Maladies infectieuses

Estimation de l'exhaustivité de la surveillance du sida par la méthode capture-recapture, France, 2004-2006

UNIVERSITE D'AUVERGNE CLERMONT-FD I

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

InVS INSTITUT DE VEILLE SANITAIRE

Sommaire

Abréviations 2

1. Introduction 3

2. Matériels et méthodes 4

2.1 Population d'étude 4

2.2 La méthode capture-recapture à deux sources 4

2.2.1 Principe 4

2.2.2 Estimation du nombre total de cas avec deux sources 4

2.2.3 Conditions d'applications 5

2.3 Les bases de données 5

2.3.1 La déclaration obligatoire du sida 5

2.3.2 La French Hospital Database on HIV 6

2.3.3 La base du Groupe d'épidémiologie clinique du sida en Aquitaine 6

2.3.4 Recherche d'une troisième base 7

2.4 Évaluation de la dépendance entre les sources 7

2.5 Travail préalable et harmonisation des bases 7

2.6 Identifications des cas communs 8

2.6.1 Préambule 8

2.6.2 Algorithme pour la recherche des cas communs 9

3. Résultats 10

3.1 Identification des cas communs 10

3.2 Taux d'exhaustivité pour la période 2004-2006 13

3.3 Taux d'exhaustivité par année de diagnostic 13

3.4 Taux d'exhaustivité par centre 13

3.5 Taux d'exhaustivité par région 13

3.6 Évaluation qualitative de la dépendance entre les sources 17

4. Discussion 18

5. Conclusion 21

Références bibliographiques 22

Annexes 24

Estimation de l'exhaustivité de la surveillance du sida par la méthode capture-recapture, France, 2004-2006

Rédacteurs

Guillaume Spaccaferri[1], Françoise Cazein[1], Laurence Lièvre[2], Stéphane Geffard[3], Anne Gallay[1], Josiane Pillonel[1]

1/ Institut de veille sanitaire, Saint-Maurice
2/ Institut national de la santé et de la recherche médicale, Unité 943, Paris, UPMC Université Paris 06, UMR S943, Paris
3/ Institut national de la santé et de la recherche médicale, Unité 897, Bordeaux

Nous remercions Pascale Bernillon pour son soutien en biostatistique et Florence Lot pour son expérience sur une étude du même type.

Ce rapport a été rédigé par Guillaume Spaccaferri dans le cadre du master 2 Science du risque dans le domaine de la santé de l'Université d'Auvergne. Il s'est déroulé de mars à août 2009 au sein de l'Unité "VIH-IST-Hépatites C et B chronique" du département des maladies infectieuses de l'Institut de veille sanitaire et a été soutenu courant septembre 2009. La validation de ce rapport par l'Université d'Auvergne a permis l'obtention du diplôme de master 2.

Abréviations

AES	Accident d'exposition au sang
ALD	Affection de longue durée
CIM	Classification internationale des maladies
Cisih	Centres d'information et des soins de l'immunodéficience humaine
Cnil	Commission nationale de l'informatique et des libertés
Corevih	Comités de coordination régionale de la lutte contre le virus de l'immunodéficience humaine
Ddass	Direction départementale des affaires sanitaires et sociales*
Dmac	Dossier médical minimum anonyme commun
DMI2	Dossier médical épidémiologique et économique de l'immunodéficience humaine
DO	Déclaration obligatoire
FHDH	French Hospital Database on HIV
Gesca	Groupe d'épidémiologie clinique du sida en Aquitaine
Inserm	Institut national de la santé et de la recherche médicale
InVS	Institut de veille sanitaire
Misp	Médecin inspecteur de santé publique
OMS	Organisation mondiale de la santé
PMSI	Programme de médicalisation des systèmes d'information
TEC	Technicien d'étude clinique

** Depuis le 1er avril 2010, les Ddass ont été intégrées dans les Agences régionales de santé (ARS), sous le nom de Délégation territoriale de l'ARS.*

1. Introduction

Le sida est sans conteste la maladie qui a le plus marqué le dernier quart de siècle. Cette pandémie, plus qu'aucune autre, a fortement impacté la société, en exacerbant notamment la pauvreté et l'exclusion sociale. En 2007, on estimait à 33 millions le nombre de personnes séropositives à travers la planète (2,7 millions de nouvelles contaminations en 2007) et à 2 millions le nombre de décès imputables au sida [1]. Bien qu'aucune région du monde ne soit épargnée, on constate que l'Afrique subsaharienne demeure le continent le plus touché, comptant pour les deux tiers du nombre total mondial de personnes séropositives au VIH (22 millions) et les trois quarts de tous les décès dus au sida. L'Asie constitue le second foyer le plus important, avec 5 millions de personnes vivant avec le VIH en 2007, dont 380 000 ont été infectées dans l'année.

En France, en 2007, le nombre de personnes ayant découvert leur séropositivité a été estimé à 6 500 et le nombre de nouveaux cas de sida à 1 200 [2]. Les homosexuels masculins et les personnes d'Afrique subsaharienne sont les deux sous-groupes les plus touchés. En 2007, les homosexuels représentaient 38 % des cas de séropositivité et 26 % des cas de sida, et les rapports homosexuels demeurent le seul mode de contamination pour lequel le nombre de découverte de séropositivité a augmenté depuis le début de la surveillance du VIH en 2003. En 2007, les personnes de nationalité étrangère représentent 40 % des découvertes de séropositivité, dont 71 % sont de nationalité d'un pays d'Afrique subsaharienne (ce pourcentage est de 82 % pour les femmes et de 52 % des hommes).

Dans le but de définir des stratégies efficaces de santé publique reposant sur la prévention et la lutte, il est nécessaire de disposer de nombreuses données. L'épidémiologie est la science qui étudie, à l'intérieur d'une population donnée, les tendances évolutives temporelles et spatiales d'une maladie et les facteurs susceptibles d'en influencer l'apparition, la fréquence, la distribution et l'évolution [3]. Dans cette optique, le recours aux systèmes de surveillance semble pérenne et nécessaire. Quel que soit le système de surveillance, la question de la qualité des données recueillies se pose, et notamment celle de l'exhaustivité. La non-exhaustivité peut être source de biais, à la fois pour suivre l'évolution temporo-spatiale de la maladie, mais aussi pour l'estimation de l'incidence et de la prévalence de l'infection que l'on surveille, ainsi que pour l'analyse des caractéristiques des cas. La méthode capture-recapture est une méthode qui permet d'estimer le taux d'exhaustivité d'un système de surveillance. Son intérêt repose essentiellement sur l'utilisation de plusieurs sources de données non exhaustives pour une maladie et se substitue ainsi aux systèmes de recueil exhaustif, lourd à mettre en œuvre et le plus souvent onéreux.

La méthode capture-recapture permet d'estimer la taille d'une population étudiée et le niveau de couverture du système de surveillance. La méthode de base a une longue histoire et a été appliquée dans de nombreux domaines scientifiques. Tout d'abord, cette méthode fut utilisée en zoologie, les écologistes en ayant fait une référence dans l'estimation des populations fauniques [4-6]. La méthode consiste à capturer, marquer et relâcher un échantillon aléatoire d'une population animale, puis de capturer un second échantillon. L'estimation de la taille totale de la population se fait grâce au pourcentage d'animaux marqués, présent dans le second échantillon. La méthode capture-recapture a également été utilisée en démographie, afin d'estimer les sous-dénombrements dans les recensements, ou encore de déterminer les taux de natalité et de mortalité dans les pays en développement [7]. Aux États-Unis, cette méthode a été utilisée dès 1950 pour évaluer le sous-dénombrement [8] et adoptée dans le cadre du recensement de 1991 [9]. Depuis environ 25 ans, l'épidémiologie a repris cette méthode pour évaluer la prévalence et l'incidence d'un certain nombre de maladies : le cancer [10], l'infarctus du myocarde [11], l'infection par le VIH [12,13], le sida [14], les infections à méningocoque [15], les méningites bactériennes [16]. Cette application nécessite la présence d'au moins deux sources de données indépendantes où chaque base est assimilée à un échantillon aléatoire de capture animale.

En France, le système de surveillance du VIH/sida a été profondément modifié en 2003 avec le renforcement de la protection de l'anonymat des personnes et la mise en place de la déclaration obligatoire (DO) de l'infection à VIH, s'ajoutant à celle du sida qui existait depuis 1982. La dernière estimation de l'exhaustivité (sur la période 1990-1993) de la DO du sida a été réalisée en 1995 [14]. Cette étude a permis d'estimer la couverture de la DO du sida à 83,6 % (IC 95 % : 82,9-84,3). Il est probable que celle-ci ait évolué depuis, notamment en 2003 à l'occasion des modifications de la surveillance du VIH/sida.

Notre étude a été réalisée en collaboration avec l'Unité 943 de l'Institut national de la santé et de la recherche médicale (Inserm), responsable de la French Hospital Database on HIV (FHDH-ANRS-C04) et le Groupe d'épidémiologie clinique du sida en Aquitaine (Gesca-ANRS-C03), responsable du suivi de la cohorte hospitalière d'Aquitaine, coordonnée par l'Unité U897 de l'Inserm.

L'objectif de notre étude était de pouvoir fournir une nouvelle estimation de la sous-déclaration des cas de sida au sein de la DO du sida et de pouvoir estimer la couverture des deux bases de données hospitalières, la FHDH et le Gesca.

Cette étude a reçu un avis favorable de la Commission nationale de l'informatique et des libertés (Cnil – dossier n° 902305).

2. Matériels et méthodes

2.1 Population d'étude

Elle est constituée par tous les nouveaux cas de sida selon la définition de 1993 [17], chez des adultes de 18 ans et plus, diagnostiqués entre le 01/01/2004 et le 31/12/2006. Pour entrer dans la population d'étude, il faut que le diagnostic de sida ait été recensé par au moins un des trois systèmes d'informations suivant :

- la DO du sida (InVS) ;
- la FHDH-ANRS-C04 (Inserm U943) ;
- le Gesca-ANRS-C03 (Gesca).

La période d'étude s'arrête à 2006 afin de permettre l'inclusion des cas déclarés dans les deux ans suivant leur diagnostic, soit la quasi-totalité des cas inclus avec délai dans chacune des bases.

La DO du sida couvre l'ensemble du territoire français, mais la FHDH ne comprend pas les hôpitaux de la région Aquitaine. Pour avoir une homogénéité de capture par rapport à la zone géographique étudiée, nous avons inclus conjointement à la base FHDH les données de la cohorte du Gesca qui s'intéresse spécifiquement à la région Aquitaine.

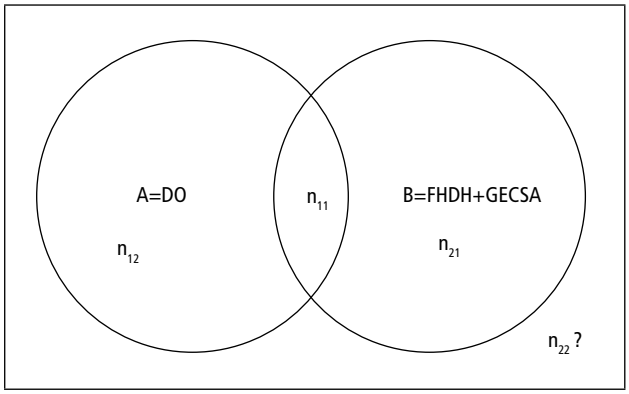
2.2 La méthode capture-recapture à deux sources

2.2.1 Principe

Cette méthode permet, sous certaines conditions (*cf.* paragraphe 2.2.3), en croisant les cas d'une maladie recensés par deux systèmes A et B (figure 1), et après avoir identifié les cas communs entre les deux systèmes, d'estimer le nombre de cas identifiés par aucun des systèmes (n_{22}), le nombre de cas total (N) de la maladie et l'exhaustivité de chaque système.

| Figure 1 |

Répartition des cas d'une maladie recensés par deux systèmes, A et B



Les indices 1 et 2 correspondent à la présence (1) et à l'absence (2) dans les sources A et B.

Soit pour notre étude :

- une population fermée de taille N inconnue constituée par les cas de sida sur la période du 01/01/2004 au 31/12/2006 ;
- N_A cas de sida déclarés dans le cadre de la DO ;
- N_B cas de sida recensés par la base constituée par la FHDH et le Gesca ;
- n_{11} cas communs aux deux bases ;
- n_{22} cas de sida non recensés par les deux systèmes.

2.2.2 Estimation du nombre total de cas avec deux sources

Le croisement des cas des deux sources de données est illustré par un tableau de contingence 2x2 (tableau 1). Les cas sont répartis en fonction de leur présence (1) ou de leur absence (2) dans l'une ou l'autre base.

| Tableau 1 |

Croisement des cas issus des deux sources de données

		Cas recensés par la DO		
		Oui	**Non**	**Total**
Cas recensés par la FHDH et le Gecsa	**Oui**	n_{11}	n_{21}	N_B
	Non	n_{12}	n_{22}=?	?
	Total	N_A	?	N=?

Sous l'hypothèse d'indépendance des sources, Sekar et Deming [4] proposent des estimateurs (estimateurs du maximum de vraisemblance) pour :

- le nombre de cas qui ne sont identifiés par aucune des deux sources (n_{22}) ;
- le nombre total de cas (N) ;
- la variance de N et son intervalle de confiance à 95 %.

$$n_{22} = \frac{n_{12} \times n_{21}}{n_{11}} \quad (1) \qquad \text{Var}(N) = \frac{N_A \times N_B \times n_{12} \times n_{21}}{n_{11}^3} \quad (2)$$

$$N = \frac{N_A \times N_B}{n_{11}} \quad (3) \qquad \text{IC } 95\% \; N = N \pm 1{,}96 \times \sqrt{\text{Var}(\hat{N})} \quad (4)$$

Chapman et Seber [18,19] ont montré que ces estimateurs pouvaient être biaisés lorsque les effectifs sont faibles et que n_{11} a une probabilité non nulle de tendre vers zéro. Les estimateurs non biaisés qu'ils proposent sont notamment appropriés dans le cas où un des systèmes ne recense qu'une partie de la population couverte par le premier système. Ceci correspond à notre situation, puisque certains hôpitaux de la base DO ne sont pas présents dans la base FHDH et c'est d'autant plus le cas lorsque les estimations sont réalisées par centre ou par région.

Les estimateurs non biaisés qui seront utilisés dans notre étude sont les suivants :

$$N = \left[\frac{(N_A + 1) \times (N_B + 1)}{n_{11} + 1} \right] - 1 \quad (5)$$

$$\text{Var (N)} = \frac{(N_A + 1) \times (N_B + 1) \times n_{12} \times n_{21}}{(n_{11} + 1)^2 \times (n_{11} + 2)} \quad (6)$$

Les taux d'exhaustivité (E) des sources sont :

$$E_A = \frac{N_A}{N} \quad \text{et} \quad E_B = \frac{N_B}{N} \quad (7)$$

où N_A et N_B sont le nombre de cas notifiés dans chacune des sources et N le nombre total de cas estimés.

Le nombre de cas distincts recensés par l'ensemble des systèmes (noté *n*) est obtenu par la formule :

$$n = n_{11} + n_{12} + n_{21} = N_A + N_B - n_{11} \quad (8)$$

2.2.3 Conditions d'applications

La validité des résultats obtenus est conditionnée au respect de certaines conditions d'application de la méthode [20-21].

• Tous les cas identifiés sont des vrais cas
L'identification de faux cas par une des sources induira une surestimation du nombre de cas total (N) et donc une sous-estimation de l'exhaustivité de toutes les sources. Ce problème peut se rencontrer lorsque la définition de cas au sein des différentes sources n'est pas identique.

• Les cas sont recensés sur la même période et la même zone géographique pour toutes les sources
Si une source identifie des cas dans une zone géographique ou au cours d'une période différente de l'autre source, le nombre de cas communs identifiables sera plus faible, ce qui entraînera une surestimation du nombre de cas total (N) et une sous-estimation de l'exhaustivité.

• La population étudiée est close
La population doit être close, c'est-à-dire qu'il n'y a pas de mouvement de population, ni d'entrée (ex : naissance) ni de sortie (ex : décès). Le non-respect de cette condition équivaut à un défaut d'homogénéité de capture et pourra entraîner un biais dans l'estimation du nombre de cas total (N) et donc de l'exhaustivité.

• Tous les vrais cas communs et seulement les vrais cas communs aux sources sont identifiés
Une surestimation des cas communs (n_{11}) induirait une sous-estimation du nombre total de cas (N) et réciproquement.

• Les sources sont indépendantes
Cette condition suppose que la probabilité qu'un cas soit recensé dans une source ne dépende pas de la probabilité que ce cas soit recensé dans une autre source. On évoquera une dépendance positive lorsque l'identification d'un cas par un système augmente la probabilité pour ce cas d'être recensé par l'autre système. La dépendance positive entraîne une sous-estimation de N. À l'inverse, on évoquera une dépendance négative entre deux bases lorsque l'identification d'un cas par un système diminue la probabilité de recensement de ce cas par l'autre système. La dépendance négative induit une surestimation de N.

• L'homogénéité de capture des cas
Tous les cas de la population doivent avoir la même probabilité d'être identifiés au sein d'une même source. La capture des cas dans une source ne doit donc pas être liée à certaines de leurs caractéristiques.

Les possibles hétérogénéités de capture peuvent être prises en compte en stratifiant sur ces variables, afin de créer des strates de probabilité de capture homogène et réduire les biais pour l'estimation de N.

