Estimation de prévalences et d'incidences à partir d'enquêtes épidémiologiques transversales répétées auprès de populations difficiles d'accès : Application au virus de l'hépatite C chez les usagers de drogues en France.

Lucie Léon

NNT : 2016 SACLS440

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Université Paris-Sud

Ecole doctorale n°570
EDSP Santé publique
Spécialité de doctorat : Épidémiologie

par

Lucie LÉON

Estimation de prévalences et d'incidences à partir d'enquêtes épidémiologiques transversales répétées auprès de populations difficiles d'accès. Application au virus de l'hépatite C chez les usagers de drogues en France

Thèse présentée et soutenue à Santé publique France, le 6 décembre 2016.

Composition du Jury :

M. Loic Josseran	Professeur des Universités-Praticien Hospitalier Université de Versailles	Président du jury
M. Ahmadou Alioum	Professeur des universités Université de Bordeaux	Rapporteur
M. René Ecochard	Professeur des Universités-Praticien Hospitalier Université Claude Bernard Lyon 1	Rapporteur
M. Viet Chi Tran	Maître de conférences Université Sciences et Technologies de Lille	Examinateur
M. Yann Le Strat	Biostatisticien Santé publique France	Directeur de thèse
Mme. Marie Jauffret-Roustide	Chargé de recherche Santé publique France	Encadrante de thèse

Remerciements

Je souhaite remercier ici toutes les personnes qui, de près ou de loin, ont contribué à l'aboutissement de cette thèse réalisée à la Direction des Maladies Infectieuses au sein de Santé publique France.

Mes remerciements s'adressent, en premier lieu, à Yann Le Strat qui m'a donné l'opportunité de m'inscrire en thèse en me confiant ce projet passionnant et en me faisant l'honneur de diriger cette thèse. Le partage d'expériences, son expertise scientifique et ses qualités pédagogiques ont été pour moi un enrichissement permanent.

Je remercie également Marie Jauffret-Roustide, co-encadrante de la thèse, pour la mise à disposition des données nécessaires à la réalisation de ce travail et son expertise en tant que sociologue.

Je remercie vivement tous les membres du jury pour avoir accepté d'évaluer ce travail. A Monsieur Loic Josseran qui me fait l'honneur de présider ce jury. A Messieurs Ahmadou Alioum et René Ecochard qui me font l'honneur d'être rapporteurs de cette thèse. A Monsieur Viet Chi Tran pour l'honneur qu'il me fait d'être examinateur de cette thèse.

Je voudrais également remercier Josiane Pillonel pour son aide et sa collaboration.

Je remercie particulièrement Christine Larsen pour son écoute, sa réactivité, ses relectures attentives et son expertise sur l'épidémiologie de l'hépatite C.

Je tiens aussi à remercier Audrey Bourgeois, assistante de l'EDSP, pour sa gentillesse, sa disponibilité, sa réactivité et pour toutes les questions et procédures administratives liées à la réalisation de cette thèse. A Jean Bouyer, directeur de l'EDSP, j'exprime ici mes remerciements pour sa réactivité et ses conseils.

Mes remerciements s'adressent également à Bertrand Xerri qui m'a soutenu lors de la présentation initiale de ce projet à Santé publique France et, Jean-Claude Desenclos pour son écoute et ses conseils avisés.

Je remercie chaleureusement Elisabeth Couturier et Dieter Van Cauteren pour leur soutien et leurs encouragements.

J'exprime toute ma gratitude à tous les collègues de mon unité (Biostatistiques, Appui à la Surveillance et aux Emergences), pour leur soutien, leur implication, les moments de détente et de rires : Julien Durand, Camille Pelat, Etienne Lucas, Adèle Grembombo, Anne-sophie Barret, Daniel Dubois, Denise Gmazel et particulièrement, Didier Che pour son regard critique et son écoute, et Cécile Sommen pour sa relecture et nos échanges constructifs.

Je voudrais également remercier Bruno Coignard et Muriel Lacoste pour leur implication sur les aspects logistiques dans les derniers mois. Cela a été une grande aide pour moi.

Je remercie affectueusement Elise, Laure, Nico, Manu, Claire et Delphine pour nos moments privilégiés et pour votre soutien. Merci à Olivier, Christophe, Yanna, Sandrine et Laetitia, pour votre présence constante et pour avoir subi mon indisponibilité durant plusieurs mois avec votre réplique toute faite « *Ah oui! C'est vrai.....la thèse* ». Merci à Vanessa pour ton soutien, ta relecture minutieuse et ton aide précieuse. Merci à Claude, pour ton soutien, ta présence et nos rires en toutes circonstances. *Pou ri nou ka ri.*

Je témoigne toute ma reconnaissance à mes parents, mon frère et ma soeur, pour leur présence, leur confiance et leur soutien constant, plus particulièrement à mon père pour ses précieux conseils et ses relectures méticuleuses jusqu'au bout du bout.

A tous mes proches avec leur question récurrente et inévitable « *Alors c'est quand ? Tu as une date?*» qui, de cette façon, s'associaient complètement à mon objectif final, je vous dis merci.

Enfin, je renouvelle mes remerciements à Yann le Strat, à la fois mon directeur de thèse, le coordonnateur du programme statitisques de l'unité et mon collègue de bureau avec qui j'ai pu passer ces années de thèse dans les meilleures conditions. Je peux notamment noter la technique du *changement de casquette*, les pauses café/thés, les échanges culturels, les rires... Je te remercie pour ta confiance, tes qualités humaines et ton soutien au cours de ce travail.

Merci - *Mèsi on pil*

Rézimé la an kréyol

Viris a épatit C (VHC), sé on pwoblem majè a santé piblik, é contaminasyon an Fwans ka vinn plis dè sé izajé a dwòg (ID) la. Lè ou ka fè sa yo ka kriyé dé "ankèt épidémiolojik" owa sé popilasyon la sa, pou ou pé vwè évolisyon a VHC la, sa rèd, dabò pou yonn, padavwa pratik a yo ilégal. Adan popilasyon la sa ni dé kalté moun, moun ou pé jwenn (ou ka fasadé) padavwa yo kay adan sa yo ka kriyé an fwansé dé "*lieux d'enquêtes*" (ki vlé di *koté yo pé ankété*) épi moun ki kaché padavwa yo pa kay pon koté ki répèwtorié. Pou nou pé ankété si yo, nou konsidéré on échantiyonaj *koté-moman* ("*lieux-moments*" an fawnsé é TLS an anglé), èvè on échantiyonaj kè sé répondan la ka kondui yo menm. Lè ou vwè nou fòwmalizé TLS la adan on kad a sondaj indirekt, nou pwopozé on èstimatè pou on total é on propòwsyon ki ka tchen'n kont dè frékantasyon miltip é étérojèn a sé *lieux d'enquêtes* la. Nou ka rèkomandé métòd la sa pou estimé prévalans a on maladi adan étid ki fèt owa popilasyon ki ka frékanté dé sèwvis, menm si ni érè anlè sa sé pawtisipan la ka déclaré kom frékantasyon. On ankèt kè yo ka kriyé ANRS-Coquelicot té fèt an 2004 owa sé ID la ki té ka frékanté sant dédyé, i woufèt an 2011, é i tè pèwmèt kè yo èstimé prévalans a VHC la a 43,7%. Pawtan dè sé dé anket la sa, nou estimé insidans a VHC la silon aj é silon tan èvè konstriksyon a on modèl matématik ki té ka rèpozé asi fòwmilasyon a on rèlasyon ant prevalans épi insidans. Modèl ta la, sé té on konbinézon a on modèl konpawtimantal èvè on modèl dè régrésyon. Insidans a VHC la touvé-y èstimé an 2011 a 4,4/100 moun asi on lanné. Apwòch la sa, sé on bon altèwnativ pou nou pé èstimé insidans a on maladi, asi la baz dè ankèt épidémiolojik transvèwsal é lè ou vwè pa ni "*cohorte*" ou byen pon test biolojik ki té ké pèwmèt dè identifyé on infèksyon ki fré. Kanmenm sé izajé a dwòg la ka sanm ni plis kompòtman a risk, nou pé atann nou a vwè on bès a insidans a VHC la ki ka kontinyé padavwa bès a prévalans la, mèzi kè yo mèt an plas pou rédui sé risk la, èvè lé avansé térapétik.

Résumé

Le virus de l'hépatite C (VHC) est un problème majeur de santé publique dont les usagers de drogues (UD) constituent la principale source de contamination en France. Réaliser des enquêtes séro-épidémiologiques auprès de cette population pour suivre la dynamique du VHC s'avère difficile notamment en raison de leurs pratiques illicites. Cette population est en partie accessible par les lieux d'enquêtes et en partie "cachée" car ne fréquentant aucun lieu répertorié. Pour enquêter chaque partie, nous avons considéré l'échantillonnage lieux-moments (TLS) puis l'échantillonnage conduit par les répondants. Après avoir formalisé le TLS dans le cadre d'un sondage indirect, nous avons proposé un estimateur pour un total et une proportion, qui tient compte de la fréquentation multiple et hétérogène des lieux d'enquêtes. Nous recommandons cette méthode pour estimer la prévalence d'une maladie dans des études auprès de populations fréquentant des services, même en cas d'erreurs sur les fréquentations déclarées par les participants. L'enquête ANRS-Coquelicot réalisée en 2004 auprès des UD fréquentant des centres dédiés, puis répétée en 2011, a permis d'estimer la prévalence du VHC à 43,7%. A partir des deux enquêtes, nous avons ensuite estimé l'incidence de l'infection à VHC par âge et en fonction du temps en construisant un modèle mathématique reposant sur la formulation d'une relation entre la prévalence et l'incidence. Ce modèle consistait en la combinaison d'un modèle compartimental et d'un modèle de régression. L'incidence de l'infection à VHC a ainsi été estimée à 4,4/100 personnes-années en 2011. Cette approche est une alternative satisfaisante pour estimer l'incidence d'une maladie à partir d'enquêtes épidémiologiques transversales en l'absence de cohorte ou de tests biologiques permettant d'identifier les infections récentes. Compte tenu de la baisse de la prévalence, des mesures de réduction des risques et des avancées thérapeutiques, une diminution de l'incidence de l'infection à VHC devrait se poursuivre malgré une potentielle augmentation des comportements à risque des UD.

Mots clés : Populations difficiles d'accès, Usagers de drogues, Virus de l'hépatite C, Techniques de sondages, Prévalence, Incidence.

Abstract

Hepatitis C virus (HCV) is a public-health issue that drug users (DU) remain the major source of contamination in France. Conducting seroepidemiological surveys among this population to assess the HCV dynamic is difficult particularly due to their illicit practices. This population can be accessible through survey locations or can be hidden (who does not visit any location). To survey each part, we presented time-location sampling (TLS) and respondent-driven sampling. We presented TLS in the context of an indirect sampling and proposed a design-based inference taking into account the frequency of venue attendance (FVA) to estimate a total or a proportion. We recommend this method for estimating the prevalence of a disease in surveys among hard-to-reach populations, even if errors occur in the FVA reported by the participants. The ANRS-Coquelicot survey carried out in 2004 among DU attending centres providing services to drug users, then repeated in 2011, allowed us to estimate the HCV prevalence at 43.7%. Using these two surveys, we estimated age- and time-dependent HCV incidence from a mathematical model linking prevalence and incidence. This model consisted in combining a compartmental model with a regression model. The HCV incidence was thus estimated at 4.4/100 person-years in 2011. This method is an alternative approach to estimate incidence of a disease from cross-sectional epidemiological data in the absence of cohort or biological tests to identify acute infections. The decline in HCV incidence is to be expected given decreasing prevalence, recent developments in harm reduction measures and new therapeutic approaches despite a potential increase of at-risk behaviors.

Keywords: Hard-to-reach populations, Drug users, Hepatitis C Virus, Design surveys, Prevalence, Incidence.

Liste des productions scientifiques

Articles publiés

Léon L., Kasereka S., Barin F., Larsen C., Weill-Barillet L., Pascal X., Chevalliez S., Pillonel J., Jauffret-Roustide M., Le Strat Y. Age- and time-dependent prevalence and incidence of hepatitis C virus infection among drug users in France, 2004-2011 : Model-based estimation from two national cross-sectional serosurveys. *Epidemiology and Infection*, (2016), 1-13.

Léon L., Des Jarlais D., Jauffret-Roustide M., Le Strat Y. Update on Respondent-Driven Sampling : Theory and Practical considerations for studies of persons who inject drugs. *Methodological Innovations 9* (2016), 1-9.

Weill-Barillet L., Pillonel J., Semaille C., **Léon L.**, Le Strat Y., Pascal X., Barin F., Jauffret-Roustide, M. Hepatitis C virus and HIV seroprevalences, sociodemographic characteristics, behaviors and access to syringes among drug users, a comparison of geographical areas in France, ANRS-Coquelicot 2011 survey. *Revue d'Épidémiologie et de Santé Publique 64*, 4 (2016), 301-312.

Léon L., Jauffret-Roustide M., Le Strat Y. Design-based inference in time-location sampling. *Biostatistics 16*, 3 (2015), 565-579.

Jauffret-Roustide M., Pillonel J., Weill-Barillet L., **Léon L.**, Le Strat Y., Brunet S., Benoit T., Chauvin C., Lebreton M., Barin F., Semaille, C. Estimation de la séroprévalence du VIH et de l'hépatite C chez les usagers de drogues en France - premiers résultats de l'enquête ANRS-COQUELICOT 2011. *Bulletin Épidémiologie Hebdomadaire 39-40*, (2013), 504-509.

Communications orales

Jauffret-Roustide M., Molinier M., **Léon L.**, Le Strat Y., Barin F., Pillonel J. Facteurs de vulnérabilité individuels et structurels de l'exposition au VIH et au VHC chez les usagers de drogues en France, enquête ANRS-Coquelicot 2013. 8ième Conférence Internationale Francophone VIH/Hépatites AFRAVIH Avril 2016.

Jauffret-Roustide M., **Léon L.**, Pascal X., Weill-Barillet L., Le Strat Y., Semaille C., Barin F., Chevalliez S., Pillonel J. High biological-based HCV incidence and increasing frequency of high-risk practices among IDUs in France : What are the implications for harm reduction models ? Lisbon Addiction. Rapid Communication at the First European Conference on addictive

behaviours and dependencies, 23-25 September 2015, Lisbon.

Léon L., Jauffret-Roustide M.,Le Strat Y. Inférence pour l'échantillonnage lieux-moments. 8ième Colloque Francophone sur les Sondages. Novembre 2014.

Jauffret-Roustide M., Cazein F., Pillonel J., Weill-Barillet L, **Léon L.**, Le Strat Y., Lot F., Brunet S., Semaille C. Situation épidémiologique du VIH/Sida en France : enquête ANRS Coquelicot et DO VIH. AFRAVIH, 29-31 avril 2014, Montpellier.

Léon L. Prise en compte de la fréquentation multiple des lieux d'enquêtes - Illustration sur Coquelicot 2011, une enquête auprès des usagers de drogues. Journée scientifique Epiter. Septembre 2013.

Table des matières

Liste des abréviations

AAD	:	Antiviraux d'action directe
AFEF	:	Association pour l'étude du foie
ANRS	:	Agence Nationale de Recherche sur le Sida
ARN	:	Acide RiboNucléique
CAARUD	:	Centre d'Accueil et d'Accompagnement à la Réduction des Risques pour les Usagers de Drogues
CSAPA	:	Centres de Soins, d'Accompagnement et de Prévention en Addictologie
CNR	:	Centre National de Référence
DBS	:	*Dried Blood Spot* (papier buvard)
ELISA	:	*Enzyme-Linked Immunosorbent Assays*
HSH	:	Homme ayant des relations Sexuelles avec d'autres Hommes
InVS	:	Institut de Veille Sanitaire
MGPP	:	Méthode Généralisée du Partage des Poids
NIBSC	:	*National Institute for Biological Standards and Control*
OFDT	:	Observatoire Français des Drogues et des Toxicomanies
OMS	:	Organisation Mondiale de la Santé
OR	:	Odds Ratio
PCR	:	*Polymerase Chain Reaction*
PES	:	Programmes d'Échanges de Seringues
PF	:	Période Fenêtre
RDS	:	*Respondent-Driven Sampling*
RIBA	:	*Recombinant Immuno Blot Assay*
RR	:	Risque Relatif
TLS	:	*Time Location Sampling*
TSO	:	Traitements de Substitution aux Opiacés
TROD	:	Tests Rapides d'Orientation Diagnostic
UD	:	Usager de Drogues
UDI	:	Usager de Drogues Injecteur
UI	:	Unité Internationale
VHB	:	Virus de l'Hépatite B
VHC	:	Virus de l'Hépatite C
VIH	:	Virus de l'Immunodéficience Humaine

Introduction

Les enquêtes épidémiologiques transversales sont réalisées afin d'évaluer l'état de santé des populations (ampleur, fréquence, répartition, gravité, dynamique de maladies, détermination des facteurs associés, etc.). Ces enquêtes reposent généralement sur un échantillon aléatoire d'individus et font appel à des outils statistiques en développement constant depuis plusieurs décennies, aussi bien pour construire l'échantillon que pour réaliser l'inférence. L'inférence consiste à estimer des indicateurs épidémiologiques tels que le nombre de personnes infectées, une prévalence (proportion de personnes infectées) ou une incidence (nouvelles personnes infectées sur une période donnée) au sein d'une population. D'autres indicateurs comme des forces d'associations dans les analyses étiologiques peuvent être estimés.

Enquêtes épidémiologiques auprès de populations difficiles d'accès

D'un point de vue de santé publique, ces enquêtes sont essentielles pour étudier des populations particulièrement exposées et vulnérables vis-à-vis de maladies transmissibles de personne à personne [115,136]. Les usagers de drogues (UD), les hommes ayant des relations sexuelles avec d'autres hommes (HSH), les travailleurs du sexe, les personnes sans domicile ou certains migrants sont des exemples de populations particulièrement susceptibles d'être infectées par des maladies infectieuses (VIH, hépatites, infections sexuellement transmissibles, etc.) [10, 17, 140, 159].

Ces populations sont par ailleurs considérées comme étant difficiles d'accès ou difficiles à enquêter pour plusieurs raisons [147]. En effet, certaines populations représentent une part très faible ou très rare de la population générale, ce qui nécessite la mise en oeuvre de moyens coûteux pour les atteindre. Pour d'autres populations, il est difficile de localiser ou de contacter des personnes mobiles ou cachées. Parfois, ce sont des populations où les personnes sont réticentes à participer à une enquête parce qu'elles sont socialement vulnérables ou ne parlent

pas le français. Certaines personnes redoutent également une rupture de l'anonymat qui pourrait affecter leur vie quotidienne et restent donc peu accessibles compte tenu de leurs caractéristiques (pratiques illicites, précarité, etc.). Enfin, des personnes ne souhaitent pas révéler certaines informations prévues dans le protocole d'enquête (dormir dans la rue, consommer des drogues, leur nationalité, etc.) ou encore ne s'identifient pas à la population définie par le protocole. D'un point de vue pratique, il est nécessaire de distinguer deux groupes d'individus appartenant à ces populations : les populations dites visibles et les populations dites invisibles ou cachées. Une population visible peut être accessible par les lieux qu'elle fréquente (des centres de santé, des lieux de convivialité, des points soupe, etc.). Une population invisible est celle qui ne fréquente aucune structure, aucun lieu répertorié ou listé.

Dans des populations difficiles d'accès notamment en raison du caractère illicite de certaines pratiques (recours à la prostitution, utilisation de drogues), de la stigmatisation de certaines d'entre elles et de la difficulté à les approcher (conditions de vie, insécurité), la réalisation d'études séro-épidémiologiques est délicate. C'est un réel défi d'un point de vue méthodologique, d'une part par la réalisation d'enquêtes complexes résultant de la difficulté à atteindre ces populations et d'autre part par le développement de méthodes statistiques permettant d'estimer correctement des prévalences et des incidences tout en tenant compte des spécificités de ces populations. Ces deux indicateurs, prévalence et incidence, sont pourtant indispensables pour mesurer le niveau d'une infection dans la population et sa dynamique au cours du temps.

Cependant, lorsque la population est vulnérable et peu accessible, il est d'autant plus difficile d'estimer la prévalence voire l'incidence qui nécessite d'intégrer des informations épidémiologiques mais aussi biologiques et comportementales pour une infection dont les modes de transmission sont souvent complexes, notamment en l'absence de tests d'infection récente. L'essentiel de cette thèse porte donc sur l'approche méthodologique permettant d'estimer la prévalence et l'incidence de l'infection par le virus de l'hépatite C (VHC) parmi des usagers de drogues en France à partir d'enquêtes transversales.

L'hépatite C chez les usagers de drogues

L'hépatite C est une maladie répandue dans le monde entier avec 3% de la population mondiale touchée selon l'OMS. Ses caractéristiques en termes de complications graves et de mortalité

élevée font d'elle un problème majeur de santé publique. Certains auteurs ont montré que le nombre de décès liés aux hépatites était plus élevé que le nombre de cas de sida [133, 156]. Moins médiatisée que le VIH, l'hépatite C continue de faire parler d'elle depuis plus de 20 ans et intéresse la communauté scientifique et internationale (biologie, épidémiologie, médecine, politique, pharmacologie, presse écrite). A ce jour, aucun vaccin n'est disponible. Jusqu'à récemment, les personnes infectées par le VHC bénéficiaient de traitements contraignants (lourdeur des effets secondaires, durée de traitement, etc.) et relativement coûteux. En 2015, l'arrivée de nouveaux traitements thérapeutiques (les antiviraux d'action directe (AAD)) donne beaucoup d'espoir en termes de guérison et de prévention. En France, la transmission du VHC se fait principalement par l'usage de seringues ou tout autre matériel contaminé. De plus, les mesures de réduction des risques (programmes d'échanges de seringues (PES), traitement de substitution aux opiacés (TSO), etc.), efficaces ces dernières années contre le VIH, n'ont pas eu le même impact contre le VHC [32]. C'est la raison pour laquelle la prévalence reste élevée chez les UD, faisant d'eux la source essentielle de nouvelles contaminations.

La population des UD est une population vulnérable car particulièrement à risque d'être infectée par des maladies infectieuses et peu accessible compte tenu de leurs activités illégales. Déjà considérée comme difficile d'accès, elle est aussi souvent difficile à définir car c'est une population aux profils sociaux et aux rapports à l'usage de drogues diversifiés. On peut cependant définir deux sous-populations : (1) celle qui fréquente les centres spécialisés et qui représente une part particulière des usagers ayant un "rapport problématique aux drogues" et plutôt socialement précaire [66], et (2) celle qui est cachée car inaccessible via des enquêtes épidémiologiques réalisées dans les lieux d'enquêtes. Cette dernière sous-population comprend les usagers de drogues ayant un usage "régulé" des drogues, les plus socialement insérés, les plus jeunes et les femmes [66].

En France, deux enquêtes épidémiologiques transversales avec un recueil de prélèvements biologiques ont pu être réalisées auprès de la population des UD fréquentant des centres dédiés [67, 69]. Dans cette thèse, nous nous sommes attachés à développer des méthodes qui puissent être appliquées de manière générale à des populations composées d'individus accessibles par les lieux qu'ils fréquentent et pouvant avoir des fréquentations hétérogènes dans ces lieux.

L'utilisation d'enquêtes épidémiologiques transversales répétées dans le temps auprès de populations difficiles d'accès pour estimer des indicateurs épidémiologiques nous invite à rappeler quelques termes de la théorie des sondages avant de présenter le plan de la thèse.

Théorie des sondages - termes introductifs

Pour diverses raisons, notamment de coût et d'accès aux individus, il n'est généralement pas envisageable d'interroger l'ensemble des individus d'une population pour lesquels on souhaite établir un état de santé. Cela nécessite l'utilisation d'enquêtes par sondage qui, par définition, ont pour objectif de fournir des estimations sans biais et les plus précises possibles dans une population à partir d'un échantillon de celle-ci.

On appelle **population d'intérêt**, la population pour laquelle on souhaite produire des estimations. Elle est définie par sa nature (une personne, un établissement, etc.), ses caractéristiques intrinsèques (sexe, statut sérologique, facteurs de risques, etc.), sa localisation (ville, région, etc.) et la date à laquelle on s'y intéresse.

