules infectées était similaire pour les deux premiers modèles (0.37 et 0.40 jours) et plus faible pour les autres (1.45, 1.48 et 1.85 jours).

Tab. 5.7 : Estimations des demi-vies et de la dynamique pour les différents modèles.

Paramètre	Modèle de base	Modèle cellules productrices	Modèle cellules activées	Modèle cellules effectrices	Modèle complet
CD4 totaux avant infection					
(en cellules/mm ³)	857	837	692	1189	540
Demi-vie moyenne des cellules cibles (en jours)					
	18.06	20.69	13.07	36.15	17.45
Demi-vie moyenne des cellules infectées (en jours)					
	0.37	0.40	1.45	1.58	1.85
Temps moyen du pic viral (en jours)					
	16	38	2	20	12
Charge virale moyenne au pic viral (en log₁₀ copies/mL)					
	6.03	5.65	5.48	6.10	5.07
CD4 total au pic viral					
(en cellules/mm ³)	457	480	659	818	542
Nombre total de cellules infectées au pic viral (en cellules/mm³)					
	110	55	17	368	13
Charge virale moyenne à l'équilibre (en log₁₀ copies/mL)					
	4.91	4.75	4.44	4.96	4.64
CD4 total à l'équilibre					
(en cellules/mm ³)	434	483	467	404	464
Nombre total de cellules infectées à l'équilibre (en cellules/mm³)					
	8	7	2	27	5
<i>R</i> ₀	2.01	1.76	25.37	4.21	11.97
AIC	15010	14300	14157	14117	13983

Nous pouvons déterminer la dynamique moyenne de chaque compartiment inclus dans les modèles et pas seulement ceux observés. Ainsi, nous pouvons calculer le nombre total de CD4 avant infection, la dynamique des cellules infectées, les états d'équilibre de chaque compartiment,... Le nombre prédit de CD4 avant infection était proche des connaissances actuelles (Bofill et al., 1992; Maini et al., 1996) : entre 540 et 1189 cellules/mm³. De plus, les modèles “cellules activées” et “complet” prédisaient un nombre de cellules activées (parmi les CD4 totaux) similaire à ceux de la littérature (Sachsenberg et al., 1998) : 6 et 2%. La distinction de cellules infectées non-productrices et productrices a comme principale conséquence de retarder le pic viral (38 jours après l'infection) et ainsi d'être

plus en accord avec les études publiées (Clark et al., 1991; Daar et al., 1991; Kaufmann et al., 1998; Kahn et Walker, 1998; Little et al., 1999). En effet, ils l'observaient deux à trois semaines après le début des symptômes. Lorsque nous avons inclus une réponse immunitaire spécifique au VIH dans le modèle, nous avons également observé un pic viral plus tardif (Fig. 5.8). Le nombre de cellules infectées au moment du pic viral est élevé pour les modèles “de base”, “cellules productrices” et “cellules effectrices” (entre 11 et 45% des CD4 totaux) et diminue à un nombre plus faible à l’équilibre (entre 2 et 7%). Pour les deux autres modèles, ce nombre semble constant (2 et 1%).

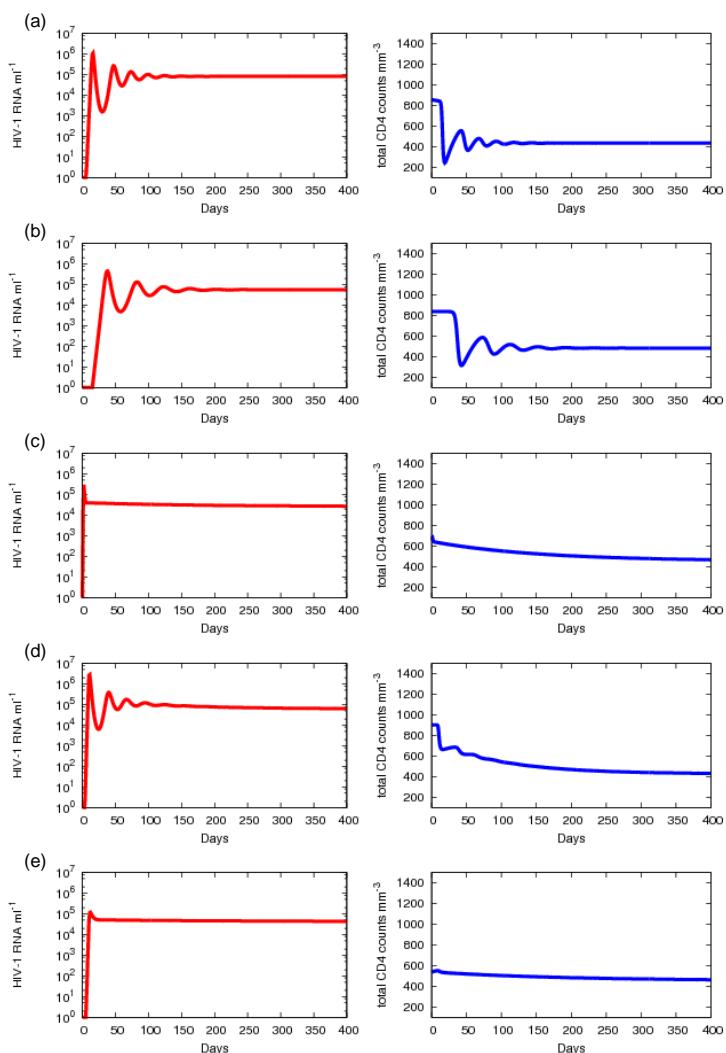


Fig. 5.8 : Trajectoires moyennes de charge virale plasmatique et des CD4 totaux suivant le modèle : (a) “modèle de base”, (b) “modèle cellules productrices”, (c) “modèle cellules activées”, (d) “modèle cellules effectrices” et (e) “modèle complet”.

5.4 Discussion

Dans ce chapitre, nous avons étudié la dynamique de la primo-infection à VIH en utilisant des modèles biomathématiques. Le “modèle de base”, qui est souvent utilisé, n'est pas le meilleur modèle pour décrire simultanément la dynamique virale et celle des CD4 durant cette phase précoce de l'infection. Nous avons modifié ce modèle pour prendre en compte plus d'hypothèses biologiques de l'interaction du VIH et du système immunitaire : seulement une partie des CD4 est cible du virus, présence d'une réponse immunitaire spécifique au VIH et un délai entre l'infection de cellules et la production de nouveaux virions. Cependant, aucun modèle n'était satisfaisant. En effet, les trajectoires prédites ont soit trop d'oscillations (le “modèle de base” par exemple) ou aucune (le “modèle complet”) et ces modèles ne peuvent pas expliquer les dynamiques à long terme : tous prédisent un état d'équilibre stable pour la charge virale et les CD4 survenant dans la première année de l'infection. De plus, les valeurs estimées de CD4 avant infection du “modèle complet” sont trop faibles comparées aux connaissances actuelles, de même que la différence de niveau de la charge virale entre le moment du pic et celui du setpoint : nous savons que la charge virale plasmatique décroît après ce pic (Mellors et al., 1995; Kaufmann et al., 1998; Schacker et al., 1998; Sterling et al., 2001). Nous connaissons également un ordre de grandeur de la demi-vie des cellules productrices : entre 0.6 et 1.4 jours (Markowitz et al., 2003) ; estimations obtenues à partir d'études à la mise sous traitement. Les estimations du “modèle de base” et du “modèle cellules productrices” semblent ainsi trop faibles.

Nous avons ainsi montré comment cette approche pourrait permettre de répondre à des questions cliniques/biologiques telles qu'avoir une idée de la dynamique précoce en regardant l'effet mécanistique de facteurs pronostiques connus et le lien entre paramètres biologiques et progression clinique.

Cependant, l'estimation peut être améliorée en prenant en compte certains *a priori* sur les paramètres biologiques en utilisant la méthode de Maximum A Posteriori présenté dans le chapitre 2. Par exemple, nous pourrions utiliser un prior sur la valeur des CD4 avant infection pour obtenir des valeurs prédites plus plausibles. Une autre source d'information possible serait d'utiliser des données à la mise sous traitement. Dans la base de données CASCADE, nous disposons du suivi des patients inclus dans notre étude et certains d'entre

eux commencent un traitement antirétroviral peu de temps après la primo-infection. Nous pourrions utiliser cette information pour étendre le modèle aux dynamiques à la mise sous traitement.

