



Estimation de la sous-déclaration des cas de Sida en France par la méthode de capture-recapture : p. 19.

Épidémie de grippe 1996-1997 comparable à celle de l'an passé : p. 21.

N° 5/1997

28 janvier 1997

ÉTUDE

REÇU LE

19 FEV 1997

ESTIMATION DE LA SOUS-DÉCLARATION DES CAS DE SIDA EN FRANCE PAR LA MÉTHODE DE CAPTURE-RECAPTURE

P. BERNILLON*, L. LIÈVRE*, J. PILLONEL**, A. LAPORTE**, D. COSTAGLIOLA*,
Groupe d'épidémiologie clinique des CISH***

INTRODUCTION

Un système de surveillance épidémiologique est rarement exhaustif à 100 %, ce qui ne constitue pas un inconvénient majeur si on est capable d'évaluer quantitativement sa couverture.

Les méthodes de capture-recapture permettent d'estimer la taille d'une population non observée dans sa totalité et d'en déduire ainsi le niveau de couverture des systèmes de surveillance qui recensent cette population. Le terme capture-recapture fait référence au procédé qui consiste à « capturer-marquer-relâcher » un échantillon aléatoire d'une espèce animale, puis à capturer un second échantillon indépendant. Le nombre d'animaux marqués présents dans le second échantillon est utilisé pour estimer la taille totale de la population. En épidémiologie, l'application de cette technique nécessite un recensement de cas à partir d'au moins deux sources de données indépendantes; dans ce cas, on assimile chaque base (provenant d'un système de surveillance) à l'échantillon issu d'une capture animale.

Nous proposons une estimation de la sous-déclaration des cas de Sida en France, par application d'une méthode de capture-recapture.

Nous avons étudié la concordance entre les cas de Sida déclarés au Réseau national de Santé publique (RNSP) dans le cadre de la déclaration obligatoire et les cas de Sida présents dans la base clinico-épidémiologique du sujet séropositif suivi à l'hôpital constituée au Service commun n° 4 de l'INSERM (SC4) à partir du DMI2. Cette étude a obtenu l'autorisation de la Commission nationale de l'Informatique et des Libertés.

MATÉRIEL ET MÉTHODES

Les bases de données

La base de la déclaration obligatoire

Le Sida est une maladie à déclaration obligatoire (DO) depuis juin 1986, et le RNSP centralise les données du système national de surveillance mis en place peu après le début de l'épidémie (1982). La première base est extraite du fichier du RNSP et contient les 19625 cas de Sida âgés de 18 ans ou plus diagnostiqués entre le 1^{er} janvier 1990 et le 31 décembre 1993 et déclarés avant le 31 décembre 1995 au RNSP. Les variables utilisées sont les suivantes : date de naissance, sexe, département de résidence, groupe de transmission, date de Sida, pathologie(s) inaugurale(s) [4 au maximum], lieu de diagnostic (hôpital ou département si la déclaration est faite par un médecin de ville), statut (vivant ou décédé). Par ailleurs, on dispose de la date de décès (sauf pour 24 cas) pour les 13 024 (66,4 %) sujets connus comme étant décédés, et de la date de transfusion unique pour 460 des 678 patients dont le groupe de transmission correspond à une transfusion.

La base constituée à partir des données du DMI2

Le dossier médical, épidémiologique et économique de l'immunodéficience humaine (DMI2), est un logiciel mis en place dans les centres d'information et de soins de l'immunodéficience humaine (CISH), issu d'une collaboration entre le SC4 de l'INSERM et la Mission Sida de la Direction des hôpitaux. Ce logiciel est la propriété du ministère de la Santé. Le SC4 a en charge la gestion et l'analyse de la base de données clinico-épidémiologiques du sujet séropositif suivi à l'hôpital. Elle contient des informations sur des patients séropositifs infectés par le VIH-1 et/ou le VIH-2, suivis dans un CISH, ayant donné leur consentement écrit [1].

Les données utilisées pour cette étude constituent un sous-ensemble des données de la base disponible en décembre 1995, et concerne 11 172 patients séropositifs de 18 ans et plus ayant présenté un diagnostic de Sida entre janvier 1990 et décembre 1993. Les variables utilisées sont les suivantes : date de naissance (mois et année), sexe, groupe de transmission, date de Sida, date de premier suivi dans un CISH, date de dernier suivi, la (ou les) pathologies inaugurales (entre 0 et 8), l'hôpital où le patient est suivi, la date de décès pour les 5 963 (53,4 %) sujets qui sont connus comme étant décédés. Le département de domicile est disponible pour 8 829 patients, et la date de transfusion (lorsqu'elle est unique) est disponible pour 123 des 328 transfusés de la base.