La population d'intérêt est composée d'un ensemble d'individus appelés **unités d'intérêt**. On peut citer par exemple la patientèle ayant consulté un médecin, la population des patients suivis dans les services hospitaliers dans un département, la population des usagers de drogues par injection dans une ville ou les habitants vivant dans un périmètre autour d'une usine d'incinération d'ordures ménagères.

Une fois la population d'intérêt définie, des unités d'intérêt appartenant à cette population sont tirées au sort selon une méthode d'échantillonnage donnée (ou **plan de sondage**) qui permet de construire un **échantillon** composé des répondants (ou participants) à l'enquête.

Pour chaque unité de l'échantillon, on recueille ensuite une ou plusieurs caractéristiques, par exemple le fait d'être vacciné ou non contre la rougeole pour un échantillon d'élèves, le nombre de consultations pour syndromes grippaux pour un échantillon de praticiens, être porteur ou non de l'antigène AgHbs pour un échantillon d'individus, l'occurrence d'un épisode de gastro-entérites aiguës dans les 3 dernières semaines pour un échantillon de personnes âgées ou le statut sérologique VIH pour un échantillon d'hommes ayant des relations sexuelles avec des hommes. On appelle ces caractéristiques les **variables d'intérêt**. Notons ici qu'elles n'ont rien de variable car ce sont des caractéristiques propres de la personne enquêtée. C'est l'inclusion ou non des

individus qui est aléatoire.

Soit une population finie U composée de N unités indexées par i $(i = 1, ..., N)$ et S un échantillon de taille n issu de cette population d'intérêt (Figure 1). Si on s'intéresse à l'infection au VHC (présence des anticorps anti-VHC), la variable d'intérêt notée y_i pour toute unité $i \in U$ est définie par :

$$y_i = \begin{cases} 1 & \text{si présence d'anticorps anti-VHC} \\ 0 & \text{sinon} \end{cases}$$

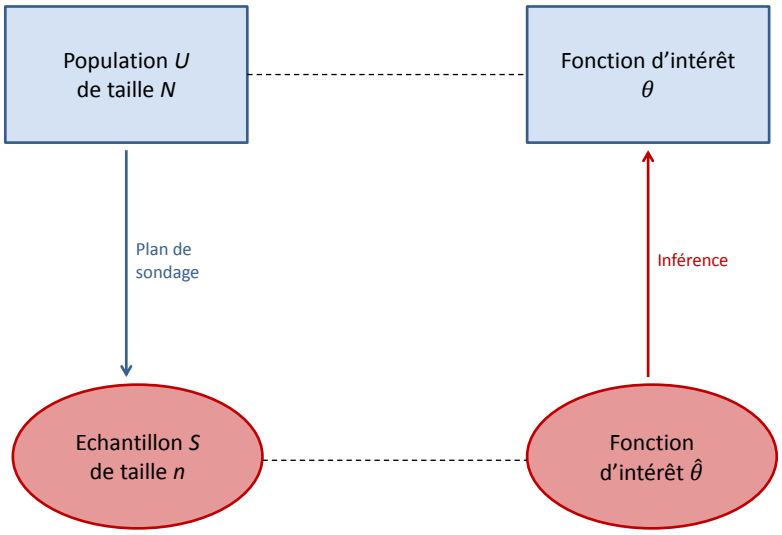


FIGURE 1 – Échantillonnage et inférence

En épidémiologie, l'intérêt se porte sur des indicateurs épidémiologiques qui sont fonction de ces variables d'intérêt. Ces indicateurs que l'on souhaite estimer dans la population d'intérêt sont appelés **fonctions d'intérêt** : un total (par exemple, le nombre de porteurs de l'antigène AgHbs en France, le nombre d'usagers de drogues par injection à Paris), une moyenne (pression artérielle moyenne dans une population à risque de maladies cardio-vasculaires), une proportion (couverture vaccinale de la rougeole, prévalence des infections nosocomiales en France en 2012), un ratio (exhaustivité d'un système de surveillance, sexe ratio) ou une mesure d'association (odds ratio, rapports de prévalences, etc.).

Si on s'intéresse au nombre de personnes anti-VHC positives (séropositives au VHC) dans la population U, la fonction d'intérêt θ associée à la variable d'intérêt y s'écrit $\theta = \sum_{i \in U} y_i$. Elle correspond au total des valeurs prises dans la population.

Pour construire un échantillon, on dispose dans la situation la plus simple, d'une liste exhaustive, appelée **base de sondage**, des identifiants de toutes les unités d'intérêt de la population d'intérêt. Ces identifiants permettent de contacter les unités une fois tirées au sort.

A chaque unité de l'échantillon est affecté un **poids de sondage** permettant d'estimer les fonctions d'intérêt à estimer dans la population. Le poids de sondage est défini par l'inverse de la probabilité d'appartenir à l'échantillon (**probabilité d'inclusion**). Dans certaines enquêtes, la probabilité d'inclusion est relativement facile à calculer. C'est le cas lorsqu'on dispose d'une base de sondage identifiant toutes les unités de la population d'intérêt. On est alors dans le cadre d'un **sondage aléatoire simple** où l'on peut construire un échantillon d'unités d'intérêt directement à partir de la base de sondage. Par exemple, on tire au sort selon un sondage aléatoire simple 25 individus dans une population de 100 individus. La probabilité pour un individu tiré au sort d'appartenir à l'échantillon est égale à 25/100 et son poids de sondage est donc égal à 4.

Soit π_i la probabilité d'inclusion correspondante à chaque unité de l'échantillon S. Le nombre estimé de séropositifs au VHC dans la population est donnée par : $\hat{\theta} = \sum_{i \in S} \frac{1}{\pi_i} y_i$ selon l'estimateur d'Horvitz-Thompson [62].

Pour d'autres enquêtes, la probabilité d'inclusion est difficile à calculer pour des raisons qui sont liées à une structure complexe ou particulière de la base de sondage ou à un comportement particulier des individus appartenant à la population d'intérêt. Ce contexte est celui notamment des enquêtes où un individu se retrouve rattaché à au moins deux identifiants distincts de la base de sondage. Ainsi, un individu joignable à partir de plusieurs identifiants a une probabilité plus élevée d'appartenir à l'échantillon qu'un individu rattaché à un unique identifiant. C'est le cas, par exemple, des enquêtes téléphoniques où une personne peut être joignable à partir de plusieurs numéros de téléphones et a une probabilité plus élevée d'appartenir à l'échantillon qu'une personne qui n'a qu'un seul numéro de téléphone. De même, pour des enquêtes réalisées

auprès de personnes fréquentant des centres pour bénéficier de prestations ou de services, une personne venant tous les jours pour une prestation a une probabilité plus élevée d'appartenir à l'échantillon qu'une personne ne venant pour une prestation qu'un jour sur deux ou une fois par semaine au cours de la période d'enquête considérée.

En pratique, il est généralement impossible de connaître à l'avance l'identité des personnes qui se rendent dans les lieux d'enquêtes et donc impossible de disposer d'une base de sondage correspondant à cette population pour construire aisément un échantillon (par exemple, les personnes se rendant dans un point-soupe ou dans un centre d'hébergement).

En l'absence de base de sondage des unités d'intérêt, des techniques d'enquêtes classiques sont utilisées [9,144]. Il s'agit le plus souvent des **sondages à plusieurs degrés** (*i.e.* une combinaison de plusieurs sondages aléatoires simples). Un **degré** est défini par une base de sondage et un tirage aléatoire. Par exemple, on souhaite réaliser une enquête auprès d'enfants scolarisés en France. Si l'on ne dispose pas d'une base de sondage d'enfants mais seulement d'une liste d'écoles, on commence par construire un échantillon d'écoles. Les écoles tirées au sort constituent les **unités primaires**, c'est le premier degré. Ensuite, pour chaque unité primaire, on construit un échantillon de classes, si chaque école dispose d'une liste de classes. Les classes tirées au sort constituent les **unités secondaires**, c'est le second degré. Et enfin, chaque classe a une liste d'élèves qui permet de construire un échantillon d'élèves. Les élèves sélectionnés constituent les **unités tertiaires**, c'est le troisième et dernier degré.

Cependant, pour des populations spécifiques, il n'existe parfois aucune base de sondage identifiant, même indirectement, les unités d'intérêt. En l'absence de base de sondage et les techniques d'enquête classiques ne pouvant pas être utilisées, des techniques d'enquête spécifiques pour des populations difficiles d'accès ont été proposées et mises en pratique [96,132]. On peut citer l'échantillonnage boule de neige [50], en réseau [15,52,137], ciblé [157], en marche aléatoire [78], adaptatif en grappes [142], lieux-moments [95,111], par liens traçants [25,40,143] ou conduit par les répondants [56,71]. Néanmoins, d'un point de vue statistique, certaines de ces techniques ne reposent pas sur un échantillon aléatoire d'individus. Cela peut entraîner un biais dans les estimations des indicateurs épidémiologiques, car certains individus de la population d'intérêt ont une probabilité nulle d'appartenir à l'échantillon, et une impossibilité d'estimer correctement

leurs variances. D'autres techniques doivent tenir compte des caractéristiques des populations étudiées (comme par exemple, la fréquentation multiple voire hétérogène des lieux d'enquêtes) et nécessitent d'être évaluées en cas de violation des hypothèses sous-jacentes aux propriétés des estimateurs utilisés.

Plan de la thèse

Le plan de cette thèse s'articule de la façon suivante :

- Le premier chapitre rappelle l'évolution clinique et biologique de l'hépatite C, présente les outils de prélèvements et d'analyses biologiques utilisés dans les enquêtes épidémiologiques notamment auprès des usagers de drogues et fait un focus sur la prévalence et l'incidence de l'infection à VHC chez les usagers de drogues dans le monde.

- Le second chapitre présente les données des deux enquêtes transversales ANRS-Coquelicot réalisées en 2004 et 2011 dont l'objectif principal était d'estimer la prévalence du virus de l'hépatite C chez les usagers de drogues fréquentant des structures dédiées en France. Ces données illustreront les chapitres suivants, car l'ensemble des résultats reposent sur ces deux enquêtes.

- Le troisième chapitre propose un estimateur de la prévalence dans les populations fréquentant les lieux d'enquêtes et ses propriétés en termes de biais et de variance. Si l'estimation de la prévalence d'une maladie en population générale peut être traitée de façon assez directe par la théorie des sondages, cette estimation est plus délicate lorsqu'on s'intéresse à des populations composées d'individus pouvant être interrogés plusieurs fois au cours de l'enquête.

- Le quatrième chapitre traite de l'estimation de l'incidence de l'hépatite C en développant un modèle mathématique à partir d'enquêtes répétées et intégrant des plans de sondage complexes. En effet, en l'absence d'un test biologique permettant de distinguer les infections récentes des infections anciennes, l'estimation de l'incidence peut être réalisée en construisant un modèle mathématique reposant sur la formulation d'une relation entre la prévalence et l'incidence. Le modèle utilisé nécessite des données externes aux enquêtes concernant la mortalité associée à la maladie, des informations cliniques et comportementales (proportion de nouveaux usagers de drogues, taux de séroconversion au VHC). Ce

modèle mathématique doit également intégrer les plans de sondages mis en place lors des enquêtes épidémiologiques.

- Enfin, le cinquième chapitre présente une autre technique d'enquêtes auprès de populations difficiles à joindre mais ne fréquentant aucun lieu d'enquête. Il s'agit de l'échantillonnage conduit par les répondants, plus connu sous le nom anglophone RDS pour *Respondent-Driven Sampling*. En effet, dans le chapitre 3 nous présentons une technique d'enquête ciblant uniquement une population d'individus fréquentant les lieux d'enquêtes (*i.e.* la population dite visible). Dans ce chapitre, il s'agit de pouvoir atteindre également la population d'usagers de drogues cachée ou invisible (*i.e.* non captée par les enquêtes réalisées dans les lieux d'enquêtes).

Chapitre 1

L'hépatite C

Le virus de l'hépatite C (VHC) a été identifié pour la première fois en 1989 par l'équipe de Michael Houghton, grâce à des techniques de biologie moléculaire [63]. L'hépatite C est une maladie transmissible qui résulte d'une infection par le VHC. Cette infection se caractérise par une atteinte du foie qui, dans sa phase aiguë, peut se manifester cliniquement par un ictère (ou jaunisse) mais reste souvent asymptomatique (*i.e.* sans signe clinique). Une personne peut guérir spontanément de l'infection (*i.e.* sans traitement) ou devenir chronique entraînant une hépatite C dite chronique. Cette forme chronique de l'infection évolue sur une longue période (environ 20 à 30 ans) de manière silencieuse avant la survenue d'une cirrhose et/ou d'un cancer primitif du foie.

Ce chapitre vise à rappeler l'évolution clinique et biologique de l'hépatite C, à décrire les outils de prélèvements et d'analyses biologiques utilisés dans les enquêtes épidémiologiques pour la recherche et la détection du virus et enfin à présenter quelques données sur l'état de santé des usagers de drogues vis-à-vis du VHC en termes de prévalence et d'incidence.

1.1 A l'échelle mondiale

Chaque année, 3 à 4 millions de personnes s'infectent par le VHC dans le monde. Les régions les plus touchées sont l'Asie centrale et orientale et l'Afrique. Environ 80 millions de personnes ont une hépatite C chronique et encourent le risque que leur atteinte hépatique évolue vers la cirrhose et/ou le cancer primitif du foie (carcinome hépatocellulaire) [51]. Environ 500 000 personnes meurent chaque année de pathologies hépatiques liées à l'hépatite C [89]. Plusieurs

génotypes du VHC existent et leur répartition varie selon les régions [108]. Selon les pays, l'épidémie d'hépatite C peut toucher la population générale ou certaines populations, comme la population des usagers de drogues [51].

1.2 Histoire naturelle et évolution clinique

En France, environ 232 000 personnes étaient porteuses du virus de l'hépatite C (0.53% de la population) en 2004 dont 40% ignoraient leur statut [107]. Aujourd'hui, la principale source de contamination est l'usage de drogues par injection [7].

1.2.1 Modes de transmission, signes cliniques et réponse thérapeutique

La transmission du virus de l'hépatite C s'effectue le plus souvent par exposition à du sang contaminé par le VHC après injections réalisées avec du matériel d'injection partagé (*e.g.* utilisation de drogues injectables), blessures par piqûre d'aiguille en milieu de soins ou autres dispositifs médicaux mal stérilisés ou lors d'une naissance chez une mère infectée par l'hépatite C. Depuis 2001, le risque résiduel de transmission du VHC par transfusion sanguine en France est rarissime (estimé à moins d'un don tous les 10 ans) suite à la mise en place du dépistage sérologique des dons de sang.

L'histoire naturelle de cette infection se déroule en trois étapes :

- L'étape 1 est la contamination entraînant, après une période d'incubation de 7 semaines en moyenne (de 2 semaines à 6 mois), une hépatite aiguë (infection très récente). Dans cette phase aiguë, l'hépatite est le plus souvent asymptomatique (environ 80% des personnes infectées par le VHC) [37]. Elle peut parfois se traduire par un syndrome pseudogrippal ou un ictère (jaunisse). Entre 60 et 80% des personnes infectées restent porteuses chroniques de ce virus. Dans la phase chronique, l'infection VHC reste longtemps asymptomatique.

- L'étape 2 est la persistance de l'infection virale qui entraîne l'apparition de lésions hépatiques telle que la fibrose hépatique qui évolue lentement et se propage au niveau de l'ensemble du foie.

- L'étape 3 correspond à l'installation de la cirrhose (fibrose hépatique étendue) qui peut se compliquer par un carcinome hépatocellulaire, ces deux pathologies étant responsables de la mortalité liée à cette infection.

Environ 5-20% des personnes présentant une hépatite C chronique évoluent vers une cirrhose. 1-5% d'entre elles en meurent. En 2008, en France, le nombre estimé de décès attribuables au VHC était de 4.5 décès pour 100 000 habitants (95% d'entre eux présentant une cirrhose) [98]. Par ailleurs, pour 25% des cas de cancers primitifs du foie, la cause sous-jacente du cancer est l'hépatite C [33, 37].

La première cause de cirrhose hépatique en France est la consommation excessive d'alcool. En 2013, les cirrhoses post-hépatite C représentent environ 10% des inscriptions en France et près de 28% des transplantations hépatiques sont dues à des cirrhoses alcooliques (première indication de la transplantation hépatique post-cirrhose) [1].

Depuis 2012, les nouveaux traitements (AAD) sont pangénotypiques c'est-à-dire actifs quel que soit le type de génotype. Il existe 7 génotypes pour le VHC [108], le génotype 3 étant le génotype qui répond le moins bien à ces nouveaux traitements (taux de guérison $\leq$ 80%) [7]. Selon les recommandations de juin 2015 émises par l'Association Française pour l'Étude du Foie (AFEF) [7], plusieurs schémas thérapeutiques sont proposés selon le génotype du VHC et ceux qui ont été privilégiés ont un taux de réponse virologique soutenue (c'est-à-dire indétectabilité de l'ARN du VHC six mois après la fin d'un traitement complet) supérieur à 90%. Autrement dit, une réponse virologique soutenue marque le succès du traitement. La durée de traitement est d'environ 12 semaines actuellement.

Jusqu'en 2011, le traitement associait l'administration d'interféron-alpha pégylé et de ribavirine [8, 108]. Ce traitement, dont les effets secondaires étaient importants, n'était pas préconisé dans la phase aiguë de l'infection, durant laquelle l'évolution vers une guérison de manière spontanée pouvait se produire dans 15 à 30% des cas. Dans la phase chronique de l'infection, la durée du traitement (6 à 12 mois) dépendait de plusieurs facteurs, dont la nature du génotype viral. L'efficacité globale du traitement atteignait 80% dans les cas d'infection par les génotypes 2 ou 3, mais était d'environ 45% en cas d'infection par les autres génotypes qui étaient plus résistants au traitement (en particulier, le génotype 1 majoritaire) [3].

Il n'existe pas actuellement de vaccin contre l'hépatite C. Les moyens les plus efficaces de lutter contre l'hépatite C résident en la maîtrise du dépistage des porteurs chroniques et de leur prise en charge pour suivi et traitement, la politique de réduction des risques chez les usagers de drogues injectables et la maîtrise du risque de transmission nosocomiale du VHC (transmission du VHC en milieu de soins).

1.2.2 Définitions de cas à partir des marqueurs biologiques spécifiques de l'infection

Les marqueurs biologiques utilisés pour le diagnostic de l'hépatite C sont l'ARN du VHC et les anticorps totaux anti-VHC [8]. L'ARN du VHC est le premier marqueur de l'infection et est détectable une semaine après le contage (c'est-à-dire, après la transmission d'agents pathogènes par exposition) comme illustré en Figure 1.1.

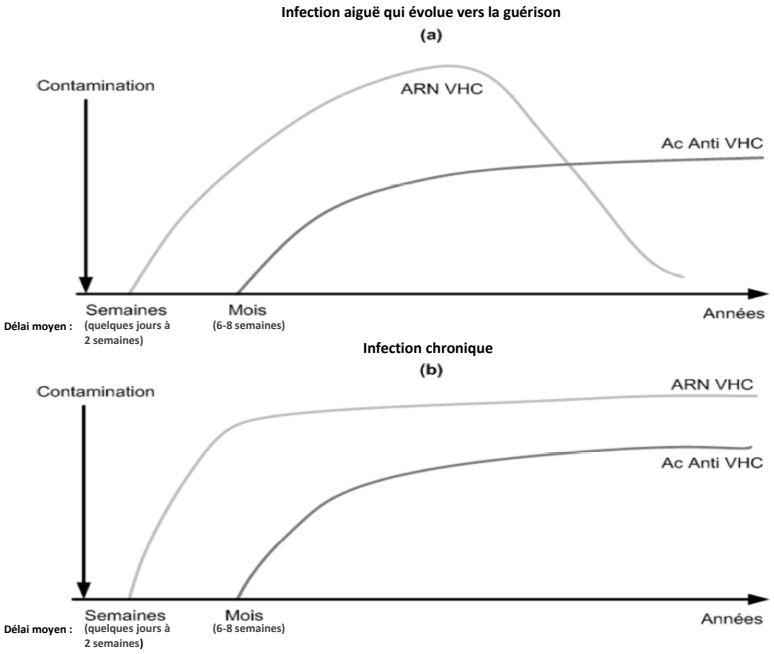


FIGURE 1.1 – Évolution des marqueurs de l'ARN du VHC et des anticorps anti-VHC en cas d'infection aiguë qui évolue vers la guérison (a) ou d'infection chronique (b)

Néanmoins, le diagnostic de l'hépatite C à partir de l'ARN du VHC est rarement posé au début de la maladie car, au cours de la phase aiguë de l'infection, la majorité des personnes infectées ne présente aucun symptôme apparent pouvant laisser suspecter une hépatite aiguë et motiver la recherche de ce marqueur. Le diagnostic se fait généralement dans le cadre d'un dépistage avec la recherche des anticorps anti-VHC. Sachant que les anticorps anti-VHC restent indétectables durant la phase aiguë de la maladie, la notion de période fenêtre a été définie comme la période entre l'exposition au VHC et le moment où le résultat du dépistage est suffisamment fiable (détection des anticorps anti-VHC) (Figure 1.2). Elle se mesure en jours et peut varier d'un individu à l'autre. Ici, on définit la période fenêtre comme étant la durée moyenne entre

la détection de l'ARN du VHC et la détection des anticorps anti-VHC (Figure 1.2). Elle est estimée en moyenne à 50 jours [12,61,113].

La présence des anticorps anti-VHC, indique qu'une personne est ou a été contaminée par le VHC. En effet, les anticorps anti-VHC demeurent présents dans le sang même quand les personnes qui ont été contaminées sont guéries.

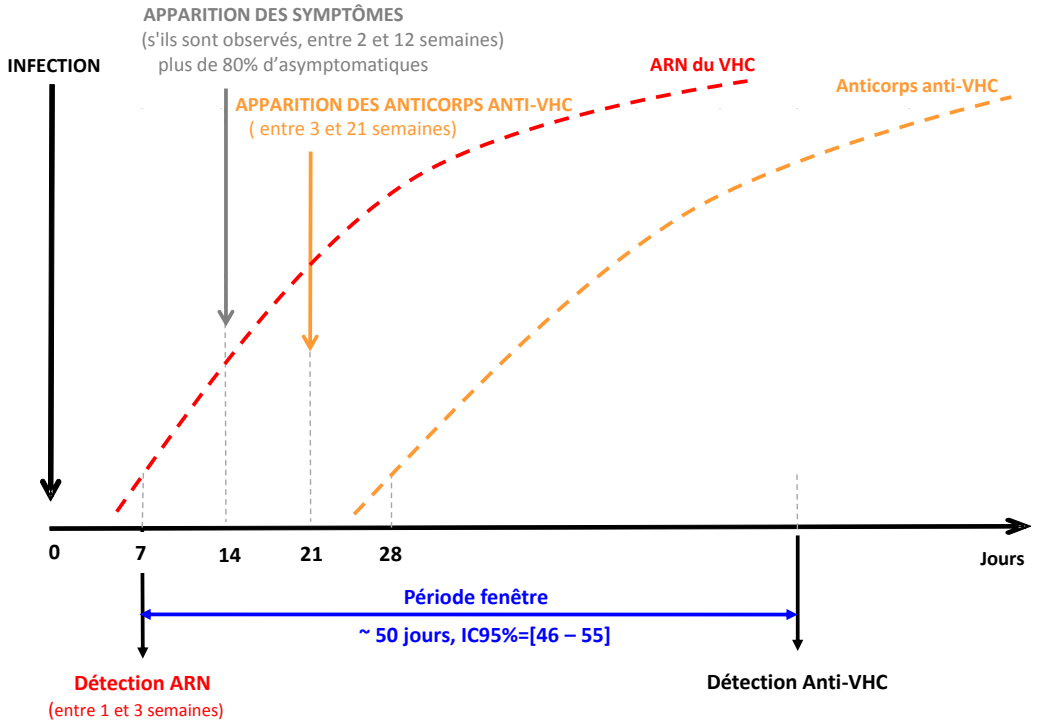


FIGURE 1.2 – Évolution clinique et biologique des personnes infectées par le VHC. Selon Pageshafer, la période fenêtre est estimée à 50 jours en moyenne [113].