Dans notre étude, deux variables ont été identifiées comme pouvant être des variables d'hétérogénéité :
- l'année de diagnostic : la dernière année de notre période d'étude (2006) étant plus proche de la date à laquelle ont été constituées les bases de données, il est possible que l'année 2006 soit incomplète (délai de notification) ;
- la région de notification des cas de sida : dans la DO du sida, il existe une disparité de la notification des cas de sida en fonction de la zone géographique.

Une stratification a été donc effectuée sur les variables, "année de diagnostic" et "région de notification des cas de sida". L'estimation de l'exhaustivité globale de chacune des sources a tenu compte de l'hétérogénéité de capture selon l'année de diagnostic. L'estimation du nombre total de cas était égale à la somme des estimations faites pour chacune des trois strates d'année de diagnostic du sida. La stratification combinée année et région n'a pas été possible compte tenu des effectifs insuffisants dans la plupart des strates.

2.3 Les bases de données

Nous appliquons dans cette étude une méthode capture-recapture à deux sources. La première source est composée des données issues de la DO et la seconde source est composée de deux bases : la French Hospital Database on HIV (FHDH) et le Gecsa.

2.3.1 La déclaration obligatoire du sida

Le système de DO, revu en 2003 pour renforcer l'anonymat des personnes, permet de détecter et de déclarer 30 maladies, afin de suivre l'évolution de leur incidence dans le temps et leurs caractéristiques épidémiologiques (annexe 1). Ce système revêt également un intérêt important dans l'évaluation des politiques de santé publique relatives à ces maladies.

Ces 30 maladies sont à DO car elles nécessitent, soit une intervention urgente (ex : méningite), soit un suivi nécessaire à la mise en place des stratégies de prévention et de prise en charge (ex : sida ou virus de l'hépatite B).

Le système de DO s'articule autour de trois acteurs :
- les déclarants : les cliniciens et les biologistes ;
- les médecins inspecteurs de santé publique (Misp) des Directions départementales des affaires sanitaires et sociales (Ddass) ;
- les épidémiologistes de l'Institut de veille sanitaire (InVS).

Le sida fait partie de ces 30 maladies à DO. Le notification des cas de sida existe depuis 1982 et a été rendue obligatoire en 1986.

Les fiches de DO sont remplies par les cliniciens en quasi-totalité hospitaliers, lors du diagnostic du sida. Le clinicien établit une fiche de DO après avoir calculé un code d'anonymat unique pour chaque patient. En pratique, les techniciens d'études cliniques (TEC) remplissent parfois les DO.

Les fiches de DO sont transmises *via* les Ddass à l'InVS où avant leur saisie, les données sont soumises à différentes procédures de validation visant à corriger les informations incohérentes et éliminer les doublons.

La fiche de déclaration comporte des données sociodémographiques et des informations médicales. Les variables de la DO du sida utilisées dans notre étude sont les suivantes : année de naissance, sexe, nationalité, département de domicile au moment du diagnostic de sida, hôpital où a eu lieu le diagnostic, date de première sérologie positive (mois et année), date de diagnostic du sida (mois et année), pathologies inaugurales, mode probable de contamination, date de mesure des CD4 (mois et année), nombre de CD4 et, le cas échéant, la date de décès (jour, mois et année).

La répartition des cas de sida par mode de contamination se fait selon huit catégories hiérarchisées : 1 – Homo/bisexuels masculins, 2 – Toxicomanes, 3 – (1) et (2), 4 – Hémophiles et troubles de la coagulation, 5 – Hétérosexuels, 6 – Transfusés, 7 – Transmission materno-fœtale, 8 – Autres, inconnu. Un cas ne peut être classé qu'à l'intérieur d'un seul mode de contamination. Dans le cas où un sujet présenterait plusieurs modes possibles, il sera classé dans celui situé le premier dans la hiérarchie. La catégorie 5 rassemble les sujets hétérosexuels ayant un partenaire sexuel à risque, qu'il soit bisexuel, hémophile, transfusé, originaire d'Afrique subsaharienne, des Caraïbes ou d'Asie, les sujets ayant un partenaire séropositif sans autre précision et les personnes dont le seul mode d'exposition au virus sont les rapports hétérosexuels, les autres modes de contamination étant exclus. La catégorie 8 regroupe les sujets pour lesquels le groupe de transmission n'est pas connu, les sujets où aucune situation à risque n'a pu être classée dans les catégories 1 à 6, les sujets dont le mode de contamination est en cours d'investigation et les sujets contaminés suite à un accident d'exposition au sang (AES).

Pour notre étude, afin de tenir compte des délais de déclaration qui peuvent aller jusqu'à 2 ans, la base utilisée est celle des cas de sida déclarés jusqu'au 30 septembre 2008 pour les cas diagnostiqués sur la période 2004-2006 (soit un recul de presque 2 ans).

2.3.2 La French Hospital Database on HIV

La FHDH a été créée en 1992 avec l'ensemble des données recueillies dans le dossier médical épidémiologique et économique de l'immunodéficience humaine (DMI2), faisant lui-même suite à la fusion du DMI et du dossier médical minimum anonyme commun (Dmac). La base est gérée au sein de l'Unité Inserm U943. Son objectif est de contribuer à la surveillance épidémiologique de l'infection VIH en France en étudiant l'histoire clinique de la maladie en mettant en relation les résultats biologiques et cliniques. Les Comités de coordination régionale de la lutte contre le virus de l'immunodéficience humaine (Corevih), qui font suite aux Centres d'information et des soins de l'immunodéficience humaine (Cisih), ont la responsabilité d'assurer la complétude et le suivi des données. La base comprend les données de 29 Corevih. Le Corevih d'Aquitaine dispose de sa propre base de données dénommée Gecsa.

La FHDH est une cohorte hospitalière ouverte et multicentrique. Pour faire partie de la base, les patients doivent être infectés par le VIH-1 ou le VIH-2, être suivis dans un Corevih et avoir donné leur consentement éclairé par écrit.

Les données sont recueillies à partir des dossiers médicaux des patients et saisies par les TEC en poste dans les hôpitaux. La saisie se fait *via* le logiciel DMI2. Dans le cas où les hôpitaux utilisent un autre logiciel (NADIS ou DIAMMG), ils doivent extraire leurs données afin qu'elles puissent être ajoutés dans la FHDH.

Lors de la première consultation dans un Corevih, un dossier patient est établi comprenant les données sociodémographiques invariables (nom, prénom, date de naissance, sexe, département de naissance…). Puis, à chaque consultation ou hospitalisation (classique ou de jour), les informations cliniques, biologiques et thérapeutiques sont recueillies dans une fiche de suivi datée. Deux fois par an, les TEC procèdent à une extraction des données. Chaque patient est indexé par un numéro d'anonymat. Les informations sont cryptées et transmises à l'U943. Avant leur incorporation, les données sont soumises à différentes procédures de validation visant à corriger les informations incohérentes.

Pour notre étude, portant sur les cas de sida chez les adultes de 18 ans et plus diagnostiqués sur la période 2004-2006, seul un sous-ensemble (utilisé pour la recherche des cas communs à la DO du sida) des données renseignées dans la FHDH a été utilisé. Les variables retenues sont les suivantes : année de naissance, sexe, département de domicile, date de première sérologie positive (mois et année), séjour de plus de six mois hors de France, groupe de transmission, pathologies inaugurales, date de sida (mois et année), date de mesure des CD4 (mois et année), nombre de CD4, hôpital et Corevih où a eu lieu le diagnostic et, le cas échéant, la date de décès (jour, mois et année). La variable "séjour hors de France de plus de six mois" a été retenue en raison de son rapprochement possible avec la variable "nationalité", présente dans les deux autres bases de données. En effet, lorsque pour un patient de la FHDH la variable "séjour hors de France" était saisie, cela pouvait correspondre à sa nationalité. En revanche, lorsque cette variable était vide, nous ne pouvions pas considérer que le patient était de nationalité française.

En ce qui concerne les modes de contamination, la classification est la même que celle de la DO du sida.

Pour notre étude, la base utilisée est celle qui nous a été donnée par la FHDH au 31 mars 2009.

2.3.3 La base du Groupe d'épidémiologie clinique du sida en Aquitaine

Il s'agit d'un système d'information crée en 1987, commun au Gecsa et au Corevih d'Aquitaine, qui a permis la constitution d'une base de données régionale et la création de la cohorte ANRS-C03 Aquitaine coordonnée par l'Inserm U897.

Les critères d'inclusion dans le système d'information Gecsa sont aux nombres de quatre : le patient doit avoir au moins 13 ans à l'inclusion, il doit être infecté par le VIH-1, avoir été vu au moins une fois en consultation ou en hospitalisation dans un des services participants, et avoir donné son consentement éclairé.

À l'inclusion, le dossier comprend les variables de type unique : caractéristiques sociodémographiques, mode de contamination, date de sérologie... À chaque suivi, les données évolutives (thérapeutique, biologique, sérologique…) sont de nouveau renseignées. La collecte des données, leur codage et leur vérification se fait en collaboration entre les médecins cliniciens, les TEC, les attachés de recherche clinique et les internes de santé publique. Une recherche de doublons est effectuée systématiquement, une recherche des perdus de vue et un contrôle de qualité sont réalisés chaque année sur un échantillon tiré au sort.

Pour notre étude, les variables retenues sont les mêmes que pour la base FHDH, à l'exception de la variable "séjour hors de France de plus de six mois" puisque la variable "nationalité" existe dans le Gecsa. Les modes de contamination sont hiérarchisés de la même manière que pour la FHDH et la DO du sida.

La base utilisée pour notre étude est celle qui nous a été donnée par le Gecsa au 30 avril 2009.

2.3.4 Recherche d'une troisième base

La méthode capture-recapture à deux sources ne permet pas de prendre en compte la dépendance des sources dans les estimateurs. Nous avons tenté de trouver une troisième source, qui aurait permis d'évaluer statistiquement la dépendance des bases [22,23] et d'en tenir compte dans l'estimation du nombre total de cas (N).

Différentes possibilités ont été étudiées :

- la base de la DO du VIH a été envisagée. La DO du VIH a été mise en place en 2003 suite à la refonte du système de surveillance des maladies à DO. Malheureusement, cette base est beaucoup trop liée à la DO sida, ce qui ne respecte pas la condition d'indépendance des sources. En effet, chacune des bases peut être complétée ou modifiée à partir de l'autre, ce qui a rendu son utilisation impossible pour notre étude ;
- la cohorte Seroco a également été étudiée. Il s'agit d'une enquête épidémiologique consistant à recueillir, dans 17 centres hospitaliers (région parisienne, Nice et Marseille), les données cliniques et biologiques provenant de consultations de patients séropositifs pour le VIH et inclus dans l'enquête sur la base du volontariat. Hélas, la base de données comportait trop peu de cas de sida (une vingtaine par an) pour pouvoir être utilisée dans notre étude. De plus, sa couverture n'est pas nationale ;
- la base de l'Assurance maladie sur les affections de longue durée (ALD). L'ALD est défini à l'article L-324 du code de la sécurité sociale. Ce système permet à 7 millions d'assurés d'être pris en charge à 100 % par l'Assurance maladie. Il existe trois catégories d'ALD : ALD 30, ALD hors liste et les polypathologies. Les infections par le VIH sont traitées au sein de l'ALD 30 et sont regroupées au sein du code n° 7. Malheureusement, il n'est pas fait de distinction dans cette base entre les patients séropositifs et les patients au stade sida. Il ne nous était donc pas possible de constituer une base de données de patients ayant contracté un sida sur notre période d'étude ;
- dernière source de données envisagée : le Programme de médicalisation des systèmes d'information (PMSI). Son objectif principal est d'analyser l'activité des établissements pour en déterminer leurs ressources. Chaque patient fait l'objet d'un dossier administratif précisant le diagnostic principal et les actes médicaux réalisés pendant le séjour. Les pathologies sont codées grâce à la classification internationale des maladies (CIM-10) publiée par l'Organisation mondiale de la santé (OMS). Pour notre étude, il ne nous a pas été possible de procéder à une extraction des données relatives aux patients diagnostiqués sida, puisque cette base ne dispose pas de la date de la pathologie inaugurale classante sida. Il ne nous était donc pas possible de savoir la date à laquelle le patient est passé au stade sida.

2.4 Évaluation de la dépendance entre les sources

En l'absence d'une troisième source de données, nous n'avons pas pu évaluer de manière quantitative la dépendance entre les sources. À défaut, nous avons tenté de l'évaluer qualitativement, en faisant une enquête auprès des services chargés de renseigner les bases FHDH et Gecsa. Dans ce but, un questionnaire a été diffusé dans ces services (annexe 2). Il s'agissait notamment de savoir si une même personne renseignait à la fois la DO et la FHDH (ou le Gecsa) ou si la notification dans l'un des systèmes entraînait la notification dans l'autre. Le questionnaire a été envoyé aux 29 Corevih (56 hôpitaux qui avaient déclaré des cas sur la période d'étude aux services hospitaliers participant à la FHDH) et aux 7 hôpitaux participant au Gecsa. Les réponses recueillies ont été saisies et analysées à l'InVS.

2.5 Travail préalable et harmonisation des bases

Avant de pouvoir croiser les bases pour l'identification des cas communs, il a été nécessaire de retravailler ces bases pour obtenir des variables et des caractéristiques homogènes d'une base à l'autre. Les différentes étapes de ce travail, qui a constitué une part importante de l'étude, sont listées ci-après :

a) dans la base FHDH, 311 cas de la base initiale qui n'avaient pas de pathologies classantes sida et qui ne répondaient donc pas à la définition de cas pour l'étude ont été exclus. En effet, dans la DO, lorsque des fiches de notification arrivent à l'InVS sans pathologie, elles ne sont pas saisies dans la base ;

b) entre les bases, le codage des pathologies classantes n'était pas homogène. Il nous a donc fallu harmoniser ce codage (annexe 3). Par exemple, dans la FHDH, il y avait une distinction entre les rétinites à cytomégalovirus et les autres infections à cytomégalovirus, alors que dans la DO les infections à cytomégalovirus sont regroupées sous le même libellé, sans distinction de localisation. Nous avons donc réuni les deux pathologies de la FHDH au sein d'une seule variable.

   Pour les infections à mycobactérie, la FHDH distingue celles dues aux mycobactéries *Avium* et *Kansasii*, des infections causées par des mycobactéries identifiées (autres que *Avium* et *Kansasii*) ou non identifiées. Dans la DO, il existe une variable renseignant le fait d'avoir ou non contracté une infection à mycobactérie puis une seconde variable où est codé le type de mycobactéries à l'origine de l'infection. Grâce à ce code, nous avons pu créer deux variables dans la DO correspondant à celles de la FHDH.

   Dans la DO, il existe une variable "code lymphome" où est renseignée la localisation du lymphome non hodgkinien. Cette précision n'est pas renseignée dans la FHDH, la localisation du lymphome n'a donc pas été prise en compte ;

c) dans la FHDH, nous disposions de deux dates pour la date de sida : une date locale et une date calculée. La date locale correspond à la date renseignée par le TEC au niveau du Corevih et la date

calculée à celle validée par l'U943 de l'Inserm. Ces deux dates sont identiques dans la majorité des cas (83 %). La date calculée peut cependant être antérieure à la date locale lorsque l'U943 de l'Inserm constate, à l'occasion de la fusion des bases des différents Corevih, que le patient avait déjà une pathologie classante sida. Dans ce cas, la date calculée est celle de la pathologie retrouvée dans les antécédents du patient.

La date calculée peut à l'inverse être postérieure à la date locale lorsque le patient avait été inclus sans pathologie dans la base (à la date locale), donc lors de l'apparition d'une pathologie classante sida, la date calculée a été complétée par la date de survenue de la pathologie.

N'ayant qu'une seule date dans la DO et dans la base du Gecsa, nous ne pouvions garder ces deux dates. La date calculée dans la FHDH est plus proche de la date de diagnostic dans la DO que la date locale : en cas d'antécédent de pathologies classantes dans la DO, c'est la date de la première pathologie qui est prise en compte ; par ailleurs, aucun cas n'est pris en compte dans la DO avant que la première date de pathologie classante ne soit renseignée. Nous avons donc retenu la date calculée et seulement dans le cas où celle-ci n'était pas renseignée (2 %), nous avons pris la date locale ;

d) en ce qui concerne la date de mesure des CD4, le même problème a été rencontré. Dans certains cas, nous disposions d'une date de mesure correspondant à la date de sida locale différente de celle de la date de sida calculée. Pour rester homogène avec le choix fait précédemment, nous avons retenu la date de mesure des CD4 correspondant à la date de sida calculée et, si celle-ci était manquante, nous avons gardé la date de mesure locale ;

e) la FHDH procède à un suivi des patients infectés par le VIH. À chaque visite du patient dans l'un des centres, les informations relatives à sa maladie sont renseignées. Nous disposions donc des différentes pathologies contractées par le patient suite à son passage au stade sida. Dans le cas où le patient était suivi dans plusieurs centres, nous disposions des informations de chacun des hôpitaux. Dans la DO, seules les pathologies contractées au maximum dans le mois suivant le diagnostic inaugural de sida sont renseignées. Dans la DO, si un cas est déclaré successivement par deux services, seule la DO du service ayant déclaré le premier est renseigné dans la base (la seconde DO peut néanmoins servir à compléter les informations manquantes de la première DO).

   Trois cas de figure se sont alors présentés :
   - pour rester homogène, les pathologies avec un écart de plus d'un mois par rapport à la date de diagnostic n'ont pas été utilisées pour l'identification des cas communs ;
   - lorsque les pathologies étaient observées avec un écart inférieur à un mois mais dans un centre différent, nous avons décidé de conserver le centre ayant renseigné la pathologie à la date la plus ancienne. Si les pathologies étaient différentes entre les centres, elles ont été ajoutées à celles diagnostiquées dans le centre retenu ;
   - lorsqu'un même cas est renseigné à la même date, mais dans deux centres différents, nous avons choisi de retenir le centre le plus important en termes de nombre de cas de sida. Si les pathologies étaient différentes, celles du centre non retenu ont été ajoutées.