Chapitre 6

Conclusion générale et perspectives

Dans cette thèse nous nous sommes principalement intéressés à l'étude de la dynamique de la primo-infection à VIH. Cette étude est complexe pour plusieurs raisons : (i) elle est asymptomatique chez la majorité des sujets donc difficilement détectable et (ii) la date d'infection est inconnue. D'autre part, les dynamiques viro-immunologiques sont rapides et complexes et nécessitent des données très riches afin de pouvoir les observer. Notre étude se base sur une approche mécanistique de l'interaction VIH-système immunitaire. Les modèles biologiques proposés sont représentés par des systèmes d'équations différentielles ordinaires qui peuvent capturer la nature complexe des interactions qui existent entre les compartiments.

Nous avons développé une méthode d'estimation basée sur un algorithme de maximisation de la vraisemblance permettant à chaque patient de contribuer à l'estimation quelque soit son nombre de mesures ; nos critères de sélection de patients sont donc plus larges que les études publiées jusqu'à présent (Kaufmann et al., 1998; Little et al., 1999; Stafford et al., 2000; Ciupe et al., 2006). De plus, notre méthode prend en compte l'incertitude sur la date d'infection utilisant les dates de dernière sérologie négative (d_-^i) et première sérologie positive (d_+^i) ; nous n'imputons pas de valeurs à l'origine des temps du modèle (i.e. la date d'infection). Nous avons défini une distribution *a priori* pour la date d'infection en nous basant sur les travaux de Fiebig et al. (2003). Nous avons supposé que la probabilité d'avoir été infecté entre d_-^i -130 jours et d_-^i est linéaire croissante puis constante entre d_-^i et d_+^i . Nous avons également testé d'autres distributions (par exemple exponentielle) et nos résultats étaient stables. Cependant, la borne inférieure de notre intervalle pourrait

être améliorée en utilisant la dynamique des anticorps ou une information sur la date de début des symptômes ou une date d'exposition au VIH par exemple.

Nous avons proposé cinq modèles dynamiques de primo-infection faisant différentes hypothèses sur l'interaction entre le VIH et le système immunitaire. D'après nos résultats, le meilleur modèle est celui distinguant des cellules quiescentes des cellules activées, mais également des cellules infectées non-productrices des cellules infectées productrices et incluant une réponse immunitaire spécifique au VIH. Cependant, les paramètres estimés ne sont pas en accord avec les connaissances biologiques actuelles. Prendre en compte ces connaissances semble être la prochaine étape pour améliorer notre méthode d'estimation. Ainsi, nous pourrions estimer certains paramètres qui étaient fixés pour des questions d'identifiabilité. En fait, il est très difficile de trouver des données plus riches en primo-infection.

Il s'agit à notre connaissance de la première étude de la dynamique de la primo-infection à VIH sur un échantillon de grande taille (761 patients) et utilisant toutes les informations disponibles (mesures de charge virale et du nombre de CD4). Comme notre approche est basée sur un modèle mécanistique, les paramètres du modèle devraient avoir une signification biologique. Nous pouvons estimer les demi-vies des cellules et prédire le temps du pic viral ainsi que les niveaux d'équilibre. A partir des paramètres estimés, nous pouvons également calculer le nombre de reproduction de base R_0 qui reflète la capacité du virus à se répandre et à se maintenir dans l'organisme. De manière surprenante, nous trouvons des valeurs inférieures à celles précédemment publiées en utilisant les mêmes modèles (Little et al., 1999; Stafford et al., 2000). Ces différences semblent être dues aux hypothèses biologiques faites dans les modèles et aux méthodes de calcul utilisées par les auteurs (nombre constant de cellules cibles, valeurs de paramètres fixés,...). Avoir une estimation du R_0 permet de déterminer l'efficacité nécessaire d'un traitement pour éradiquer l'infection. Cependant, cette estimation n'est qu'une borne inférieure car notre modèle ne prend pas en compte la constitution d'un réservoir de cellules latentes infectées ni les phénomènes de mutations (éventuellement transmises) : le réservoir de cellules latentes étant constitué durant la phase de primo-infection (Chun et al., 1998; Finzi et al., 1999). Même avec un niveau de charge virale indétectable, on peut observer une nouvelle propagation du virus due à ces cellules latentes infectées.

Bien que nos modèles semblent s'ajuster aux données, ils sont une représentation imparfaite de l'interaction entre le VIH et le système immunitaire. Cette interaction ne peut être entièrement modélisée à cause de sa complexité et du fait que les phénomènes impliqués ne sont pour l'heure actuelle pas encore tous connus. De plus, nous n'observons pas tous les compartiments inclus dans le modèle biologique (par exemple les cellules activées ou les cellules infectées) et nous devons fixer certains paramètres à des valeurs issues de la littérature (taux d'infectiosité du virus, clairance virale,...). Des mesures plus variées telles que les lymphocytes cytotoxiques ou encore des marqueurs d'activation (HLA-DR, CD38,...) permettraient d'améliorer nos estimations.

Enfin, l'étude de la primo-infection à VIH pourrait être complétée en considérant l'effet de certains facteurs génétiques tels que le groupe HLA, le sous-type viral ou le tropisme du virus. En effet, un lien entre la progression clinique et le groupe HLA des patients a été mis en évidence (Altfeld et al., 2003; Fellay et al., 2007). Les patients exprimant HLA-B57 semble avoir moins souvent une primo-infection symptomatique et ont un meilleur contrôle du virus sans la mise en place de traitement antirétroviral.

L'estimation de modèles dynamiques peut être complexe et longue. Nous avons donc développé des tests du score pour comparer différent modèles. Cet outil permet de sélectionner les effets aléatoires et les variables explicatives sans estimer tous les modèles possibles ; chose que nous devons faire si on utilise un test de rapport de vraisemblance ou un test de Wald. Cependant, pour le test du score d'homogénéité, nous utilisons un bootstrap paramétrique pour éviter le calcul de la matrice d'information et le nombre de répliques doit être choisi pour chaque modèle. Pour le modèle de base proposé par Perelson et al. (1996), la simulation de 10 jeux de données semble suffisante pour avoir de bonnes propriétés statistiques. Mais pour des modèles plus complexes comme ceux proposés dans le chapitre précédent, un plus grand nombre de répliques est nécessaire (entre 100 et 500). La multiplicité des tests doit également être prise en compte dans une telle approche de sélection de modèles. En effet, les tests du score proposés ne prennent pas en compte la corrélation entre les tests et les p-values devraient être corrigées (Liquet et Commenges, 2001; Liquet et Commenges, 2005). Nous pourrions également étendre ces tests à des variables dépendantes du temps pour tester différentes fonctions du temps pour le taux de décès des cellules infectées pour modéliser l'établissement de la réponse

immunitaire spécifique au VIH comme proposé par Stafford et al. (2000).

Comme nous l'avons déjà expliqué, il y a plusieurs méthodes d'estimation disponibles pour les modèles dynamiques mais aucune comparaison formelle n'a pour le moment été réalisée. Du fait de l'utilisation croissante de ce type de modélisation et de la complexification des modèles, il semble pertinent de réaliser une telle comparaison en terme de biais, de précision et de temps de calculs. Parmi toutes les approches proposées, celle du maximum *a posteriori* (MAP) semble être la plus prometteuse car elle combine l'utilisation d'information *a priori* et les propriétés du maximum de vraisemblance telle que les critères de convergence.

D'autre part, l'approche de modélisation proposée est basée sur les connaissances biologiques et les paramètres doivent avoir une signification biologique. En comparaison avec les modèles descriptifs, les modèles dynamiques sont censés apporter plus d'informations et d'explications sur les phénomènes étudiés. Il serait par conséquent pertinent de comparer ces deux types de modélisation en terme d'adéquation aux données mais aussi en terme de prédictions. De même, une comparaison des paramètres estimés par un même modèle dynamique utilisé dans deux phases distinctes de l'infection (primo-infection et mise sous traitement par exemple) permettrait de confirmer le caractère explicatif de ces modèles.