Les tests biologiques permettant de détecter ces marqueurs sont présentés dans la section 1.3.

A ce jour, il n'existe aucun marqueur sérologique spécifique d'une infection aiguë au VHC ou chronique. Seule la persistance au delà de six mois de l'ARN du VHC signe la chronicité de l'infection VHC. Le Tableau 1.1 présente les 4 définitions de cas selon la détection ou non des deux marqueurs biologiques ARN du VHC et anticorps anti-VHC : une infection aiguë, une infection chronique, une guérison ou une absence de contamination (la personne n'a jamais été

en contact avec le virus).

TABLE 1.1 – Classification selon la détection ou non des marqueurs ARN du VHC et anticorps anti-VHC

Détection	ARN +	ARN −
Anti-VHC +	Infection chronique	Guérison (spontanée ou après traitement)
Anti-VHC −	Infection aiguë	Non contaminé

Infection aiguë : La détection dans le sérum de l'ARN du VHC se fait 7 à 21 jours après la date de contamination (voir Figure 1.1(a)). L'augmentation des transaminases sériques, généralement supérieure à 10 fois la limite supérieure de la normale survient au-delà du 15ème jour et souvent au-delà de 4 semaines. Des symptômes cliniques (ictère, fatigue, douleurs abdominales), s'ils sont observés, apparaissent 2 à 12 semaines après la date de contamination et disparaissent rapidement. Les anticorps anti-VHC peuvent apparaître dans le sérum 20 à 150 jours après la contamination (50 jours en moyenne). La séroconversion correspond à l'apparition d'anticorps anti-VHC chez un sujet connu antérieurement comme négatif vis-à-vis des anticorps anti-VHC. La période comprise entre les deux tests (dernier test anti-VHC négatif connu et premier test anti-VHC positif) correspond à la période probable de contamination ($\pm$ 50 jours avant le dernier test anti-VHC négatif). Une infection VHC récente est définie par la période probable de contamination quand la date du premier test anticorps anti-VHC positif est inférieure à 51 jours par rapport à la date du dernier test anticorps anti-VHC négatif. Une hépatite C aiguë peut également être définie pour un sujet présentant une virémie à VHC isolée (*i.e.* ARN du VHC positif et anti-VHC négatif) suivie secondairement d'une apparition d'anticorps anti-VHC.

Infection chronique : L'hépatite C chronique est caractérisée par la présence simultanée d'anticorps anti-VHC et de l'ARN viral, et la persistance d'une ARN du VHC détectable au delà de 6 mois après la contamination chez des sujets ayant des signes cliniques et/ou biologiques d'hépatopathie chronique (voir Figure 1.1(b)) [5]. L'ARN du VHC reste constamment détectable tout au long de l'évolution de la maladie chronique en l'absence de traitement et sa quantification est essentielle au suivi du traitement du VHC chronique pour déterminer la réponse virologique au traitement.

Guérison : La guérison spontanée de l'infection aiguë est définie par la disparition sans traitement de l'ARN du VHC dans le sérum parmi les sujets ayant des anticorps anti-VHC détectables. La guérison de l'infection chronique résulte d'un traitement anti-VHC et se définit par l'absence de virémie (ARN du VHC négatif), 6 mois après l'arrêt du traitement.

1.3 Utilisation pratique des outils biologiques

1.3.1 Matériel de prélèvements biologiques

Les examens biologiques s'effectuent généralement sur sérum ou plasma à partir d'un prélèvement veineux. Plus récemment, des prélèvements alternatifs (prélèvements salivaires ou de sang total capillaire sur papier buvard) ont été développés, notamment pour cibler des populations difficilement accessibles au dépistage [18, 24]. Des prélèvements sur papier buvard à partir de sang total capillaire obtenu par simple piqûre au bout du doigt sont, par exemple, utilisés auprès de populations difficiles d'accès [60, 69, 76]. Le papier buvard (DBS, l'acronyme anglais pour *dried blood spot* est sa dénomination la plus courante) permet de recueillir du sang et de conserver ce prélèvement sous forme desséchée. Une fois séchés à température ambiante, les prélèvements peuvent être acheminés par voie postale, puis conservés à $-20^{\circ}C$ et testés avec des techniques sérologiques et moléculaires adaptées.

Indépendamment de l'aspect pratique, l'utilisation des DBS chez les usagers de drogues est une bonne alternative à la ponction veineuse [35]. En effet, les prélèvements veineux restent problématiques chez les usagers de drogues injecteurs du fait d'un accès veineux parfois rendu difficile par les injections répétées. Ces difficultés peuvent être un frein au dépistage dans cette population à haut risque de contamination vis-à-vis du VHC. Plus largement, lors d'enquêtes réalisées auprès de populations difficilement accessibles, les individus sont interrogés dans des lieux (la rue, lieu de convivialité, etc.) où les prélèvements veineux sont difficilement réalisables.

1.3.2 Tests diagnostiques : du dépistage individuel à leur utilisation dans des enquêtes épidémiologiques

Les tests sérologiques ont évolué depuis les années 1990 avec les tests ELISA (*Enzyme-Linked ImmunoSorbent Assays*) de première génération jusqu'aux tests rapides d'orientation diagnostic

(TROD) pour la recherche des marqueurs biologiques (Ac) du VHC [18].

Actuellement, la détection des anticorps anti-VHC s'effectue à l'aide de tests ELISA de 3ème ou 4ème génération (valeur seuil supérieure ou égale à 1.0 : test positif), et la détection-quantification de l'ARN du VHC à l'aide d'une technique de PCR (*Polymerase Chain Reaction*) en temps réel avec un seuil de détection de 10-15 UI/mL [24]. Les résultats quantitatifs des tests PCR sont donnés en nombre de copies, en unités internationales (UI) ou en log_{10} UI, par mL (Tableau 1.2).

TABLE 1.2 – Facteurs de conversion entre unités internationales (UI) et nombre de copies

Référence (Marqueur ARN du VHC)	Copies/mL	UI/mL	log_{10} UI/mL
Eurohep group[(1)]	2.7	1	
NIBSC[(2)]		100000	5

(1) European Concerted Action on Viral Hepatitis : http://www.bioqcontrol.com/standnat
(2) National Institute for Biological Standards and Control :
http://www.nibsc.org/documents/ifu/14-150.pdf
Auparavant exprimée en nombre de copies par mL, cette mise en place d'un standard international par l'OMS permet d'uniformiser le rendu des résultats en UI/mL.

Dans le cadre d'un dépistage individuel, après une recherche des anticorps anti-VHC par ELISA, le recours au test RIBA (*Recombinant Immuno Blot Assay*) et à la détection de l'ARN viral permet de confirmer le diagnostic.

Le diagnostic d'infection chronique est posé lorsque les anticorps anti-VHC sont présents dans le sang pendant plus de six mois. Une fois le diagnostic d'hépatite C chronique confirmé, le génotype viral doit être déterminé. En effet, actuellement, la détermination du génotype du VHC et la quantification de la charge virale du VHC (ARN) sont indispensables avant l'initiation du traitement. Dans des études épidémiologiques, la prévalence du VHC est généralement estimée comme la proportion de personnes ayant des anticorps anti-VHC. Il s'agit donc, la plupart du temps, de l'estimation de la prévalence des anticorps anti-VHC et non de l'infection chronique seule.

Le diagnostic d'infection aiguë est rarement posé du fait de son caractère paucisymptomatique/asymptomatique. Ce diagnostic est pourtant essentiel en épidémiologie, notamment pour l'estimation de l'incidence de la maladie. Dans les enquêtes épidémiologiques en populations spécifiques, à haut risque de transmission ou à forte prévalence VHC, l'ARN du VHC est généralement recherché parmi les anti-VHC négatifs pour identifier les infections récentes.

Afin de diagnostiquer des infections récentes (ou aiguës), certaines études ont utilisé des

tests biologiques qui mesurent l'avidité de l'anticorps anti-VHC, un autre marqueur biologique qui peut être lié à une infection virale récente [30, 42, 127]. Ainsi, en 2010, Gaudy-Graffin *et al.* ont évalué un test d'avidité anti-VHC, dérivé du test ELISA de 3ème génération, à partir de 117 échantillons de sérum de patients présentant des anticorps anti-VHC, dans le cadre d'une infection aiguë connue, d'une infection chronique ou chez des patients guéris [42]. Les auteurs ont conclu que ce test ne pouvait être utilisé chez des personnes guéries (donc non virémiques) et qu'il était également difficile de différencier les infections aiguës des infections chroniques.

En 2013, Shepherd *et al.* avancent des résultats satisfaisants sur l'utilisation d'un test d'avidité anti-VHC pour distinguer les infections chroniques des infections récentes sur deux types de supports de prélèvements (plasma ou DBS) [127]. En 2015, Cullen *et al.* ont déterminé les facteurs associés à une infection récente au VHC dans une enquête nationale auprès d'usagers de drogues injecteurs, intégrant l'utilisation d'un test d'avidité des anticorps anti-VHC pour identifier les infections récentes [30].

En France, le test d'avidité ne fait actuellement pas partie des recommandations pour le diagnostic et le suivi des personnes prises en charge pour une infection par le VHC [8].

1.3.3 Fiabilité des tests biologiques réalisés sur papier buvard (DBS)

Dans les enquêtes séro-épidémiologiques, un protocole de prélèvement basé sur le support, le conditionnement, le transport, le stockage ainsi que les trousses commerciales retenues pour les tests biologiques est généralement défini selon des objectifs de l'enquête.

Plusieurs trousses commerciales pour la recherche des anticorps anti-VHC et de l'ARN du VHC sur sérum ou plasma sont disponibles sur le marché. Les performances analytiques des tests biologiques issus de ces trousses sont satisfaisantes (très spécifiques et très sensibles) si les recommandations des fabricants sont respectées. Toutefois, il est possible d'observer des résultats faussement positifs, de fréquence variée, selon les trousses et des résultats faussement négatifs chez certaines personnes (*i.e.* les personnes hémodialysés, les immunodéprimés) [8].

Un changement de protocole de prélèvement sur l'une de ces trousses peut donc modifier les performances des tests biologiques [73]. Pour une enquête épidémiologique, les supports de prélèvements retenus peuvent être différents de ceux préconisés dans les recommandations des fabricants pour l'utilisation des différents tests biologiques, de même que les populations sur lesquelles ces tests sont réalisés.

Plusieurs auteurs ont montré la faisabilité et la pertinence de l'utilisation des DBS à des fins épidémiologiques ou cliniques, même si une légère perte de sensibilité a été constatée par comparaison au test effectué sur sérum ou plasma dans les conditions standards (conditionnement, stockage, valeur seuil du test, etc.) [14, 35, 131, 148]. Différentes conditions de stockage des prélèvements (en termes de température ou de durée de conservation) et différentes valeurs seuils ont été testées sur les DBS et comparées au sérum/plasma. Les performances analytiques sur DBS restent relativement stables dans la plupart des scénarii. Lors d'une étude en France auprès de 500 patients, la valeur seuil du test ELISA de 3ème génération sur DBS a été établie à 0.135 au lieu de 1 comme le stipule le fabricant sur des prélèvements sérum ou plasma [131]. Sur DBS, la sensibilité associée est de 99.1% $[IC95\% : 97.4\% - 99.8\%]$ et la spécificité est de 98.2% $[IC95\% : 94.9\% - 99.6\%]$ [131].

Comme souligné plus haut, des résultats faussement négatifs peuvent être observés chez certaines personnes dans des conditions standards. Dans d'autres conditions d'utilisation (par exemple des populations à haut risque de maladies infectieuses prélevées sur un autre matériel de prélèvement que celui préconisé par le fabricant du test utilisé), les résultats de ces tests doivent être interprétés avec précaution. A cela s'ajoute la qualité du prélèvement recueilli sur DBS. Par exemple, le diamètre minimal (5-6 mm) de chaque goutte de sang à recueillir est indiqué sur le papier buvard. Cet élément est indispensable pour assurer un recueil minimal correct. Un volume inférieur chez certains participants pourrait conduire à des résultats ininterprétables, et ne permettrait pas d'effectuer de manière fiable certains tests biologiques (par exemple la recherche de l'ARN).

1.3.4 Utilisation des résultats quantitatifs des tests biologiques pour un classement en séropositif/séronégatif vis-à-vis des anticorps anti-VHC

La prévalence et l'incidence sont généralement estimées à partir de données d'enquêtes ad hoc. Les prélèvements réalisés au cours de ces enquêtes séro-épidémiologiques sont analysés puis classés selon la valeur seuil établie et fournie par le fabricant du test biologique, spécifique pour chaque marqueur biologique. Pour rappel, les seuils fournis par les fabricants sont généralement à visée diagnostique et non épidémiologique. Un individu est diagnostiqué séropositif (ou infecté) si le résultat de son test est supérieur à la valeur seuil et séronégatif (ou susceptible) si le résultat de son test est inférieur à la valeur seuil. Le choix de ce seuil a donc un impact direct sur le

classement séropositif/séronégatif et de fait un impact sur l'estimation de la séroprévalence. En effet, un seuil trop sensible aura tendance à classer plus d'individus en séropositif et donc à surestimer la proportion de séropositifs. Certains auteurs ont d'ailleurs pointé des limites quant à l'interprétation des données biologiques en se basant sur des seuils fournis par les fabricants [43, 55].

Par ailleurs, lorsque deux valeurs seuils sont fournies par le fabricant (un seuil pour les positifs et un autre seuil pour les négatifs) et que le prélèvement analysé a un résultat compris entre les 2 valeurs, cela conduit à un classement en "équivoque". Dans ce cas, il est nécessaire de (1) retester le prélèvement concerné, ou (2) de le considérer comme positif, ou (3) comme négatif ou (4) de le supprimer de l'analyse épidémiologique ou (5) de le considérer comme ininterprétable. Les personnes ayant un prélèvement classé "ininterprétable" sont généralement conservés dans les analyses et comparés aux autres en termes de caractéristiques épidémiologiques.

Si la valeur seuil est inconnue (par exemple, dans le cas d'un support de prélèvement alternatif qui n'a pas encore fait l'objet d'une analyse de sensibilité), le statut sérologique va être inconnu ou sujet à des erreurs de classement selon le choix arbitraire de la valeur seuil.

Afin d'éviter le classement de certains prélèvements en "équivoque" ou l'utilisation de valeurs seuils arbitraires ou à visée diagnostique, une autre approche consiste à utiliser des mélanges de lois pour modéliser la distribution des résultats quantitatifs des tests biologiques. On l'appelle l'approche directe par opposition à la méthode classique des seuils [16, 74]. Il s'agit ici d'utiliser directement l'ensemble des résultats quantitatifs des tests qui s'expriment sur une échelle continue. En effet, dans cette approche, toute l'information disponible (toutes les valeurs brutes) est utilisée sans passer par une étape de classement binaire (séropositif/séronégatif selon le seuil).

Le modèle le plus simple, que nous noterons F, considère deux états : l'état séropositif et l'état séronégatif comme illustré dans la Figure 1.3. C'est un mélange de deux lois (ou mélange de lois à deux composantes) f_i, indexées par l'état i ($i = 1, 2$) dont l'expression mathématique est donnée par la formule $F(x) = p_1 f_1(x) + p_2 f_2(x)$ où x est le résultat quantitatif du test utilisé (i.e la concentration ou le niveau d'anticorps) et les termes p_1 et $p_2 = 1 - p_1$ correspondent à la probabilité d'être respectivement séronégatif ou séropositif. Par exemple, chaque distribution $f_i(x)$ peut être une distribution normale de moyenne μ_i et de variance σ_i^2. Les prélèvements dont le résultat quantitatif du test appartient à la première distribution f_1 sont considérés séronégatifs et ceux dont le résultat quantitatif du test appartient à la seconde distribution f_2 sont considérés

séropositifs.

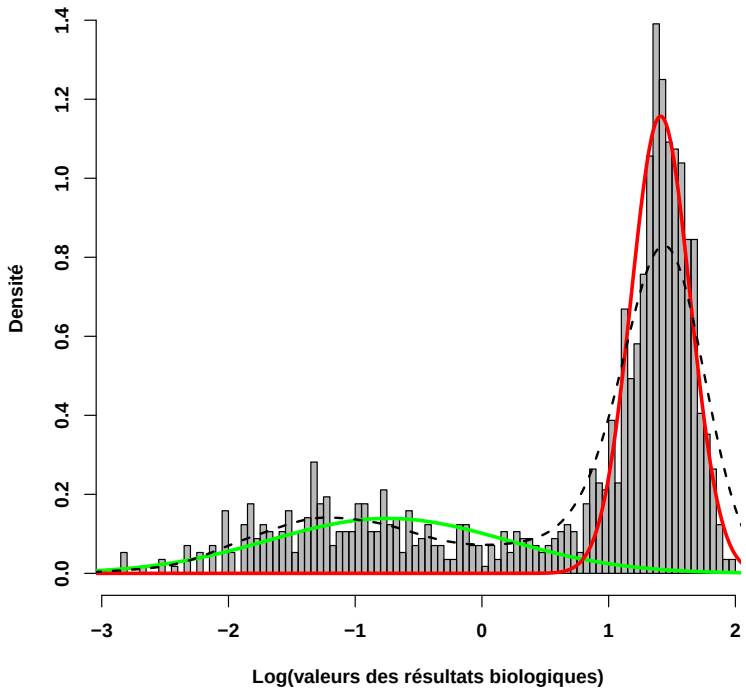


FIGURE 1.3 – Exemple d'un mélange de deux lois normales (ligne en pointillés) classant les individus ayant un résultat biologique négatif (première composante, en vert) et les individus ayant un résultat biologique positif (seconde composante, en rouge).

Autrement dit, les deux composantes renvoient aux deux sous-populations : les séronégatifs (susceptibles) et les séropositifs (infectés), et p_2 est la prévalence. Cependant, il est parfois nécessaire d'utiliser un mélange de lois à plus de deux composantes lorsque la distribution des résultats quantitatifs des tests biologiques varie selon un certain niveau de réactivité des anticorps [55,119]. Dans un modèle de mélange de lois à c-composantes, la distribution des résultats quantitatifs des tests biologiques réalisés est modélisée par $F(x) = \sum_i^c p_i f_i(x)$, avec $f_i(x)$ la distribution de la ième composante et p_i la proportion d'échantillons biologiques testés appartenant à cette ième composante.

Désormais, l'approche directe est une alternative aux méthodes classiques de seuils [16, 74]. Nous avons appliqué cette approche sur les prélèvements réalisés sur DBS lors de l'enquête Coquelicot (Chapitre 2).

1.4 Éléments épidémiologiques chez les usagers de drogues

Les usagers de drogues par injection sont une population particulièrement à risque de transmission du virus l'hépatite C [11]. En France, l'observatoire français des drogues et des toxicomanies (OFDT) a estimé en 2006 entre 210 000 et 250 000 usagers de drogues, dont 81 000 usagers injecteurs actifs dans le dernier mois [26]. La France se situe dans la moyenne de l'Union européenne (voir Figure 1.4).

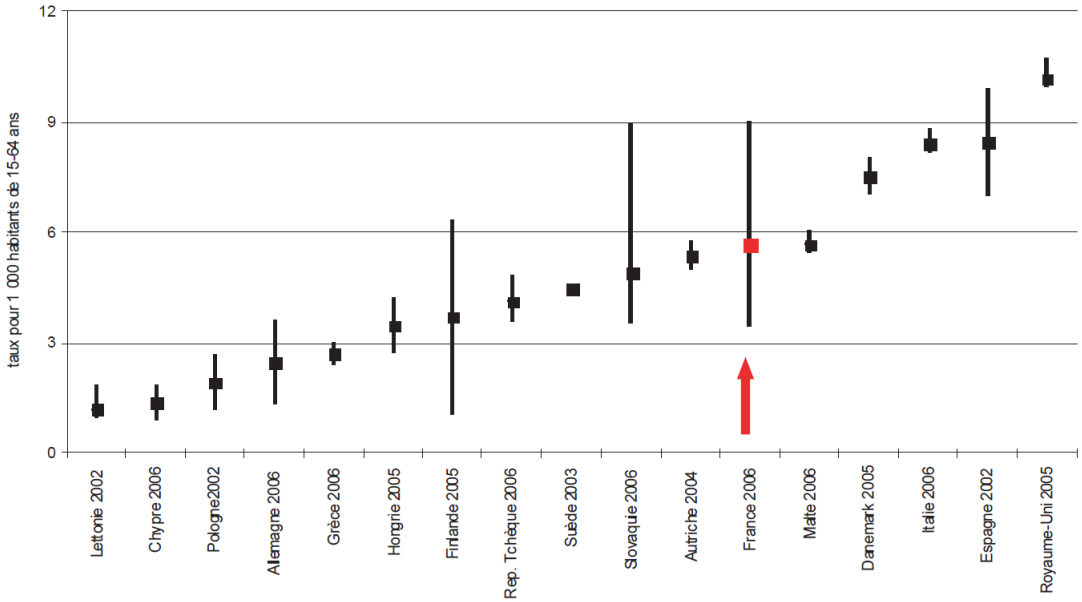


FIGURE 1.4 – Estimations du nombre d'usagers problématiques de drogues pour 1 000 habitants, Pays de l'Union européenne, 2002-2006 [26].

La seringue empruntée et le matériel de préparation partagé sont des facteurs de risque majeur de transmission de ce virus. Les aiguilles et les seringues partagées sont le plus grand vecteur de contamination du VHC dans cette population en raison de leur contact direct avec le sang pendant l'injection intraveineuse. En effet, le VHC a un pouvoir contaminant 10 fois supérieur au VIH : une seringue utilisée par une personne infectée peut rester contaminante 3 semaines, contrairement au VIH qui résiste peu à l'air libre [4].

Ainsi, tout le matériel utilisé pour l'injection (récipient, coton, garrot, mains, etc.) peut être contaminé et resté pendant un certain temps une source de contamination [141]. A cela, peuvent s'ajouter l'ignorance vis-à-vis des prises de risque, la méconnaissance de son statut sérologique

vis-à-vis du VHC et une perception erronée des complications de l'infection VHC [106].

1.4.1 Prévalence

La prévalence de l'hépatite C dans cette population est encore aujourd'hui très élevée dans la majorité des pays pour lesquels des données de surveillance au sein de cette population sont disponibles [112] : 22.6% – 90.5% en Europe de l'Est, 21.1% – 86.2% en Europe de l'Ouest, 36.0% – 84.0% en Asie du Sud, 9.8% – 97.4% en Amérique Latine, 22.2% – 97.3% en Afrique Sub Saharienne (Figure 1.5).

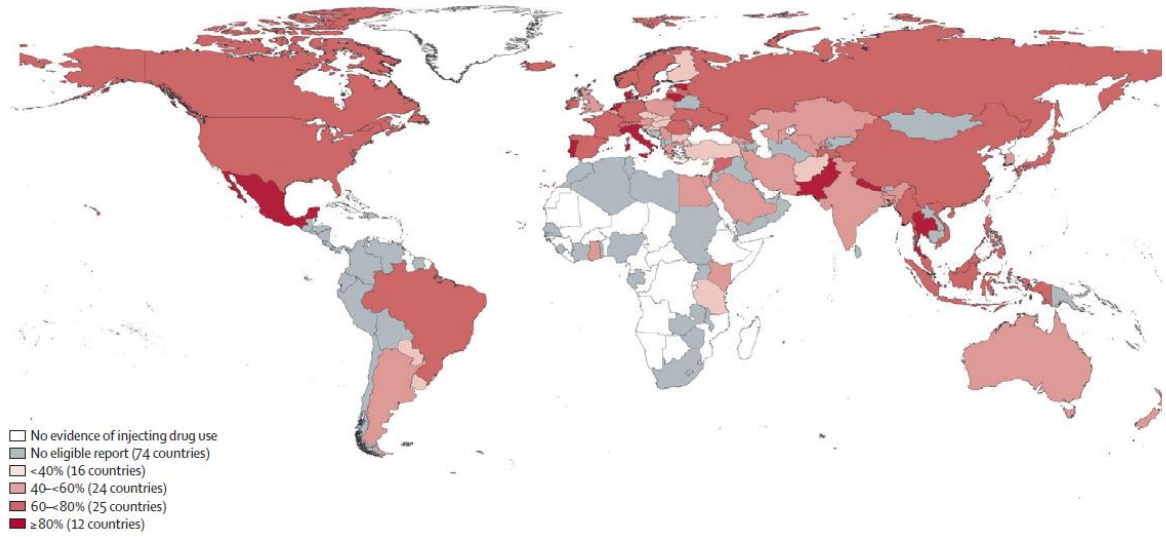


FIGURE 1.5 – Prévalence des anticorps anti-VHC chez les usagers des drogues injecteurs [112].