   Pour les deux derniers cas de figure, conscient du caractère arbitraire du choix qui a été fait, nous avons conservé les informations non retenues. Ces deux cas de figure ne concernaient cependant que 16 patients de la base FHDH ;

f) en ce qui concerne les hôpitaux dans la base FHDH, il arrivait que nous ne disposions pas du nom de l'hôpital, mais uniquement de la ville. Pour ces villes, nous avons donc effectué un recodage dans la DO, en regroupant les hôpitaux d'une même ville pour être homogène avec la base FHDH ;

g) bien que n'étant pas exactement superposable, l'hypothèse a été faite de rapprocher la variable "séjour hors de France de plus de 6 mois" de la FHDH à la variable "nationalité" de la DO et du Gecsa. Cependant, il existait des disparités de codage entre ces variables. Par exemple, la base FHDH possède un code unique pour la Belgique et le Luxembourg, ce qui n'est pas le cas dans la DO. Nous avons donc dû recoder certains pays au sein de la base DO ;

h) pour les modes de contamination, les différentes bases possèdent une variable "autre" et une variable "inconnu", mais avec un classement non totalement superposable entre les bases. Nous avons choisi de regrouper ces deux variables pour éviter les problèmes liés à un classement différent selon les bases.

L'ensemble de ces modifications et de ces recodages a été effectué avec le logiciel Stata® 9.0.

2.6 Identifications des cas communs

2.6.1 Préambule

Notre étude porte sur les cas de sida diagnostiqués entre le 01/01/2004 et le 31/12/2006. Cependant, un même cas pouvait être déclaré à des dates différentes dans chaque base. Pour tenter de retrouver le maximum de cas communs entre ces bases, nous avons pris en compte les cas diagnostiqués au dernier semestre 2003 et au premier semestre 2007. Ainsi, des cas "hors période" ont pu être comptabilisés et, inversement, des cas compris dans la période de l'une des bases ont pu être exclus de l'étude car finalement considérés hors période. Deux exemples permettent d'expliquer ces deux cas de figure :
- un cas diagnostiqué au cours du second semestre 2003 dans la base DO, couple avec un cas de 2004 dans la base FHDH : nous avons considéré que ce cas de 2004 était en fait un cas incident de 2003 et, à ce titre, il n'est plus pris en compte dans notre étude ;
- un cas diagnostiqué au cours du premier semestre 2007 dans la base DO, couple avec un cas de 2006 dans la base Gecsa : nous avons considéré que ce cas de 2007 était en fait un cas incident de 2006 et, à ce titre, il est pris en compte dans notre étude.

Pour évaluer l'exhaustivité de la base de la DO du sida, nous avons dû identifier les cas communs entre les cas de la base DO et ceux des bases FHDH et Gecsa.

En revanche, pour évaluer l'exhaustivité de la base FHDH, le croisement des cas ne s'est effectué qu'entre la base DO et la base FHDH.

Enfin, pour l'exhaustivité de la base Gecsa, nous avons croisé les cas de cette base avec les cas de la base DO ayant été diagnostiqués dans un hôpital d'Aquitaine. Par ailleurs, la base du Gecsa n'inclut pas les

patients infectés par le VIH-2. Pour l'estimation de l'exhaustivité du Gecsa, nous avons choisi d'exclure les cas infectés par le VIH-2 et recensés par un hôpital d'Aquitaine au sein de la base DO. Cela ne concerne cependant que 2 cas.

Initialement, entre le 01/07/2003 et le 31/06/2007, les bases comprenaient 5 009 cas pour la DO, 4 112 cas pour la FHDH et 216 cas pour le Gecsa.

Une étape d'identification des doublons intrasource a été mise en œuvre. Nous avons retrouvé 7 doublons dans la base DO, 8 dans la base FHDH et aucun dans la base Gecsa.

La taille des bases après élimination des doublons intrasources était la suivante :
- 5 002 cas pour la base de la DO ;
- 4 104 cas pour les bases FHDH ;
- 216 cas pour la base Gecsa.

L'ajout des données du Gecsa à celles de la FHDH pour estimer l'exhaustivité de la DO a nécessité une identification des cas communs entre ces deux bases. Nous disposions du mois de naissance pour ces deux bases, ce qui a considérablement facilité la recherche. Seuls 4 cas de la base Gecsa ont été retrouvés au sein de la base FHDH : les bases FHDH et Gecsa réunies avaient donc un total de 4 316 cas (4 320-4).

2.6.2 Algorithme pour la recherche des cas communs

La mise au point de l'algorithme a constitué la part la plus importante du travail. En effet, nous ne disposions pas d'identifiants communs uniques entre les différents systèmes, ce qui a rendu l'identification des cas communs plus complexe. Le détail de l'algorithme avec l'ensemble des requêtes informatiques (R1 à R20) ayant permis l'identification des cas communs est présenté dans l'annexe 4, nous ne présentons ici que le principe général de la démarche.

En premier lieu (requête R1), nous avons décidé de déterminer un ensemble de couples potentiels à partir de deux variables de base : le sexe et l'année de naissance. Tous les couples potentiels étaient donc caractérisés par une égalité stricte sur le sexe et l'année de naissance. C'est à partir de ces 368 080 couples potentiels que s'est déroulée l'identification des cas communs. À ce stade, un patient de la base DO pouvait être apparié avec plusieurs patients de la base composée par les données de la FHDH et du Gecsa, et réciproquement. L'objectif a ensuite été d'identifier, parmi ces couples potentiels, les vrais cas communs, s'ils existaient. L'identification d'un vrai cas commun parmi un ensemble de couples potentiels permettait d'éliminer les autres couples potentiels de ce groupe. La répétition de ce processus a abouti à diminuer progressivement le "pool" des couples potentiels. L'identification des cas communs reposait sur une définition évoluant d'une définition très spécifique vers une définition de plus en plus sensible.

Pour identifier les cas communs, une variable a été traitée à part car considérée comme très spécifique. Il s'agit de la date de décès, puisque nous disposions de l'année, du mois et du jour de décès. La première sélection de cas communs a donc porté sur les couples potentiels ayant une égalité stricte sur cette variable (R2).

Dans une seconde sélection (R3), les couples potentiels ont été considérés comme cas communs lorsqu'ils présentaient une égalité stricte sur l'ensemble des variables. Puis, au fur et à mesure (R4 et suivantes), nous avons élargi les critères d'identification des cas communs. Par exemple, nous avons autorisé une différence sur la date de mesure des CD4 ainsi que sur leur nombre. L'autre variable pour laquelle nous avons rapidement toléré des différences est la variable "nationalité". En effet, dans la base FHDH, il ne s'agit pas de la nationalité, mais d'un séjour hors de France de plus de six mois, pouvant donner une indication sur la nationalité. Lorsque "séjour hors de France" concordait avec la variable "nationalité" de la DO, cela a permis de conforter certains cas communs, mais des différences n'ont jamais permis d'en exclure.

Une des variables les plus spécifiques concernait les pathologies inaugurales du sida. Nous disposions de la première pathologie classante ainsi que de celles éventuellement déclarées le mois suivant. Dans notre étude, aucun patient n'avait plus de 5 pathologies inaugurales. Pour permettre l'identification des cas communs, nous nous sommes intéressés en premier lieu aux couples potentiels ayant une égalité stricte sur toutes les pathologies. Puis, nous avons élargi notre sélection aux couples potentiels ayant au moins une pathologie commune. Pour faciliter cette sélection, les pathologies étaient ordonnées par ordre de fréquence. Par exemple, un patient présentant plusieurs pathologies aura toujours en première pathologie celle dont la fréquence rencontrée au sein de nos bases était la plus grande (annexe 3).

L'algorithme devenant de plus en plus sensible plus les cas communs sélectionnés avaient des disparités, certaines vérifications ont alors été réalisées.

Tout d'abord, pour la date de décès, il arrivait qu'au sein d'un couple, un patient soit décédé dans une base et pas dans l'autre. Dans la base FHDH, nous disposions de la date de dernier suivi, ce qui nous a permis de vérifier qu'un patient décédé dans la base DO n'ait pas une date de suivi dans la base FHDH postérieure à la date de décès indiquée dans la DO. Cela ne s'est jamais produit, ce qui a conforté la fiabilité de notre algorithme.

Ensuite, nous avons comparé les différences existantes sur l'ensemble des variables pour les cas communs identifiés à partir de la date de décès identique (R2) et, pour les cas communs identifiés, à la fin de l'algorithme (entre R3 et R20). Cette analyse de sensibilité sur les variables nous a permis de faire le choix de l'algorithme permettant d'identifier au mieux les cas communs.

Les distributions de fréquence ont été comparées avec le test du χ^2 de Pearson en plaçant le seuil de significativité (p) à 5 %. Par ailleurs, afin de déterminer l'existence de tendances sur la période d'étude, le test du χ^2 d'Armitage pour l'analyse de tendance linéaire a été utilisé.

L'ensemble de l'algorithme et de la recherche de cas communs s'est effectué à partir du logiciel Access® version 2003.

3. Résultats

3.1 Identification des cas communs

DO *vs* FHDH-Gecsa

Le croisement de la base DO du sida avec les bases FHDH-Gecsa a permis l'identification de 2 824 cas communs, dont 2 204 (n_{11}) sur notre période d'étude 2004-2006. Cette différence de 620 cas communs correspond aux cas communs hors période.

Le nombre total *n* de cas distincts recensés est de 4 940 sur la même période.

La base DO comprenait 3 816 cas sur la période d'étude (N_A) et la base composée des données de la FHDH et du Gecsa comprenait 3 328 cas (N_B). Les nombres de cas présents dans les bases diffèrent de ceux présentés précédemment (*cf.* paragraphe 2.6.1), ce qui s'explique par l'exclusion des cas "hors période".

DO *vs* FHDH

Le croisement de la base DO avec la base FHDH a permis l'identification de 2 124 cas communs sur la période d'étude 2004-2006 (n_{11}). Le nombre total n de cas distincts recensés est de 4 873 sur cette même période.
La base DO comprenait 3 816 cas et la base FHDH 3 181 cas.

DO *vs* Gecsa

Pour le croisement spécifique de la base Gecsa, nous avons sélectionné, au sein de la base DO, les cas diagnostiqués dans les hôpitaux d'Aquitaine.
Cela a permis l'identification de 83 cas communs, les bases DO et Gecsa étant composées respectivement de 92 et 151 cas sur la période d'étude.
Le nombre total n de cas distincts recensés est de 160.

Concordance et discordance des cas communs

Pour évaluer la fiabilité de notre algorithme, nous avons comparé les concordances et discordances pour chacune des variables entre les cas communs identifiés grâce à la date de décès, variable considérée comme très spécifique (requête R2), et les cas communs identifiés à la fin de l'algorithme (entre les requêtes R3 et R20). Cette analyse de sensibilité a été effectuée sur la totalité des 2 824 cas communs identifiés sur la période allant du 01/07/2003 au 30/06/2007 (tableau 2), nous permettant ainsi de tester les discordances retrouvées avec un effectif plus important que les seuls cas de la période d'étude (2004-2006), ce qui augmente la puissance statistique. Les variables analysées sont le département de domicile (tableau 2A), le mode de contamination (tableau 2B), l'hôpital (tableau 2C), la date de sérologie positive (tableau 2D), la date de sida (tableau 2E) et les pathologies inaugurales. En ce qui concerne les pathologies, nous avons regardé les cas communs ayant une égalité stricte sur l'ensemble des pathologies (tableau 2F) et ceux ayant au moins une pathologie commune (tableau 2G).

| Tableau 2A |

Concordances et discordances observées entre les cas communs des bases DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "département de domicile"

	Département de domicile concordant		Département de domicile discordant		Département de domicile inconnu		Total N
	N	%	N	%	N	%	
Cas communs après R2	236	84,0	26	9,3	19	6,8	281
Cas communs entre R3 et R20	1 959	77,0	368	14,5	216	8,5	2 543
Total	**2 195**	**77,7**	**394**	**14,0**	**235**	**8,3**	**2 824**

χ^2 global, p=0,02.

Sur l'ensemble des cas communs, 77,7 % ont le même département de domicile, pour 14 % des cas communs le département est différent et pour 8,3 % le département était inconnu (tableau 2A). Cette répartition est significativement différente après la requête R2 et la requête R20 (p=0,02). Ainsi, la proportion de cas pour lesquels le département de domicile est différent est plus élevée après la requête 20 (14,5 %) qu'après la requête 2 (9,3 %).

Le fait de trouver un département de domicile différent (14,0 %) peut s'expliquer par la différence sur le moment où est saisie l'information relative au département de domicile au sein des différentes sources. Pour la FHDH et le Gecsa, les données sociodémographiques sont saisies à l'entrée du patient au sein de la base, soit au moment du diagnostic du VIH, alors que pour la DO, ces données sont saisies au moment du passage au stade sida et donc en général beaucoup plus tard. Le patient a donc pu changer de département entre le moment du diagnostic VIH et celui du sida. L'Inserm estime à 15 % le nombre de patients ayant changé de département de domicile au sein de la base FHDH, ce qui nous a encouragé à être moins restrictifs sur cette variable.

| Tableau 2B |

Concordances et discordances observées entre les cas communs de la base DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "mode de contamination"

	Mode de contamination concordant		Mode de contamination discordant		Mode de contamination inconnu		Total N
	N	%	N	%	N	%	
Cas communs après R2	214	76,2	11	3,9	56	19,9	281
Cas communs entre R3 et R20	2016	79,3	123	4,8	404	15,9	2543
Total	**2 230**	**79,0**	**134**	**4,7**	**460**	**16,3**	**2 824**

χ^2 global, p=0,19.

Sur l'ensemble des cas communs, seuls 4,7 % ont un mode de contamination discordant (tableau 2B). Il n'existe pas de différence significative entre les cas communs identifiés après la requête R2 (3,9 %) et ceux identifiés à la fin de l'algorithme (4,8 %).

| Tableau 2C |

Concordances et discordances observées entre les cas communs de la base DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "hôpital"

	Hôpital concordant		Hôpital discordant		Total N
	N	%	N	%	
Cas communs après R2	270	96,1	11	3,9	281
Cas communs entre R3 et R20	2318	91,2	225	8,8	2543
Total	**2 588**	**91,6**	**236**	**8,4**	**2 524**

χ^2 global, p=0,005.

Les cas communs possèdent une égalité sur l'hôpital dans 91,6 % (tableau 2C). L'égalité sur cette variable est significativement plus importante (p=0,005) pour les cas communs identifiés à partir de l'égalité sur la date de décès.

| Tableau 2D |

Concordances et discordances observées entre les cas communs de la base DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "date de sérologie positive"

	Date de sérologie positive concordante		Date de sérologie positive discordante								Date de sérologie positive inconnue		Total N
			<1 mois		1-3 mois		3-6 mois		>6 mois				
	N	%	N	%	N	%	N	%	N	%	N	%	
Cas communs après R2	205	73	16	5,7	7	2,5	12	4,3	36	12,8	5	1,8	281
Cas communs entre R3 et R20	1 785	70,2	209	8,2	96	3,8	111	4,4	312	12,3	30	1,2	2 543
Total	**1 990**	**70,5**	**225**	**8**	**103**	**3,6**	**123**	**4,4**	**348**	**12,3**	**35**	**1,2**	**2 824**

χ^2 global (sans les inconnus), p=0,48.

Plus de deux tiers (70,5 %) des cas communs ont une égalité stricte sur la date de sérologie positive (tableau 2D). Si l'on étend à une différence inférieure à six mois, on passe à 86,5 %. Pour la date de sérologie positive, la faible concordance (70,5 %) peut s'expliquer par le délai séparant la date de sérologie de la date de diagnostic sida. En effet, pour les cas de sida diagnostiqués entre 2004 et 2006, le diagnostic VIH peut remonter à de nombreuses années, il est donc possible que la date indiquée pour la DO sida soit différente de celle renseignée dans la FHDH ou le Gecsa. Cette précision sur la date de sérologie est d'autant plus difficile que l'événement rapporté est lointain.

La répartition des cas communs selon une égalité stricte ou une discordance allant de moins de 1 mois à plus de 6 mois n'était pas significative (p=0,48) pour cette variable entre la requête R2 et la requête R20.

| Tableau 2E |

Concordances et discordances observées entre les cas communs de la base DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "date de sida"

	Date de sida concordante		Date de sida discordante						Total N
			<1 mois		1-3 mois		3 mois		
	N	%	N	%	N	%	N	%	
Cas communs après R2	195	69,4	56	19,9	17	6	13	4,6	281
Cas communs entre R3 et R20	1 947	76,6	395	15,5	116	4,6	85	3,3	2 543
Total	**2 142**	**75,8**	**451**	**16**	**133**	**4,7**	**98**	**3,5**	**2 824**

χ^2 global, p=0,07.

Sur l'ensemble des cas communs, 75,8 % ont une égalité stricte sur la date de sida et ce chiffre s'élève à 96,5 % pour une différence inférieure à trois mois (tableau 2E).