Principes de la méthode de capture-recapture

La méthode de capture-recapture consiste à croiser les informations provenant de deux systèmes de surveillance afin d'identifier le nombre de cas communs appelés « doublons ». Le nombre total de cas est estimé à l'aide du nombre de cas dans chacun des 2 systèmes et du nombre de doublons en prenant en compte (après estimation) le nombre de cas omis par les 2 systèmes [2, 3].

Croisement des 2 bases

Les doublons sont les cas de Sida enregistrés dans les 2 systèmes ayant même mois et année de naissance, même sexe et même groupe de transmission pour lesquels toutes les autres informations disponibles sont proches (même date de Sida à 3 mois près, même date de décès à 1 mois près) ou compatibles (décédé dans une seule des 2 bases et bonne concordance sur les autres variables).

Hypothèses du modèle

Les estimateurs de la méthode de capture-recapture sont valides sous les hypothèses suivantes :

- tous les cas identifiés sont des cas réellement survenus ;
- la procédure de croisement des 2 bases identifie tous les doublons et uniquement les « vrais doublons » ;
- les 2 systèmes de surveillance sont indépendants, c'est-à-dire que le fait qu'un individu soit recensé par un des systèmes de surveillance n'augmente ni ne diminue la probabilité qu'il soit identifié par l'autre système.

Enquête réalisée auprès des techniciens d'études cliniques

Ce sont les écarts par rapport à l'hypothèse d'indépendance qui introduisent les plus grands biais dans les estimations [4]. Il n'est pas possible, dans le cas de 2 bases de données, de tester statistiquement l'indépendance.

Pour explorer ce phénomène, nous avons élaboré un questionnaire à destination des techniciens d'études cliniques des 54 centres participant à la base en décembre 1995 afin d'évaluer si l'un des systèmes de surveillance (DO ou DMI2) était utilisé pour déclarer les cas dans l'autre système.

* B3E, INSERM SC4, Institut fédératif Saint-Antoine de recherche en santé, faculté de médecine Saint-Antoine, université Pierre et Marie-Curie, 27, rue Chaligny, 75571 Paris Cedex 12.

** Réseau national de Santé publique, 14, rue du Val-d'Osne, 94415 Saint-Maurice Cedex.

*** Centres d'information et de soins de l'immunodéficience humaine (CISH) : voir liste en fin d'article.

RÉSULTATS

Identification des doublons

Le croisement des 2 bases de données a permis de déterminer 9 337 doublons. Le nombre total de cas distincts recensés par les 2 bases est 21 460. La détermination des doublons a mis en évidence une sous-déclaration des décès d'au moins 2,5 % pour la base du RNSP et 21,8 % pour la base du SC4.

Taux global d'exhaustivité

L'application de la méthode de capture-recapture a permis d'estimer le nombre total de cas de Sida diagnostiqués entre 1990 et 1993 à 23 482 (IC_{95 %} : 23 342 - 23 621). La couverture du RNSP est estimée à 83,6 % (IC_{95 %} : 82,9-84,3), celle du DMI2 à 47,6 % (IC_{95 %} : 46,9-48,3) pour la période 1990-1993. Le regroupement des 2 systèmes de surveillance permet de recenser 91,4 % du nombre total des cas.

Taux d'exhaustivité selon l'année de Sida

La stratification par rapport à l'année de diagnostic du Sida (tabl. 1) met en évidence une amélioration de la couverture du DMI2 au cours de la période étudiée, avec 39,2 % en 1990 et 53,5 % en 1993.

Taux d'exhaustivité selon le centre de déclaration

Quand on se limite aux hôpitaux participant à la base du SC4, la couverture de la base du RNSP est estimée à 85,8 % (IC_{95 %} : 85,2-86,4), celle de la base du SC4 est estimée à 70,4 % (IC_{95 %} : 69,7-71,1).

Indépendance

49 centres (sur 54) ont répondu au questionnaire envoyé aux techniciens d'études cliniques des hôpitaux participant à la base du DMI2. Nous avons pris comme critère d'indépendance le fait que les déclarations dans chaque système soient faites par des personnes différentes et à partir de dossiers différents. Il ressort de cette enquête que les déclarations sont faites de manière indépendante dans 30 centres et que les déclarations sont faites par la même personne et/ou à partir du même dossier dans 6 centres; 13 ont mentionné une dépendance partielle entre les 2 systèmes.

En stratifiant par rapport à cette classification (indépendance, dépendance partielle, non-indépendance) et en se limitant donc là encore aux hôpitaux participant à la base du SC4, la couverture de la base du RNSP est estimée à 85,7 % (IC_{95 %} : 83,4-88,0), celle de la base du SC4 à 70,7 % (IC_{95 %} : 67,9-73,4).

DISCUSSION-CONCLUSION

La méthode de capture-recapture a permis d'estimer la couverture de la DO à 83,6 % (IC_{95 %} : 82,9-84,3), et celle du DMI2 à 47,6 % (IC_{95 %} : 46,9-48,3) en ce qui concerne la surveillance des cas de Sida diagnostiqués chez des adultes en France entre 1990 et 1993. Le nombre total de cas est estimé à 23 481 (IC_{95 %} : 23 342 - 23 621) pour la période 1990-1993.