Pour conclure, nous avons développé une méthode d'estimation pour étudier la primo-infection à VIH prenant en compte l'incertitude sur la date d'infection. Nous avons également amélioré la sélection de modèles et la méthode d'estimation (MAP). Il y a plusieurs perspectives à ce travail. D'un point de vue statistique, une comparaison des différentes méthodes semble prioritaire pour trouver le meilleur outil pour estimer les modèles dynamiques. De nombreuses applications pourraient être envisagées suite à ce travail. Par exemple, nous pourrions utiliser les estimations de la date d'infection pour estimer l'incidence due aux patients récemment infectés en appliquant notre méthode après des analyses phylogénétiques. Ces analyses permettent de reconstruire la phylogénie d'échantillons vitaux à partir de séquences génétiques pol pour identifier des réseaux potentiels de transmission. Combiner ces deux estimations (date d'infection et réseaux de transmission) pourrait donner une idée du nombre de patients récemment infectés impliqués dans des

réseaux de transmission pour une période de temps donnée. Ceci fait l'objet d'un projet en collaboration avec Stéphane Hué (Centre for Medical Molecular Virology, University College London). Une autre application est la comparaison de l'infection à VIH-1 et VIH-2. Nous savons que la progression clinique est différente pour les deux infections ainsi que les états d'équilibre viral et des CD4 (Drylewicz et al., 2008) mais également en terme de réponse à un traitement antirétroviral (Drylewicz et al., 2008; Drylewicz et al., 2009). Mieux comprendre comment les états d'équilibre s'établissent pourrait être un moyen de poser des hypothèses sur les différences observées entre ces deux virus et de manière plus générale, trouver les facteurs pouvant expliquer les différentes progressions cliniques observées (progresseurs rapides vs. non-progresseurs à long-terme).

Chapitre 7

Matériels complémentaires

Modèles biomathématiques

Le “modèle de base” peut s’écrire sous la forme du système d’équations différentielles ordinaires suivant :

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dT^*}{dt} = \gamma VT - \mu_{T^*} T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{cases}$$

On suppose que les compartiments sont à l’équilibre avant l’infection, ce qui peut se traduire par : $T_{0-} = \lambda / \mu_T$, $T_{0-}^* = 0$ et $V_{0-} = 0$. Après l’introduction de virus ($V_0 = 10^{-6}$ virions/mm³), l’état d’équilibre initial est perturbé et le système converge vers un nouvel état d’équilibre si R_0 ($R_0 = \frac{\gamma\lambda\pi}{\mu_T\mu_{T^*}\mu_V}$) est supérieur à 1 :

$$\begin{cases} \bar{T} = \frac{\mu_{T^*}\mu_V}{\gamma\pi} \\ \bar{T}^* = \frac{\lambda\gamma\pi - \mu_T\mu_{T^*}\mu_V}{\gamma\pi\mu_{T^*}} \\ \bar{V} = \frac{\lambda\gamma\pi - \mu_T\mu_{T^*}\mu_V}{\gamma\mu_{T^*}\mu_V} \end{cases}$$

Le “modèle cellules productrices” peut s’écrire sous la forme suivante :

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dI}{dt} = \gamma VT - \mu_I I - \delta I \\ \frac{dP}{dt} = \delta I - \mu_I P \\ \frac{dV}{dt} = \pi P - \mu_V V \end{cases}$$

Le taux pour qu’une cellule infectée devienne productrice est noté δ . L’équilibre avant infection pour ce modèle peut s’écrire : $T_{0-} = \lambda / \mu_T$, $I_{0-} = 0$, $P_{0-} = 0$ et $V_{0-} = 0$. Le

système se stabilise à un nouvel équilibre si R_0 ($R_0 = \frac{\lambda\gamma\pi\alpha}{\mu_T\mu_I\mu_V(\mu_I+\alpha)}$) est supérieur à 1 :

$$\left\{ \begin{array}{l} \bar{T} = \frac{\mu_I\mu_V(\mu_I+\alpha)}{\gamma\pi\alpha} \\ \bar{I} = \frac{\lambda}{\mu_I+\alpha} - \frac{\mu_T\mu_V\mu_I}{\gamma\pi\alpha} \\ \bar{P} = \frac{\alpha}{\mu_I} \left(\frac{\lambda}{\mu_I+\alpha} - \frac{\mu_T\mu_V\mu_I}{\gamma\pi\alpha} \right) \\ \bar{V} = \frac{\pi\alpha}{\mu_V\mu_I} \left(\frac{\lambda}{\mu_I+\alpha} - \frac{\mu_T\mu_V\mu_I}{\gamma\pi\alpha} \right) \end{array} \right.$$

Le “modèle cellules activées” peut s’écrire sous la forme suivante :

$$\left\{ \begin{array}{l} \frac{dQ}{dt} = \lambda - \alpha Q - \alpha' Q V + \rho A - \mu_Q Q \\ \frac{dA}{dt} = \alpha Q + \alpha' Q V - \gamma V A + \rho A - \mu_A A \\ \frac{dT^*}{dt} = \gamma V A - \mu_{T^*} T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{array} \right.$$

L’équilibre avant infection s’écrit sous la forme suivante :

$$Q_{0-} = \frac{\lambda(\rho + \mu_A)}{\mu_Q\rho + \mu_A\alpha + \mu_A\mu_Q}, A_{0-} = \frac{\lambda\alpha}{\mu_Q\rho + \mu_A\alpha + \mu_A\mu_Q}, T_{0-}^* = 0, V_{0-} = 0.$$

L’introduction de virions conduit à un nouvel équilibre si R_0 est supérieur à 1 : $R_0 = \frac{\gamma\pi\lambda\alpha}{\mu_{T^*}\mu_V(\rho\mu_Q + \mu_A\alpha + \mu_A\mu_Q)}$.

Le “modèle cellules effectrices” s’écrit sous la forme suivante :

$$\left\{ \begin{array}{l} \frac{dT}{dt} = \lambda - \gamma V T - \mu_T T \\ \frac{dT^*}{dt} = \gamma V T - \left[\mu_{T^*} - \kappa \frac{T^*}{T^* + \theta} \right] T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{array} \right.$$

Avant infection, l’hypothèse d’équilibre conduit aux mêmes équations que pour le “modèle de base”, le nombre de reproduction de base R_0 est également identique.

Le “modèle complet”, combinant la distinction entre cellules infectées productrices et non-productrices, cellules quiescentes et activées et incluant des cellules effectrices peut s’écrire sous la forme :

$$\left\{ \begin{array}{l} \frac{dQ}{dt} = \lambda - \alpha Q - \alpha' Q V + \rho A - \mu_Q Q \\ \frac{dA}{dt} = \alpha Q + \alpha' Q V - \gamma V A + \rho A - \mu_A A \\ \frac{dI}{dt} = \gamma V A - \left[\mu_I - \kappa \frac{I+P}{I+P+\theta} \right] I - \delta I \\ \frac{dP}{dt} = \delta I - \left[\mu_I - \kappa \frac{I+P}{I+P+\theta} \right] P \\ \frac{dV}{dt} = \pi P - \mu_V V \end{array} \right.$$

Avant infection, ce système est supposé être à l'équilibre :

$$Q_{0-} = \frac{\lambda(\rho + \mu_A)}{\mu_Q\rho + \mu_A\alpha + \mu_A\mu_Q}, A_{0-} = \frac{\lambda\alpha}{\mu_Q\rho + \mu_A\alpha + \mu_A\mu_Q}, I_{0-} = 0, P_{0-} = 0, V_{0-} = 0.$$

Il converge vers un autre état d'équilibre si R_0 est supérieur à 1 : $R_0 = \frac{\gamma\pi\lambda\alpha\delta}{\mu_I\mu_V(\delta+\mu_I)(\rho\mu_Q+\alpha\mu_A+\mu_A\mu_Q)}$.