En Europe, l'épidémie de l'infection par le VHC en population générale est en perpétuelle évolution en termes de prévalence, incidence, modes de transmission, distribution géographique des génotypes, depuis ces 20 dernières années [38]. Entre 14% et 84% des usagers de drogues injecteurs sont infectés par le VHC [37]. En France, la prévalence reste très élevée chez les usagers de drogues avec 44% [$IC95\% : 39 - 48$] d'usagers porteurs d'anticorps anti-VHC en 2011 [69].

1.4.2 Incidence

Chaque année, environ 70% des nouvelles contaminations sont associées à l'usage de drogues [100, 103, 106]. Les contaminations interviennent souvent lors des premières injections [106]. En

Europe, l'incidence de l'infection à VHC chez les usagers de drogues injecteurs varie entre 2.7 et 66 pour 100 personnes-années [163] comme l'illustre la Figure 1.6.

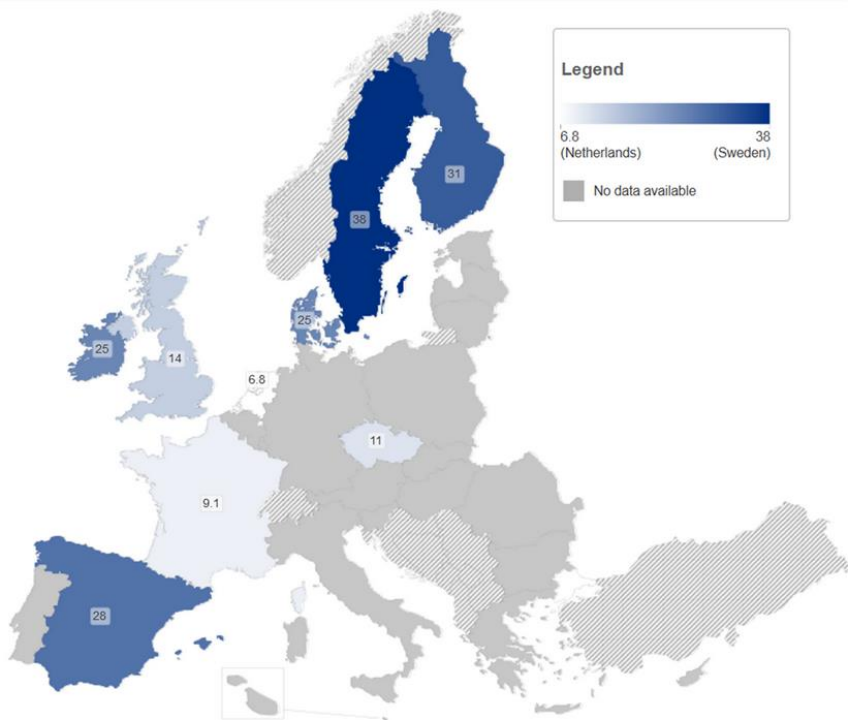


FIGURE 1.6 – Incidence (pour 100 personnes-années) de l'infection au VHC chez les personnes qui injectent des drogues [163].

Dans le Nord-Est de la France, une étude longitudinale sur l'infection au VHC conduite entre 1999 et 2001 auprès d'usagers de drogues a permis d'estimer l'incidence de l'infection à VHC à 9 pour 100 personnes-années et, parmi ceux qui se sont injectés des drogues au moins une fois dans les 6 mois précédant l'enquête, à 11 pour 100 personnes-années [93].

Comme évoqué précédemment dans ce chapitre, l'usage de drogues par injection est connu comme étant la principale source de contamination par le VHC dans les pays industrialisés. Dans un objectif de réduction de la diffusion de cette maladie, il est donc nécessaire de disposer de données concernant la population des usagers de drogues afin, d'évaluer la dynamique du VHC dans cette population d'une part, et de mettre en place de nouvelles mesures de réduction des risques et de nouveaux moyens de prévention d'autre part, en amont des avancées sur les traitements thérapeutiques.

En France, différentes enquêtes transversales ont été menées auprès d'usagers de drogues fréquentant des structures (ENa-CAARUD) [21]. Cependant, l'enquête ANRS-Coquelicot est la première enquête à introduire un recueil de données biologiques. Elle a été réalisée en 2004 puis en 2011 afin de décrire les profils socio-démographiques des usagers de drogues, d'identifier les principaux déterminants de la prise de risque et de mesurer l'état de santé de cette population vis-à-vis du VIH et du VHC à l'échelle nationale. Nous présentons maintenant ces deux enquêtes.

Chapitre 2

Enquêtes ANRS-Coquelicot auprès d'usagers de drogues

Afin de suivre la dynamique du virus de l'hépatite C chez les usagers de drogues en France, Santé publique France (anciennement Institut de Veille Sanitaire (InVS)) et l'Inserm ont été chargés en 2002, dans le cadre d'un projet soutenu par l'Agence Nationale de Recherche sur le Sida (ANRS), de mener un étude de séroprévalence auprès de cette population. La population cible de cette étude, généralement en situation de précarité et "cachée", était à la fois difficile à définir et à atteindre, notamment en raison de profils sociaux, de rapports à l'usage de drogues diversifiés et de la stigmatisation de la pratique d'usage (la consommation de drogues étant une pratique illégale en France). Deux enquêtes, dénommées ANRS-Coquelicot, ont donc été réalisées en France en 2004[1] et 2011[2] avec la collaboration de l'Institut National d'Etudes Démographiques (INED), le Centre National de Référence (CNR) du VIH de Tours et le Centre de recherche Santé Mentale, Psychotropes et Société (CESAMES). La population d'intérêt a été définie par celle des UD fréquentant les dispositifs spécialisés en France métropolitaine. Cette population représente une sous-population ayant un "rapport problématique aux drogues" et plutôt socialement en situation de précarité.

[1] Gouvernance par InVS-Inserm (Cesames), en collaboration avec l'INED et le CNR VIH : J. Emmanuelli et M. Jauffret-Roustide (responsables scientifiques), M. Quaglia, C. Lefevre , G. Vivier, M. Jauffret-Roustide (coordination terrain), Y Le Strat et N. Razandratsima (plan de sondage), F. Barin (analyses biologiques), M. Jauffret-Roustide, M. Rondy et E. Couturier (analyses épidémiologiques), Y Le Strat (soutien biostatistisques)

[2] Gouvernance par InVS-Inserm (Cermes3), en collaboration avec le CNR VIH et le CNR Hépatites :M. Jauffret-Roustide (responsable scientifique), T. Benoit, G. Guibert et M. Jauffret-Roustide (coordination terrain), L. Léon et Y Le Strat (plan de sondage et soutien biostatistiques), F. Barin et S. Chevaliez (analyses biologiques), L. Weill-Barillet, M. Jaufret-Roustide, J. Pillonel, X. Pascal, M. Molinier (analyses épidémiologiques).

Ce chapitre présente ces deux enquêtes épidémiologiques transversales qui ont été utilisées pour l'ensemble des travaux réalisés dans le cadre de cette thèse.

2.1 Objectifs

Les enquêtes ANRS-Coquelicot sont des enquêtes multicentriques menées pour décrire les profils et pratiques des UD, estimer la séroprévalence du VIH et du virus de l'hépatite C et évaluer la politique de réduction des risques [67, 69].

En 2004, le recrutement des UD s'est effectué entre septembre et décembre, dans l'ensemble des services issus de la chaîne thérapeutique spécifique aux UD (centres de soins spécialisés pour toxicomanes, centres de post-cure, appartements thérapeutiques, boutiques, programmes d'échange de seringues, équipes de rue) et dans des cabinets de médecins généralistes prescripteurs de traitements de substitution aux opiacés. Cette première enquête a été menée auprès de 1462 UD fréquentant 136 centres participants (dont 36 cabinets de médecins généralistes) et ayant sniffé ou injecté de la drogue au moins une fois dans leur vie.
En 2011, l'enquête a été rééditée. Le recrutement des UD s'est effectué, en mai-juin, dans les centres participants, excepté les cabinets de médecins généralistes prescripteurs de traitements de substitution aux opiacés. L'enquête a été menée durant 11 semaines auprès de 1568 UD fréquentant les 121 centres participants. Cette seconde enquête a permis d'actualiser les connaissances sur la dynamique du virus de l'hépatite C et du VIH, en déterminer les facteurs associés, évaluer les comportements à risque des UD fréquentant des structures en France et mesurer la prévalence de l'Ag HBs (un marqueur de l'hépatite B).
Ces enquêtes épidémiologiques transversales ont été réalisées dans 5 villes (Lille, Strasbourg, Paris, Bordeaux et Marseille) en 2004 et élargies à la notion d'agglomération et à 2 départements (Seine-et-Marne et Seine-Saint-Denis) en 2011. Le choix de ces villes intégrait notamment des contraintes de faisabilité (budget et durée de l'étude, accessibilité logistique, etc.), une certaine diversité du point de vue de la localisation géographique et des modalités de consommation, et un nombre minimal de personnes éligibles.

2.2 Méthode d'échantillonnage

Dans chaque ville, un inventaire complet de tous les centres dédiés aux UD a été effectué : services d'hébergement (incluant les centres d'hébergement, des chambres d'hôtel, les *sleep-in* centres (centres d'hébergement de services sociaux français)), les centres de traitement de la toxicomanie (y compris ceux fournissant des traitements de substitution aux opiacés, du sevrage ou de la psychothérapie), les services dits à bas-seuil d'exigence (boutiques, programmes d'échange de seringues et des équipes de travail de sensibilisation). Les premiers sont regroupés sous le terme de CSAPA (Centres de soins, d'Accompagnement et de Prévention en Addictologie) et les derniers sont regroupés depuis 2008 sous le nom de CAARUD (Centre d'Accueil et d'Accompagnement à la Réduction des Risques pour les Usagers de Drogues) [8, 21]. Parallèlement, dans chaque centre, les file actives journalières d'UD éligibles pour l'enquête étaient recueillies, une file active étant définie par le nombre de visites dans la demi-journée.

Ensuite, pour chaque centre, une base de sondage listant l'ensemble des demi-journées d'ouverture pendant les périodes d'enquête a été construite.

Les personnes ont été tirées au sort selon un échantillonnage "lieux-moments" que nous décrivons en détails dans la section 3.1. La quasi-totalité des centres a participé à l'enquête avec un taux de participation très élevé (95% en 2004 et de 100% en 2011) [66]. Puis, dans tous les centres, des demi-journées d'ouverture ont été tirées au sort en utilisant un sondage aléatoire proportionnel aux files actives déclarées par les centres. Un calendrier de visites des centres a alors été établi.

Enfin, dans chaque "centre/demi-journée" tiré au sort, les UD ont été sélectionnés selon un sondage systématique adapté (sauf pour les centres résidentiels où ils ont tous été invités à participer à l'enquête). La file active de la demi-journée tirée au sort (N_i personnes) a de nouveau été recueillie auprès des responsables des centres quand les enquêteurs s'y rendaient. Le tirage des n_i personnes de la file active a été assimilé à un sondage aléatoire simple, avec une probabilité d'inclusion d'un usager égale à n_i/N_i.

2.3 Population d'étude et questionnaire

Les critères d'inclusion de l'enquête étaient d'avoir au moins 18 ans, d'avoir injecté des drogues ou sniffé "au moins une fois au cours de sa vie", de parler le français et d'accepter de participer à l'enquête en fournissant un consentement oral (avec remise d'une note d'information écrite). Les participants ont donc été inclus après consentement oral et ont fourni un auto-prélèvement de sang sur papier buvard pour la détection des anticorps anti-VHC, de l'ARN du VHC, des anticorps VIH et de l'antigène HBs.

Le questionnaire proposé était anonyme, confidentiel et passé en face-à-face. Il durait environ 45 minutes et était administré par des enquêteurs professionnels n'ayant aucun lien avec le recrutement des centres. Les enquêteurs recrutés étaient habitués à travailler auprès des populations difficiles et ont été formés sur les profils et les pratiques de la population des usagers de drogues.

Le questionnaire portait entre autres sur les items suivants :

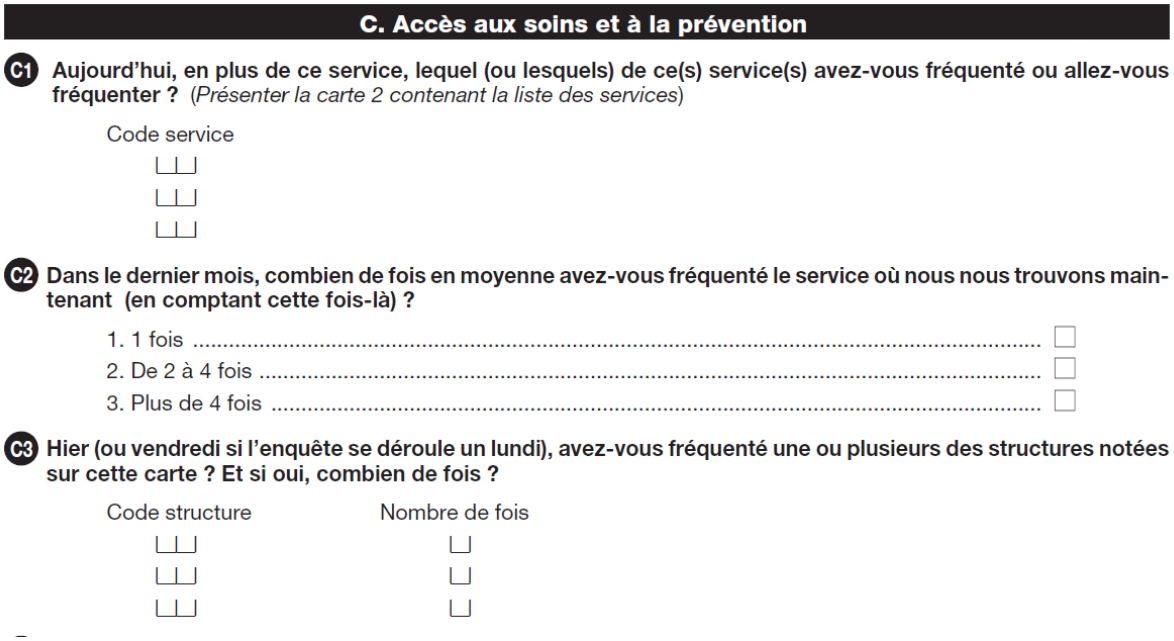
- Données socio-démographiques, état de santé général, accès aux soins et à la prévention,
- Représentations et connaissances vis-à-vis des risques d'exposition virale,
- Histoire de la toxicomanie et consommation de produits,
- Pratiques à risque vis-à-vis de la consommation de drogues et de la sexualité.

En plus des items liés aux objectifs des deux enquêtes, des questions sur la fréquentation des centres ont été posées et ont évolué entre les deux vagues d'enquêtes (Figures 2.1 et 2.2).

Le questionnaire se terminait par la remise d'un jeu de brochures de prévention, d'une lettre d'information invitant au dépistage et d'un ticket-service d'une valeur de 10 euros.

Un usager de drogue injecteur (UDI) a été défini comme un usager de drogues s'étant injecté des drogues au moins une fois au cours de sa vie.

Un UDI actif a été défini comme un usager de drogues s'étant injecté des drogues dans le dernier mois précédent l'entretien.



C. Accès aux soins et à la prévention

C1 **Aujourd'hui, en plus de ce service, lequel (ou lesquels) de ce(s) service(s) avez-vous fréquenté ou allez-vous fréquenter ?** (*Présenter la carte 2 contenant la liste des services*)

Code service
⌊_⌋
⌊_⌋
⌊_⌋

C2 **Dans le dernier mois, combien de fois en moyenne avez-vous fréquenté le service où nous nous trouvons maintenant (en comptant cette fois-là) ?**

1. 1 fois ☐
2. De 2 à 4 fois ☐
3. Plus de 4 fois ☐

C3 **Hier (ou vendredi si l'enquête se déroule un lundi), avez-vous fréquenté une ou plusieurs des structures notées sur cette carte ? Et si oui, combien de fois ?**

Code structure	Nombre de fois
⌊_⌋	⌊_⌋
⌊_⌋	⌊_⌋
⌊_⌋	⌊_⌋

FIGURE 2.1 – Extrait du questionnaire Coquelicot 2004 - Partie fréquentations.

2.4 Recueil et analyse des données biologiques

Le prélèvement de sang était effectué par l'UD lui-même avec le matériel remis par l'enquêteur. Ce prélèvement restait anonyme et ne donnait lieu à aucun retour d'information. L'enquêteur remettait à chaque usager volontaire une compresse alcoolisée et un pansement adhésif. Après une auto-piqûre par micro-lancette sur la pulpe du doigt, plusieurs gouttes de sang étaient recueillies sur un buvard apparié au questionnaire par un même numéro (Figure 2.3). Un minimum de 6 spots de sang capillaire sur papier buvard (DBS) était donc recueilli pour chaque UD. Afin d'optimiser la qualité de la collecte d'échantillons, le diamètre minimal (5-6 mm) de chaque goutte de sang à recueillir était indiqué sur le DBS. Cet élément était indispensable pour que le même volume de prélèvement soit recueilli chez chaque participant et pour avoir du matériel biologique suffisant pour les analyses biologiques.

Les analyses des DBS pour la recherche des anticorps vis-à-vis du VIH, du VHC et de l'antigène HBs ont été effectuées par le centre national de référence (CNR) du VIH, à l'aide de tests ELISA (respectivement Genscreen HIV Ac/Ac Biorad®, anti-HCV 3.0 Ortho® assay et

J1- Hier et les trois jours précédents, avez-vous fréquenté un ou plusieurs services spécialisés pour les usagers de drogues ? Et si oui, lesquels et combien de fois ?

Enquêteur : Par services spécialisés chez les usagers de drogue, on entend centres de soins pour toxicomanes, les PES, les boutiques, les centres d'hébergements, les équipes de rue ...
Indiquer le nombre de fréquentation par structure et par jour.
Si l'enquêté n'arrive pas à répondre avec les jours de la semaine, comptabiliser le nombre de fois où il a fréquenté une structure dans les quatre derniers jours ouvrés sans tenir compte des jours de la semaine.
Être le plus précis possible sur le nom de la structure.

Nom de la structure	Type de la structure (PES, centre métha...)	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche	Ou	Nombre de fois dans les quatre derniers jours ouvrés
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|
		|__|	|__|	|__|	|__|	|__|	|__|	|__|		|__|

J2- Aujourd'hui, en comptant l'endroit où nous sommes, quel(s) service(s) avez-vous fréquenté ou allez-vous fréquenter?

Enquêteur : il doit toujours y avoir dans le tableau au moins le service dans lequel l'usager répond au questionnaire, donc au moins une réponse. Être le plus précis possible sur le nom de la structure.

Nom de la structure	Type de la structure (PES, centre métha...)	Nombre de fréquentation
		|__|
		|__|
		|__|
		|__|

FIGURE 2.2 – Extrait du questionnaire Coquelicot 2011 - Partie fréquentations.

Monolisa HBs Ag ULTRA Biorad®) puis par le CNR des Hépatites virales B, C et delta pour des analyses moléculaires (recherche de l'ARN du VHC et du génotype) à l'aide de la trousse commerciale (QIAamp® MinElute Virus, Qiagen) et d'une analyse phylogénétique.

Les deux sections suivantes traitent du classement des prélèvements en séropositif/séronégatif vis-à-vis des anticorps anti-VHC.

2.4.1 Classement en séropositif/séronégatif vis-à-vis des anticorps anti-VHC : approche biologique

Le seuil de détection des anticorps anti-VHC, analysé sur DBS, a été abaissé à une valeur de 0,64 en se basant sur des données expérimentales, par rapport à la valeur unité lorsqu'on utilise

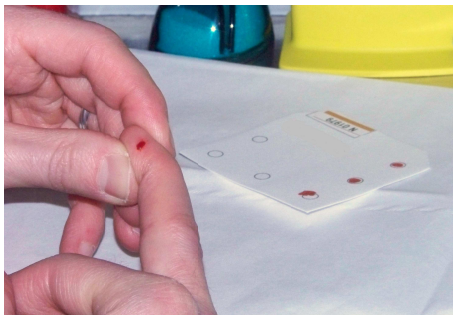


FIGURE 2.3 – Exemple d'autoprélèvement de sang.

le sérum ou le plasma pour un classement des individus en séropositifs ou séronégatifs pour le VHC. Pour l'enquête de 2011, les résultats biologiques pour la recherche de l'ARN du VHC étant disponibles, les individus classés anti-VHC négatifs dans un premier temps ont pu être classés en infection récente (*i.e.* présence de l'ARN du VHC) à partir des critères suivants : (1) une charge virale comprise entre 1.6 et 2.5 log_{10} UI/mL (Tableau 1.2) et (2) un repassage pour la recherche de l'ARN dont le résultat était codé "OK" ou "QI" (*i.e.* quantité insuffisante) ou "génotypé".

2.4.2 Classement en séropositif/séronégatif vis-à-vis des anticorps anti-VHC : approche par mélange de lois

Pour le classement des individus prélevés et identifiés séropositifs ou séronégatifs pour le VHC, nous avons comparé la distribution des résultats quantitatifs des tests d'anticorps anti-VHC stratifiés selon le statut VIH, le sexe, la qualité du DBS (*i.e.* taille des gouttes suffisantes pour réaliser les analyses biologiques) et l'année de l'enquête. Graphiquement, les distributions différaient uniquement selon l'année d'enquête. Nous avons donc appliqué un mélange de lois par année d'enquête. A cause d'un mauvais ajustement par un mélange de deux lois, nous avons modélisé les résultats biologiques de l'enquête en utilisant des modèles de mélange de plusieurs lois normales (jusqu'à 6 lois) qui décrivaient le mieux la distribution observée pour chaque année. Nous n'avons pas jugé nécessaire d'introduire plus de 6 composantes dans le modèle de mélange pour éviter un sur-ajustement. Le modèle ayant le critère d'information bayésien (BIC) le plus faible a été conservé (Tableau 2.1).

Les modèles s'ajustant le mieux aux données sont présentés en Figure 2.4.