Il n'y a pas de différence significative (p=0,07) sur la répartition des cas communs, selon une égalité stricte ou une discordance allant de moins de un mois à plus de trois mois, à la requête R2 et à la requête R20.

| Tableau 2F |

Concordances et discordances observées entre les cas communs de la base DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "pathologies"

	Pathologies identiques		Pathologies non identiques		Total
	N	%	N	%	
Cas communs après R2	213	75,8	68	24,2	281
Cas communs entre R3 et R20	1 960	77,1	583	22,9	2 543
Total	**2 173**	**76,9**	**651**	**23,1**	**2 824**

χ^2, p=0,63.

Dans près de 77 % des cas, les cas communs ont des pathologies identiques (tableau 2F).

On ne note pas de différence significative (p=0,63) entre les cas communs de la requête R2 et les cas communs identifiés à la fin de l'algorithme pour cette variable.

| Tableau 2G |

Concordances et discordances observées entre les cas communs de la base DO et de la base FHDH-Gecsa réunie obtenus à la requête 2 et entre les requêtes 3 à 20 pour la variable "au moins une pathologie commune"

	Au moins une pathologie commune		Pas de pathologie commune		Total
	N	%	N	%	
Cas communs après R2	252	89,7	29	10,3	281
Cas communs entre R3 et R20	2 371	93,2	172	6,8	2 543
Total	**2 623**	**92,9**	**201**	**7,1**	**2 824**

χ^2, p=0,04.

Près de 93 % des cas communs possèdent au moins une pathologie commune (tableau 2G). La proportion de cas communs avec au moins une pathologie en commun est significativement supérieure (p=0,04) pour les cas communs identifiés à la fin de l'algorithme.

3.2 Taux d'exhaustivité pour la période 2004-2006

En tenant compte de l'hétérogénéité de capture concernant l'année de diagnostic, l'application de la méthode capture-recapture a permis d'estimer le nombre total de cas de sida diagnostiqués entre 2004 et 2006 à 5 770 (IC 95 % : 5 679-5 861) en France. Ce nombre est obtenu en croisant la base DO avec les bases FHDH et Gecsa réunies (tableau 3).

L'exhaustivité de la DO est estimée à 66,1 % (IC 95 % : 65,1-67,2) (tableau 3). Pour la base FHDH, l'exhaustivité est estimée à 55,6 % (IC 95 % : 54,7-56,5) (tableau 4) ; pour la base du Gecsa, elle est de 90,3 % (IC 95 % : 86,4-94,5) (tableau 5).

Le regroupement des trois systèmes de surveillance permet de recenser 85,7 % (IC 95 % : 84,4-87,1) du nombre total de cas.

3.3 Taux d'exhaustivité par année de diagnostic

La stratification sur l'année de diagnostic du sida (tableaux 3, 4 et 5) mettait en évidence une diminution significative de l'exhaustivité de la DO du sida au cours de notre période d'étude (χ^2 de tendance, p=0,032), passant de 67,5 % (IC 95 % : 66,0-69,2) en 2004 à 64,2 % (IC 95 % : 62,1-66,4 %) en 2006. Cette diminution était plus marquée pour la FHDH (χ^2 de tendance, $p<10^{-4}$), avec une exhaustivité de 60,0 % (IC 95 % : 58,6-61,5) en 2004 et de 49,7 % (IC 95 % : 48,1-51,6) en 2006.

En ce qui concerne le Gecsa, l'exhaustivité est stable au cours des trois années (χ^2 de tendance, p=0,3) autour de 90 %.

3.4 Taux d'exhaustivité par centre

L'analyse de l'exhaustivité par centre n'est présentée que pour la FHDH, pour laquelle elle est pertinente. En effet, nous avons pris comme référence les hôpitaux participant à la base, ce qui a exclu certains cas de la base DO. De plus, lorsqu'un cas commun était identifié dans des centres différents dans la DO et la FHDH, nous avons choisi de toujours prendre pour référence le centre de la FHDH (tableau 6).

Les taux de couverture par centre varient entre 16,8 % et 97,5 %. Les centres présentant les meilleurs taux de couverture sont Grenoble (97,5 %), Seine-Saint-Denis (95,0 %), la Réunion (93,9 %) et Rennes (90,5).

3.5 Taux d'exhaustivité par région

Pour la DO du sida, l'analyse de l'exhaustivité par région est beaucoup plus pertinente que celle par centre, puisque tous les cas de la DO peuvent être ainsi pris en compte. La référence est donc ici la DO. Les taux d'exhaustivité varient de 51,2 % à 94,7 % selon les régions. Pour certaines régions, le nombre de cas présents dans les bases étant trop faible, nous avons dû procéder à un regroupement géographique de régions (tableau 7). Après stratification sur la région, le nombre total de cas a été estimé à 5 730 pour la période 2004-2006 ; ce nombre était inclus dans l'intervalle de confiance à 95 % du nombre total de cas estimés en ne prenant en compte que la stratification par année.

Pour la région Île-de-France, les résultats sont également présentés par département (tableau 7 bis). Cependant, cela ne concerne que Paris et les départements de la Petite Couronne, puisque le nombre de cas présents au sein de la Grande Couronne n'est pas suffisant pour pouvoir appliquer la méthode. Le département des Hauts-de-Seine présente la meilleure exhaustivité (77,1 %). Les départements de Seine-Saint-Denis et du Val-de-Marne ont des résultats légèrement plus faibles, respectivement 73,0 % et 71,1 %, mais le département de Paris a une exhaustivité largement inférieure aux autres départements (61,7 %).

Pour les DOM, l'exhaustivité est également présentée par département. Les départements de la Réunion et de la Martinique ont les meilleurs taux de couverture, avec 90,0 % et 89,5 %, loin devant la Guadeloupe et la Guyane (72,8 % et 64,1 %).

| Tableau 3 |

Taux d'exhaustivité de la déclaration obligatoire (DO) du sida selon l'année de diagnostic, France, 2004-2006

Année	Nombre de cas observés			Nombre de cas estimés[a]		Taux d'exhaustivité de la DO*[b]	
	DO	FHDH Gecsa	Cas communs	N	IC 95 %	%	IC 95 %
Global	3 816	3 328	2 204	5 762	5 671-5 853	66,2	65,2-67,3
2004	1 373	1 263	853	2 033	1 985-2 080	67,5	66,0-69,2
2005	1 317	1 152	765	1 983	1 930-2 036	66,4	64,7-68,2
2006	1 126	913	586	1 754	1 695-1 813	64,2	62,1-66,4
Total[c]	**3 816**	**3 328**	**2 204**	**5 770**	**5 679-5 861**	**66,1**	**65,1-67,2**

* χ^2 *de tendance, p=0,032.*
[a] *Estimation de la population totale à l'aide de la formule 5 présentée dans le paragraphe matériel et méthodes. Les intervalles de confiance sont basés sur une approximation normale de N (formule 4) et la variance de N est proposée dans la formule 6.*
[b] *Estimation de l'exhaustivité des bases à l'aide de la formule 7. Les intervalles de confiance sont obtenus avec les bornes inférieures et supérieures de N.*
[c] *Sommes des estimations obtenues pour chaque strate.*

| Tableau 4 |

Taux d'exhaustivité de la base FHDH-ANRS CO4 selon l'année de diagnostic, France, 2004-2006

Année	Nombre de cas observés			Nombre de cas estimés[a]		Taux d'exhaustivité de la FHDH[b]	
	DO	FHDH	Cas communs	N	IC 95 %	%	IC 95 %
Global	3 816	3 181	2 124	5 715	5 621-5 808	55,7	54,8-56,6
2004	1 374	1 209	824	2 016	1 967-2 065	60,0	58,6-61,5
2005	1 316	1 103	740	1 961	1 908-2 015	56,2	54,7-57,8
2006	1 126	869	560	1 747	1 686-1 808	49,7	48,1-51,6
Total[c]	**3 816**	**3 181**	**2 124**	**5 724**	**5 630-5 817**	**55,6**	**54,7-56,5**

* χ^2 *de tendance, p<10-4.*
[a] *Estimation de la population totale à l'aide de la formule 5 présentée dans le paragraphe matériel et méthodes. Les intervalles de confiance sont basés sur une approximation normale de N (formule 4) et la variance de N est proposée dans la formule 6.*
[b] *Estimation de l'exhaustivité des bases à l'aide de la formule 7. Les intervalles de confiance sont obtenus avec les bornes inférieures et supérieures de N.*
[c] *Sommes des estimations obtenues pour chaque strate.*

| Tableau 5 |

Taux d'exhaustivité de la base Gecsa-ANRS CO3 selon l'année de diagnostic, Aquitaine, 2004-2006

Année	Nombre de cas observés			Nombre de cas estimés[a]		Taux d'exhaustivité de la Gecsa[b]	
	DO	Gecsa	Cas communs	N	IC 95 %	%	IC 95 %
Global	92	151	83	167	160-175	90,3	86,4-94,5
2004	32	56	30	60	56-63	93,8	88,6-99,7
2005	28	45	25	50	46-54	89,4	82,7-97,4
2006	32	50	28	57	52-62	87,7	80,9-95,6
Total[c]	**92**	**151**	**83**	**167**	**160-175**	**90,3**	**86,4-94,5**

* χ^2 *de tendance, p=0,3.*
[a] *Estimation de la population totale à l'aide de la formule 5 présentée dans le paragraphe matériel et méthodes. Les intervalles de confiance sont basés sur une approximation normale de N (formule 4) et la variance de N est proposée dans la formule 6.*
[b] *Estimation de l'exhaustivité des bases à l'aide de la formule 7. Les intervalles de confiance sont obtenus avec les bornes inférieures et supérieures de N.*
[c] *Sommes des estimations obtenues pour chaque strate.*

| Tableau 6 |

Taux d'exhaustivité de la base FHDH-ANRS CO4 par centre, France, 2004-2006

Centre (ville ou hôpital)	Nombre de cas observés			Nombre de cas estimés[a]		Taux d'exhaustivité[b] de la FHDH	
	DO	FHDH	Cas communs	N	IC 95 %	%	IC 95 %
Clermont-Ferrand	39	30	19	61	50-72	49,2	41,6-60,2
Caen	41	41	29	58	52-64	70,9	64,4-79,0
Grenoble	40	46	39	47	46-48	97,5	95,7-99,4
Guadeloupe	152	169	123	209	200-217	81,0	77,8-84,3
Guyane	166	167	109	254	238-270	65,7	61,8-70,2
Tourcoing-Lille	101	96	87	111	109-114	86,2	84,2-88,2
Bourgogne-Franche Comté	39	50	34	57	54-61	87,3	81,9-93,4
Martinique	71	55	49	80	76-84	69,1	65,8-72,7
Montpellier	67	91	51	119	109-130	76,3	70,2-83,5
Nancy	66	12	11	72	62-81	16,8	14,7-19,5
Nantes	67	76	60	85	82-88	89,6	86,4-93,0
Nice	96	162	73	213	195-230	76,2	70,4-83,0
Pitié-Salpetrière	90	161	75	193	180-206	83,4	78,2-89,4
Saint-Louis	346	308	257	415	404-425	74,3	72,5-76,2
Paris Ouest	5	21	3	32	17-47	65,6	44,6-100
Bichat Claude Bernard	117	103	55	218	190-246	47,2	41,8-54,2
Rennes	31	53	28	59	54-63	90,5	84,1-97,9
Rouen	55	54	41	72	67-78	74,7	69,5-80,6
Réunion	49	51	46	54	53-56	93,9	91,9-96,0
Toulouse	80	129	69	149	141-158	86,3	81,5-91,7
Tours	30	41	26	47	43-51	86,8	80,3-94,5
Strasbourg	69	58	44	91	83-98	63,9	58,9-69,8
Paris Centre	132	169	107	208	198-219	81,1	77,3-85,3
Paris Est	246	230	120	471	429-512	48,9	44,9-53,6
Paris Sud	202	221	163	274	264-283	80,7	78,0-83,6
Corevih (92)	90	77	55	126	115-137	61,2	56,4-67,0
Corevih (93)	139	178	132	187	184-191	95,0	93,2-96,8
Lyon	138	122	98	172	164-180	71,1	67,9-74,5
Marseille	151	210	121	262	248-275	80,2	76,3-84,5

[a] *Estimation de la population totale à l'aide de la formule 5 présentée dans le paragraphe matériel et méthodes. Les intervalles de confiance sont basés sur une approximation normale de N (formule 4) et la variance de N est proposée dans la formule 6.*
[b] *Estimation de l'exhaustivité des bases à l'aide de la formule 7. Les intervalles de confiance sont obtenus avec les bornes inférieures et supérieures de N.*

| Tableau 7 |

Taux d'exhaustivité de la déclaration obligatoire (DO) du sida par région ou regroupement de région, France, 2004-2006

Région	Nombre de cas observés			Nombre de cas estimés[a]		Taux d'exhaustivité de la DO[b]	
	DO	FHDH + Gecsa	Cas communs	N	IC 95 %	%	IC 95 %
Global	3 816	3 328	2 204	5 762	5 671-5 853	66,2	65,2-67,3
Alsace	69	56	42	92	83-100	75,2	68,8-82,8
Aquitaine	104	137	70	203	184-222	51,2	46,9-56,4
Auvergne – Limousin*	71	29	18	113	87-139	63,0	51,2-81,8
Bourgogne – Franche-Comté*	64	51	35	93	82-104	68,9	61,5-78,4
Bretagne	70	55	30	127	105-150	55,0	46,8-66,7
Centre	62	43	28	95	80-109	65,5	56,7-77,6
Île-de-France	1 695	1 470	969	2 571	2 509-2 633	65,9	64,4-67,6
Languedoc-Roussillon	87	88	48	159	139-179	54,8	48,7-62,6
Lorraine – Champagne-Ardenne*	89	18	17	94	85-103	94,7	86,6-100
Midi-Pyrénées	102	126	66	194	175-213	52,5	47,9-58,2
Nord-Pas-de-Calais – Picardie*	133	96	87	147	141-152	90,7	87,4-94,2
Basse-Normandie	42	40	28	60	53-67	70,2	63,1-79,1
Haute-Normandie	75	57	44	97	88-106	77,4	71,7-84,8
Pays de la Loire – Poitou-Charentes*	125	73	57	160	146-174	78,2	71,9-85,8
Provence-Alpes-Côte d'Azur – Corse*	306	370	192	589	554-624	51,9	49,0-55,2
Rhône-Alpes	289	182	151	348	332-364	83,0	79,4-86,9
Département d'outre-mer	433	437	322	588	571-604	73,7	71,7-75,8
Total[c]	**3 816**	**3 328**	**2 204**	**5 730**	**5 639-5821**	**66,6**	**65,5-67,7**

** Régions pour lesquelles le nombre de cas au sein des bases ne permet pas d'estimer l'exhaustivité et pour lesquelles nous avons du effectuer un regroupement.*
Pour la région Auvergne, le nombre de cas de sida estimés sur la période 2004-2006 est de 70 (IC 95 % : 55-85) et l'exhaustivité est estimée à 61,5 % (IC 95 % : 50,6-78,5).
Pour la région Bourgogne, le nombre de cas de sida estimés sur la période 2004-2006 est de 70 (IC 95 % : 62-77) et l'exhaustivité est estimée à 66,2 % (IC 95 % : 59,5-74,6).
Pour la région Lorraine, le nombre de cas de sida estimés sur la période 2004-2006 est de 49 (IC 95 % : 41-56) et l'exhaustivité est estimée à 90,7 % (IC 95 % : 78,2-100).
Pour la région Nord-Pas-de-Calais, le nombre de cas de sida estimés sur la période 2004-2006 est de 129 (IC 95 % : 125-133) et l'exhaustivité est estimée à 90,6 % (IC 95 % : 87,7-93,6).
Pour la région Pays de Loire, le nombre de cas de sida estimés sur la période 2004-2006 est de 108 (IC 95 % : 100-117) et l'exhaustivité est estimée à 76,6 % (IC 95 % : 71,0-83,1).
Pour la région Provence-Alpes-Côte d'Azur, le nombre de cas de sida estimés sur la période 2004-2006 est de 569 (IC 95 % : 536-603) et l'exhaustivité est estimée à 51,8 % (IC 95 % : 49,0-55,0).

| Tableau 7 bis |

Taux d'exhaustivité de la déclaration obligatoire (DO) du sida pour certains départements d'Île-de-France et pour les DOM, France 2004-2006

Région/département	Nombre de cas observés			Nombre de cas estimés[a]		Taux d'exhaustivité de la DO[b]	
	DO	FHDH + Gecsa	Cas communs	N	IC 95 %	%	IC 95 %
Paris	937	982	606	1 518	1 474-1 562	61,7	60,0-63,6
Hauts-de-Seine	163	122	94	211	198-224	77,1	72,6-82,2
Seine-Saint-Denis	150	170	124	206	198-213	73,0	70,3-75,8
Val-de-Marne	244	176	125	343	321-366	71,1	66,7-76,1
Guadeloupe	152	169	123	209	200-217	72,8	70,0-75,8
Martinique	73	57	51	82	78-85	89,5	85,4-94,0
Guyane	160	161	103	250	233-267	64,1	60,0-68,8
Réunion	48	50	45	53	52-55	90,0	88,8-92,1

[a] *Estimation de la population totale à l'aide de la formule 5 présentée dans le paragraphe matériel et méthodes. Les intervalles de confiance sont basés sur une approximation normale de N (formule 4) et la variance de N est proposée dans la formule 6.*
[b] *Estimation de l'exhaustivité des bases à l'aide de la formule 7. Les intervalles de confiance sont obtenus avec les bornes inférieures et supérieures de N.*

3.6 Évaluation qualitative de la dépendance entre les sources

Un questionnaire permettant d'évaluer qualitativement la dépendance positive entre la FHDH et la DO du sida a été envoyé aux 56 hôpitaux participant à la FHDH pour lesquels la base contenait des patients sur la période d'étude. Ces 56 hôpitaux sont regroupés en 29 centres (annexe 5). Nous avons recueilli 64 réponses provenant de 45 hôpitaux, soit un taux de réponse de 80,4 %. Nous disposions de la totalité des questionnaires pour 21 centres (72,4 %), de données incomplètes pour 5 centres (17,3 %) et d'aucun questionnaire pour 3 autres centres (10,3 %).