Correspondances entre modèles

Pour comparer les paramètres des différents modèles, il est nécessaire de prendre en compte les différents compartiments et leur correspondance. Ainsi, il est possible de passer du “modèle cellules productrices” au “modèle de base” en posant $T^* = I + P$. On a alors $\frac{dT^*}{dt} = \frac{dI}{dt} + \frac{dP}{dt}$, on obtient le système suivant :

$$\begin{cases} \frac{dT}{dt} = \lambda - \gamma VT - \mu_T T \\ \frac{dT^*}{dt} = \gamma VT - \mu_{T^*} T^* \\ \frac{dV}{dt} = \check{\pi} T^* - \mu_V V \end{cases}$$

où $\check{\pi} = \pi \frac{P}{P+I}$ et $\mu_{T^*} = \mu_I$. Ces deux modèles devraient donc nous fournir des estimations similaires pour tous les paramètres A l'exception du nombre de virions produit par une cellule infectée par jour. Si on connaît l'ordre de grandeur du rapport $\frac{P}{P+I}$, on peut connaître le nouvel ordre grandeur de ce paramètre.

De la même manière, il est possible de passer du “modèle cellules activées” au “modèle de base” en posant $T = Q + A$. On a alors $\frac{dT}{dt} = \frac{dQ}{dt} + \frac{dA}{dt}$ et on obtient le système suivant :

$$\begin{cases} \frac{dT}{dt} = \lambda - \check{\gamma} VT - \mu_Q Q - \mu_A A \\ \frac{dT^*}{dt} = \check{\gamma} VT - \mu_{T^*} T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{cases}$$

Or on sait que $\mu_Q \ll \mu_A$ car les cellules quiescentes sont des cellules à longue vie. On peut donc supposer que pour tout t , $\mu_Q Q \ll \mu_A A$ donc :

$$\begin{cases} \frac{dT}{dt} = \lambda - \check{\gamma} VT - \check{\mu}_T T \\ \frac{dT^*}{dt} = \check{\gamma} VT - \mu_{T^*} T^* \\ \frac{dV}{dt} = \pi T^* - \mu_V V \end{cases}$$

où $\check{\gamma} = \gamma \frac{A}{Q+A}$ et $\check{\mu}_T = \mu_A \frac{A}{Q+A}$. On s'attend donc à avoir des valeurs différentes pour le taux d'infectiosité du virus et le taux de décès des cellules cibles. A cause de problèmes

d'identifiabilité, nous avons été obligé de fixer certains paramètres dont γ , il est donc important de prendre en compte cette correspondance entre modèles pour savoir à quelle valeur il est nécessaire de le fixer. Ainsi, dans le “modèle de base”, nous avons pris $\gamma = 0.00027 \text{ virions}^{-1} \text{ jour}^{-1} \mu\text{L}^{-1}$ (Wilson et al., 2007). Si on considère que chez un individu sain, la proportion de cellules activées est de 1% (Sachsenberg et al., 1998) alors on doit prendre $\gamma = 0.05 \text{ virions}^{-1} \text{ jour}^{-1} \mu\text{L}^{-1}$.

On peut passer du “modèle cellules effectrices” au “modèle de base” en posant $\mu_{T^*} + \kappa \frac{T^*}{T^* + \theta} = \check{\mu}_{T^*}$; on s'attend donc à un taux de décès des cellules infectées différents dans ces deux modèles. Enfin, pour passer du “modèle complet” au “modèle de base”, on pose $T^* = I + P$ et $T = Q + A$, on obtient le système suivant :

$$\begin{cases} \frac{dT}{dt} = \lambda - \check{\gamma}VT - \check{\mu}_T T \\ \frac{dT^*}{dt} = \check{\gamma}VT - \check{\mu}_{T^*}T^* \\ \frac{dV}{dt} = \check{\pi}T^* - \mu_V V \end{cases}$$

où $\check{\pi} = \pi \frac{P}{P+I}$, $\check{\mu}_{T^*} = \mu_I + \kappa \frac{I+P}{I+P+\theta}$, $\check{\gamma} = \gamma \frac{A}{Q+A}$ et $\check{\mu}_T = \mu_A \frac{A}{Q+A}$.

Bibliographie

- Altfeld, M., Addo, M., Rosenberg, E., Hecht, F., Lee, P., Vogel, M., Yu, X., Draenert, R., Johnston, M., Strick, D., Allen, T., Feeney, M., Kahn, J., Sekaly, R., LEvy, J., Rockstroh, J., Goulder, P. et Walker, B. (2003). Influence of HLA-B 57 on clinical presentation and viral control during acute HIV-1 infection. *AIDS* **17**, 2581–2591.
- Appay, V., Papagno, L., Spina, C., Hansasuta, P., King, A., Jones, L., Ogg, G., Little, S., McMichael, A., Richman, D. et Rowland-Jones, S. (2002). Dynamics of T cell responses in HIV infection 1. *The Journal of Immunology* **168**, 3660–3666.
- Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R. et Puren, A. (2005). Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: The ANRS 1265 trial. *PLoS Medicine* **2**, e298.
- Ball, C., Gilchrist, M. et Coombs, D. (2007). Modeling within-host evolution of HIV: mutation, competition and strain replacement. *Bulletin of Mathematical Biology* **69**, 2361–2385.
- Barbosa, P., Charneau, P., Dumey, N. et Clavel, F. (1994). Kinetic analysis of HIV-1 early replicative steps in a coculture system. *AIDS Research and Human Retroviruses* **10**, 53–59.
- Barre-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W. et Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868–871.
- Bock, P. et Markovitz, D. (2001). Infection with HIV-2. *AIDS(Supplement)* **15**, 35–45.
- Bofill, M., Janossy, G., Lee, C., MacDonald-Burns, D., Phillips, A., Sabin, C., Timms, A., Johnson, M. et Kernoff, P. (1992). Laboratory control values for CD4 and CD8 T lymphocytes. Implications for HIV-1 diagnosis. *Clinical and Experimental Immunology*

- 88, 243.
- Bonhoeffer, S., May, R., Shaw, G. et Nowak, M. (1997). Virus dynamics and drug therapy. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6971–6976.
- Borrow, P., Lewicki, H., Hahn, B., Shaw, G. et Oldstone, M. (1994). Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *Journal of Virology* **68**, 6103–6110.
- Boscardin, W., Taylor, J. et Law, N. (1998). Longitudinal models for AIDS marker data. *Statistical Methods in Medical Research* **7**, 13–27.
- Brown, E., Mawhinney, S., Jones, R., Kafadar, K. et Young, B. (2001). 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* **20**, 2489–2504.
- Burg, D., Rong, L., Neumann, A. et Dahari, H. (2009). Mathematical modeling of viral kinetics under immune control during primary HIV-1 infection. *Journal of Theoretical Biology* **259**, 751–759.
- Burton, D., Stanfield, R. et Wilson, I. (2005). Antibody vs. HIV in a Clash of Evolutionary Titans. *Proceedings of the National Academy of Sciences of the United States of America* pages 14943–14948.
- CASCADE, C. (2003). Determinants of survival following HIV-1 seroconversion after the introduction of HAART: results from CASCADE. *The Lancet* **362**, 1267–1274.
- Chun, T., Engel, D., Berrey, M., Shea, T., Corey, L. et Fauci, A. (1998). Early establishment of a pool of latently infected, resting CD4 (1) T cells during primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 8869–8873.
- Ciupe, M., Bivort, B., Bortz, D. et Nelson, P. (2006). Estimating kinetic parameters from HIV primary infection data through the eyes of three different mathematical models. *Mathematical Biosciences* **200**, 1–27.
- Clark, S., Saag, M., Decker, W., Campbell-Hill, S., Roberson, J., Veldkamp, P., Kappes, J., Hahn, B. et Shaw, G. (1991). High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. *New England Journal of Medicine* **324**,