TABLE 2.1 – Critères de sélection des modèles de mélange, France, Enquêtes Coquelicot 2004 et 2011

Nombre de composantes	2004		2011	
normales dans le modèle	AIC	BIC	AIC	BIC
2	1855,17	1878,67	-223,83	-199,53
3	990,33	1072,93	-422,76	-383,88
4	879,90	931,61	-538,52	-485,07
5	800,35	866,16	**-554,71**	**-486,67**
6	**753,35**	**833,26**	-557,14	-474,52

AIC : Akaike information criterion ; BIC : Bayesian information criterion

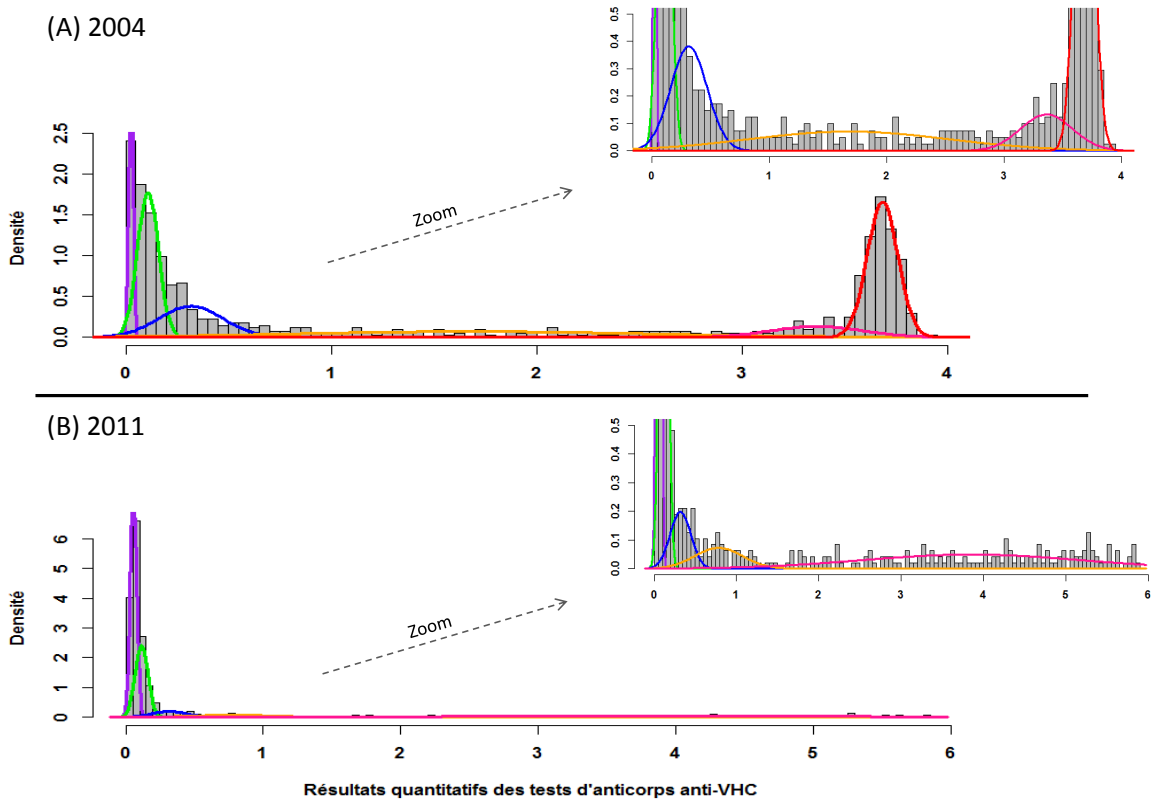


FIGURE 2.4 – (A) Distribution des résultats quantitatifs des tests biologigues anti-VHC par un mélange de 6 lois normales, 2004. (B) Distribution des résultats quantitatifs des tests biologigues anti-VHC par un mélange de 5 lois normales, 2011.

Nous avons utilisé un mélange de 6 lois normales pour l'enquête de 2004 et un mélange de 5 lois normales pour l'enquête en 2011, afin de distinguer les personnes séronégatives des personnes séropositives pour le VHC, en fonction de leur niveau de réactivité anti-VHC. Nous

avons observé que l'apport d'une sixième composante en 2011 était négligeable. Nous avons fait l'hypothèse que les niveaux 1-3 (les trois premières composantes) correspondant aux plus faibles réactivités, représentaient les tests anti-VHC classés négatifs. Et les niveaux 4-6 étaient supposés représenter les tests anti-VHC classés positifs.

Dans les échantillons, une discordance de 1.5% a été observée dans le classement des UD en séropositif/séronégatif au VHC selon les deux approches. En 2004, 1.6% des UD ont été classés séropositifs par l'approche biologique et séronégatifs par l'approche par mélange des lois. En 2011, 1.4% des UD présentaient un classement discordant selon la méthode de classement dont la moitié a été classée en séropositif par l'approche biologique et en séronégatif par l'approche par mélange des lois.

2.5 Statistiques descriptives

En 2004, parmi les participants :

- 79% avaient fourni des échantillons biologiques par auto-prélèvements.
- La majorité d'entre eux étaient essentiellement masculins (74%) et l'âge moyen était de 35 ans.
- 11% étaient séropositifs au VIH et 60% étaient séropositifs au VHC.
- 71% avaient reçu un traitement de substitution aux opiacés dans les 6 derniers mois.
- L'injection par voie intraveineuse au cours de la vie avait été pratiquée par 70% d'entre eux, à un âge moyen de 24 ans.

En 2011, parmi les participants :

- 92% avaient fourni des échantillons biologiques par auto-prélèvements.
- La majorité d'entre eux étaient essentiellement masculins (79%) et l'âge moyen était de 39 ans.
- 10% étaient séropositifs au VIH et 44% étaient séropositifs au VHC.

- 77% avaient reçu un traitement de substitution aux opiacés dans les 6 derniers mois.
- L'injection avait été pratiquée par 65% des UD au moins une fois dans leur vie et par 36% dans le dernier mois.

La distribution des fréquentations déclarées par les UD est représentée selon le statut sérologique de l'infection par le VHC (détection ou non des anticorps anti-VHC) en Figure 2.5. En 2011, près de 86% des personnes ont déclaré avoir fréquenté une seule fois un des centres dédiés durant la période d'enquête.

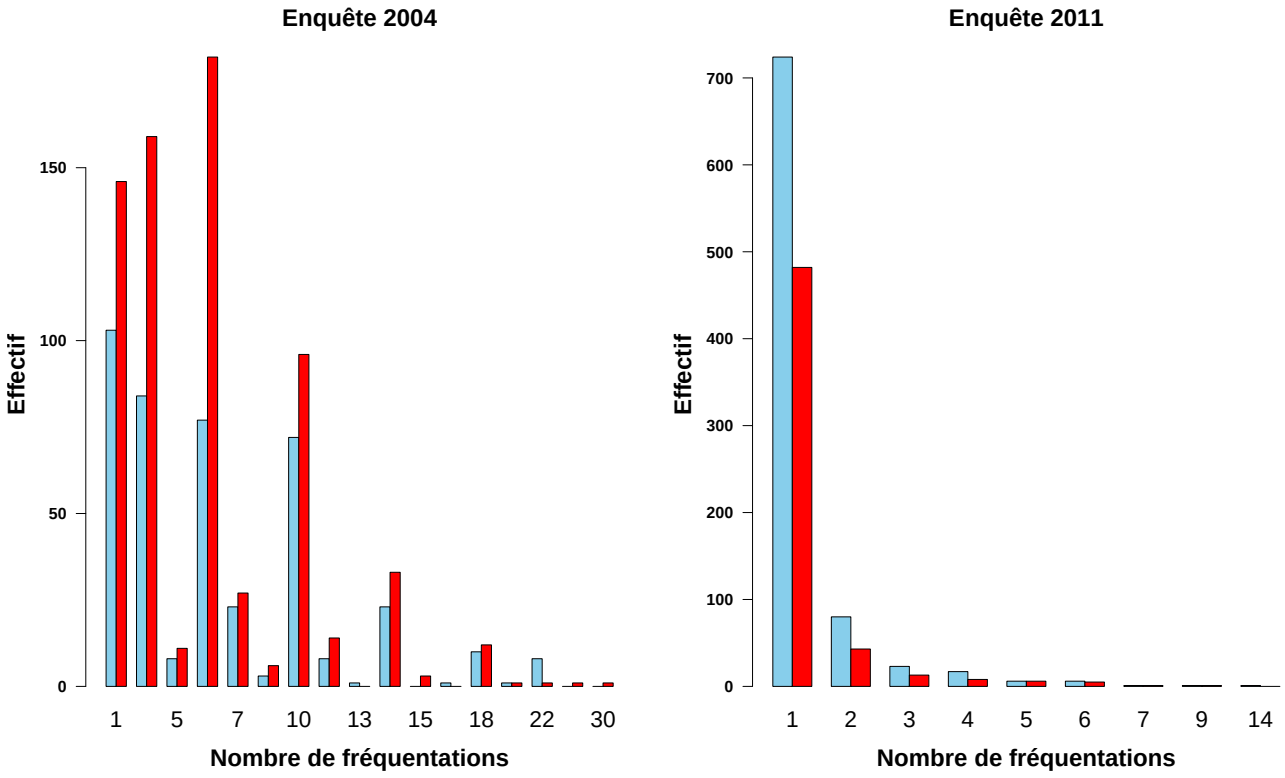


FIGURE 2.5 – Distribution du nombre de fréquentations déclarées par les participants à l'enquête Coquelicot 2004 (à gauche) et 2011 (à droite). Les barres rouges représentent les individus infectés au VHC et les barres bleus représentent les individus non infectés au VHC.

A partir de ces deux enquêtes, nous avons estimé la prévalence et l'incidence de l'infection à VHC dans la population des usagers de drogues fréquentant des centres spécialisés. Ces deux estimations font l'objet des deux chapitres suivants.

Chapitre 3

Estimation de la prévalence pour des populations fréquentant les lieux d'enquêtes

L'objectif de ce chapitre est de décrire l'échantillonnage lieux-moments puis de le formaliser dans un cadre théorique. Cela est, à notre connaissance, novateur dans le champ épidémiologique pour fournir des estimations sans biais d'indicateurs épidémiologiques dans une population d'individus fréquentant des lieux d'enquêtes. Cette technique d'enquête est utilisée pour atteindre une population d'individus fréquentant des lieux d'enquêtes à des moments précis, en l'absence de base de sondage d'individus comme cela est le cas pour les usagers de drogues ou les hommes ayant des relations sexuelles avec d'autres hommes.

Bien que quelques auteurs aient présenté l'échantillonnage lieux-moments comme une procédure en plusieurs étapes [117, 136], ce n'est que récemment que cette technique a été vue comme un sondage aléatoire à deux ou trois degrés [75]. Cependant, certains auteurs considèrent encore qu'il s'agit d'une technique d'enquête non aléatoire [109] et d'autres se posent la question de la nécessité de pondérer les estimations dans le cadre de ce type d'échantillonnage [70, 118, 166]. C'est dans ce contexte controversé que nous avons voulu clarifier cet échantillonnage en abordant à la fois la prise en compte ou non des poids de sondage et la fréquentation multiple des lieux d'enquêtes. Plus précisément, nous traiterons les points suivants :

- proposer un cadre théorique pour l'estimation d'une prévalence ou de facteurs d'association (odds ratio, rapports de prévalences, etc.) pour des enquêtes auprès de personnes bénéficiant de services (hébergement, consultation médicale, etc.) et ayant des fréquentations hétérogènes des lieux d'enquêtes,

- décrire la Méthode Généralisée du Partage des Poids (MGPP), reposant sur des travaux menés par des instituts de la statistique publique (Insee, Statistique Canada) ; notamment son principe, son application et son intérêt.

- présenter un estimateur qui tient compte des fréquentations des lieux et le comparer à un estimateur qui ignore ces fréquentations.

- appliquer cet estimateur sur les données réelles de l'enquête Coquelicot afin de prendre en compte les spécificités de la population étudiée, notamment en ce qui concerne les biais de mémoire des fréquentations au cours de l'enquête Coquelicot.

- évaluer la pertinence de cet estimateur par une étude de simulation, notamment pour décrire l'influence de la fréquentation et des erreurs potentielles sur ces fréquentations déclarées dans l'estimation d'indicateurs épidémiologiques.

Ce chapitre a fait l'objet d'un papier publié dans la revue *Biostatistics* en 2015.

3.1 Échantillonnage lieux-moments (TLS)

Plus connu sous le terme anglais time-location sampling (TLS), l'échantillonnage lieux-moments est utilisé pour collecter de l'information dans des populations difficiles à joindre en échantillonnant les personnes dans les lieux qu'elles fréquentent. Son principe est d'échantillonner des lieux (centres spécialisés, places, etc.) et des moments (jours, demi-journées, etc.) puis d'échantillonner des personnes fréquentant ces lieux-moments.

Soit une population d'individus, nommée B, qui fréquentent des centres (lieux) durant des horaires d'ouverture (moments). Pour simplifier, mais sans perte de généralité, on considère que l'unité de temps d'ouverture des centres est la demi-journée. Tout ce qui va suivre est valable pour des populations d'individus qui fréquentent des centres, quels que soient le type de centres (structures, rue, bus, etc.), le nombre de centres et l'unité de temps.

L'échantillonnage lieux-moments peut être vu comme un sondage aléatoire à 3 degrés, comme illustré en Figure 3.1.

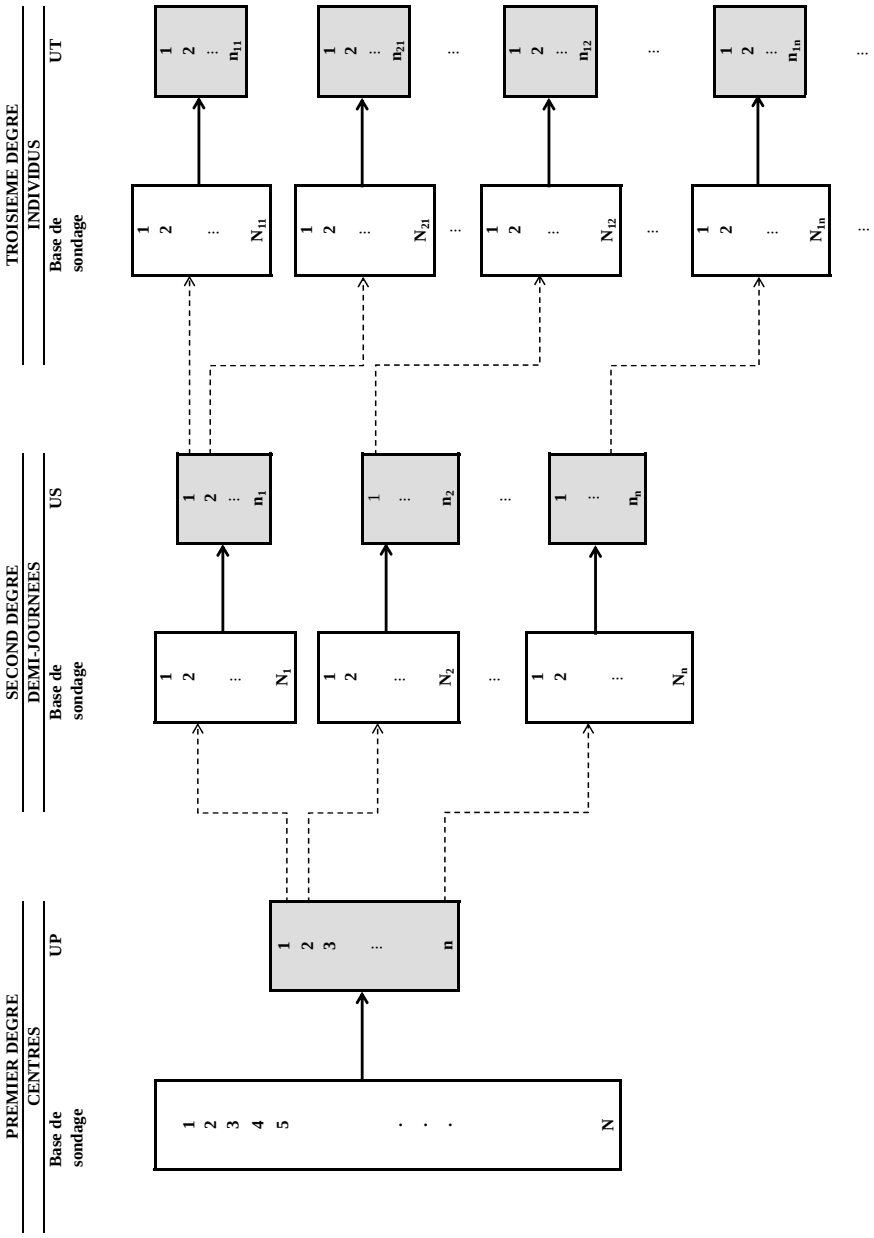


FIGURE 3.1 – Sondage aléatoire à 3 degrés. Les flèches noires représentent les échantillons construits et les flèches en pointillés représentent la construction des bases de sondages à partir desquelles des unités ont été tirées au sort au premier et au second degrés. UP : unité primaire, US : unité secondaire, UT : unité tertiaire.

Au premier degré, n centres sont tirés au sort parmi N centres, indexés par l ($l = 1, ..., N$). On utilise soit un sondage aléatoire simple sans remise, soit un sondage aléatoire à probabilités inégales sans remise pour ces tirages. Pour le dernier type de tirage, la probabilité d'inclusion du centre est proportionnelle à une variable auxiliaire (par exemple, le nombre moyen de personnes se rendant quotidiennement dans le centre).

Au second degré, pour chaque centre l ($l = 1, ..., n$) tiré au sort et appelé unité primaire (UP), on construit une base de sondage listant les N_l demi-journées d'ouverture du centre sur la période d'enquête, indexées par k ($k = 1, ..., N_l$).

On tire au sort n_l demi-journées parmi les N_l demi-journées recensées pour chaque unité primaire l. On les appelle les unités secondaires (US). On établit alors un calendrier de visites représentant chaque centre sélectionné et chaque demi-journée tirée au sort pour l'enquête. Pour illustrer ceci, un calendrier fictif de visites de 5 centres durant 4 semaines d'enquête est représenté en Figure 3.2.

Au troisième degré, un ou plusieurs enquêteurs se rendent dans les centres selon le calendrier de visites établi. Pour chaque centre l durant une demi-journée k du calendrier, les enquêteurs sélectionnent aléatoirement n_{kl} personnes parmi N_{kl} personnes éligibles qui arrivent dans les centres. Ces individus représentent les unités tertaires (UT).

		semaine 1							semaine 2							semaine 3							semaine 4						
centre	demi-jounrée	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche
1	matin	X												X															
	après-midi							X															X						
2	matin																X												
	après-midi			X						X							X												
3	matin																										X		
	après-midi																												
4	matin	X				X								X															
	après-midi																												
5	matin												X																
	après-midi			X																						X			

FIGURE 3.2 – Calendrier de visites de 5 centres durant 4 semaines d'enquête. Ouverture des centres (en blanc), fermeture des centres (en gris) et demi-journées tirées au sort (croix).

Au second et troisième degrés, ce sont des sondages aléatoires simples qui sont généralement utilisés. Dans la plupart des cas, l'enquêteur ne dispose d'aucune liste de personnes qui vont arriver dans le centre, ce qui justifie l'utilisation d'un sondage systématique : l'enquêteur sélectionne

aléatoirement une personne qui arrive au centre puis sélectionne les suivantes en fonction de leur ordre d'arrivée en utilisant un pas de sondage défini *a priori* (par exemple, toutes les 10 personnes). Parfois, un sondage stratifié est appliqué : les individus sont stratifiés par sexe, tranche d'âge, nationalité ou tout autre caractéristique d'intérêt. Les tirages aléatoires des unités à chaque degré (centre, demi-journée, individu) ont pour objectif de réduire les biais de sélection.

Dans un TLS, la probabilité d'inclusion d'un individu dépend de la probabilité de se rendre dans le lieu d'enquête à un moment donné et de la probabilité d'être interrogé parmi les individus éligibles. Nous verrons par la suite que cette probabilité d'inclusion d'un individu dépend aussi de sa fréquentation des lieux d'enquêtes. En effet, un individu qui fréquente beaucoup les lieux d'enquêtes a plus de chance d'appartenir à l'échantillon qu'un individu qui les fréquente peu. Cet élément nous a conduit à proposer une définition de l'échantillonnage lieux-moments comme un sondage indirect en représentant les liens entre les différents lieux d'enquêtes et les individus.

3.2 Sondage indirect

3.2.1 Définition

Soient une population A de taille N^A où chaque unité est indexée par j ($j = 1, ..., N^A$) et la population d'intérêt B qui contient N^B unités d'intérêt indexées par i ($i = 1, ..., N^B$) dans laquelle on souhaite estimer une fonction d'intérêt (total, proportion).
Un lien entre ces deux populations A et B désigne la correspondance entre n'importe quelle unité de $j \in A$ avec au moins une unité de $i \in B$ et permet donc le va-et-vient entre ces deux populations.
Un sondage indirect définit un sondage pour lequel :

- des unités j sont tirées au sort dans une population A afin d'accéder à des unités i appartenant à la population d'intérêt B et,
- les unités i dans la population B sont liées à au moins une unité j appartenant à la population A.

La correspondance entre les deux populations peut être représentée par une matrice de liens L de taille $N^A \times N^B$. Chaque élément l_{ji} de L, non nul, définit le lien entre $i \in B$ et $j \in A$ et, s'il n'existe pas de lien entre les deux unités, $l_{ji} = 0$.

Si on représente par exemple, une population A de 5 unités et une population B de 3 unités (Figure 3.3) alors la matrice de liens est donnée par :

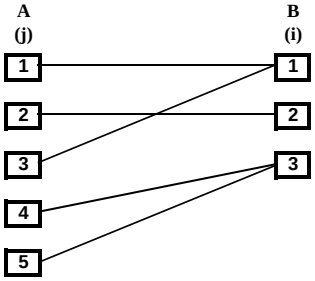


FIGURE 3.3 – Population A composée de 5 unités liée à une population B composée de 3 unités.

$$L = \begin{pmatrix} l_{11} & 0 & 0 \\ 0 & l_{22} & 0 \\ l_{31} & 0 & 0 \\ 0 & 0 & l_{43} \\ 0 & 0 & l_{53} \end{pmatrix}$$

Un lien l_{ji} est :

- bijectif si une unité $i \in B$ a un lien et un seul avec une unité $j \in A$,
- injectif si une unité $i \in B$ a au plus un lien avec une unité $j \in A$,
- surjectif si une unité $i \in B$ a au moins un lien avec une unité $j \in A$.

Plusieurs exemples prenant en compte le type de liens selon qu'ils soient bijectifs, injectifs ou surjectifs, comme illustré dans la Figure 3.4, sont présentés dans la section suivante.

3.2.2 Exemples

Liens bijectifs : Une enquête transversale (Saturn-inf) a été réalisée en 2009 afin d'estimer la prévalence du saturnisme parmi les enfants âgés de 1 à 6 ans résidant en France métropolitaine [39]. La variable d'intérêt est le dosage de la plombémie dans le sang pour chaque enfant. Ne disposant pas d'une liste exhaustive d'enfants en France et afin de réaliser un prélèvement de sang dans de bonnes conditions avec un taux d'acceptation élevé, une enquête à l'hôpital a été

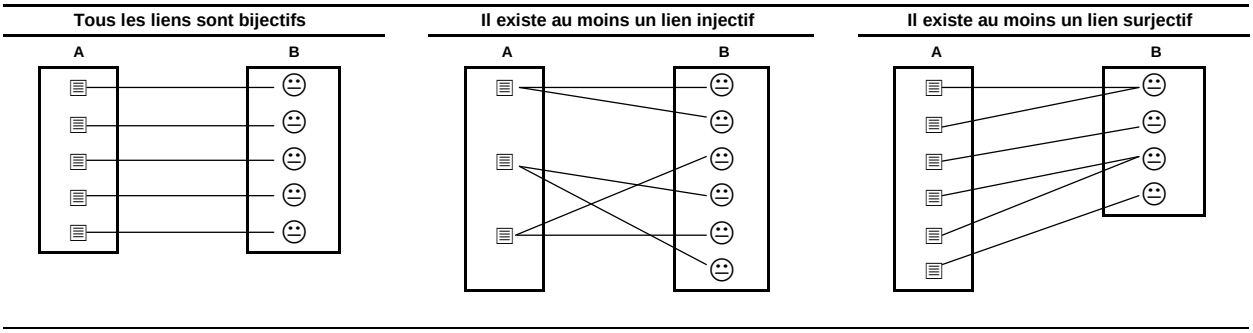


FIGURE 3.4 – Représentation des liens bijectifs, injectifs et surjectifs.

réalisée auprès d'enfants hospitalisés pour une pathologie sans lien avec le saturnisme. La liste des consultations dans les services hospitaliers a été utilisée comme base de sondage. Les liens entre la liste des consultations (population A) et les enfants (population B) sont bijectifs car une consultation concerne un unique enfant qui ne peut pas avoir plusieurs consultations au cours de l'enquête celle-ci ayant une durée limitée (Figure 3.4, colonne 1).

Liens injectifs : On réalise une enquête en milieu scolaire avec un sondage à deux degrés. La liste exhaustive des écoles est établie afin de construire un échantillon d'enfants répondant à l'enquête. Au premier degré, on tire au sort des écoles (unités primaires) et au second degré, on tire au sort des élèves (unités secondaires) qui sont les unités d'intérêt. Les liens entre la liste des écoles (population A) et les élèves (population B) sont injectifs car plusieurs enfants sont scolarisés dans une même école mais un enfant ne peut pas fréquenter plusieurs écoles au même moment (Figure 3.4, colonne 2).