Nous avons regroupé les 21 centres pour lesquels nous disposions des données complètes en deux groupes :

- 9 centres pour lesquels la majorité des participants déclaraient établir des DO suite à la saisie ou à la consultation de leur base locale. Cette situation plaide en faveur d'une dépendance positive entre les deux sources ;
- 12 centres pour lesquels la saisie des fiches de DO n'est pas liée à la saisie des cas dans leur base locale. Dans ce cas, les bases de données sont considérées comme indépendantes.

Pour pouvoir interpréter ces données, nous les avons pondérées par rapport au nombre de cas de sida de chaque centre au sein de la base FHDH, en excluant les centres pour lesquels nous n'avions pas de données.

Les 12 centres pour lesquels les sources de données ont été considérées comme indépendantes représentent 56,3 % des cas de sida contre 43,7 % pour les 9 centres qui semblent être en situation de dépendance positive.

Le taux d'exhaustivité calculé pour les 12 centres "indépendants" est de 70,8 % contre 76,4 % pour les autres centres et de 71,6 % pour les centres dont nous n'avons pas les données de l'enquête.

Pour la base du Gecsa, dont les établissements ont été interrogés avec le même questionnaire, nous disposons des données des 7 hôpitaux participant à cette base. Pour 5 d'entre eux, les personnes saisissant les cas dans la base du Gecsa ne remplissent pas de DO, ce qui plaide pour une situation d'indépendance. Pour les deux autres hôpitaux, les personnes saisissant dans la base Gecsa remplissent également des DO, soit suite à la demande du clinicien, soit grâce à une fiche d'aide à la déclaration des cas incidents qui est transmise aux différents services participants. Dans ce cas, il est possible que nous nous trouvions dans une situation de dépendance positive.

4. Discussion

Une estimation de la sous-déclaration des cas de sida a été menée en 1995 sur la période 1990-1993 en appliquant la méthode capture-recapture à deux sources [14]. Quatorze ans après cette première estimation, période au cours de laquelle le système de surveillance du VIH/sida a beaucoup évolué, il était important de réévaluer cette sous-déclaration qui a pu être modifiée par la mise en place de la DO du VIH en 2003.

L'application sur la période 2004-2006 de la méthode capture-recapture à deux sources a permis d'estimer l'exhaustivité de la DO du sida en France à 66,1 % (IC 95 % : 65,1-67,2). Cette étude a également permis d'estimer la couverture des deux bases utilisées comme seconde source : celle de la FHDH est estimée à 55,6 % (IC 95 % : 54,7-56,5) sur l'ensemble du territoire et celle du Gecsa à 90,3 % (IC 95 % : 86,4-94,5) dans la région Aquitaine.

Par ailleurs, le nombre de cas de sida diagnostiqués sur cette période de trois ans a été estimé à 5 770 (IC 95 % : 5 679-5 861).

La validité et la précision de ces différentes estimations dépend du respect de plusieurs conditions.

La première d'entre elles est que tous les cas soient des vrais cas. Pour être respectée, cette condition nécessite une définition de cas commune aux différentes bases. Pour notre étude, la définition d'un cas de sida utilisée, commune aux trois bases, était celle de 1993 [17].

Pour la base du Gecsa, seuls les patients infectés par le VIH-1 sont inclus, contrairement à la DO et à la FHDH qui incluent tous les cas de sida, qu'ils soient infectés par le VIH-1 ou le VIH-2. Cependant, la proportion de VIH-2 est faible en France : parmi les découvertes de séropositivité en 2007, elle a été estimée à 2,2 % [24]. Au sein de la base de la DO sur notre période d'étude, seuls 2 cas étaient infectés par le VIH-2. Afin d'avoir une définition de cas strictement identique et d'éviter un biais dans l'estimation de l'exhaustivité du Gecsa, ces 2 cas ont été exclus.

Par ailleurs, sur notre période d'étude, 311 patients de la base FHDH n'avaient pas de pathologies nécessaires à la classification d'un cas au stade sida. Ces patients ont donc été exclus de l'analyse pour éviter un biais dans les estimations.

La seconde condition est que la période d'étude et la zone géographique soient identiques. Notre période d'étude concerne les cas de sida diagnostiqués du 1er janvier 2004 au 31 décembre 2006 et ce, pour les trois sources.

En ce qui concerne la zone géographique, les bases de la DO et de la FHDH englobent les départements d'outre-mer et l'ensemble de la France métropolitaine, à l'exception de l'Aquitaine, dont les données ne sont pas comprises dans la base FHDH. Afin de travailler sur la même zone géographique, nous avons ajouté aux données de la base FHDH les données de la base du Gecsa, qui ne concernent que la région Aquitaine. Pour l'estimation de l'exhaustivité de la FHDH, le choix a été fait de ne pas retirer de la base DO du sida les cas diagnostiqués dans les hôpitaux d'Aquitaine. En effet, la FHDH est une base nationale, il était donc important, comme cela avait été fait lors de la précédente étude, de donner une estimation de la couverture sur l'ensemble du territoire. De plus, si les cas diagnostiqués dans les hôpitaux d'Aquitaine au sein de la DO du sida avaient été retirés, cela aurait empêché l'identification de 8 cas communs diagnostiqués en Aquitaine dans la DO du sida et dans une autre région dans la FHDH. Enfin, d'autres hôpitaux censés participer à la FHDH n'avaient pas envoyé leurs données sur la période d'étude et donc d'autres zones géographiques, en plus de l'Aquitaine, étaient concernées par ce problème. Pour contourner cette difficulté, l'exhaustivité a également été estimée par centre, c'est-à-dire sur des zones géographiques strictement identiques.

La troisième condition d'application est que la population d'étude soit close. Pour respecter cette condition, il est nécessaire de ne pas avoir de mouvement de population (ni entrée, ni sortie), pour pouvoir assurer une équiprobabilité de capture des cas dans chaque source. Ce problème peut se poser pour les personnes migrantes si, durant leur séjour en France, elles n'ont été incluses que dans l'une des sources. Mais dans notre étude, quelle que soit la source, il n'y a pas nécessité d'être suivi plusieurs fois après un diagnostic de sida pour être intégré dans une des bases, la probabilité de capture semble donc identique pour chacune des sources.

La quatrième condition d'application est que tous les cas communs et seuls les vrais cas communs soient identifiés. L'identification des cas communs est rendue complexe par l'absence d'identifiant unique. De ce fait, l'identification des cas communs repose sur une combinaison de plusieurs critères. Le risque d'identification de faux positifs et de faux négatifs a une conséquence sur l'estimation du nombre total de cas de sida [25], puisque les faux positifs induisent une sous-estimation de ce nombre, et inversement.

Une grande majorité des cas communs identifiés par l'algorithme présente une bonne concordance sur la plupart des variables (tableau 2). Pour s'assurer de l'absence de faux positifs, il serait nécessaire de retourner aux dossiers médicaux des patients mais, étant donné les effectifs, le temps employé à confirmer les cas communs ferait perdre le bénéfice de la méthode capture-recapture. De plus, l'anonymisation très stricte mise en place en 2003 ne permet pas de remonter aux dossiers médicaux à partir de la DO.

La principale question posée par la mise en œuvre de l'algorithme d'identification des cas communs est de savoir à quel moment cesser d'élargir les critères d'identification. Cette question a été le sujet de nombreuses discussions. Pour pouvoir nous aider à sélectionner le meilleur algorithme, nous avons comparé les différences existantes sur l'ensemble des variables pour les cas communs identifiés à partir d'une date de décès identique (R2) et pour les cas communs identifiés à la fin de l'algorithme (R20). En effet, une date complète de décès (jour, mois, année) permet d'identifier de façon quasi-certaine les cas communs. Cette analyse de sensibilité nous a permis de faire le choix de l'algorithme permettant d'identifier au mieux les cas communs. Globalement, pour quatre variables (mode de contamination, date de sérologie VIH, date de diagnostic du sida et pathologies), la proportion de discordances n'est pas significativement différente

entre la requête R2 (identification des cas communs grâce à la date de décès) et la requête R20 (dernière requête de l'algorithme), ce qui est en faveur de la robustesse de l'algorithme choisi pour détecter les cas communs. Pour les variables "département de domicile" et "hôpital", la proportion de concordance est plus élevée lors de la requête R2 que la requête R20. Cependant, nous pouvons considérer que la date de décès a plus de chance d'être rigoureusement la même dans les deux bases si l'inclusion dans ces bases se fait dans le même hôpital, et qu'il est plus probable que le département de domicile renseigné par le déclarant soit alors le même. Malgré cela, ce constat pourrait également amener à penser que nous avons une proportion de faux cas communs plus importante à la fin de l'algorithme, ce qui induirait une sous-estimation du nombre total de cas de sida. *A contrario*, la proportion de cas communs avec "au moins une pathologie commune" est significativement plus élevée à la requête R20 qu'à la requête R2, ce qui tendrait à démontrer l'inverse. D'après Brenner [25], l'effet antagoniste des faux positifs et des faux négatifs s'annule en pratique, ce qui semble être le cas dans notre étude.

L'algorithme mis en place a permis de détecter la présence de doublons intrabase et ce, malgré les procédures de recherche de doublons mis en place au sein de chaque système, ce qui est également en faveur de la robustesse de notre algorithme. Enfin, au vu de la démarche adoptée pour la recherche des cas communs, ceux qui n'auraient pas été identifiés (faux négatifs) présenteraient des discordances importantes et non négligeables sur de nombreuses variables, ce qui traduirait un problème de qualité des données.

La cinquième condition d'application est que tous les cas de la population doivent avoir la même probabilité d'être identifiés par une source, c'est-à-dire que la capture ne doit pas être liée à certaines caractéristiques du système de surveillance ou du cas. Dans notre étude, l'hétérogénéité de capture a été étudiée pour deux variables : l'année de diagnostic et la région de notification des cas. La stratification par rapport à l'année de diagnostic met en évidence une diminution de l'exhaustivité sur la période d'étude pour la DO et la FHDH. Nous avons pris en compte l'hétérogénéité de capture pour l'année de diagnostic, qui montre cependant que l'estimation est proche de celle estimée globalement. La stratification par rapport à la région de notification montre des disparités importantes d'exhaustivité d'une région à l'autre. L'idéal aurait été de stratifier sur les deux variables (année de diagnostic et région) pour avoir une estimation prenant en compte à la fois ces deux variables d'hétérogénéité, mais la faiblesse des effectifs dans un grand nombre de strate ne nous a pas permis de le réaliser.

Il existe probablement aussi une hétérogénéité de capture en fonction des centres de la FHDH. Cependant, il n'a pas été possible de réaliser une stratification par centre, car un nombre important de cas notifiés dans la DO (n=901) ne pouvait pas être affecté à un centre de la FHDH.

Ces deux hétérogénéités (année de diagnostic et région) de capture sont principalement liées aux systèmes de surveillance, mais il est possible que d'autres hétérogénéités de capture liée aux caractéristiques des cas soient présentes. Cela pourrait être le cas pour la nationalité, mais l'absence d'homologie stricte sur cette variable entre la FHDH ("séjour à l'étranger de plus de six mois") et la DO ("nationalité") ne nous permet pas d'apprécier cette possible hétérogénéité.

La dernière condition d'application est certainement la plus importante : les bases doivent être indépendantes, c'est-à-dire que le fait d'être inclus dans une des bases ne doit pas influer sur la probabilité d'inclusion au sein d'une autre base. Les résultats de l'enquête permettant d'évaluer qualitativement la dépendance positive entre les sources ont montré que, pour plus de la moitié des cas de sida, il n'y avait pas de dépendance alors que les autres présentaient probablement une dépendance positive. L'exhaustivité moyenne calculée sur les centres de la FHDH susceptibles de présenter une dépendance positive était plus importante (76,4 %) que celle des centres en situation d'indépendance (70,8 %), ce qui tendrait à montrer que, pour les premiers, l'exhaustivité est surestimée. Cependant, si l'on applique le taux d'exhaustivité des "centres indépendants" (70,8 %) aux centres en situation de dépendance, le nombre total de cas (N) n'augmente que de 46, ce qui ne représente qu'une augmentation de 0,8 % du nombre total estimé de cas de sida sur notre période d'étude. De plus, le fait que, sur la totalité des cas communs identifiés (n=2 824), seuls 211 (7,5 %) avaient une homologie stricte sur l'ensemble des variables, n'est pas en faveur d'une situation de forte dépendance positive.

Parallèlement à une dépendance positive, il est possible qu'il existe une dépendance négative pour certains centres, c'est-à-dire que des déclarants ayant signalé un cas dans l'un des deux systèmes oublient ou jugent superflu de le faire à nouveau dans l'autre. Cela pourrait notamment être le cas pour la région Aquitaine, où le taux d'exhaustivité de la DO du sida (51,2 %) apparaît très faible. Il est possible que le fait de déclarer un cas au sein de la base Gecsa, très implantée sur la région, ait un impact négatif sur l'établissement d'une fiche de DO du sida et que, par conséquent, l'implication des centres d'Aquitaine dans la DO du sida soit moins forte. Cette dépendance négative pourrait entraîner une sous-estimation de l'exhaustivité, mais l'appréciation qualitative de cette dépendance apparaît très difficile à étudier.

Globalement, il semble qu'il existe une faible dépendance positive entre la DO et la FHDH, mais sans écarter également une dépendance négative pour certains centres. Tous ces éléments nous amène à penser que l'impact de la dépendance est probablement limité sur nos estimations. La dépendance positive avait déjà été constatée lors de l'étude capture-recapture réalisée en 1995 [14]. Au regard des résultats de l'enquête d'évaluation de la dépendance réalisée à cette époque, il ne semble pas qu'il y ait eu une évolution sur ce critère entre les deux études. Il est donc possible de comparer les estimations de ces deux études.

Malgré ces limites, cette étude a permis de donner une nouvelle estimation du nombre de cas de sida en France, d'actualiser les estimations d'exhaustivité pour la DO du sida et la FHDH, et de fournir une estimation d'exhaustivité de la base de données du Gecsa.

Pour la base du Gecsa, nous ne disposions pas du taux d'exhaustivité avant cette étude nous permettant de comparer les chiffres obtenus. Cependant, il n'est pas surprenant de retrouver un taux de couverture très élevé (90,3 %) pour cette base. En effet, la base du Gecsa ne se concentre que sur une région : l'Aquitaine. Une recherche des perdus de vue est réalisée chaque année. Cette recherche active n'est pas réalisée pour la DO du sida et la FHDH, ce qui peut expliquer la différence importante d'exhaustivité entre le Gecsa et ces deux bases.

L'exhaustivité de la DO du sida et de la FHDH avait été estimée sur la période 1990-1993 avec une méthode comparable à celle utilisée dans notre étude [14]. Il est possible que l'identification des cas communs dans notre étude ait été plus sévère que lors de l'étude de 1995 car nous disposions de certaines variables absentes dans la précédente

étude (nationalité, nombre de CD4 et date de mesure des CD4). Si l'identification des cas communs lors de l'étude précédente avait été faite avec notre algorithme, il est possible que les estimations d'exhaustivité de la période 1990-1993 aient été légèrement plus faibles.

Pour la base FHDH, l'exhaustivité a progressé puisqu'elle est passée de 47,6 % (IC 95 % : 46,9-48,3) en 1900-1993 à 55,6 % (IC 95 % : 54,7-56,5) en 2004-2006. Cela s'explique notamment par l'inclusion dans la base de centres qui n'étaient pas présents auparavant. Cependant, sur notre période d'étude, certains centres n'avaient pas envoyé leurs données à la FHDH, notamment pour l'année 2006, ce qui peut expliquer la diminution de la couverture entre 2004 et 2006 (60,0 % à 49,7 %).

Les résultats d'exhaustivité par centre ont été comparés avec ceux obtenus dans l'étude précédente sur la période 1990-1993 (annexe 7). Pour les centres communs entre les deux études, il existe une bonne concordance. En effet, les centres qui avaient une bonne exhaustivité sur la période 1990-1993 l'ont toujours actuellement et réciproquement. Dans l'ensemble, la quasi-totalité des centres ont vu leur couverture augmenter.

Contrairement à la FHDH, l'exhaustivité de la DO du sida a diminué puisqu'elle est passée de 83,6 % (IC 95 % : 82,9-84,3) pour la période 1990-1993 [14] à 66,1 % (IC 95 % : 65,1-67,2) en 2004-2006. Entre ces deux estimations, le système de surveillance a évolué, pouvant expliquer cette diminution. En effet, un nouveau système de surveillance des maladies à DO a été mis en place en 2003. Les changements pouvant avoir un impact négatif sur l'exhaustivité de la DO du sida sont, d'une part, la mise en place de la DO du VIH et, d'autre part, l'instauration d'une procédure d'anonymisation obligeant les médecins déclarants à calculer un code d'anonymat. En effet, le nombre de déclarations de cas de sida chez l'adulte reçues dans le cadre de la DO (*cf.* figure en annexe 6) a nettement diminué en 2003 (-10,5 % par rapport à l'année 2002). Certains cliniciens ont probablement délaissé la DO du sida au profit de celle sur le VIH, soit parce qu'ils n'ont pas compris l'utilité des deux déclarations, soit parce que l'anonymisation de la DO du sida a rendu cette surveillance plus complexe.