- 954–960.
- Commenges, D. (2009). Statistical models: Conventional, penalized and hierarchical likelihood. *Statistics Surveys* **3**, 1–17.
- Commenges, D., Jacqmin-Gadda, H., Proust, C. et Guedj, J. (2006). A newton-like algorithm for likelihood maximization: The robust-variance scoring algorithm. *eprint arXiv: math/0610402*.
- Commenges, D. et Rondeau, V. (2006). Relationship between derivatives of the observed and full loglikelihoods and application to Newton-Raphson algorithm. *International Journal of Biostatistics* **2**, 1010–1010.
- Crum-Cianflone, N., Eberly, L., Zhang, Y., Ganesan, A., Weintrob, A., Marconi, V., Barthel, R., Fraser, S., Agan, B. et Wegner, S. (2009). Is HIV Becoming More Virulent? Initial CD4 Cell Counts among HIV Seroconverters during the Course of the HIV Epidemic: 1985–2007. *Clinical Infectious Diseases* **48**, 1285–1292.
- Daar, E., Moudgil, T., Meyer, R. et Ho, D. (1991). Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection. *New England Journal of Medicine* **324**, 961.
- Davidian, M. et Giltinan, D. (1995). Hierarchical Non-linear models. *Non-linear models for repeated measurement data*. 2nd ed. London: Chapman & Hall/CRC pages 63–95.
- Deeks, S., Kitchen, C., Liu, L., Guo, H., Gascon, R., Narvaez, A., Hunt, P., Martin, J., Kahn, J., Levy, J., McGrath, M. et Hecht, F. (2004). Immune activation set point during early HIV infection predicts subsequent CD4+ T-cell changes independent of viral load. *Blood* **104**, 942–947.
- Dempster, A., Laird, N. et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.
- Annals of Statistics **14**, 1–26.

Dixit, N. et Perelson, A. (2004). Complex patterns of viral load decay under antiretroviral therapy: influence of pharmacokinetics and intracellular delay. *Journal of Theoretical Biology* **226**, 95–109.

Donnet, S. et Samson, A. (2008). Parametric inference for mixed models defined by

- stochastic differential equations. *ESAIM: Probability and Statistics* **12**, 196–218.
- Dorrucci, M., Rezza, G., Porter, K. et Phillips, A. (2007). Temporal trends in postseroconversion CD4 cell count and HIV load: the Concerted Action on Seroconversion to AIDS and Death in Europe Collaboration, 1985–2002. *The Journal of Infectious Diseases* **195**, 525–534.
- Drylewicz, J., Commenges, D. et Thiébaut, R. (2009). Score tests for exploring complex models: Application to HIV dynamical models. *Biometrical Journal* sous presse.
- Drylewicz, J., Eholie, S., Maiga, M., Zannou, M., Sow, P., Ekouevi, D., Peterson, K., Bissagnene, E., Dabis, F. et for IeDEA West Africa Study Group, T. R. (2009). Lymphocyte T CD4+ response to antiretroviral therapy according to the HIV type in IeDEA West Africa Collaboration. *AIDS* accepted.
- Drylewicz, J., Guedj, J., Commenges, D., Thiébaut, R. et on behalf of the CASCADE Collaboration (2009). Modelling the dynamics of biomarkers during primary HIV infection taking into account the uncertainty of infection date. *submitted*.
- Drylewicz, J., Matheron, S., Lazaro, E., Damond, F., Bonnet, F., Simon, F., Dabis, F., Brun-Vezinet, F., Chêne, G. et Thiébaut, R. (2008). Comparison of viro-immunological marker changes between HIV-1 and HIV-2-infected patients in France. *AIDS* **22**, 457–468.
- Fellay, J., Shianna, K., Ge, D., Colombo, S., Goldstein, D., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., DeLuca, A., P., E., Francioli, P. et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944–947.
- Fidler, S., Fox, J., Porter, K. et Weber, J. (2008). Primary HIV infection: to treat or not to treat?. *Current Opinion in Infectious Diseases* **21**, 4–10.
- Fiebig, E., Wright, D., Rawal, B., Garrett, P., Schumacher, R., Peddada, L., Hellebrant, C., Smith, R., Conrad, A., Kleinman, S. et Busch, M. (2003). Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS* **17**, 1871–1879.
- Finzi, D., Blankson, J., Siliciano, J., Margolick, J., Chadwick, K., Pierson, T., Smith, K., Lisziewicz, J., Lori, F., Flexner, C. et al. (1999). Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective

- combination therapy. *Nature Medicine* **5**, 512–517.
- Geman, S. et Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration 24 of images'. *IEEE Trans. PAMI* **6**, 721–741.
- Gibaldi, M. et Perrier, D. (1982). Pharmacokinetics (2nd edn).
- Grabar, S., Selinger-Leneman, H., Abgrall, S., Pialoux, G., Weiss, L. et Costagliola, D. (2009). Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. *AIDS* **23**, 1163–1169.
- Grant, R., Buchbinder, S., Cates, W., Clarke, E., Coates, T., Cohen, M., Delaney, M., Flores, G., Goicochea, P., Gonsalves, G., Harrington, M., Lama, J., MacQueen, K., Moore, J., Peterson, L., Sanchez, J., Thompson, M. et Wainberg, M. (2005). Promote HIV chemoprophylaxis research, don't prevent it. *Science* **309**, 2170–2171.
- Graziosi, C., Pantaleo, G., Butini, L., Demarest, J., Saag, M., Shaw, G. et Fauci, A. (1993). Kinetics of human immunodeficiency virus type 1 (HIV-1) DNA and RNA synthesis during primary HIV-1 infection. *Proceedings of the National Academy of Sciences* **90**, 6405–6409.
- Grossman, Z., Meier-Schellersheim, M., Sousa, A., Victorino, R. et Paul, W. (2002). CD4+ T-cell depletion in HIV infection: are we closer to understanding the cause? *Nature Medicine* **8**, 319–323.
- Guedj, J., Thiébaut, R. et Commenges, D. (2007). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63**, 1198–1206.
- Han, C. et Chaloner, K. (2004). Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. *Biometrics* **60**, 25–33.
- Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- Hecht, F., Wang, L., Collier, A., Little, S., Markowitz, M., Margolick, J., Kilby, J., Daar, E., Conway, B. et Holte, S. (2006). A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute HIV infection. *The Journal of Infectious Diseases* **194**, 725–733.
- Henrard, D., Daar, E., Farzadegan, H., Clark, S., Phillips, J., Shaw, G. et Busch, M. (1995). Virologic and immunologic characterization of asymptomatic primary HIV-1 infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*

- 9, 305–310.
- Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. et Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–126.
- Huang, Y., Liu, D. et Wu, H. (2006). Hierarchical bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62**, 413–423.
- Huang, Y., Rosenkranz, S. et Wu, H. (2003). Modeling HIV dynamics and antiviral response with consideration of time-varying drug exposures, adherence and phenotypic sensitivity. *Mathematical biosciences* **184**, 165–186.
- Huang, Y. et Wu, H. (2006). A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *Journal of Applied Statistics* **33**, 155–174.
- Hunt, P., Martin, J., Sinclair, E., Bredt, B., Hagos, E., Lampiris, H. et Deeks, S. (2003). T cell activation is associated with lower CD4+ T cell gains in human immunodeficiency virus-infected patients with sustained viral suppression during antiretroviral therapy. *The Journal of Infectious Diseases* **187**, 1534–1543.
- InVS (2008). L'infection à VIH/Sida en France. *BEH* **45–46**, 434–443.
- Kahn, J. et Walker, B. (1998). Acute human immunodeficiency virus type 1 infection. *New England Journal of Medicine* **339**, 33–39.
- Kassutto, S., Rosenberg, E. et Mayer, K. (2004). Primary HIV type 1 infection. *Clinical Infectious Diseases* **38**, 1447–1453.
- Kaufmann, D., Lichterfeld, M., Altfeld, M., Addo, M., Johnston, M., Lee, P., Wagner, B., Kalife, E., Strick, D., Rosenberg, E. et Walker, B. (2004). Limited durability of viral control following treated acute HIV infection. *PLoS Medicine* **1**, 137–148.
- Kaufmann, G., Cunningham, P., Kelleher, A., Zaunders, J., Carr, A., Vizzard, J., Law, M. et Cooper, D. (1998). Patterns of viral dynamics during primary human immunodeficiency virus type 1 infection. *The Journal of Infectious Diseases* **178**, 1812–1815.
- Kaufmann, G., Cunningham, P., Zaunders, J., Law, M., Vizzard, J., Carr, A. et Cooper, D. (1999). Impact of early HIV-1 RNA and T-lymphocyte dynamics during primary HIV-1 infection on the subsequent course of HIV-1 RNA levels and CD4+ T-lymphocyte counts in the first year of HIV-1 infection. *Journal of Acquired Immune Deficiency Syndromes* **22**, 437–444.