L'augmentation observée de la couverture de la base du SC4 (d'année en année de diagnostic du Sida) peut en partie être expliquée par la participation progressive des centres à la base.

La première hypothèse du modèle, qui concerne la pertinence des cas recensés par au moins l'un des systèmes, ne semble pas sujette à caution puisque les 2 systèmes adoptent la même définition. Il est vrai cependant qu'il existe probablement dans chacun des systèmes une petite fraction de sujets qui présentent une pathologie « classant Sida » tout en n'étant pas nécessairement séropositifs pour le VIH.

Concernant la deuxième hypothèse, les « vrais doublons » que notre procédure n'aurait pas identifiés sont nécessairement liés à des discordances non négligeables sur plusieurs variables, ce qui pose le problème de la qualité des données dans chacune des 2 bases. Un plus petit nombre de doublons entraînerait une surestimation de la taille de la population totale, et donc une sous-estimation des couvertures [5]. Pour évaluer « quantitativement » cette hypothèse, il faudrait retourner aux dossiers médicaux. Cette solution est impossible pour des raisons de confidentialité.

L'hypothèse d'indépendance est la plus délicate dans le cadre d'une méthode de capture-recapture car, d'une part, elle est difficile à évaluer lorsqu'on ne dispose que de 2 sources de données et, d'autre part, la validité des estimations obtenues peut être remise en cause si on s'écarte trop de cette hypothèse. D'une manière générale, une dépendance positive entre les 2 systèmes, c'est-à-dire que le fait d'être recensé dans une base augmente la probabilité de l'être dans l'autre, entraîne une surestimation de la couverture obtenue par la méthode. Inversement, une dépendance négative conduit à une sous-estimation de la couverture.

Dans le cas de notre application, les différents résultats obtenus nous font soupçonner la présence d'une faible dépendance positive entre les deux bases. En effet, les résultats de l'enquête montrent que les déclarations sont faites de manière indépendante dans un peu moins de deux tiers des cas. De plus, on retrouve une couverture différente pour la base du RNSP, sur la population totale (tabl. 1) et lorsque l'on se limite aux centres qui participent à la base du SC4 (83,6 % vs 85,8 %, $p < 10^{-3}$). Cependant, Brenner a montré que lorsqu'une des 2 bases a une couverture importante, la surestimation due à une dépendance positive reste modérée [4]. Un moyen d'évaluer précisément la dépendance entre les systèmes serait d'utiliser une troisième base (par exemple, une base issue de celle du Service commun n° 8 de l'INSERM qui recense les données relatives à toutes les causes de décès en France).

RÉFÉRENCES

- [1] INSERM Service commun n° 4 et Groupe d'épidémiologie clinique des CISH. – Base de données clinico-épidémiologiques du sujet séropositif suivi à l'hôpital. Manifestations cliniques et traitements à partir de 1989. – BEH, 1994 : 35.
- [2] SEKAR C.C., DEMING W.E. – On a method of estimating birth and death rates and the extent of registration. – J. Amer. Statist. Assn., 1949; 101-115.
- [3] HUBERT B., DESENCLOS J.-C. – 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 1990. – Rev. épidémiol. Santé publique, 1993; 41 (3) : 241-9.
- [4] BRENNER H. – Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. – Epidemiology, 1995; 6 (1) : 42-8.
- [5] BRENNER H. – Application of capture-recapture methods for disease monitoring : potential effects of imperfect record linkage. – Method. Inform. Med., 1994; 33 : 502-6.

Tableau 1. – Taux d'exhaustivité de la déclaration obligatoire (DO) et de la base des données collectées via le logiciel DMI2 selon l'année de diagnostic du Sida France (1990-1993)

	Nombre de cas observés			Estimation du nombre total de cas		Taux d'exhaustivité			
	DO	DMI2	Doublons	N	IC _{95 %}	DO		DMI2	
						%	IC _{95 %}	%	IC _{95 %}
Global.....	19 625	11 172	9 337	23 482	23 342-23 621	83,6	82,9-84,3	47,6	46,9-48,3
Année de Sida :									
1990.....	4 379	2 042	1 717	5 208	5 427-5 672	84,1	82,5-85,7	39,2	37,8-40,7
1991.....	4 596	2 488	2 017	5 669	5 826-6 044	81,1	79,5-82,6	43,9	42,4-45,3
1992.....	5 121	3 154	2 646	6 104	5 919-6 073	83,9	82,6-85,2	51,7	50,3-53,0
1993.....	5 529	3 488	2 957	6 522	6 148-6 319	84,8	83,6-86,0	53,5	52,2-54,8
Total.....	19 625	11 172	9 337	23 502	23 514-23 914	83,5	80,7-86,3	47,5	44,7-50,3