L'exhaustivité de la DO du sida estimée dans notre étude est comparable à celle de la DO du VIH qui a été estimée à environ 65 % sur la période 2004-2006 [2]. Cependant, et contrairement à la DO du sida, l'exhaustivité de la DO du VIH a augmenté sur cette période, passant de 63 % en 2004 à 66 % en 2006, ce qui renforce l'idée que la DO du sida soit quelque peu délaissée au profit de la DO du VIH. Pour remédier à ce phénomène, de nouvelles fiches de DO du VIH ont été mises en place en 2007, permettant aux médecins de déclarer un sida sur le même formulaire que l'infection VIH. Il est encore trop tôt pour dire si ces nouvelles fiches ont un impact positif sur la DO du sida.

5. Conclusion

Cette étude a permis d'estimer à 5 770 (IC 95 % : 5 679-5 861) le nombre de nouveaux cas de sida sur la période 2004-2006 en France. L'exhaustivité au niveau national a été estimée à 66,1 % (IC 95 % : 65,1-67,2) pour la déclaration obligatoire du sida et à 55,6 % (IC 95 % : 54,7-56,5) pour la FHDH. En région Aquitaine, l'exhaustivité du Gecsa a été estimée à 90,3 % (IC 95 % : 86,4-94,5) sur la même période.

La validité de ces différentes estimations repose sur certaines conditions, dont la plus importante est l'indépendance entre les sources. L'évaluation de cette condition, réalisée à partir d'une enquête spécifique, a montré qu'il existait une faible dépendance positive entre la FHDH et la DO, dont l'impact est vraisemblablement limité sur l'estimation du nombre total de cas, d'autant plus qu'on ne peut pas exclure une dépendance négative pour certains centres.

L'exhaustivité de la DO du sida est en recul depuis l'étude réalisée sur la période 1990-1993, ce qui peut en partie s'expliquer par la mise en place de la DO du VIH en 2003. Cependant, elle est proche de celle du VIH qui a été estimée à environ 65 % sur la même période 2004-2006 [2]. Même si ces taux d'exhaustivité semblent relativement faibles, ils restent supérieurs à ceux d'autres maladies à DO, comme par exemple l'hépatite B, dont l'exhaustivité a été estimée en 2005 à 23,4 % (IC 95 % : 21,2-26,0) [26].

Références bibliographiques

[1] Onusida. Rapport sur l'épidémie mondiale de sida 2008.

[2] Cazein F, Pillonel J, Le Strat Y, Lot F, Pinget R, David D *et al.* Surveillance de l'infection à VIH-sida en France, 2007. Bull Epidemiol Hebd 2008;45-46:434-43.

[3] Lillienfield A, Lillienfield DE. Foundations of epidemiology. NY: Oxford university press; 1980.

[4] Cormack RM. The statistics of capture-recapture methods. Oceanogr Mar Biol Ann Rev 1968;6:455-506.

[5] Chapman DG. Some properties of the hypergeometric distribution with applications to the zoological sample censuses. Uni Calif Public Stat 1951;26:13-22.

[6] Pollock KH. Modelling capture, recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations; past, present and future. J Amer Stat Assoc 1991;86:225-38.

[7] Sekar CC, Deming WE. On a method of estimating birth and death rates and extent of registration. Amer Stat Assoc J 1949;44:101-15.

[8] Himes CL, Clogg CC. An overview of demographic analysis as a method for evaluating census coverage in the United States. Popul Index 1992 Winter;58(4):87-607.

[9] Wolker KM. Accounting for America's uncounted and miscounted. Science 1991 Jul;253:12-5.

[10] Robles S, Marret LD, Clarke EA, Rish A. An application of capture-recapture methods to estimation of the completeness of cancer registration. J Clin Epidemiol 1988;41:495-501.

[11] Laporte RE, Stull E, McCarty D. Monitoring the incidence of myocardial infarctions: applications of capture-mark-recapture technology. Intern J Epidemiol 1992;21:258-63.

[12] Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York state AIDS registry and a hospital discharge file. Stat Med 1995 Mar;14(5-7):499-509.

[13] Hall HI, Song R, Gerstle JE 3rd, Lee LM. HIV/AIDS reporting system evaluation group. Assessing the completeness of reporting of human immunodeficiency virus diagnoses in 2002-2003: capture-recapture methods. Am J Epidemiol 2006;164:391-7.

[14] Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D. Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990-1993. The Clinical Epidemiology Group from Centre d'information et de soin de l'immunodéficience humaine. Int J Epidemiol 2000;41:495-501.

[15] Hubert B, Desenclos JC. Évaluation de l'exhaustivité et de la représentativité d'un système de surveillance par la méthode de capture-recapture. Application à la surveillance des infections à méningocoque en France en 1989 et en 1990. Rev Epidemiol Sante Publique 1993;41:241-9.

[16] Faustini A, Fano V, Sangalli M, Ferro S *et al.* Estimating incidence of bacterial meningitis with capture-recapture method, Lazio region. Italy. Eur J Epidemiol 2000;16:843-8.

[17] Révision de la définition du sida en France. Bull Epidemiol Hebd 1993;11:47-8.

[18] Chapman DG. Some properties of the hypergeometric distribution with applications to the zoological sample censuses. Uni Calif Public Stat 1951;26:13-22.

[19] Seber GAF. The effect of trap response on tag recapture estimates. Biometrics 1970;26:13-22.

[20] Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol reviews 1995;17:243-64.

[21] Gallay A, Nardone A, Vaillant V, Desenclos JC. La méthode capture-recapture appliqué à l'épidémiologie : principes, limites et applications. Rev Epidemiol Sante Publique 2002;50:219-32.

[22] Cormarck RM. Log linears model for capture-recapture. Biometrics 1989;45:395-413.

[23] Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information source. J Chron Dis 1974;27:25-36.

[24] Situation du VIH/sida en France. Institut de veille sanitaire. Département des maladies infectieuses. Données du 30 septembre 2008.

[25] Brenner H. Application of capture-recapture methods for disease monitoring : potential effects of imperfect record linkage. Methods Inf Med 1994;33(5):502-6.

[26] Antona D, Letort MJ, Lévy-Bruhl D. Estimation du nombre annuel de nouvelles infections par le virus de l'hépatite B en France, 2004-2007. Bull Epidémiol Hebd 2009;20-21:196-9.

Annexe 1 – Liste des 30 maladies à déclaration obligatoire

- Botulisme
- Brucellose
- Charbon
- Chikungunya
- Choléra
- Dengue
- Diphtérie
- Fièvres hémorragiques africaines
- Fièvre jaune
- Fièvres typhoïde et paratyphoïdes
- Hépatite aiguë A
- Infection aiguë symptomatique par le virus de l'hépatite B
- Infection par le VIH
- Infection invasive à méningocoque
- Légionellose
- Listériose
- Orthopoxviroses dont la variole
- Paludisme autochtone
- Paludisme d'importation dans les départements d'outre-mer
- Peste
- Poliomyélite
- Rage
- Rougeole
- Saturnisme de l'enfant mineur
- Suspicion de la maladie de Creutzfeldt-Jakob et autres encéphalopathies subaiguës spongiformes transmissibles humaines
- Tétanos
- Toxi-infection alimentaire collective
- Tuberculose
- Tularémie
- Typhus exanthématique

Annexe 2 – Questionnaire d'évaluation de la dépendance entre les sources

Évaluation de la dépendance entre les bases de données FHDH-ANRS C04 et DO Sida

Afin de suivre l'évolution du VIH/sida, plusieurs systèmes de surveillance existent et se complètent au niveau national. Il s'agit de la déclaration obligatoire du sida, mise en place depuis 1986 et gérée par l'Institut de veille sanitaire (InVS), et de la FHDH-ANRS C04, gérée par l'Unité 943 de l'Institut national de la santé et de la recherche médicale (Inserm) qui recueille les informations des patients infectés par le VIH et suivis à l'hôpital depuis 1989. L'intérêt de ces bases de données est incontestable et il apparaît important de pouvoir connaître l'exhaustivité de chacune d'elle et d'estimer le nombre total de cas de sida.

L'objectif de l'étude capture-recapture menée actuellement par l'InVS et l'Inserm U943 est d'estimer l'exhaustivité de ces deux bases de données sur la période 2004-2006.

Cette méthode nécessite d'apprécier la dépendance qui pourrait exister entre ces deux bases. Nous vous faisons donc parvenir un questionnaire visant à évaluer la dépendance entre la DO et la FHDH. Nous vous serions très reconnaissants d'accorder quelques minutes de votre temps pour remplir ce questionnaire.

À la fin de l'étude, les résultats relatifs à l'exhaustivité de ces deux bases de données vous seront communiqués.

Merci par avance pour votre collaboration.

Remplir un questionnaire par personne.

Nom et prénom : …… Nom du Corevih : ……

Nom de l'hôpital : …… Nom du service : ……

Fonction : ……

1) Dans votre service, utilisez-vous un logiciel de dossier médical (Nadis, Diammg, …) pour le suivi des patients ?
   - OUI ☐
   - NON ☐

Si OUI, répondez aux questions 2 à 5 pour la période **avant** l'utilisation de ce logiciel de dossier médical et aux questions 6 à 11 pour la période **après** la mise en place de ce dossier médical informatisé.

Si NON, répondez aux questions 2 à 5.

Questions si pas de dossier médical informatisé :

2) Vous arrive t-il de remplir (en totalité ou en partie) des fiches de déclaration obligatoire du **sida** ?
   - OUI ☐
   - NON ☐

   Si OUI, combien en remplissez-vous en moyenne sur une année ?

3) S'il vous arrive de remplir des fiches de DO sida, quel est l'élément déclencheur du remplissage de cette fiche ?
   - La saisie ou consultation du DMI2 : ☐
   - La demande du médecin qui a diagnostiqué le sida : ☐
   - Le diagnostic fait par moi-même : ☐
   - Autre, précisez : ……
     ……
     ……

4) D'autres personnes remplissent-elles les DO sida dans le(s) service(s) ?
   - OUI ☐
   - NON ☐

   Si OUI, précisez leur(s) fonction(s) : ……
   ……
   ……

5) Sachant que l'étude porte sur la période 2004-2006, savez-vous si les modalités décrites ci-dessus étaient semblables à cette époque.
   - OUI ☐
   - NON ☐

   Si NON, précisez : ……
   ……
   ……

Questions si dossier médical informatisé :

6) Quel logiciel utilisez-vous ? ……

7) Depuis quelle date ? ……..

8) Vous arrive t-il de remplir (en totalité ou en partie) des fiches de déclaration obligatoire du **sida** ?
   - OUI ☐
   - NON ☐

   Si OUI, combien en remplissez-vous en moyenne sur une année ? ……

9) S'il vous arrive de remplir des fiches de DO sida, quel est l'élément déclencheur du remplissage de cette fiche ?
   - La saisie ou consultation du dossier médical informatisé : ☐
   - La demande du médecin qui a diagnostiqué le sida : ☐
   - Le diagnostic fait par moi-même : ☐
   - Autre, précisez : ……
     ……
     ……

10) D'autres personnes remplissent-elles les DO sida dans le(s) service(s) ?
   - OUI ☐
   - NON ☐

   Si OUI, précisez leur(s) fonction(s) : ……
   ……
   ……

11) Sachant que l'étude porte sur la période 2004-2006, savez-vous si les modalités décrites ci-dessus étaient semblables à cette époque.
   - OUI ☐
   - NON ☐

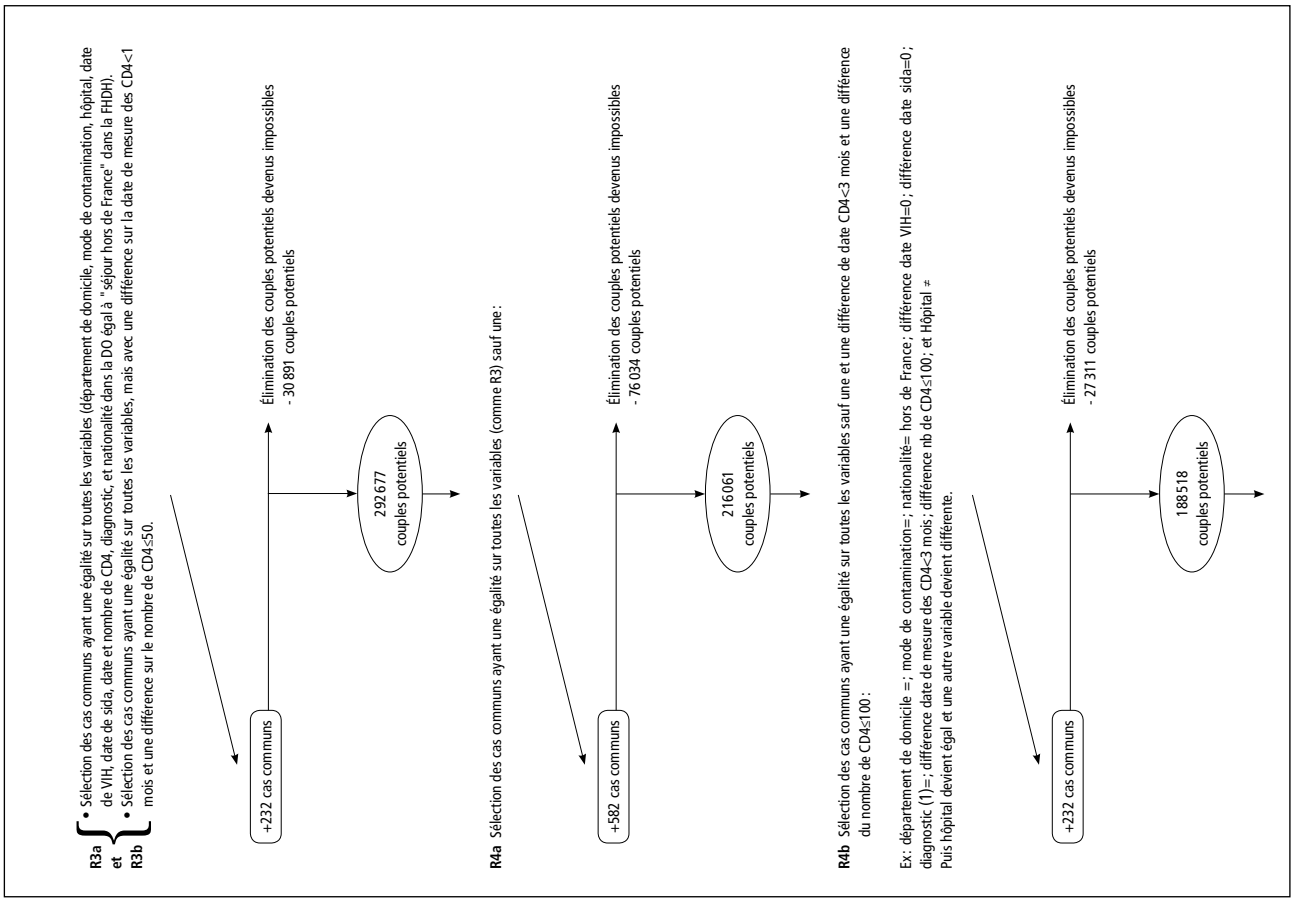
   Si NON, précisez : ……
   ……
   ……
   ……

Annexe 3 – Liste des pathologies classantes sida et de leur fréquence parmi les cas de sida de 18 ans et plus de la base DO entre le 01/07/2003 et le 30/06/2007

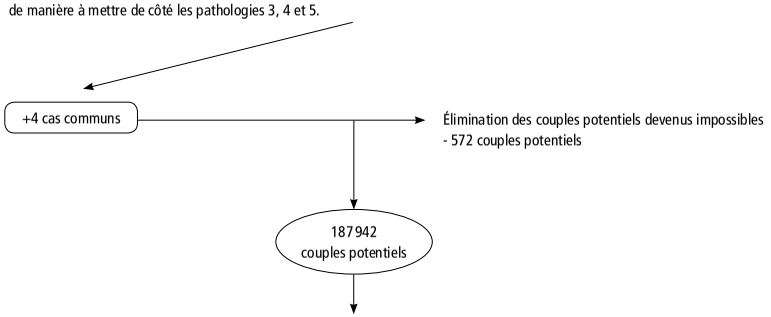
1) Pneumocystose pulmonaire (22,8 %)
2) Candidose œsophagienne (15,9 %)
3) Toxoplasmose cérébrale (12,1 %)
4) Tuberculose pulmonaire (11,2 %)
5) Tuberculose extra-pulmonaire ou miliaire (10,4 %)
6) Sarcome de Kaposi (9,3 %)
7) Lymphome non cérébrale (5,7 %)
8) Infection à cytomégalovirus (4,9 %)
9) Encéphalopathie (4,5 %)
10) Leucoencéphalite multifocale progressive (3,0 %)
11) Cryptococcose (2,8 %)
12) Syndrome cachectique (2,5 %)
13) Cryptosporidiose (2,1 %)
14) Histoplasmose (1,7 %)
15) Infection à mycobactérie *Avium* ou *Kansasii* (1,7 %)
16) Pneumopathie bactérienne récidivante (1,23 %)
17) Infection à HSV (1,1 %)
18) Infection à mycobactérie autres (1,0 %)
19) Isosporidiose (0,8 %)
20) Candidoses bronches/trachée/poumons (0,7 %)
21) Cancer invasif du col (0,6 %)
22) Lymphome cérébral primaire (0,4 %)
23) Septicémie récidivante à *Salmonella non typhi* (0,1 %)
24) Coccidioïdomycose (0,04 %)

NB : la somme des pourcentages dépasse les 100 %. Cela s'explique par le fait qu'un même patient puisse avoir plusieurs pathologies.