- Kaufmann, G., Duncombe, C., Zaunders, J., Cunningham, P., Cooper, D., Ratio, C., Blotting, W., Serodiagnosis-methods, A., Infections-virology, H., Infections-immunology, H. et al. (1998). Primary HIV-1 infection: a review of clinical manifestations, immunologic and virologic changes. *AIDS Patient Care STDS* **12**, 759–767.
- Kiernan, R., Marshall, J., Bowers, R., Doherty, R. et McPhee, D. (1990). Kinetics of HIV-1 replication and intracellular accumulation of particles in HTLV-I-transformed cells. *AIDS research and human retroviruses* **6**, 743–752.
- Kilby, J., Lee, H., Hazelwood, J., Bansal, A., Bucy, R., Saag, M., Shaw, G., Acosta, E., Johnson, V., Perelson, A., Goepfert, A. et UAB Acute Infection, E. D. R. P. A. G. (2008). Treatment response in acute/early infection versus advanced AIDS: equivalent first and second phases of HIV RNA decline. *AIDS* **22**, 957–962.
- Koup, R., Safrit, J., Cao, Y., Andrews, C., McLeod, G., Borkowsky, W., Farthing, C. et Ho, D. (1994). Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *Journal of Virology* **68**, 4650–4655.
- Kuhn, E. et Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics* **8**, 115–131.
- Kuhn, E. et Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis* **49**, 1020–1038.
- Laird, N. et Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* pages 963–974.
- Lindback, S., Karlsson, A., Mittler, J., Blaxhult, A., Carlsson, M., Briheim, G., Sonnerborg, A. et Gaines, H. (2000). Viral dynamics in primary HIV-1 infection. Karolinska Institutet Primary HIV Infection Study Group. *AIDS* **14**, 2283–2291.
- Liquet, B. et Commenges, D. (2001). Correction of the P-value after multiple coding of an explanatory variable in logistic regression. *Statistics in Medicine* **20**, 2815–2826.
- Liquet, B. et Commenges, D. (2005). Computation of the p-value of the maximum of score tests in the generalized linear model; application to multiple coding. *Statistics and Probability Letters* **71**, 33–38.
- Little, S., McLean, A., Spina, C., Richman, D. et Havlir, D. (1999). Viral dynamics of acute HIV-1 infection. *Journal of Experimental Medicine* **190**, 841–850.

- Lloyd, A. (2001). The dependence of viral parameter estimates on the assumed viral life cycle: limitations of studies of viral load data. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 847–854.
- Louis, T. (1982). Finding the observed Information matrix when using the *EM* algorithm. *Journal of the Royal Statistical Society. Series B* **44**, 226–233.
- Mackewicz, C., Ortega, H. et Levy, J. (1991). CD8+ cell anti-HIV activity correlates with the clinical state of the infected individual. *Journal Clinical Investigation* **87**, 1462–1466.
- Maini, M., Gilson, R., Chavda, N., Gill, S., Fakoya, A., Ross, E., Phillips, A. et Weller, I. (1996). Reference ranges and sources of variability of CD4 counts in HIV-seronegative women and men. *British Medical Journal* **72**, 27–31.
- Malone, J., Simms, T., Gray, G., Wagner, K., Burge, J. et Burke, D. (1990). 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* **3**, 144–151.
- Markowitz, M., Louie, M., Hurley, A., Sun, E., Di Mascio, M., Perelson, A. et Ho, D. (2003). A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *Journal of Virology* **77**, 5037–5038.
- McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.
- McLean, A., Emery, V., Webster, A. et Griffiths, P. (1991). Population dynamics of HIV within an individual after treatment with zidovudine. *AIDS* **5**, 485–489.
- McLean, A. et Michie, C. (1995). *in vivo* estimates of division and death rates of human t lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 3707–3711.
- Mellors, J., Kingsley, L., Rinaldo, C., Todd, J., Hoo, B., Kokka, R. et Gupta, P. (1995). Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion. *Annals of Internal Medicine* **122**, 573–579.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. et Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*

- 21, 1087–1092.
- Moore, J., Sattentau, Q., Wyatt, R. et Sodroski, J. (1994). Probing the structure of the human immunodeficiency virus surface glycoprotein gp120 with a panel of monoclonal antibodies. *Journal of Virology* **68**, 469–484.
- Musey, L., Hughes, J., Schacker, T., Shea, T., Corey, L. et McElrath, M. (1997). Cytotoxic-T-cell responses, viral load, and disease progression in early human immunodeficiency virus type 1 infection. *New England Journal of Medicine* **337**, 1267–1274.
- Nowak, M., Bonhoeffer, S., Shaw, G. et May, R. (1997). Anti-viral Drug Treatment: Dynamics of Resistance in Free Virus and Infected Cell Populations. *Journal of Theoretical Biology* **184**, 203–217.
- Pantaleo, G., Graziosi, C. et Fauci, A. (1997). Virologic and immunologic events in primary HIV infection. In *Springer seminars in immunopathology*, volume 18, pages 257–266. Springer.
- Pantazis, N., Touloumi, G., Vanhems, P., Gill, J., Bucher, H., Porter, K. et the CASCADE Collaboration (2008). The effect of antiretroviral treatment of different durations in primary HIV infection. *AIDS* **22**, 2441–2450.
- Parren, P., Burton, D. et Sattentau, Q. (1997). HIV-1 antibody-debris or virion? *Nature Medicine* **3**, 366–367.
- Perelson, A., Neumann, A., Markowitz, M., Leonardy, J. et Ho, D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582–1586.
- Phillips, A. (1996). Reduction of HIV concentration during acute infection: independence from a specific immune response. *Science* **271**, 497–499.
- Piatak, M., Saag, M., Yang, L., Clark, S., Kappes, J., Luk, K., Hahn, B., Shaw, G. et Lifson, J. (1993). High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science* **259**, 1749–1754.
- Putter, H., Heisterkamp, S., Lange, J. et De Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in Medicine* **21**, 2199–2214.
- Ramratnam, B., Bonhoeffer, S., Binley, J., Hurley, A., Zhang, L., Mittler, J., Markowitz, M., Moore, J., Perelson, A. et Ho, D. (1999). Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *The Lancet* **354**,

- 1782–1786.
- Ribeiro, R., Mohri, H., Ho, D. et Perelson, A. (2002). In vivo dynamics of T cell activation, proliferation, and death in HIV-1 infection: Why are CD4+ but not CD8+ T cells depleted? *Proceedings of the National Academy of Sciences* **99**, 15572–15577.
- Robbins, H. et Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics* pages 400–407.
- Rowland-Jones, S. et Whittle, H. (2007). Out of Africa: what can we learn from HIV-2 about protective immunity to HIV-1? *Nature Immunology* **8**, 329–331.
- Saah, A., Hoover, D., Weng, S., Carrington, M., Mellors, J., Rinaldo, C., Mann, D., Apple, R., Phair, J., Detels, R., O'Brien, S., Enger, C., Johnson, P. et Kaslow, R. (1998). Association of HLA profiles with early plasma viral load, CD 4+ cell count and rate of progression to AIDS following acute HIV-1 infection. *AIDS* **12**, 2107–2113.
- Sachsenberg, N., Perelson, A., Yerly, S., Schockmel, G., Leduc, D., Hirschel, B. et Perrin, L. (1998). Turnover of CD4+ and CD8+ T lymphocytes in HIV-1 infection as measured by Ki-67 antigen. *Journal of Experimental Medicine* **187**, 1295–1303.
- Samson, A., Lavielle, M. et Mentré, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Computational Statistics and Data Analysis* **51**, 1562–1574.
- Schacker, T., Hughes, J., Shea, T., Coombs, R. et Corey, L. (1998). Biological and virologic characteristics of primary HIV infection. *Annals of Internal Medicine* **128**, 613–620.
- Sedaghat, A., Siliciano, J., Brennan, T., Wilke, C. et Siliciano, R. (2007). Limits on Replenishment of the Resting CD4+; T Cell Reservoir for HIV in Patients on HAART. *PLoS Pathogens* **3**, 1165–1174.
- Seng, R., Goujard, C., Desquibet, L., Sinet, M., Rouzioux, C., Deveau, C., Boufassa, F., Delfraissy, J., Meyer, L., Venet, A., the ANRS PRIMO et Groups, S. S. (2008). Rapid CD4+ cell decrease after transient cART initiated during primary HIV infection (ANRS PRIMO and SEROCO cohorts). *Journal of Acquired Immune Deficiency Syndromes* **49**, 251–258.
- Sheiner, L., Rosenberg, B. et Melmon, K. (1972). Modelling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research, an interna-*