Liens surjectifs : Il s'agit par exemple des enquêtes auprès de populations bénéficiant de services. Dans ce type d'enquêtes, les individus peuvent être interrogés plusieurs fois au cours de l'enquête car ils peuvent fréquenter plusieurs fois des centres dédiés qui proposent différents types de services. Les trois exemples suivants sont concernés par des liens surjectifs :

- l'enquête SAMENTA auprès des personnes sans logement personnel dans laquelle l'objectif était d'estimer la prévalence des principaux troubles psychiatriques dont souffrent ces personnes [80]. Une personne sans logement personnel peut fréquenter plusieurs services

d'aide (centre d'hébergements, points soupe, etc.), un jour donné et même sur plusieurs jours.

- l'enquête PREVAGAY auprès des hommes ayant des relations sexuelles avec d'autres hommes (HSH) fréquentant des établissements gays parisiens (backrooms, lieux de drague, etc.) dont l'objectif est d'estimer le nombre de HSH séropositifs au VIH dans cette population [151].

- l'enquête ANRS-Coquelicot auprès des UD [67].

Les liens entre la population des services et les individus sont donc surjectifs (Figure 3.4, colonne 3).

Enquêtes téléphoniques en population générale : On réalise une enquête téléphonique en population générale. A partir de deux bases de sondage (une liste constituée de numéros filaires et une liste constituée de numéros mobiles), des ménages (unités primaires) sont enquêtés à partir de numéros filaires tirés au sort aléatoirement. Les liens entre les numéros de téléphone (population A) et les ménages (population C) sont surjectifs car plusieurs numéros peuvent appartenir à un même ménage (Figure 3.5).

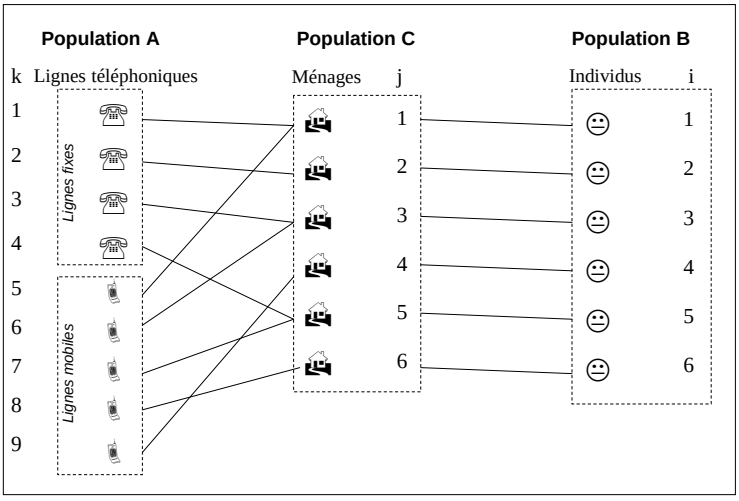


FIGURE 3.5 – Schéma d'enquête téléphonique pour un sondage indirect à 2 degrés. Les lignes droites représentent les liens entre les unités de chaque degré : les lignes téléphoniques en lien avec les ménages au premier degré, et les individus au second degré.

Un individu (unité secondaire) est ensuite tiré au sort au sein de chaque ménage. Parallèlement, à partir de l'échantillon de numéros mobiles généré aléatoirement, des individus sont également enquêtés. Les individus pouvant appartenir aux deux listes (filaire ou mobile) constituant la base de sondage complète peuvent donc être tirés au sort plusieurs fois [164].

3.3 Formalisation de l'échantillonnage lieux-moments

L'échantillonnage lieux-moments défini précédemment a été appliqué à la population des UD fréquentant des centres dédiés à leur usage. Il peut donc être considéré comme un sondage indirect à 3 degrés où, au troisième degré (Figure 3.6) :

- la population A est la population des services,
- la population B est celle des UD qui bénéficient des services proposés dans les centres et
- la fréquentation des UD constitue les liens entre les deux populations.

Une liste des services proposés par les centres (*e.g.* nombre de repas pouvant être servis dans un point-soupe) n'est généralement pas disponible, sauf dans des cas particuliers (par exemple nombre de places pour l'hébergement). L'estimateur présenté dans la section suivante, qui tient compte de la fréquentation multiple des individus, est théoriquement basé sur les services reçus par les individus interrogés. Nous verrons comment cet estimateur est calculé dans la pratique, même lorsque nous ignorons le type exact de services dont ont bénéficié les individus et lorsqu'on ne dispose que de l'identifiant et de la fréquence des centres visités.

Afin de produire une estimation d'un indicateur épidémiologique dans la population d'intérêt B, on s'appuie donc sur les liens existant entre les unités de la population A et les unités d'intérêt de la population B [34, 83].
La section suivante présente deux estimateurs (avec et sans prise en compte des liens entre les deux populations) et leurs propriétés.

3.4 Estimateurs

Pour produire des estimations dans une population à partir d'un échantillon, un poids de sondage est affecté à chaque individu interrogé. La probabilité d'inclusion (de premier ordre) pour une unité est la probabilité que cette unité appartienne à l'échantillon. Un poids de sondage,

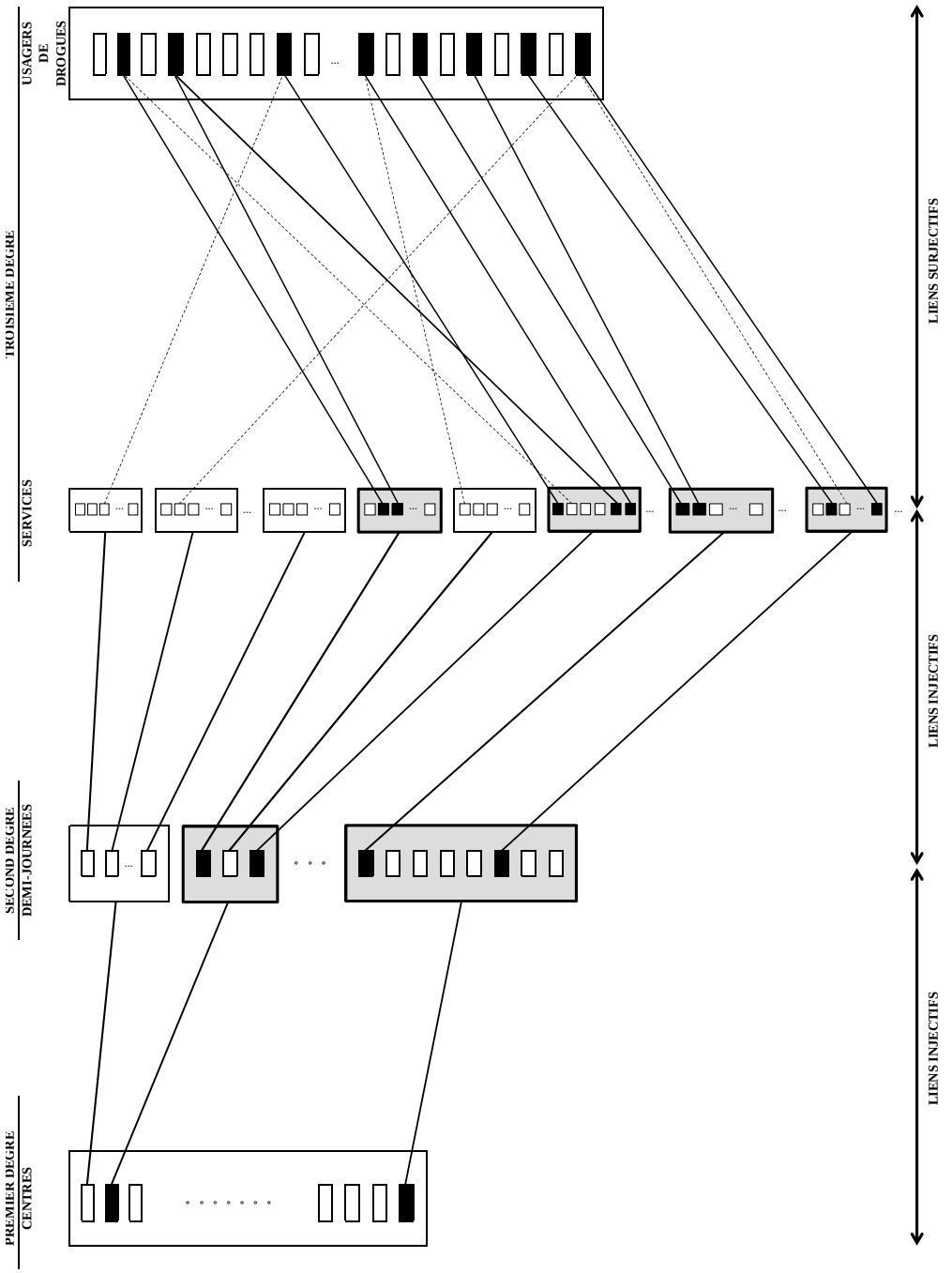


FIGURE 3.6 – Sondage indirect à 3 degrés. Premier degré : UP (carrés noirs) et centres non tirés au sort (carrés blancs). Second degré : US (carrés noirs) et demi-journées non tirées au sort (carrés blancs) parmi les centres tirés au sort (rectangles gris). Troisième degré : UT (carrés noirs) et services non tirés au sort (carrés blancs) parmi les demi-journées tirées au sort (rectangles gris). Les lignes continues représentent les liens connus et les lignes en pointillés les liens déclarés par les usagers de drogues.

défini par l'inverse de la probabilité d'inclusion, peut être exprimé par le produit des poids de sondage calculés à chaque degré du plan de sondage. Les notations de probabilités d'inclusion sont introduites dans le Tableau 3.1 (colonne 2), sous l'hypothèse d'un sondage aléatoire simple à chaque degré.

TABLE 3.1 – Expression du nombre total d'unités et des probabilités d'inclusion de ces unités sous l'hypothèse d'un sondage aléatoire simple sans remise à chaque degré d'échantillonnage dans un plan de sondage à 3 degrés.

Degré	Premier ordre	Second ordre†	Δ Quantités	Totaux
1	$\pi_l = \frac{n}{N}$	$\pi_{ll'} = \frac{n}{N}\left(\frac{n-1}{N-1}\right)$	$\Delta_{ll'} = \pi_{ll'} - \pi_l\pi_{l'}$	$T = \sum_{l=1}^{N} t_l$
2	$\pi_{k\|l} = \frac{n_l}{N_l}$	$\pi_{kk'\|l} = \frac{n_l}{N_l}\left(\frac{n_l-1}{N_l-1}\right)$	$\Delta_{kk'\|l} = \pi_{kk'\|l} - \pi_{k\|l}\pi_{k'\|l}$	$t_l = \sum_{k=1}^{N_l} t_{k\|l}$
3	$\pi_{i\|kl} = \frac{n_{kl}}{N_{kl}}$	$\pi_{ii'\|kl} = \frac{n_{kl}}{N_{kl}}\left(\frac{n_{kl}-1}{N_{kl}-1}\right)$	$\Delta_{ii'\|kl} = \pi_{ii'\|kl} - \pi_{i\|kl}\pi_{i'\|kl}$	$t_{k\|l} = \sum_{i=1}^{N_{kl}} y_i$

† $\pi_{ll} = \pi_l$; $\pi_{kk|l} = \pi_{k|l}$; $\pi_{ii|kl} = \pi_{i|kl}$

Au premier degré, le poids de sondage d'un centre l est $w_l = 1/\pi_l$. Au second degré, le poids de sondage d'une demi-journée k pour un centre l est $w_{k|l} = 1/\pi_{k|l}$. Au troisième degré, le poids de sondage d'un individu i interrogé dans le centre l durant la demi-journée k est $w_{i|kl} = 1/\pi_{i|kl}$.

La probabilité d'inclusion finale à l'enquête d'un individu i est donc $\pi_i = \pi_l \times \pi_{k|l} \times \pi_{i|kl}$ et son poids de sondage final est :

$$w_i = w_l \times w_{k|l} \times w_{i|kl} \tag{3.1}$$

3.4.1 Estimateur ignorant la fréquentation des lieux d'enquêtes

Très souvent, l'objectif principal des enquêtes transversales est d'estimer des fonctions d'intérêt telles qu'un total (par exemple le nombre de personnes infectées), une proportion (par exemple la proportion des individus infectés (appelée prévalence)) ou une moyenne (par exemple la valeur moyenne d'un biomarqueur). Pour chaque individu i dans la population B, on considère une variable d'intérêt (binaire) y_i correspondant au statut sérologique de la maladie d'intérêt : y_i est égal à 1 si i est infecté et 0 sinon.

L'estimateur d'Horvitz-Thompson [62] du nombre total d'individus infectés dans la popula-

tion $T = \sum_{i \in B} y_i$ est :

$$\hat{T} = \sum_{i \in s^B} w_i y_i, \tag{3.2}$$

où s^B est l'échantillon d'individus provenant de la population B en utilisant l'échantillonnage lieux-moments décrit plus haut. La taille de la population N^B, qui est inconnue dans la plupart des cas, en particulier celle des personnes difficiles à joindre, est estimée par $\hat{N}^B = \sum_{i \in s^B} w_i$.

La prévalence $P = \frac{T}{N^B}$ est estimée par :

$$\hat{P} = \frac{\hat{T}}{\hat{N}^B} \tag{3.3}$$

L'estimateur d'Horvitz-Thompson est sans biais pour n'importe quel plan de sondage, si $\pi_i > 0$ pour tout $i \in B$ et si bien sûr les probabilités d'inclusion sont correctement calculées. Pour une population où des individus se rendent dans plusieurs centres pour bénéficier des services proposés, le calcul des probabilités d'inclusion est plus difficile que pour une population où les individus sont plus statiques dans le temps et dans l'espace et peuvent donc n'être sélectionnés qu'une seule fois.

Dans l'échantillonnage lieux-moments, la probabilité d'inclusion d'un individu dépend de sa fréquentation des lieux. Pour collecter cette information sur leur fréquentation, un ensemble de questions est généralement posé aux participants. Une des questions pourrait être par exemple : "Combien de fois vous êtes-vous rendus dans l'un des lieux suivants au cours des cinq jours précédents ?". D'autres questions plus détaillées peuvent être posées selon le type de centre [53]. Ainsi, le nombre de centres visités pour chaque participant peut être pris en compte dans un nouvel estimateur.
Dans la section suivante, nous proposons un estimateur sans biais tenant compte de la fréquentation des individus comme une alternative aux estimateurs (équations 3.2 et 3.3) qui peuvent être biaisés quand les individus ont des fréquentations différentes.

3.4.2 Estimateur tenant compte de la fréquentation des lieux d'enquêtes

Dans le cadre du sondage indirect, nous pouvons utiliser la méthode généralisée du partage des poids (MGPP), développée par Lavallée [81], pour donner un poids de sondage à chaque indi-

vidu interrogé. Un nouveau poids de sondage est affecté à chaque individu $i \in s^B$ et est défini comme une moyenne arithmétique pondérée des poids de sondage de la population de services impliquant les liens entre i et les services j qu'il/elle a reçu.

Le poids de sondage d'un service $j \in s^A$ est $w_j = w_i$ comme défini dans l'équation 3.1.
Si une unité $j \in s^A$ est liée à une unité $i \in s^B$, $l_{ji} \geq 0$. Et si ces deux unités ne sont pas liées, $l_{ji} = 0$.

Notons que plusieurs auteurs ont souligné l'importance du choix de la valeur des liens qui influencent la précision des estimateurs issus du sondage indirect, même si, dans la plupart des applications, la valeur des l_{ji} pour des unités liées est égale à 1 [34,84]. Ainsi, pour chaque unité $i \in s^B$, on peut calculer le nombre total de liens $L_i^B = \sum_{j \in A} l_{ji}$.
Le poids de sondage final intégrant la fréquentation pour chaque unité $i \in s^B$ est défini par :

$$\tilde{w}_i = \frac{1}{L_i^B} \sum_{j \in s^A} l_{ji} w_i \tag{3.4}$$

Les estimations pour les totaux T et N^B et la prévalence P sont donc respectivement :

$$\hat{T}_G = \sum_{i \in s^B} \tilde{w}_i y_i \tag{3.5}$$

$$\hat{N}_G^B = \sum_{i \in s^B} \tilde{w}_i \tag{3.6}$$

$$\hat{P}_G = \frac{\hat{T}_G}{\hat{N}_G^B} \tag{3.7}$$

3.4.3 Propriétés des estimateurs

Biais

Le nombre total T d'individus infectés dans la population d'intérêt peut s'obtenir soit à partir des unités $i \in B$, soit à partir des unités $j \in A$. On parle donc de *dualité*. L'estimateur $\hat{T}_G$ du total T peut donc s'écrire en fonction des unités $i \in s^B$ ou en fonction des unités $j \in s^A$. En utilisant cette dualité, on montre que $\hat{T}_G$ est sans biais [82, 83]. Pour chaque unité $j \in A$, on construit la variable, $z_j = \sum_{i \in B} \frac{l_{ji}}{L_i^B} y_i$. A partir des équations 3.4 et 3.5 on obtient l'égalité

suivante :

$$\begin{aligned}
\hat{T}_G &= \sum_{i\in s^B}\sum_{j\in s^A}\frac{l_{ji}}{L_i^B}w_j y_i \\
&= \sum_{j\in s^A} w_j \left[\sum_{i\in s^B}\frac{l_{ji}}{L_i^B}y_i + \sum_{i\in B\setminus s^B}\frac{l_{ji}}{L_i^B}y_i\right] \text{ car } l_{ji}=0 \text{ pour tout } i\notin s^B, j\in s^A \\
&= \sum_{j\in s^A} w_j \sum_{i\in B}\frac{l_{ji}}{L_i^B}y_i \\
&= \sum_{j\in s^A} w_j z_j \\
&= \sum_{j\in A} w_j z_j I_j. \qquad (3.8)
\end{aligned}$$

où $I_j = 1$ si $j \in s^A$, et 0 sinon.

Le total T peut alors être estimé sans biais par $\hat{T}_G = \sum_{j\in s^A}\frac{1}{\pi_j}z_j$, qui est un estimateur d'Horvitz-Thompson. Il en est de même pour l'estimateur sans biais $\hat{N}_G^B$ de $\hat{N}^B$. A partir de l'équation 3.7, on peut donc montrer que l'estimateur $\hat{P}_G$ de P est sans biais.

L'égalité 3.8 montre qu'il n'est pas nécessaire de connaître les valeurs des liens entre les deux populations complètes A et B, mais uniquement les liens pour les unités j de s^A et les unités i de s^B afin de calculer les poids de sondage des unités i de B servant au calcul de l'estimation [34]. En pratique, pour toute unité i tirée au sort, ces liens avec des unités j de A sont généralement identifiables, soit lors de l'entretien, soit par une source administrative (patients à l'hôpital).

Mesure du biais en cas de non prise en compte de la fréquentation

Sans prise en compte des liens (ici, la fréquentation des individus), l'estimateur $\hat{T}$ du total T peut être biaisé. Les probabilités d'inclusion des unités i sont supposées égales aux probabilités des unités j qui ont permis de les repérer, ce qui n'est pas toujours exact puisque dans le cadre d'un sondage indirect, une unité i peut être liée à une ou plusieurs unités j en fonction du type de lien qui les lie. On sait que $\hat{T}_G$ est un estimateur sans biais de T. Donc $E(\hat{T}_G) = T$. Le biais de $\hat{T}$ s'exprime comme :

$$
\begin{aligned}
Biais(\hat{T}) &= T - E[\hat{T}] \\
&= E[\hat{T}_G] - E[\hat{T}] \\
&= E[\hat{T}_G - \hat{T}] \\
&= E\left[\sum_{i\in s^B}\sum_{j\in A}\frac{l_{ji}}{L_i^B}w_j y_i I_j - \sum_{i\in s^B}\sum_{j\in A} l_{ji} w_j y_i I_j\right] \\
&= E\left[\sum_{i\in s^B}\sum_{j\in A} l_{ji}(\frac{1}{L_i^B} - 1) w_j y_i I_j\right] \\
&= \sum_{i\in s^B}\sum_{j\in s^A} l_{ji}(\frac{1}{L_i^B} - 1) w_j y_i E[I_j] \\
&= \sum_{i\in s^B}\sum_{j\in s^A}(\frac{l_{ji}}{L_i^B} - l_{ji}) y_i \text{ puisque } E[I_j] = \pi_j \\
&= \sum_{i\in s^B}\sum_{j\in s^A} l_{ji}(\frac{1}{\sum_{j\in A} l_{ji}} - 1) y_i \\
&= \sum_{i\in s^B|y_i\neq 0}\sum_{j\in s^A} l_{ji}(\frac{1}{\sum_{j\in A} l_{ji}} - 1) y_i.
\end{aligned}
$$

La valeur du biais dépend alors des liens l_{ji} (leur nombre, leur type, leur valeur) et de la valeur de la variable d'intérêt y_i. Ainsi, si pour tout $i \in s^B$,

- $y_i = 0$, le biais est nul quels que soient le nombre, le type et la valeur des liens
- $y_i \neq 0$, alors
  - si les liens sont injectifs ou bijectifs, (ie. $\sum_{j\in A} l_{ji} = 1$), le biais est nul
  - si les liens sont surjectifs, (*i.e.* $\sum_{j\in A} l_{ji} > 1$), le biais est non nul et est égal à

$$\sum_{i\in s^B|y_i\neq 0}\sum_{j\in s^A} l_{ji}(\frac{1}{\sum_{j\in A} l_{ji}} - 1) y_i.$$

La MGPP n'a donc d'intérêt que dans le cas où il existe au moins un lien surjectif entre les deux populations et que la prévalence est non nulle.

Variance

La variance de l'estimation du total $\hat{T}$ [125], en respectant le plan de sondage, est estimée en utilisant les probabilités d'inclusion du second ordre (qui correspond à la probabilité d'inclusion jointe de deux unités distinctes) et d'autres notations introduites pour simplifier la formule suivante (voir Tableau 3.1, colonnes 3, 4, 5) :

$$\widehat{Var}(\hat{T}) = \sum_{l=1}^{n}\sum_{l'=1}^{n}\Delta_{ll'}\frac{\hat{t}_l}{\pi_l}\frac{\hat{t}_{l'}}{\pi_{l'}} + \sum_{l=1}^{n}\frac{\widehat{Var}(\hat{t}_l)}{\pi_l} + \sum_{l=1}^{n}\frac{1}{\pi_l}\sum_{k=1}^{n_l}\frac{\widehat{Var}(\hat{t}_{k|l})}{\pi_{k|l}} \tag{3.9}$$

où l et $l^{'}$ sont deux centres distincts, k and $k^{'}$ sont deux demi-journées distinctes, i et $i^{'}$ sont deux personnes distinctes et où $\hat{t}_l = \sum_{k=1}^{n_l}\frac{\hat{t}_{k|l}}{\pi_{k|l}}$, $\hat{t}_{k|l} = \sum_{i=1}^{n_{kl}}\frac{y_i}{\pi_{i|kl}}$,
$\widehat{Var}(\hat{t}_l) = \sum_{k=1}^{n_l}\sum_{k'=1}^{n_l}\Delta_{kk'|l}\frac{\hat{t}_{k|l}}{\pi_{k|l}}\frac{\hat{t}_{k'|l}}{\pi_{k'|l}}$ et $\widehat{Var}(\hat{t}_{k|l}) = \sum_{i=1}^{n_{kl}}\sum_{i'=1}^{n_{kl}}\Delta_{ii'|kl}\frac{y_i}{\pi_{i|kl}}\frac{y_{i'}}{\pi_{i'|kl}}$.
$\widehat{Var}(\hat{N}^B)$ est calculée de façon similaire en utilisant l'équation 3.9 et en posant que $y_i = 1$ pour tout $i \in B$.
La variance estimée de la prévalence estimée est :

$$\widehat{Var}(\hat{P}) = \widehat{Var}\left(\frac{\hat{T}}{\hat{N}^B}\right) = \frac{1}{\hat{N}^{B2}}\{\widehat{Var}(\hat{T}) - 2\hat{P}\widehat{Cov}(\hat{T},\hat{N}^B) + \hat{P}^2\widehat{Var}(\hat{N}^B)\} \tag{3.10}$$

où

$$\widehat{Cov}(\hat{T},\hat{N}^B) = \sum_{l=1}^{n}\sum_{l'=1}^{n}\Delta_{ll'}\frac{\hat{t}_l}{\pi_l}\frac{\hat{N}_l}{\pi_{l'}} + \sum_{l=1}^{n}\frac{\widehat{Cov}(\hat{t}_l,\hat{N}_l)}{\pi_l} + \sum_{l=1}^{n}\frac{1}{\pi_l}\sum_{k=1}^{n_l}\frac{\widehat{Cov}(\hat{t}_{k|l},\hat{N}_{k|l})}{\pi_{k|l}} \tag{3.11}$$

avec $\hat{N}_l = \sum_{k=1}^{n_l}\frac{\hat{N}_{k|l}}{\pi_{k|l}}$, $\hat{N}_{k|l} = \sum_{i=1}^{n_{kl}}\frac{1}{\pi_{i|kl}}$, $\widehat{Cov}(\hat{t}_l,\hat{N}_l) = \sum_{k=1}^{n_l}\sum_{k'=1}^{n_l}\Delta_{kk'|l}\frac{\hat{t}_{k|l}}{\pi_{k|l}}\frac{\hat{N}_{k'|l}}{\pi_{k'|l}}$ et
$\widehat{Cov}(\hat{t}_{k|l},\hat{N}_{k|l}) = \sum_{i=1}^{n_{kl}}\sum_{i'=1}^{n_{kl}}\Delta_{ii'|kl}\frac{y_i}{\pi_{i|kl}}\frac{1}{\pi_{i'|kl}}$.