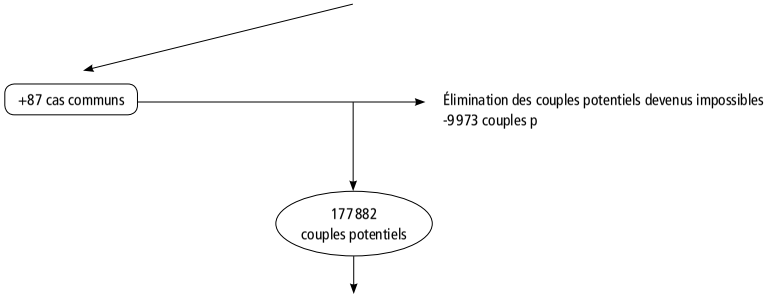
Annexe 4 – Algorithme de détermination des cas communs



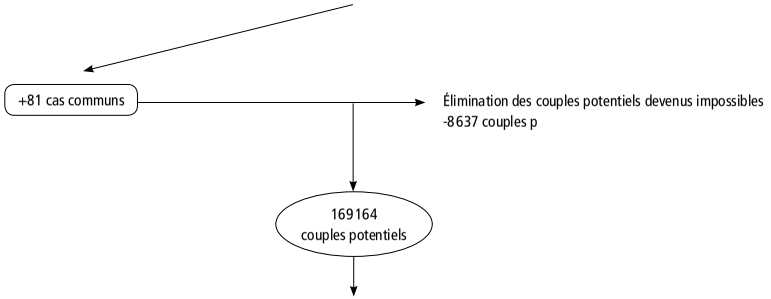
R4c Sélection des cas communs ayant une égalité sur toutes les variables sauf une, et patho1 DO=patho1 FHDH et patho2 DO=patho2 FHDH, de manière à mettre de côté les pathologies 3, 4 et 5.



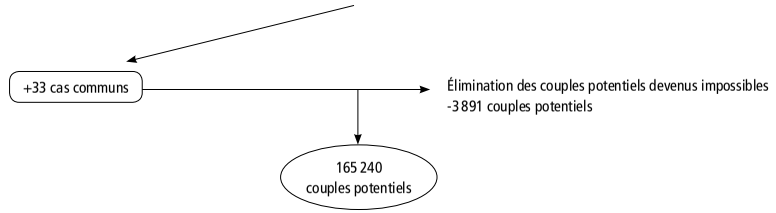
R4d Sélection des cas communs identique à R4c, sauf pour le croisement des pathologies.
Dans ce cas, patho1 DO=patho1 ou 2 FHDH ou patho1 FHDH=patho2 DO ou patho2 DO=patho2 FHDH.



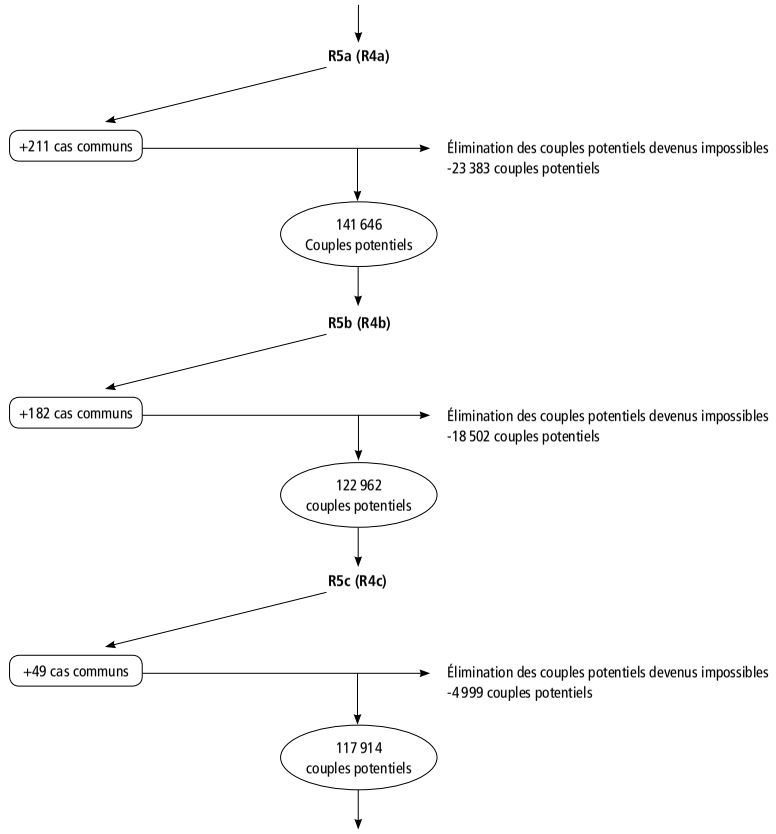
R4e Sélection des cas communs ayant une égalité sur toutes les variables sauf une (comme R4a) mais avec une différence de date VIH<1 an.

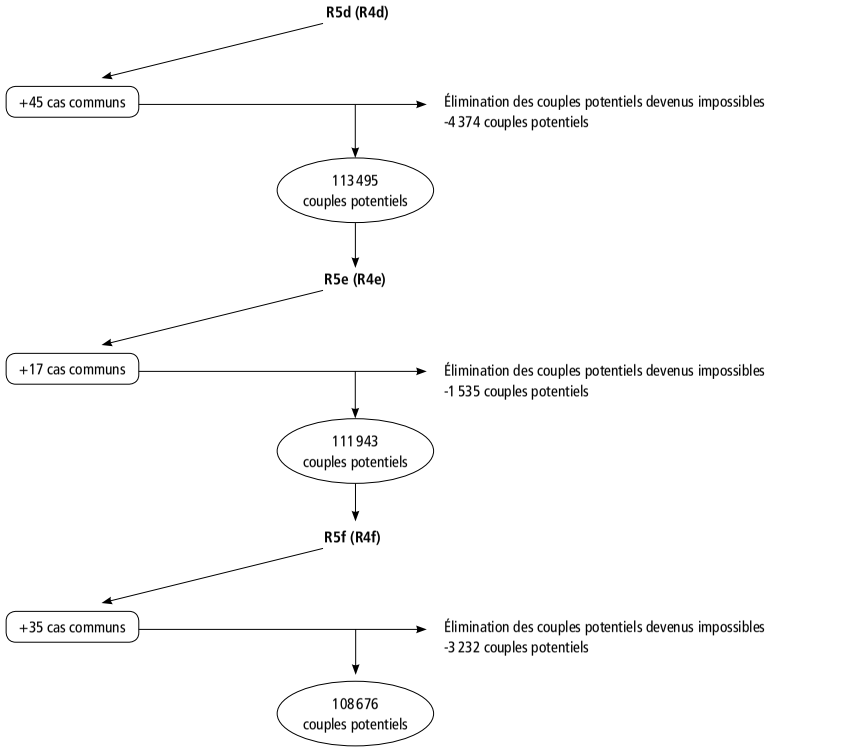


R4f Sélection des cas communs comme dans R9 sauf que la différence de date VIH est remplacée par une différence sur la date sida<6 mois.

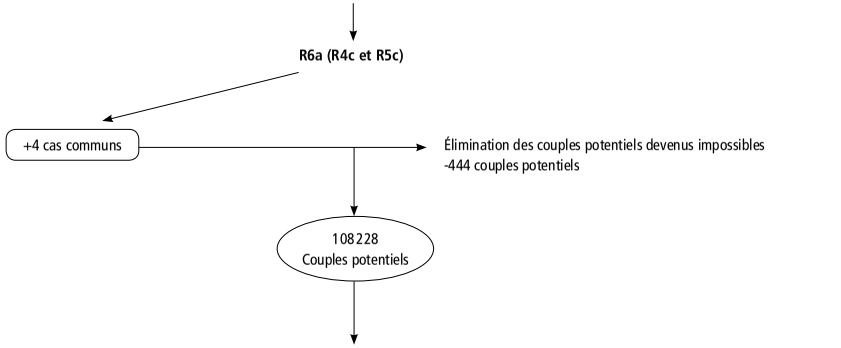


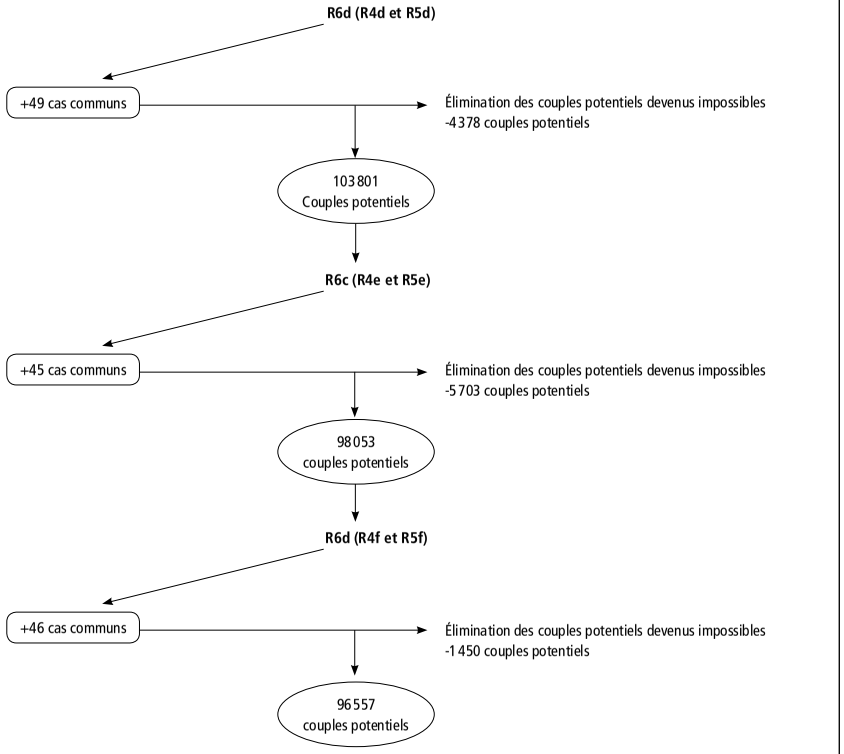
Après la requête 4, nous sommes à 1 532 cas communs et 165 240 couples potentiels restants.
Pour les requêtes suivantes, il a été fait le choix d'enlever l'égalité entre nationalité et hors de France, puisque ces variables n'expriment pas nécessairement la même chose surtout dans les cas ou la nationalité du patient est française. En effet dans ce cas la variable hors de France est très rarement remplie.
Les 6 requêtes suivantes (R5a à R5f) sont construites sur le même modèle que les requêtes R4.



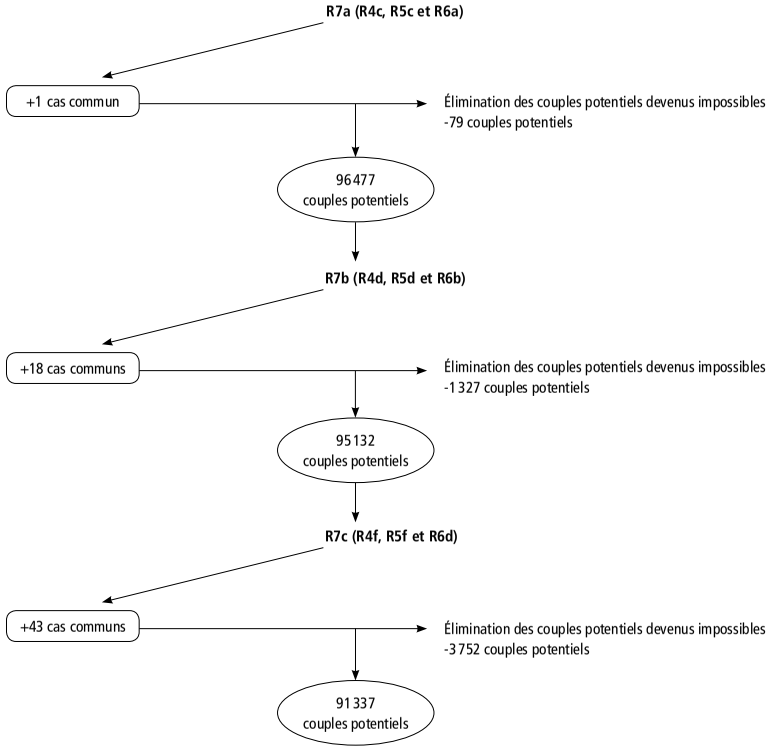


Après la requête 5, nous sommes à 2 071 cas communs et 108 676 couples potentiels restants.
Pour les requêtes suivantes, en plus de la différence entre nationalité et séjour hors de France, nous ajoutons une différence sur la date de mesure des CD4<3 mois et une différence sur le nombre de CD4≤100.
Les requêtes suivantes sont construites sur le même modèle que précédemment.

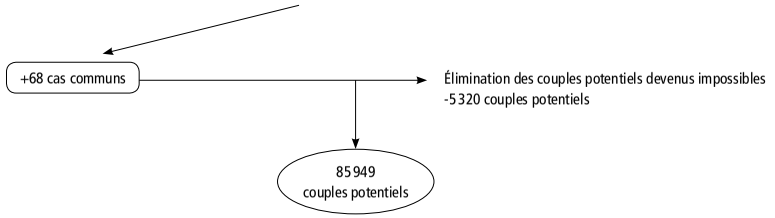




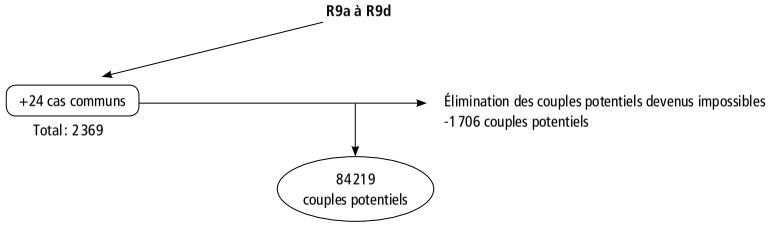
Après la requête 6, nous sommes à 2 215 cas communs et 96 625 couples potentiels restants.
Pour les requêtes suivantes, en plus de la différence entre nationalité et séjour hors de France, et des différences sur les CD4, nous ajoutons une différence d'un an sur la date de VIH.
Les requêtes suivantes sont construites sur le même modèle que précédemment.



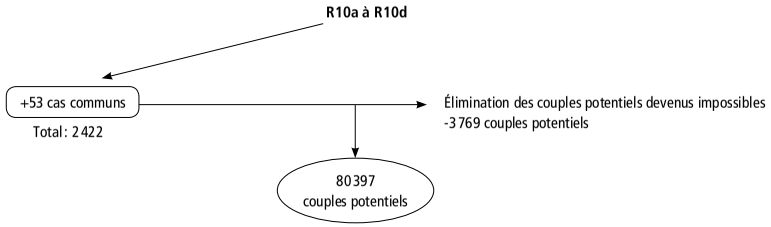
Après la requête 7, nous sommes à 2277 cas communs et 91337 couples potentiels restants.
Sélection des cas communs ayant une égalité sur la date et le nombre de CD4, des modes de contamination compatible et des pathologies communes.



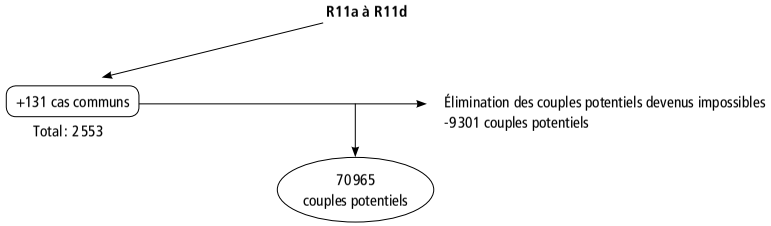
R9a Sélection des cas communs ayant une égalité sur l'hôpital, le mode de contamination, le diagnostic et la date de sida (sauf une), et une différence de date VIH <1an, une date de CD4<3 mois et de nombre de CD4≤100. Le département de domicile est égal, inconnu ou étranger.
Les requêtes suivantes (**R9b à R9d**) sont construites sur le même modèle que précédemment (les deux croisements successifs de pathologies et l'écart de la date de sida <6 mois).



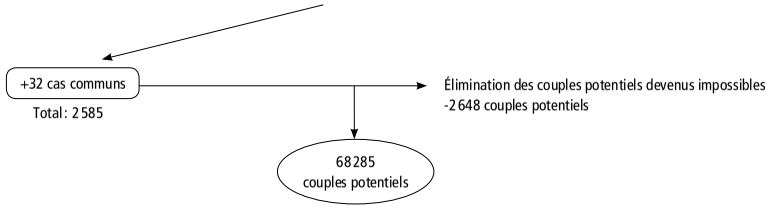
Les requêtes **R10a à R10d** sont identiques aux requêtes R9a à R9d sauf que la variable sur la différence de date de VIH est enlevée.



Les requêtes **R11a à R11d** sont identiques aux requêtes R9a à R9d sauf que la variable sur la différence de date de VIH est remise et que les variables sur les CD4 sont enlevées.

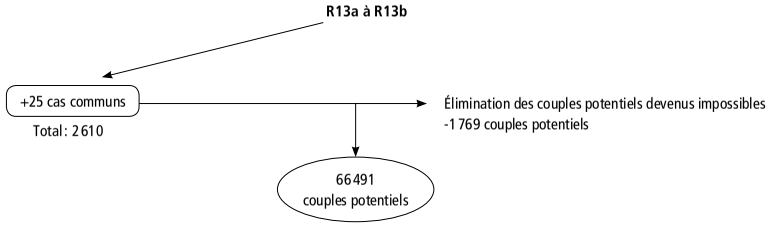


R12 Sélection des cas communs ayant une égalité sur la date de sida, le département de domicile (sauf étranger et inconnu), un mode de contamination compatible, la variable "nationalité" égale à la variable hors de France (sauf quand la nationalité est française) et au moins une pathologie commune (croisement pathologie 1 et 2).