- tional journal **5**, 441–459.
- Stafford, M., Cao, Y., Ho, D., Corey, L. et Perelson, A. (2000). Modeling Plasma Virus Concentration and CD4+ T Cell Kinetics during Primary HIV Infection. *Journal of Theoretical Biology* **203**, 285–301.
- Steingrover, R., Pogány, K., Garcia, E., Jurriaans, S., Brinkman, K., Schuitemaker, H., Miedema, F., Lange, J. et Prins, J. (2008). HIV-1 viral rebound dynamics after a single treatment interruption depends on time of initiation of highly active antiretroviral therapy. *AIDS* **22**, 1583–1588.
- Sterling, T., Vlahov, D., Astemborski, J., Hoover, D., Margolick, J. et Quinn, T. (2001). Initial plasma HIV-1 RNA levels and progression to AIDS in women and men. *New England Journal of Medicine* **344**, 720–725.
- Streeck, H., Jolin, J., Qi, Y., Yassine-Diab, B., Johnson, R., Kwon, D., Addo, M., Brumme, C., Routy, J., Little, S., Jessen, H., Kelleher, A., Hecht, F., Sekaly, R., Rosenberg, E., Walker, B.D abd Carrington, M. et Altfeld, M. (2009). HIV-1-specific CD8+ T cell responses during primary infection are major determinants of the viral set point and loss of CD4+ T cells. *Journal of Virology* **83**, 7641–7648.
- Tanner, M. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. Springer Verlag.
- Thiébaut, R., Guedj, J., Jacqmin-Gadda, H., Chêne, G., Trimoulet, P., Neau, D. et Commenges, D. (2006). Estimation of dynamical model parameters taking into account undetectable marker values. *BMC Medical Research Methodology* **6**, 1–10.
- Thiébaut, R., Jacqmin-Gadda, H., Leport, C., Katlama, C., D., C., Le Moing, V., Morlat, P., Chene, G. et the APROCO Study Group (2003). 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. *Journal of Biopharmaceutical Statistics* **13**, 271–282.
- Tuckwell, H. et Le Corfec, E. (1998). A stochastic model for early HIV-1 population dynamics. *Journal of Theoretical Biology* **195**, 451–463.
- UNAIDS (2008). Report on the global AIDS epidemic.
- Vanhems, P., Lambert, J., Cooper, D., Vizzard, J., Loes, S. et Allard, R. (1998). Severity and prognosis of acute human immunodeficiency virus type 1 illness: a dose-response

- relationship. *Clinical Infectious Diseases* **26**, 323–329.
- Vanhems, P., Lambert, J., Guerra, M., Hirschel, B. et Allard, R. (1999). Association between the rate of CD4+ T cell decrease and the year of human immunodeficiency virus (HIV) type 1 seroconversion among persons enrolled in the Swiss HIV Cohort Study. *Journal of Infectious Diseases* **180**, 1803–1808.
- Voirin, N., Routy, J., Smith, D., Baratin, D., Trepo, C., Cotte, L., Touraine, J., Livrozet, J., Cooper, D., Ritter, J., André, P. et Vanhems, P. (2008). Effect of early initiation of highly active antiretroviral therapy on CD4 cell count and HIV-RNA viral load trends within 24 months of the onset of acute retroviral syndrome*. *HIV Medicine* **9**, 440–444.
- Wei, X., Ghosh, S. K., E., T. M., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, N. A., Hahn, B. H., Saag, M. S. et Shaw, G. M. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117–122.
- Wick, D., Self, S. et Corey, L. (2002). Do scarce targets or T killers control primary HIV infection? *Journal of Theoretical Biology* **214**, 209–214.
- Wilson, D., Mattapallil, J., Lay, M., Zhang, L., Roederer, M. et Davenport, M. (2007). Estimating the Infectivity of CCR5-Tropic Simian Immunodeficiency Virus SIVmac251 in the Gut? *Journal of Virology* **81**, 8025–8029.
- Wodarz, D. et Hamer, D. (2007). Infection dynamics in HIV-specific CD4 T cells: Does a CD4 T cell boost benefit the host or the virus? *Mathematical Biosciences* **209**, 14–29.
- Wodarz, D., Klenerman, P. et Nowak, M. (1998). Dynamics of cytotoxic T-lymphocyte exhaustion. *Proceedings: Biological Sciences* **265**, 191–203.
- Wodarz, D. et Nowak, M. (2000). Immune responses and viral phenotype: Do replication rate and cytopathogenicity influence virus load? *Computational and Mathematical Methods in Medicine* **2**, 113–127.
- Wu, H., Ding, A. et De Gruttola, V. (1998). Estimation of HIV dynamic parameters. *Statistics in Medicine* **17**, 2463–2485.

Annexes



COMPARISON OF THREE DYNAMICAL MODELS DURING PRIMARY HIV INFECTION: APPLICATION TO THE CASCADE DATASET

Julia Drylewicz ^{1,2}, Rodolphe Thiébaut ^{1,2} and Daniel Commenges ^{1,2} for the CASCADE Collaboration

(1) INSERM U897 Epidemiology and Biostatistics Research Centre, Bordeaux, France (2) University of Bordeaux 2, Bordeaux, France

Background:

We aim at comparing three different mathematical models for fitting the dynamics of plasma HIV RNA and CD4 cell count during primary HIV infection in 761 HIV-1 seroconverters included in the CASCADE Collaboration.

Methods:

CASCADE is a collaboration of observational cohorts of HIV-1 infected patients in Europe, Canada and Australia. We compared a 3-compartment model (the "basic model") including target cells, infected cells and free virions (fig. 1a) with two 4-compartment models: a "productive cells model" differentiating infected non-productive cells and infected productive cells (fig. 1b) and a "activated cells model" differentiating quiescent non-infected cells and activated non-infected cells (fig. 1c). Parameters were estimated in population models with a maximum likelihood approach (Guedj et al., *Biometrics* 2007). Moreover, the uncertainty of the date of infection was taking into account.

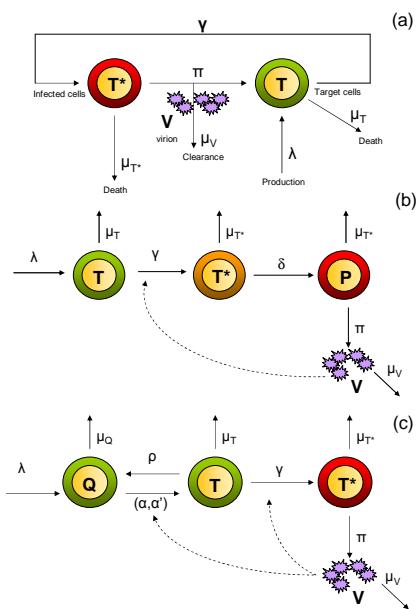


Figure 1: Biological models, basic model (a), productive cells model (b) and activated cells model (c).