Notons que si les probabilités d'inclusion de deuxième ordre sont faciles à calculer en utilisant un sondage aléatoire simple, leur calcul est plus complexe et parfois insoluble avec d'autres techniques d'échantillonnages et dépendent des algorithmes de tirage utilisés [145]. Avec des plans de sondage plus complexes, les variances peuvent être estimées en utilisant des procédures du type jackknife ou bootstrap [125].

Les variances respectives de $\hat{T}_G$, $\hat{N}_G^B$, $\hat{P}_G$ sont estimées en utilisant les mêmes expressions présentées dans les équations 3.9, 3.10 et 3.11 avec $\widehat{Var}(\hat{t}_{k|l}) = \sum_{j=1}^{n_{kl}} \sum_{j'=1}^{n_{kl}} \Delta_{ii'|kl} \frac{z_j}{\pi_{i|kl}} \frac{z_{j'}}{\pi_{i'|kl}}$, $\widehat{Cov}(\hat{t}_{k|l}, \hat{N}_{k|l}) = \sum_{j=1}^{n_{kl}} \sum_{j'=1}^{n_{kl}} \Delta_{ii'|kl} \frac{z_j}{\pi_{i|kl}} \frac{1}{\pi_{i'|kl}}$ et $z_j = \sum_{i=1}^{n_i} \frac{l_{ji}}{L_i^B} y_i$.

Nous présentons dans la section suivante une application de cette méthode en utilisant l'enquête ANRS-Coquelicot de 2011 et une étude de simulation permettant d'évaluer l'impact des liens à travers différents scénarios en faisant varier : la prévalence dans la population simulée, la distribution des fréquentations (liens) qui dépend ou non du statut sérologique et les erreurs sur les fréquentations déclarées.

3.5 Application auprès d'une population d'usagers de drogues

En pratique, il est souvent inutile de demander aux participants de quels services spécifiques ils ont bénéficié ou même leur nombre total de visites sur une période d'enquête. Premièrement, les personnes peuvent hésiter à répondre à cette question. Ils se rendent dans les centres pour des raisons particulières et ne voient pas l'intérêt de passer du temps à essayer de se rappeler ce qu'ils ont fait dans le passé, surtout après une longue entrevue. Cette question peut également être considérée par les répondants comme un contrôle de leurs pratiques illicites. Deuxièmement, les participants peuvent trouver difficile de répondre à ces questions avec précision en raison de l'oubli ou de la confusion dans l'identification du centre. C'est encore plus marqué dans les populations précaires et plus particulièrement chez les UD qui sont sous l'emprise de produits ayant un impact délétère sur les fonctions cognitives, ou bien quand il y a un très grand nombre de centres (par exemple dans une grande ville, comme Paris où une cinquantaine de centres dédiés aux UD ont été recensés).

Enfin, les conditions pratiques de l'entretien permettent rarement la collecte de ces informations détaillées, par exemple lors de l'administration d'un questionnaire dans la rue.

Pour toutes ces raisons, lors de l'entretien, quelques questions sont posées sur la fréquentation des lieux d'enquêtes sur une courte période du passé. Ce point sera développé plus en détail dans la section 3.5.1.

Toutefois, il est important de noter qu'il n'est pas nécessaire d'avoir des informations détaillées sur les services dont ont bénéficié les participants pour utiliser l'estimateur tenant compte des liens. En effet, les individus sont généralement sélectionnés aléatoirement à leur arrivée dans le centre, quel que soit le nombre de services dont ils vont bénéficier et quel que soit le temps qu'ils vont passer dans le centre. Leurs probabilités d'inclusion ne dépendent donc pas du nombre de services qu'ils reçoivent mais bien du nombre de visites dans les centres. C'est pourquoi nous avons besoin de comptabiliser le nombre de visites dans les différents centres.

Enfin, on peut dire que lister l'ensemble des services dont chaque participant aurait bénéficié ou encore l'ensemble de visites sur toute la période d'enquête n'est pas réaliste. Plusieurs chercheurs se sont alors focalisés sur quelques questions à poser sur la fréquentation des lieux d'enquêtes avec quelques restrictions : la fréquentation est renseignée comme une variable discrète à plusieurs modalités, sur une courte période passée et parfois sur un nombre limité de centres [53, 75].

3.5.1 L'enquête ANRS-Coquelicot 2011

Le plan de sondage de l'enquête peut être représenté par la Figure 3.6.

La taille de la population est estimée à $\hat{T} = 48147$, $IC95\% = [43741; 52553]$ selon l'estimateur ignorant la fréquentation des participants tandis qu'elle est estimée à $\hat{T}_G = 43710$, $IC95\% = [39667; 47753]$ lorsqu'on en tient compte.

La prévalence du VHC est estimée à $\hat{P} = 43.4\%$, $IC95\% = [39.3\%; 47.6\%]$ selon l'estimateur ignorant la fréquentation des participants tandis qu'elle est estimée à $\hat{P}_G = 43.7\%$, $IC95\% = [39.5\%, 47.9\%]$ lorsqu'on en tient compte. Ces deux dernières estimations sont très proches, probablement à cause de la faible variabilité du nombre de fréquentations déclarées par les UD (Figure 2.5, à droite). Plusieurs raisons pourraient expliquer cette faible variabilité. Premièrement, elle est réelle, auquel cas il n'est pas nécessaire de prendre en compte la fréquentation des lieux. Deuxièmement, cette variance observée est erronée, soit parce que les participants ne se rappellent pas exactement leurs visites, soit parce que les questions relatives aux fréquentations ne sont pas adéquates pour cette population d'intérêt. Dans cette étude, on suppose une fréquentation stable au cours des 11 semaines d'enquête, ce qui permet d'interroger les individus sur leur fréquentation uniquement durant les 5 jours précédents. Cependant, les

usagers de drogues peuvent tout de même être réticents à répondre à ces questions pour les raisons décrites plus haut. Des erreurs peuvent donc apparaître sur la déclaration de fréquentations et entraîner une sous-estimation de la variance.

Pour aller plus loin dans cette réflexion et mesurer l'impact de ces erreurs potentielles, nous avons conduit une étude de simulation présentée dans la section suivante.

3.5.2 Étude de simulation

Nous avons généré N centres proposant des services sur une période d'enquête fixée. Nous avons fait l'hypothèse que chaque centre l ($l = 1, \ldots, N$) était ouvert N_l demi-journées durant la période d'enquête et proposaient N_{kl} services durant une demi-journée k ($k = 1, \ldots, N_l$). Un total de $N^A = \sum_{l=1}^{N} \sum_{k=1}^{N_l} N_{kl}$ services était donc proposé par l'ensemble des N centres durant la période d'enquête. Notons N^B la taille de la population d'individus B qui bénéficient d'un ou plusieurs services.

Premièrement, nous avons construit une base de sondage des services et des individus (Étapes 1-3). Puis, nous avons échantillonné les centres, puis les demi-journées et les services (Étapes 4-6). Enfin, nous avons obtenu un échantillon d'individus (Étape 7).

Étape 1. Nous avons construit une base de sondage des services en générant une matrice de taille $N^A \times 3$, N^A lignes représentant l'ensemble des services et 3 colonnes identifiant pour chaque service le centre, la demi-journée et le service lui-même.

Étape 2. Nous avons généré une population de N^B individus. Pour chaque individu i ($i = 1, \ldots, N^B$), nous avons généré aléatoirement son nombre total de liens (*i.e.* de fréquentations) au cours de la période d'enquête, L_i^B, suivant une distribution binomiale négative de moyenne μ et de variance $\theta\mu$ avec un paramètre de dispersion $\theta \geq 1$. Douze combinaisons de paramètres μ et θ (appelées scénarios 1 à 12) présentées dans le Tableau 3.2 ont été utilisés. Pour l'enquête ANRS-Coquelicot, la fréquentation dans les lieux d'enquêtes suivait une distribution de Poisson de paramètres ($\lambda = 1$). Un statut sérologique a ensuite été affecté à chaque individu pour pouvoir générer son nombre de liens en fonction de son statut. Pour cela, nous avons généré un nombre de liens pour les individus séropositifs (respectivement séronégatifs) suivant une distribution binomiale négative de moyenne μ_1 (respectivement μ_2) et de variance $\theta_1\mu_1$ (respectivement $\theta_2\mu_2$). Quatre combinaisons de paramètres ont alors été utilisées (appelées scénarios 13 à 16, Tableau

3.2). Pour l'enquête ANRS-Coquelicot, $(\mu_1, \mu_2, \theta_1, \theta_2) = (1, 1, 1, 1)$. Pour chaque scénario, onze prévalences allant de 1% à 90% ($P = 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$) ont été générées. La distribution des liens est présentée dans l'annexe 3.

TABLE 3.2 – Paramètres associés aux 16 scénarios.

Scénario	μ	θ	μ_1	θ_1	μ_2	θ_2
1	3	1	–	–	–	–
2	3	3	–	–	–	–
3	3	5	–	–	–	–
4	3	10	–	–	–	–
5	3	20	–	–	–	–
6	3	50	–	–	–	–
7	5	1	–	–	–	–
8	5	3	–	–	–	–
9	5	5	–	–	–	–
10	5	10	–	–	–	–
11	5	20	–	–	–	–
12	5	50	–	–	–	–
13	–	–	3	1	10	1
14	–	–	3	1	10	10
15	–	–	5	1	10	1
16	–	–	5	1	10	10

Étape 3. A chaque individu, nous avons associé autant de services que de liens générés. Ainsi, à la matrice $N^A \times 3$ construite à l'étape 1, nous avons ajouté trois autres colonnes identifiant l'individu associé au service lui-même, son nombre total de liens et son statut sérologique avec la contrainte qu'un individu ne peut pas avoir plus d'un lien au cours d'une demi-journée donnée.

Étape 4. Nous avons généré un échantillon de centres en tirant au sort n nombres compris entre 1 et N selon un sondage aléatoire simple sans remise.

Étape 5. Pour chaque centre l tiré au sort, nous avons généré un échantillon de demi-journées en tirant au sort n_l nombres compris entre 1 et N_l selon un sondage aléatoire simple sans remise.

Étape 6. Pour chaque centre l et chaque demi-journée k tirés au sort, nous avons généré un échantillon de services en tirant au sort n_{kl} nombres compris entre 1 et N_{kl} selon un sondage aléatoire simple sans remise.

Étape 7. Un échantillon de n_B individus ($n_B \leq n_A$) a été obtenu en utilisant les liens entre les services tirés au sort et les individus.

Puis, nous avons généré 10000 échantillons de taille $n_A = 2000$ pour chacun des 16 scénarios et pour chacune des 11 prévalences.
Pour chaque échantillon généré, $\hat{N}^B$, $\hat{T}$, $\hat{P}$, $\hat{N}_G^B$, $\hat{T}_G$, $\hat{P}_G$ ont été calculés.

Dans la réalité, il est difficile pour des individus de se souvenir de toutes leurs fréquentations particulièrement quand elles sont multiples et variées. Chez les UD, les individus interrogés sont parfois socialement vulnérables et consomment des substances psychoactives qui peuvent détériorer les capacités des fonctions cognitives et affecter la mémoire. C'est pourquoi, nous avons généré trois sortes d'erreurs (Tableau 3.3) pour explorer les propriétés de l'estimateur quand des erreurs apparaissent dans la déclaration des liens.

TABLE 3.3 – Paramètres utilisés pour générer des liens erronés

Erreur	$L_i^{B,erreur}$	k
1	$L_i^B \times (k+1)$	$k \sim Unif(-0.5, 0.5)$
2	$L_i^B + k$	$k \in [-(L_i^B - 1); L_i^B]$
3	$L_i^B \times (k+1)$	$k \sim Unif(-0.5, 0)$

Résultats de l'étude de simulation. L'étude de simulation a montré que, quel que soit le scénario, l'estimateur qui tient compte de la fréquentation est sans biais et ce, quelle que soit la prévalence réelle dans la population étudiée. Au contraire, l'estimateur qui ne tient pas compte de la fréquentation est biaisé dans plusieurs scénarios, particulièrement pour les scénarios 13-16 où la fréquentation dépend du statut sérologique comme l'illustre la Figure 3.7.

La Figure 3.8 représente le biais relatif pour toutes les prévalences estimées. Pour les scénarios 13-16, les prévalences estimées selon l'estimateur ignorant la fréquentation des lieux d'enquêtes sont 1.05 à 2.22 supérieures à la prévalence réelle contrairement aux prévalences estimées sans biais avec l'autre estimateur.

La couverture de probabilité des prévalences estimées varient de 87% à 100% en utilisant

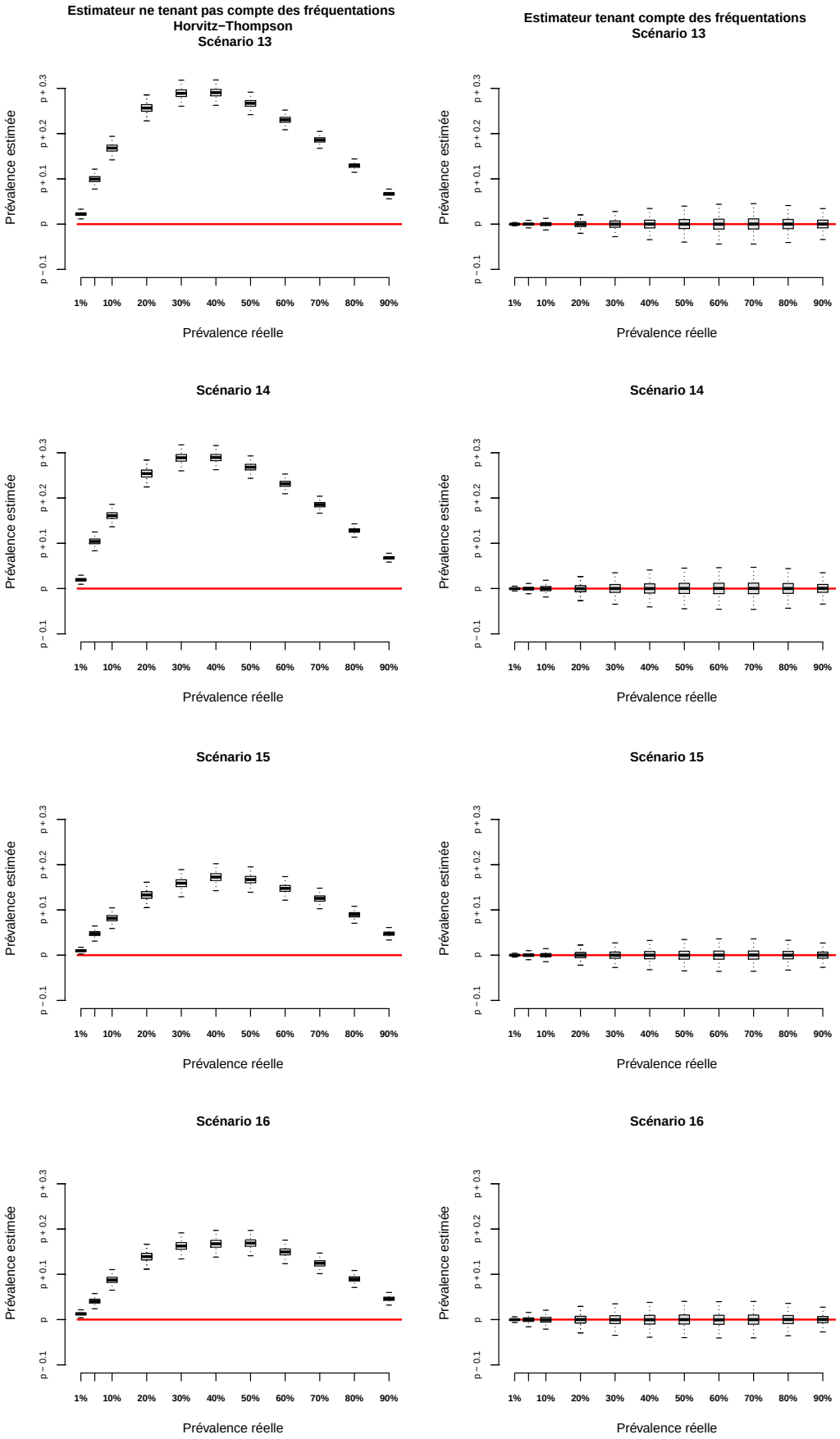


FIGURE 3.7 – Boxplots des prévalences estimées à partir des deux estimateurs avec prise en compte (à droite) ou non (à gauche) de la fréquentation des lieux d'enquêtes pour les scénarios 13-16. Pour chaque graphique, la ligne rouge représente la prévalence réelle de chaque population générée pour chaque scénario.

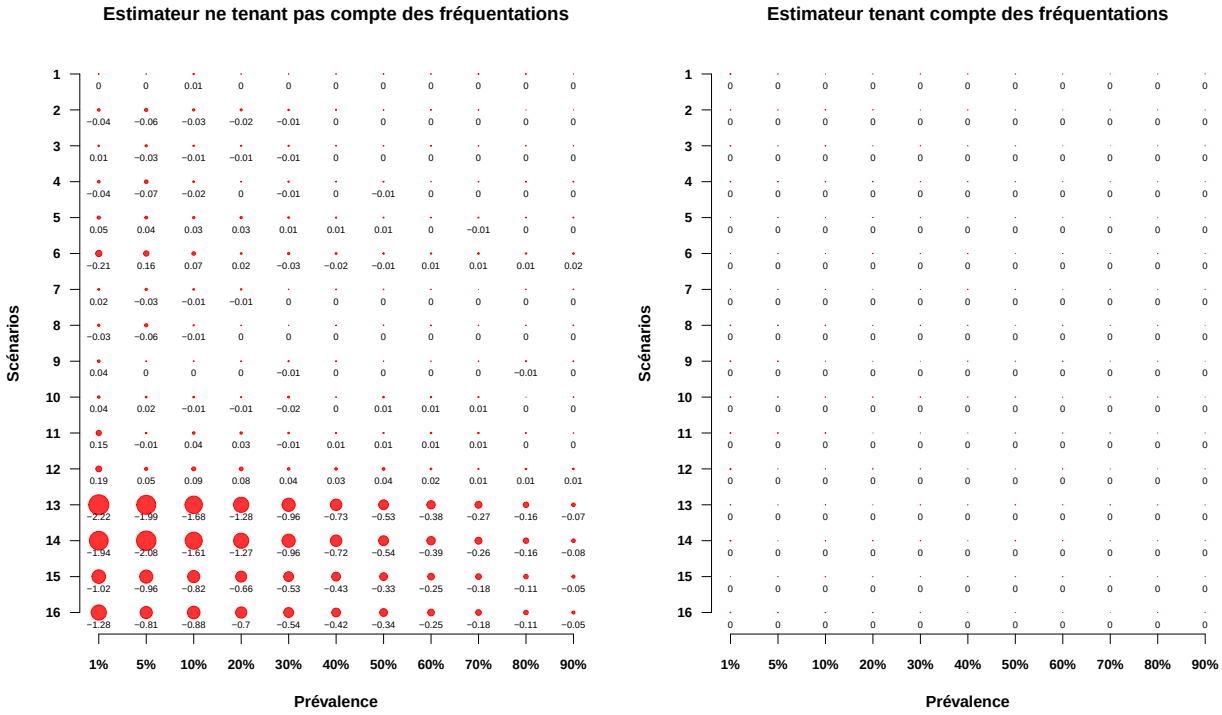


FIGURE 3.8 – Biais relatif représenté par des cercles selon les différents scénarios et les différentes prévalences pour les deux estimateurs avec prise en compte (à droite) ou non (à gauche) de la fréquentation des lieux d'enquêtes.

l'estimateur qui prend en compte la fréquentation, et de 0% (scénarios 13-16) à 95% avec l'estimateur qui n'en tient pas compte (Figure 3.9).

Quand des erreurs apparaissent dans la déclaration des fréquentations, un biais faible (scénarios 13-16) voire une absence de biais (scénarios 1-12), est observé dans les estimations de la prévalence lorsqu'on utilise l'estimateur avec prise en compte des fréquentations, et ce, quelle que soit la prévalence réelle dans la population (Figures 3.10 et 3.11). Aussi, comme attendu, le biais observé augmente en fonction du type d'erreurs introduit dans les fréquentations (erreur 1 $\leq$ erreur 2 $\leq$ erreur 3) (Tableau 3.3).

3.6 Discussion

Dans ce chapitre, nous avons présenté et formalisé le TLS dans le cadre d'un sondage indirect à plusieurs degrés. Nous avons pu proposer un estimateur utilisant la MGPP pour fournir

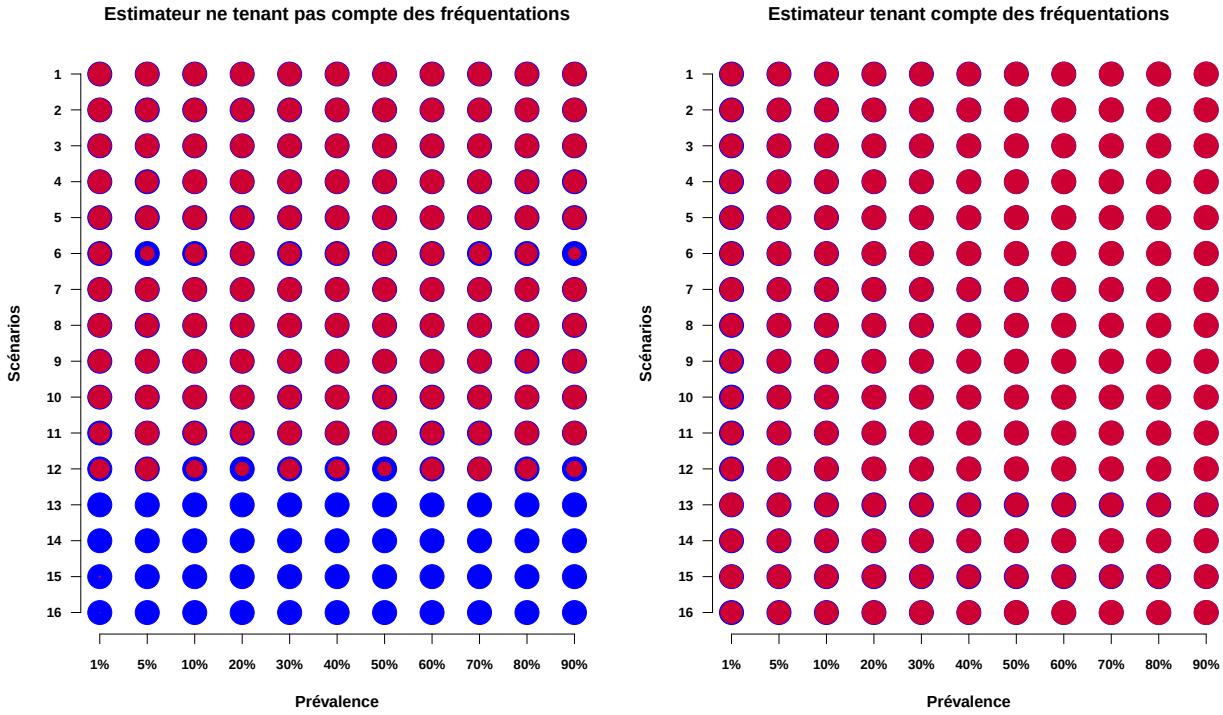


FIGURE 3.9 – Couverture de probabilité. Un cercle complètement bleu indique une couverture de probabilité nulle et un cercle complètement rouge indique une couverture de probabilité totale (i.e égale à un) pour l'estimateur ne tenant pas compte des fréquentations (à gauche) et pour l'estimateur tenant compte des fréquentations (à droite).

des estimations correctes d'indicateurs tels qu'un total ou une proportion, quand la population d'intérêt fréquente des lieux spécifiques. Cet estimateur tient compte de la fréquentation des lieux d'enquêtes, parfois hétérogène au sein d'un groupe d'individus.