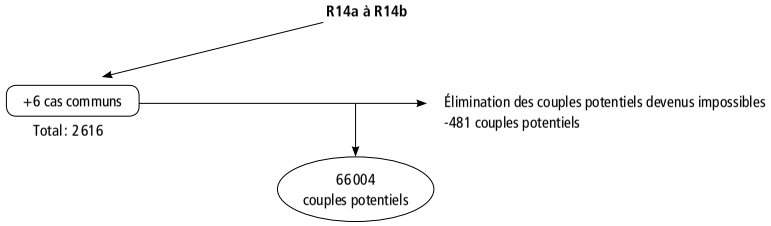


R13a Sélection des cas communs ayant une égalité sur la date de VIH et de sida, l'hôpital, un mode de contamination compatible et au moins une pathologie commune.

R13b Sélection des cas communs comme dans R38, mais l'égalité sur l'hôpital est remplacée par l'égalité sur le département de domicile (sauf étranger et inconnu).

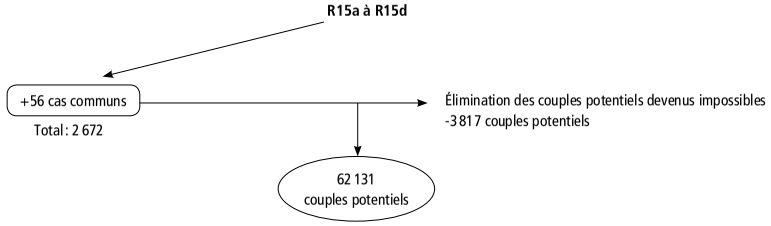


Les requêtes **R14a et R14b** sont identiques à R13a et R13b, mais une différence sur la date de VIH<1 an est autorisée.

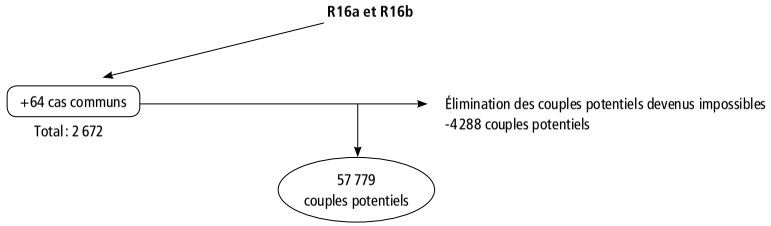


R15a Sélection des cas communs ayant une date de VIH et une date de sida identique. En plus de ces deux égalités, les patients doivent avoir une pathologie commune (croisement des pathologies 1, 2 et 3). Pour être moins restrictifs, certaines différences sur les pathologies sont admises. Cela concerne les lymphomes, les tuberculoses et les candidoses, où l'on ne s'intéresse plus à leur localisation. De plus, lorsque le patient présente un syndrome cachectique dans la base FHDH, il est toléré que cette pathologie ne soit pas renseignée dans la DO. En effet, dans la DO, le syndrome cachectique n'est saisi qu'en l'absence d'autres pathologies. Pour vérifier si le syndrome cachectique est présent, nous sommes retournés systématiquement au dossier papier du patient de la DO.

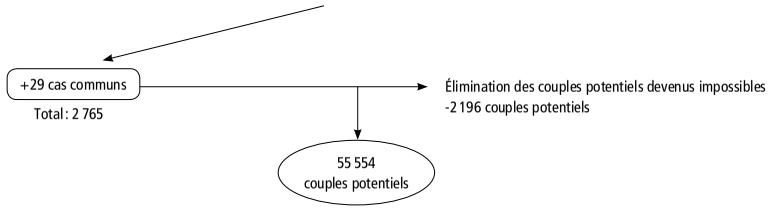
Pour la requête **R15b**, l'algorithme est identique à la R14a, mais une différence d'un mois sur la date de sida est autorisée.



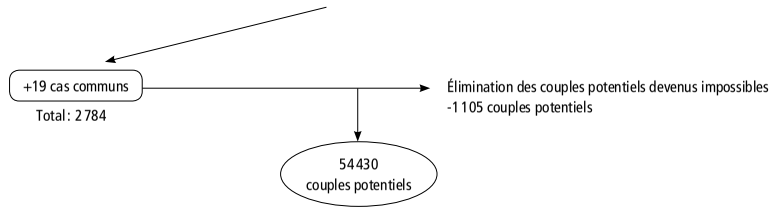
Les requêtes **R16a et R16b** sont identiques à R15a et R15b, mais une différence d'un an sur la date de VIH est tolérée.



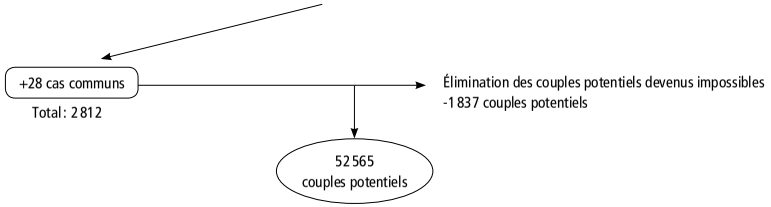
R17 Sélection des cas communs ayant une différence sur la date de sida <3 mois, le même hôpital et un diagnostic identique.



R18 Sélection des cas communs comme dans R17 mais, à la place d'un diagnostic identique, on tolère les mêmes différences sur les pathologies que dans R15.

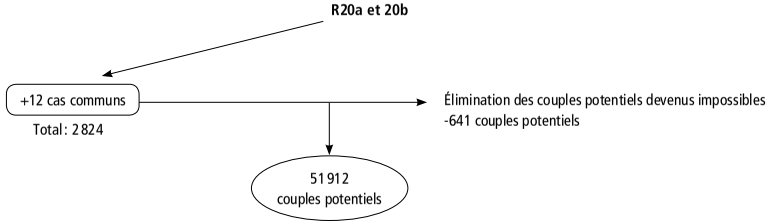


R19 Sélection des cas communs ayant un hôpital identique, une différence sur la date de sida <6 mois et une différence sur la date de VIH <1an.



R20a Sélection des cas communs ayant un hôpital identique, un mode de contamination égal ou inconnu, un département de domicile égal ou inconnu, une différence de sérologie <3 mois et la variable "nationalité" identique à la variable "hors de France" (ou nationalité française).

La requête **20b** est identique à la requête 20a, mais la variable département a été enlevée.



À ce stade, nous avons décidé d'arrêter la recherche de cas communs. Au total, l'algorithme aura permis l'identification de 2 824 cas communs. Nous disposons encore de 51 912 couples potentiels, mais leurs disparités sur un grand nombre de variables ne nous permettent plus d'identifier de cas communs.

(1) Par diagnostic égal, il est entendu une homologie stricte sur l'ensemble des pathologies classantes.

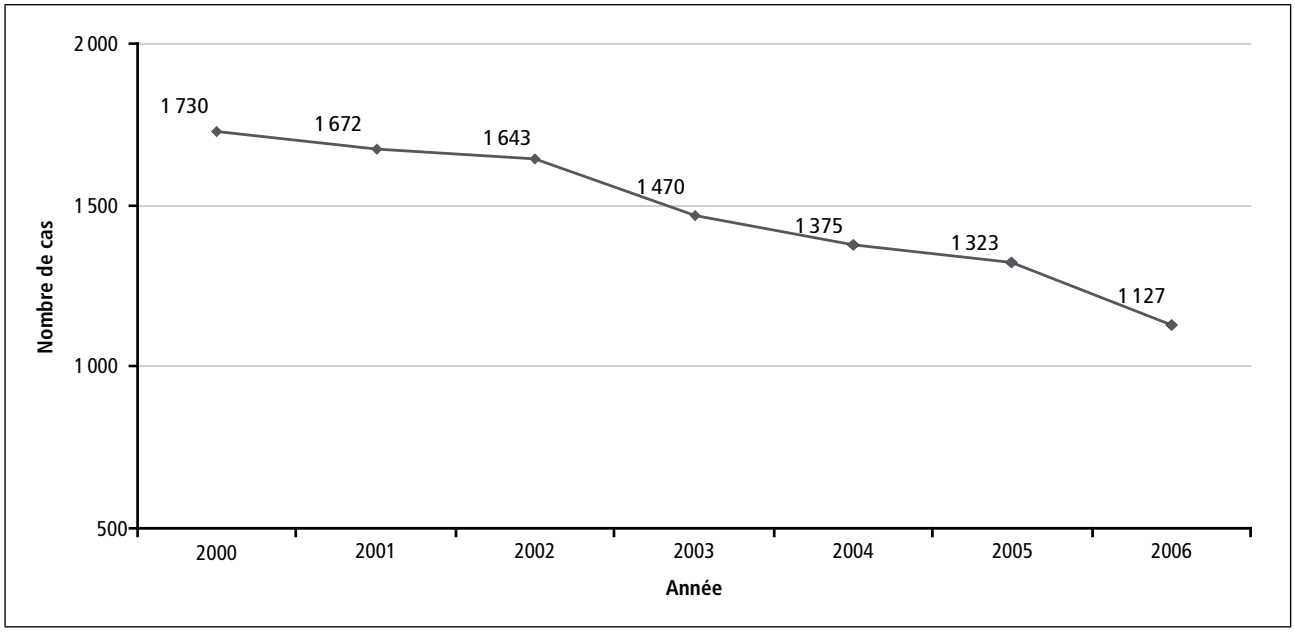
Annexe 5 – Liste des centres de la base FHDH

Centres	Ville ou hôpital
Auvergne	Clermont-Ferrand
Caen	Caen
Grenoble	Grenoble
Guadeloupe	Guadeloupe Saint-Martin
Guyane	Guyane
Lille-Tourcoing	Tourcoing
Bourgogne – Franche-Comté	Dijon Belfort
Martinique	Martinique
Montpellier	Montpellier Nîmes
Nancy	Nancy Reims
Nantes	Nantes
Nice	Nice Antibes
Pitié-Salpêtrière	Hôpital de la Pitié-Salpêtrière
Saint-Louis	Hôpital Saint-Louis Hôpital Lariboisière
Paris Ouest	Hôpital Necker
Bichat Claude Bernard	Hôpital Bichat Claude Bernard
Rennes	Rennes
Rouen	Rouen
Réunion	Réunion
Toulouse	Toulouse
Tours	Tours
Strasbourg	Strasbourg Mulhouse
Paris Centre	Hôpital Cochin Hôpital européen Georges Pompidou Hôpital Saint- Joseph
Paris Est	Hôpital Saint-Antoine Hôpital Tenon
Paris Sud	Hôpital Béclère Hôpital Henry Mondor Hôpital Kremlin Bicêtre Hôpital Paul-Brousse
Cisih 92	Hôpital Ambroise Paré Hôpital Louis Mourier Hôpital Raymond Poincaré
Cisih 93	Hôpital Avicenne Hôpital Verdier Hôpital Saint-Denis
Lyon	Hôtel Dieu Hôpital Croix Rousse Hôpital Edgar Herriot
Marseille	Hôpital Marseille-Nord Hôpital des Baumettes Hôpital Marseille-Conception Hôtel Dieu Hôpital Paoli Calmette Hôpital Saint-Marguerite Toulon Avignon Aix

Annexe 6 – Évolution du nombre de nouveaux cas de sida chez les adultes de 18 ans et plus au sein de la déclaration obligatoire du sida, France, 2000-2006 (données au 31/12/2008)

| Figure 2 |

Nouveaux cas de sida chez les adultes de 18 ans et plus déclarés au sein de la déclaration obligatoire du sida entre 2000 et 2006



Pourcentage de diminution du nombre de nouveaux cas :

- 2001 : -3,4 %
- 2002 : -1,7 %
- 2003 : -10,5 %
- 2004 : -6,5 %
- 2005 : -3,8 %
- 2006 : -14,8 %

Annexe 7 – Comparaison de l'exhaustivité de la FHDH par centre entre les périodes 1900-1993 et 2004-2006

Centres (ville ou hôpital)	FHDH	
	E % 1990-1993	E % 2004-2006
Auvergne	25,2	49,2
Caen	83,6	70,9
Grenoble	92,0	97,5
Guadeloupe	84,6	81,0
Guyane	57,8	65,7
Lille-Tourcoing	67,1	86,2
Bourgogne – Franche-Comté*	-	87,3
Martinique	78,6	69,1
Montpellier	57,3	76,3
Nancy*	-	16,8
Nantes	82,4	89,6
Nice*	-	76,2
Pitié Salpetrière	73,6	83,4
Saint-Louis*	-	74,3
Paris Ouest	42,8	65,6
Bichat Claude Bernard	45,6	47,2
Rennes	91,2	90,5
Rouen	75,6	74,7
Réunion	46,9	93,9
Toulouse	68,5	86,3
Tours	79,3	86,8
Strasbourg*	-	63,9
Paris centre*	-	81,1
Paris Est	41,9	48,9
Paris Sud	66,3	80,7
Corevih 92*	-	61,2
Corevih 93*	-	95,0
Lyon	70,1	71,1
Marseille*	-	**80,2**

** Centres pour lesquels nous ne pouvons pas comparer les résultats d'exhaustivité.*

Estimation de l'exhaustivité de la surveillance du sida par la méthode capture-recapture, France, 2004-2006

La surveillance du sida repose en France sur la déclaration obligatoire (DO), dont l'exhaustivité n'avait pas été évaluée depuis le début des années 1990. L'objectif de cette étude était de fournir une nouvelle estimation du nombre total de cas de sida en France et d'évaluer l'exhaustivité des bases de données utilisées.

La méthode capture-recapture à deux sources a été utilisée sur les cas de sida diagnostiqués de 2004 à 2006. La première source est la base de la DO du sida. La seconde source est composée de deux bases : la base de données hospitalière française sur l'infection à VIH (FHDH-ANRS-CO4), qui ne comprend pas l'Aquitaine, et la base du groupe d'épidémiologie clinique du sida en Aquitaine (Gecsa-ANRS-CO3). La comparaison de la base DO (n=3 816) avec les bases FHDH et Gecsa réunies (n=3 328) a permis d'identifier les cas communs (n=2 204).

Le taux d'exhaustivité de la DO du sida en France a été estimé à 66,1 % (IC 95 % : 65,1-67,2), celle de la FHDH à 55,6 % (IC 95 % : 54,7-56,5) et celle du Gecsa sur la région Aquitaine à 90,3 % (IC 95 % : 86,4-94,5). Le nombre total de cas de sida diagnostiqués chez les adultes de 18 ans et plus entre 2004 et 2006 a été estimé à 5 770 (IC 95 % : 5 679-5 861).

Par comparaison avec l'étude réalisée sur la période 1990-1993, l'exhaustivité de la FHDH a augmenté, passant de 47,6 % à 55,6 %, alors que celle de la DO du sida a diminué de 83,6 % à 66,1 %, diminution probablement liée à la mise en place de la DO du VIH en 2003.

Mots clés : diagnostic de sida, exhaustivité, méthode capture-recapture, sida, déclaration maladie, surveillance épidémiologique, banque données, système information, hôpital, Aquitaine

Estimation of completeness of AIDS surveillance with capture-recapture method, France, 2004-2006

AIDS surveillance relies on mandatory reporting (DO), whose completeness has not been estimated since the beginning of the 90s. The aim of this study was to supply a new estimate of the total number of AIDS cases in France and to estimate the completeness of the databases used.

The two sources capture-recapture method was used on AIDS cases diagnosed between 2004 and 2006. The first source is the database from the AIDS'DO. The second is composed of two databases: the French Hospital database on HIV (FHDH-ANRS-CO4) that does not include the Aquitaine region and the database of the AIDS Aquitaine Epidemiological group (Gecsa-ANRS-CO3).

The comparison between the DO database (n=3,816) and the FHDH and Gecsa database (n=3,328) has permitted to match common cases (n=2,204).

The completeness of the AIDS'DO in France was estimated at 66.1% (CI 95%: 65.1-67.2), the one of the FHDH at 55.6% (CI 95%: 54.7-56.5), and the one of the Gecsa in Aquitaine region at 90.3% (CI 95%: 86.4-94.5). The number of AIDS cases diagnosed in patients 18 years of age or older between 2004 and 2006 was estimated at 5,770 (CI 95%: 5,679-5,861).

Compared with the study carried out over the 1990-1993 period, the completeness of the FHDH increased from 47.6% to 55.6%, while the one of the AIDS'DO decreased from 83.6% to 66.1%. This decrease is probably linked to the implementation in 2003 of the HIV case reporting.

Citation suggérée :
Spaccaferri G, Cazein F, Lièvre L, Geffard S, Gallay A, Pillonel J. Estimation de l'exhaustivité de la surveillance du sida par la méthode capture-recapture, France, 2004-2006. Saint-Maurice (Fra) : Institut de veille sanitaire, juillet 2010, 36 p. Disponible sur : www.invs.sante.fr

INSTITUT DE VEILLE SANITAIRE
12 rue du Val d'Osne
94 415 Saint-Maurice Cedex France
Tél. : 33 (0)1 41 79 67 00
Fax : 33 (0)1 41 79 67 67
www.invs.sante.fr

ISSN : 1956-6956
ISBN-NET : 978-2-11-099274-1
Réalisé par Diadeis-Paris
Dépôt légal : juillet 2010