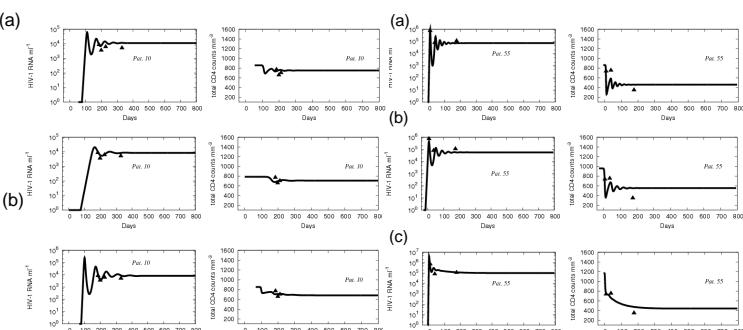


Figure 2: Individual predicted fit of HIV RNA and CD4 cell count for two patients with the three models: the "basic model" (a), the "productive cells model" (b) and the "activated cells model" (c).

Results:

The estimated half-lives of target and infected cells and the estimated R₀ were higher in the "activated cells model". The predicted day of viral peak was later for the "productive cells model" (tab. 1).

The fit of the "productive cells model" was the best (lowest AIC) (fig. 2). The improvement brought by this model w.r.t. the "basic model" can be considered as large and w.r.t. the "activated cells model" as moderate (Commenges et al., *The Annals of Applied Statistics* 2008).

Table 1: Estimates for the three models based on 761 patients.

	Basic model	Productive cells model	Activated cells model
AIC	15010	14300	14567
R ₀ (median IQR)	1.85 (1.43 ; 2.30)	1.80 (1.53 ; 2.30)	3.02 (1.89 ; 5.27)
Half-live of target cells (in days)	18.06	20.69	69.65
Half-live of infected cells (in days)	0.37	0.40	0.65
Median day of viral peak (IQR)	16 (13 ; 23)	28 (21 ; 35)	14 (10 ; 22)
Median day of setpoint* (IQR)	64 (51 ; 77)	62 (52 ; 72)	62 (46 ; 78)

* Predicted setpoint of CD4 and HIV RNA are defined by a variation of less than 50 cells and a between variation of less than $0.5\log_{10}$ between time t and time $t+1$.

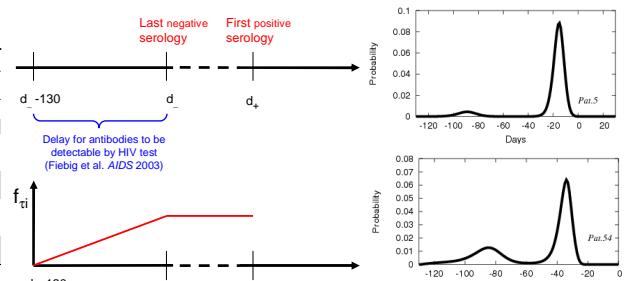


Figure 3: A priori and a posteriori distributions of the date of infection.

We determined for each patient the a posteriori distribution of the date of infection (fig. 3). We assume that the infection date is between (d_- -130 days) and d_+ . For patients with data around the time of the viral peak, we were able to reduce the window of infection date. The mode of the posterior distribution could be before the date of last negative serology (day 0).

Conclusion:

We estimated the parameters of three dynamical models in 761 patients taking into account the uncertainty of infection date. Including more compartments in the model leads to a better fit of observed dynamics but requires also to re-examine the interpretation and the values of fundamental parameters.

CASCADE collaboration. **Steering Committee:** Julia Del Amo (Chair), Laurence Meyer (Vice Chair), Hervé C. Bucher, Geneviève Chêne, Deenan Pillay, Maria Prins, Magda Rosinska, Caroline Sabin, Giota Tsoloumi. **Coordinating Center:** Khalid Porter (Project Leader), Sara Lodi, Sarah Banbury, Kate Coughlin, Sarah Walker, Abdel Sabbar, Jane Drylewicz, Clinical and Laboratory Board; Hervé Bucher, Andrea de Luca, Martin Fisher, Roberto Muggia. **Collaborators:** Australia Sydney AIDS Prospective Study and Sydney Primary HIV Infection cohort (John Kaldor, Tony Kelleher, Tim Ramachandran); Canada South Alberta clinic (John Gill); Denmark Copenhagen HIV Seminar Center cohort (Louise Brink-Jorgensen, Claus Nielsen, Claus Pedersen); France INSERM U897 Epidemiology and Biostatistics Research Centre (Khalid Porter, Sarah Banbury, Kate Coughlin, Sarah Walker, Abdel Sabbar, Jane Drylewicz, Clinical and Laboratory Board); France Paris Hospital cohort (Françoise Dabat, Robert Thivierge); Germany (Osnabrück, Hannover, Cologne, Ulm, Regensburg, Ingolstadt); Greece Greek Hemophilia cohort (George Tsiaras, Angeliki Hatzakis, Dimitris Paraskevis, Anastasia Karayannidis); Italy Italian Seroconversion Study (Giovanni Rezza, Maria Donucci, Benedetta Longo, Claudia Balotra); Netherlands Amsterdam Cohort Studies among homosexual men and drug users (Maria Prins, Janine van der Helm, Anneke Krol, Hanenke Schutte-Smeekens, Agnesse Karalus); Poland National Institute of Hygiene (Magdalena Rosinska); Portugal Universidade Nova de Lisboa (Santiago Plaza-Hoyos, Isolda Hernandez Aguado); Switzerland Swiss HIV Cohort Study (Heiner C. Bucher, Martin Rickenbäch, Patrick Fricker); Ukraine Perinatal Prevention of AIDS Initiative (Ruslan Matyush); United Kingdom Edinburgh Hospital cohort (Ray Buttoft), Health Protection Agency (Yvonne Delpont, Sam Lamerton, Gary Murphy, John Parry, Noel Gill); Royal Free Haemophilia cohort (Caroline Sabin, Christine Leek); UK Register of HIV Seroneverters (Khalid Porter, Anne Johnson, Andrew Phillips, Abdell Bakker, Janet Darbyshire, Valérie Depêche), University College London (Deenan Pillay), University of Oxford (Hans Grotto).

Autres travaux

Comparaison VIH-1 et VIH-2

Articles

- *Comparison of viro-immunological marker changes between HIV-1 and HIV-2-infected patients in France.* Drylewicz J., Matheron S., Lazaro E., Damond F., Bonnet F., Simon F., Dabis F., Brun-Vezinet F., Chêne G. and Thiébaut R., **AIDS** 2008, 22:457-468.
- *Lymphocyte T CD4+ response to antiretroviral therapy according to the HIV type in IeDEA West Africa Collaboration.* Drylewicz J., Eholie S., Maiga M., Zannou M.D., Sow P.S., Ekouevi D., Peterson K., Bissagnene E., Dabis F., Thiébaut R. for IeDEA West Africa Study Group. accepté dans **AIDS**.

Communications dans un congrès avec comité de lecture

- *Lymphocyte T CD4+ Response to ART According to the HIV Type in a Large Cohort Collaboration in West Africa.* Drylewicz J., Eholie S., Maiga M., Amani-Bosse C., Sow P.S., Ekouevi D., Peterson K., Bissagnene E., Dabis F., Thiébaut R. for IeDEA West Africa Study Group.
16th Conference on Retroviruses and Opportunistic Infections, Montréal (Canada), February 8-11 2009 (Poster)
- *The first year CD4+ response to antiretroviral treatment and its determinants in patients infected with HIV types 1 and 2: report from the IeDEA-West Africa collaboration.* Maiga M., Drylewicz J., Ba-Gomis F.O., Zannou D.M., Minga A., Traoré H., Ekouevi D., Calixte G., Peterson K., Amani-Bosse C., Sow P.S., Messou E., Etard J.F., Eholie S., Bissagnene E., Dabis F., Thiébaut R. for IeDEA West Africa Study Group.
15th International Conference on AIDS and STIs in Africa, Dakar (Sénégal), December 3-7 2008 (Poster)
- *Comparison of Viro-Immunological Marker Evolution between HIV-1- and HIV-2-infected Patients in France.* Drylewicz J., Matheron S., Lazaro E., Damond F., Bonnet F., Simon F., Dabis F., Brun-Vezinet F., Chêne G. and Thiébaut R.
14th Conference on Retroviruses and Opportunistic Infections, Los Angeles (Etats-Unis), February 25-28 2007 (Poster)