Dans l'enquête Coquelicot, l'estimateur que nous proposons a été pondéré par le nombre de visites dans les lieux d'enquêtes. Il montre des résultats similaires par rapport à l'estimateur établi par Horvitz-Thompson. Cela est peut-être dû à la faible variabilité observée dans les fréquentations déclarées par les participants à l'enquête. Bien que nous n'ayons pas examiné plus en détails la raison pour laquelle la variance observée est faible, nous pouvons avancer quelques hypothèses.

Premièrement, la variance réelle dans la population étudiée d'UD est faible. Ce n'est sans doute pas l'explication la plus probable : les UD ont des caractéristiques hétérogènes, notamment en termes de consommation de drogues (type de drogues, quantité, modes de consommation) et

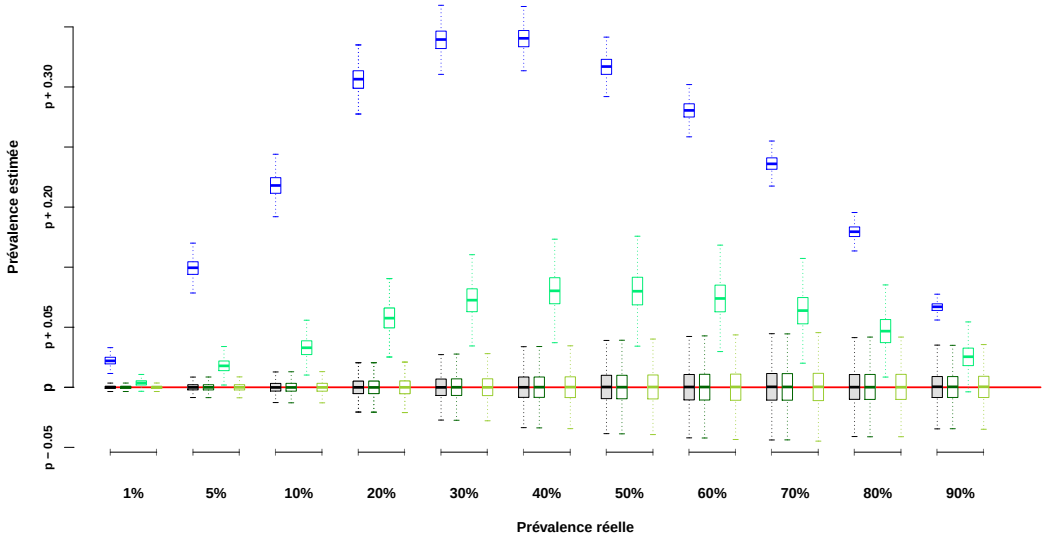


FIGURE 3.10 – Scénario 13 : Boxplots des prévalences estimées à partir d'un estimateur ignorant les fréquentations (en bleu), de l'estimateur tenant compte des fréquentations, sans erreurs sur les fréquentations déclarées d'une part (en gris) et avec erreurs sur les fréquentations déclarées d'autre part (en verts). La ligne horizontale rouge représente la prévalence réelle (variant de 1% à 90%) de chaque population générée.

donc nous nous attendions à une fréquentation hétérogène des centres [158]. Si cette première hypothèse s'avère vraie, il n'y a aucun bénéfice à utiliser un estimateur qui tienne compte de la fréquentation plutôt qu'un estimateur qui l'ignore pour estimer une proportion. En revanche, le bénéfice est réel pour l'estimation d'un total, au risque de biaiser l'estimation. En effet, si la fréquentation des lieux d'enquêtes n'est pas prise en compte et en supposant que les individus de la population étudiée fréquentent tous 5 fois les lieux d'enquêtes sur la période d'étude, la taille de la population estimée sera alors 5 fois plus importante que la taille réelle de la population.

Deuxièmement, la variance réelle est beaucoup plus élevée que celle observée à cause de la difficulté à collecter les fréquentations exactes des personnes interrogées. En effet, les participants ayant un grand nombre de visites (i.e. une fréquentation élevée) ne trouvent aucun intérêt à perdre leur temps à se rappeler et à lister l'ensemble de leurs visites. Cela conduit à une sous-estimation de la variance.

Il n'existe probablement pas de méthode parfaite pour collecter la fréquentation des lieux d'enquêtes. Cela dépend de la population d'intérêt (UD, HSH, personnes sans domicile, mi-

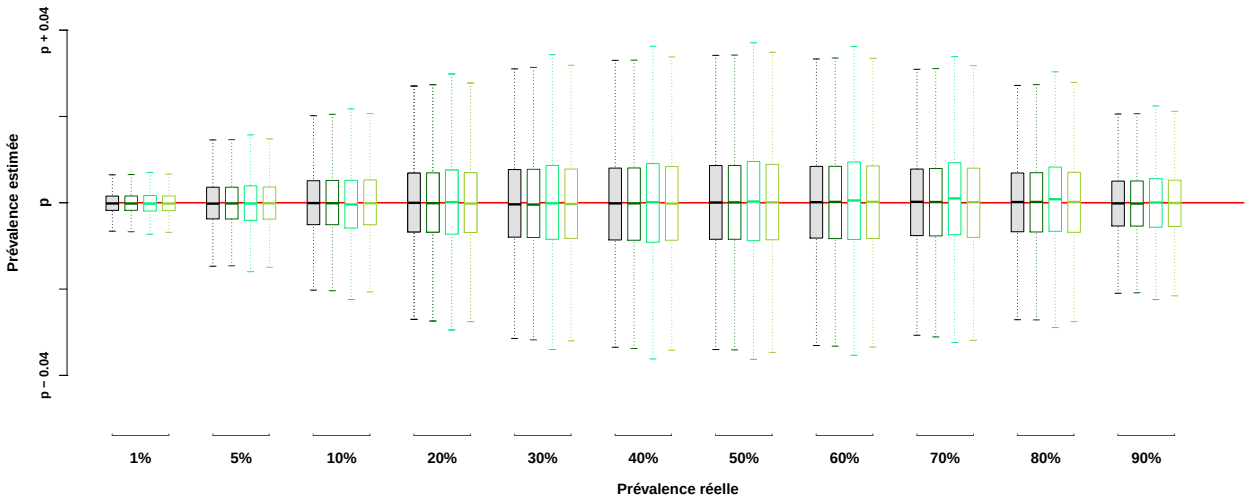


FIGURE 3.11 – Scénario 1 : Boxplots des prévalences estimées à partir de l'estimateur tenant compte des fréquentations, sans erreurs sur les fréquentations déclarées d'une part (en gris) et avec erreurs sur les fréquentations déclarées d'autre part (en verts). La ligne horizontale rouge représente la prévalence réelle (variant de 1% à 90%) de chaque population générée.

grants) et des lieux d'enquêtes (centres spécifiques, consultations médicales, lieux festifs, la rue). Des études spécifiques pourraient être envisagées selon les populations étudiées pour proposer un guide recommandant des questions précises liées à la fréquentation et les intégrer dans les questionnaires pour des enquêtes TLS.

D'ailleurs, dans l'étude de simulation, nous avons étudié différents scénarios pour illustrer plusieurs types de populations difficiles à joindre avec différentes valeurs de prévalences et différentes distributions des fréquentations individuelles dépendantes ou non du statut sérologique. Nous avons conclu que recueillir des informations sur la fréquentation dans les entretiens en face-à-face était crucial pour pouvoir modifier les poids de sondage et obtenir des estimations sans biais. Au lieu de cela, ignorer la fréquentation entraînait un biais sévère et une faible couverture de probabilité, en particulier lorsque la fréquentation dépendait du statut sérologique. De plus, même avec des erreurs dans la déclaration des fréquentations, le biais était réduit par rapport à l'omission de la fréquentation.

Cette étude de simulation s'est principalement focalisée sur l'impact de la fréquentation dans l'évaluation du biais de l'estimateur. D'autres sources de biais pourraient jouer un rôle dans la

robustesse de notre estimateur mais n'ont pas été étudiées. Cependant, même si elles existent, notre estimateur doit être privilégié par rapport à un estimateur ignorant la fréquentation lorsque un biais peut être lié à la fréquentation.

Dans une future étude, il serait également intéressant de comparer notre estimateur et un estimateur basé sur un modèle comme celui développé par Gustafson *et al.* [53], à partir de données simulées, et de discuter les avantages et les inconvénients de ces 2 estimateurs. D'après les résultats de la présente étude, nous pouvons déjà affirmer que l'utilisation du sondage indirect associé à la MGPP pourrait résoudre les problèmes rencontrés dans les enquêtes téléphoniques lorsque des personnes sont joignables à partir de plusieurs contacts téléphoniques (à la fois fixe et mobile) qui doivent être pris en compte dans les estimations.

Chapitre 4

Estimation de l'incidence de l'infection par le virus de l'hépatite C chez les usagers de drogues

Nous avons souligné la nécessité en épidémiologie de disposer de deux indicateurs pour évaluer le niveau d'une infection dans une population et sa dynamique au cours du temps. Une estimation sans biais de la prévalence d'une infection dans une population fréquentant des lieux d'enquêtes a été proposée dans le précédent chapitre.

Dans ce chapitre, nous nous intéressons à l'estimation de l'incidence de l'infection à VHC chez les usagers de drogues. Les travaux réalisés dans cette partie font l'objet d'un article accepté dans la revue *Epidemiology and Infection* en 2016.

4.1 État des lieux

4.1.1 Différentes approches pour estimer l'incidence d'une maladie

Le moyen le plus approprié pour estimer l'incidence d'une maladie dans une population est de mettre en place une cohorte prospective de personnes susceptibles - c'est-à-dire indemnes de la maladie que l'on souhaite étudier - mais à risque de contracter la maladie. Les personnes incluses dans une cohorte répondent à une définition de cas précise, discutée et validée lors de l'élaboration du protocole d'étude. Ces personnes sont suivies au cours du temps puis testées pour la maladie étudiée à intervalles de temps réguliers. Pour certaines études épidémiologiques,

cette approche peut nécessiter de suivre un grand nombre de personnes sur une période de temps parfois très longue (en fonction de la maladie étudiée) ce qui peut rendre l'étude difficile, coûteuse et chronophage. C'est le cas des études où l'on s'intéresse à une maladie avec une longue période d'incubation auprès de populations difficiles d'accès.

A titre d'exemples, l'incidence de l'infection à VHC a été estimée à partir de cohortes de personnes anti-VHC négatives à haut-risque de contracter la maladie en détectant la présence ou non d'anticorps anti-VHC à intervalles de temps réguliers : chez des prisonniers tous les 6-12 mois pendant 4 ans [91], chez des usagers de drogues au bout d'un an [28] ou chez des personnes séropositives au VIH tous les 3-4 mois [138].

En l'absence de cohorte prospective, des approches alternatives sont envisagées, soit par la mise en place d'un système de surveillance, soit à partir d'une ou plusieurs enquêtes épidémiologiques transversales. Plusieurs méthodes découlent de l'utilisation d'enquêtes épidémiologiques transversales (Tableau 4.1).

TABLE 4.1 – Méthodes d'estimation de l'incidence de l'infection à VHC chez les usagers de drogues

Type d'étude	Marqueurs biologiques recherchés	Test biologique	Méthode d'estimation	Référence
Cohorte prospective	Anti-VHC	3.0 ELISA	Proportion de séropositifs parmi les séronégatifs inclus dans la cohorte	[28]
Enquête transversale	Anti-VHC et ARN du VHC	3.0 ELISA	Détection de l'ARN du VHC parmi les séronégatifs en tenant compte de la période fenêtre	[12, 19]
Enquête transversale	Anti-VHC	3.0 ELISA modifié (Test d'avidité)	Proportion d'infections récentes (*i.e.* aiguës)	[30]
Cohorte rétrospective (construite à partir d'enquêtes transversales)	Anti-VHC	3.0 ELISA	Proportion de séropositifs parmi les séronégatifs inclus dans la cohorte	[65]
Enquêtes transversales	Anti-VHC	3.0 ELISA	Modèle mathématique	[139]

L'incidence de l'infection à VHC peut être estimée à partir de cohortes rétrospectives, généralement construites à partir de plusieurs enquêtes transversales quand seuls les anticorps anti-VHC sont recherchés [65]. Elle peut aussi être mesurée à partir de modèles mathématiques [27, 59, 101].

Enfin, l'incidence de l'infection à VHC peut être évaluée à partir d'enquêtes séro-épidémiologiques transversales où l'ARN du VHC est également recherché parmi les personnes interrogées anti-VHC négatives tout en tenant compte de la période fenêtre (Section 1.2.1). Cette dernière est définie soit à partir des données collectées lors de l'enquête [113], soit à partir de la littérature [20]. En utilisant la proportion des personnes nouvellement infectées (*i.e.* ARN du VHC positives et anti-VHC négatives) et une estimation de la période fenêtre, l'incidence de l'infection à VHC est calculée à partir de la formule suivante [20, 91] :

$$I = \frac{n^+}{n^+ + n^-} \times \frac{365}{PF} \times 100 \tag{4.1}$$

où, n^+ (respectivement n^-) est le nombre de personnes positives (respectivement négatives) pour l'ARN du VHC parmi les personnes anti-VHC négatives, et PF est la valeur de la période fenêtre. L'incidence I s'exprime pour 100 personnes susceptibles par an.

Par ailleurs, une autre approche biologique permettant d'identifier des infections primaires récentes est la mesure de l'index d'avidité des IgG anti-VHC [30, 127]. En France, cette méthode biologique est développée mais n'a pas encore été appliquée [42, 116].

La collecte d'échantillons de sang (sur plasma, sérum ou DBS) est donc nécessaire à la recherche d'anticorps anti-VHC et/ou de l'ARN du VHC quel que soit le design d'étude retenu pour estimer la prévalence ou l'incidence de l'infection à VHC.

4.1.2 Données disponibles en France

Actuellement en France, le projet de recherche sur le suivi épidémiologique et socio-comportemental d'UD a abouti à la mise en place d'une cohorte prospective d'UD (cohorte COSINUS [161]) dans quatre villes françaises (Bordeaux, Marseille, Paris et Strasbourg) dont l'objectif est d'évaluer l'impact des politiques de réduction des pratiques à risque de VHC et d'accès aux soins et à la prévention. Toutefois, dans cette cohorte dont le démarrage est programmé à l'automne 2016, il n'est pas prévu à ce jour de recueillir de données biologiques permettant de fournir des indicateurs épidémiologiques pour l'étude de l'hépatite C. Les seules données disponibles permettant de mesurer l'incidence de l'infection à VHC dans cette population sont issues des enquêtes ANRS-Coquelicot. Dans ces deux enquêtes, une recherche d'anticorps anti-VHC a été réalisée. Les résultats de recherche de l'ARN du VHC étaient disponibles uniquement pour l'enquête de

2011. Ainsi, à partir de ces données et des éléments de la section précédente, deux approches ont été envisagées :

- Une approche par modèle mathématique, basée sur la relation entre prévalence et incidence, à partir des données des 2 enquêtes Coquelicot 2004 et 2011 (Sections 4.2 et 4.3).
- Une approche biologique à partir des données de Coquelicot 2011 où les anti-VHC et l'ARN du VHC étaient recherchés (Section 4.4).

4.2 Estimation de l'incidence à partir d'enquêtes transversales répétées

Nous présentons dans cette section une estimation de l'incidence de l'infection à VHC chez les UD en construisant un modèle mathématique reposant sur la formulation d'une relation entre la prévalence et l'incidence à partir de deux enquêtes transversales. Trois étapes ont été nécessaires et sont détaillées dans les sous-sections suivantes.

4.2.1 Combinaison des deux enquêtes

Afin de travailler sur des populations similaires, nous avons exclu des analyses les UD qui ont été interrogés uniquement, dans les cabinets de médecins généralistes prescripteurs de traitements de substitution aux opiacés en 2004 d'une part, et dans les deux départements (Seine-et-Marne et Seine-Saint-Denis) ajoutés en 2011 d'autre part. Nous avons également vérifié que les données recueillies renseignaient la même information dans les deux enquêtes. L'ensemble des analyses a tenu compte du plan de sondage des deux enquêtes (stratification, unités primaires, poids de sondage, etc.).

4.2.2 Estimation de la prévalence en fonction de l'âge et du temps par modèles de régression

Pour chaque individu i, considérons la variable d'intérêt binaire Y_i telle que :

$$Y_i = \begin{cases} 1 & i \text{ est séropositif au VHC (anti-VHC positif)} \\ 0 & \text{sinon} \end{cases}$$

Notons $P(Y = 1|a, t)$, la probabilité d'être séropositif à l'âge a à l'instant t. Nous avons appliqué un modèle de régression multivariée pour estimer la prévalence du VHC par âge et en fonction du temps en incluant la variable âge en continue. Les deux approches les plus connues pour le traitement des variables continues dans une régression sont les splines ou les polynômes fractionnaires. Dans le premier cas, un modèle additif généralisé est utilisé. Dans le second cas, un modèle linéaire généralisé est appliqué. En supposant a priori une forme simple de la relation entre l'infection VHC et l'âge (pas plus de deux points d'inflexion), nous avons fait le choix des polynômes fractionnaires pour notre modèle de régression. L'équation du modèle est donnée par :

$$g(E[Y|a, t]) = g(P(Y = 1|a, t)) = \alpha + \eta(a) + ct, \tag{4.2}$$

où, g est la fonction de lien, α la constante, c le coefficient de régression associé au temps t et $\eta(a)$ le polynôme fractionnaire associé à l'âge a.

Les polynômes fractionnaires sont une extension de polynômes classiques où les puissances peuvent être réelles [120]. Il s'agit de proposer différentes combinaisons de polynômes au lieu d'une simple droite pour estimer la relation entre la variable d'intérêt et une variable continue et conserver le modèle qui s'ajuste le mieux aux données (*i.e.* celui maximisant la vraisemblance).

Le polynôme fractionnaire d'ordre m pour le prédicteur linéaire, associé à l'âge a est défini par :

$$\eta_m(a, \beta, p_1, p_2, \ldots, p_m) = \sum_{j=0}^{m} \beta_j H_j(a), \tag{4.3}$$

où m est un entier, β le vecteur des coefficients de régression, $p_1 \leq p_2 \ldots \leq p_m$ est une séquence de puissances et $H_j(a)$ est définie récursivement, pour tout $j = 0, \ldots, m$ par :

$$H_j(a) = \begin{cases} a^{p_j} & \text{si } p_j \neq p_{j-1} \\ H_{j-1}(a) \times ln(a) & \text{si } p_j = p_{j-1} \end{cases}$$

avec pour valeurs initiales $p_0 = 0$ et $H_0 = 1$.

Classiquement, les puissances sont choisies dans un ensemble restreint de valeurs {-2;-1;-0.5;0;-.5;1;2;3}. Une variable explicative continue peut être modélisée par un polynôme fraction-

naire dans un modèle de régression.

Comme utilisées dans d'autres travaux [128], deux fonctions de liens ont été testées pour modéliser la probabilité d'être séropositif au VHC : la fonction *logit*, $(log(x/(1-x)))$ et la fonction *cloglog*, $(log(-log(1-x)))$.

Avec le lien *logit*, la prévalence en fonction de l'âge et du temps s'écrit :

$$\begin{aligned} & \log\left[\frac{P(Y=1 \mid a,t)}{1-P(Y=1 \mid a,t)}\right] = \alpha + \eta(a) + ct \\ \Rightarrow \quad & \frac{P(Y=1 \mid a,t)}{1-P(Y=1 \mid a,t)} = \exp(\alpha + \eta(a) + ct) \\ \Rightarrow \quad & P(Y=1 \mid a,t) = \frac{\exp(\alpha + \eta(a) + ct)}{1+\exp(\alpha + \eta(a) + ct)} \end{aligned} \tag{4.4}$$

Avec le lien *cloglog*, la prévalence en fonction de l'âge et du temps s'écrit :

$$\begin{aligned} & log(-log(1-P(Y=1 \mid a,t))) = \alpha + \eta(a) + ct \\ \Rightarrow \quad & log(1-P(Y=1 \mid a,t)) = -\exp[\alpha + \eta(a) + ct] \\ \Rightarrow \quad & 1-P(Y=1 \mid a,t) = \exp[-\exp(\alpha + \eta(a) + ct)] \\ \Rightarrow \quad & P(Y=1 \mid a,t) = 1-\exp[-\exp(\alpha + \eta(a) + ct)] \end{aligned} \tag{4.5}$$

Le critère d'information d'Akaike (AIC) permet de sélectionner le meilleur modèle (*i.e.* celui ayant l'AIC le plus faible entre le modèle avec un lien *logit* ou celui avec un lien *cloglog*).

Il est relativement simple d'inclure ensuite d'autres covariables afin de prendre en compte d'autres caractéristiques d'intérêt comme le statut VIH , les pratiques d'injections ou la consommation de crack. Rappelons que toutes les drogues n'exposent pas de la même manière à la transmission du VHC car c'est le mode de consommation (injection, snif ou voie fumée par des objets cassables telles que les pipes à crack en verre) qui exposent.

4.2.3 Estimation de l'incidence en fonction de l'âge et du temps à partir de la relation prévalence/incidence

La transmission du VHC dans une population peut se modéliser (en simplifiant la réalité) par un modèle compartimental à deux états [27] : anti-VHC négatif (N) et anti-VHC positif (P), comme l'illustre la Figure 4.1. Suite à une exposition, des personnes dans l'état N peuvent s'infecter et passer dans l'état P à un taux $\lambda(a,t)$ dépendant de l'âge a et du temps t. Ici, $\lambda(a,t)$ est l'incidence de l'infection au VHC. Notons $N(a,t)$ la proportion de personnes anti-VHC négatives d'âge a au temps t et $P(a,t)$ la proportion de personnes anti-VHC positives (*i.e.* la prévalence des anti-VHC) d'âge a au temps t. Très rarement, des personnes anti-VHC positives peuvent redevenir anti-VHC négatives à un taux γ ; c'est la séroréversion [85]. La proportion de personnes nouvellement exposées (proportion de nouveaux usagers de drogues) intégrant le modèle est noté β. A l'inverse, des personnes peuvent sortir du modèle de transmission, avec un taux de mortalité μ_1 (respectivement μ_2) si elles se trouvent dans l'état N (respectivement P).

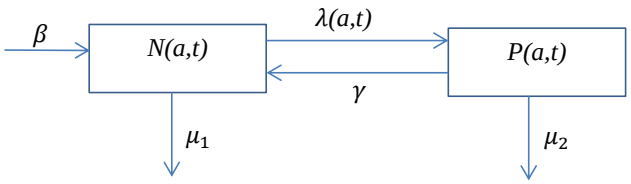


FIGURE 4.1 – Modèle compartimental à 2 états pour la transmission de l'hépatite C. β est la proportion de nouveaux usagers de drogues ; γ est le taux de séroréversion (défini par l'absence d'anticorps anti-VHC chez un individu connu auparavant anti-VHC positif) ; μ_1 est le taux de mortalité toutes causes chez les personnes anti-VHC négatives (hors infection VHC) ; μ_2 est le taux de mortalité toutes causes chez les personnes anti-VHC positives ($\mu_2 = \mu_1 + \mu_{VHC}$ où μ_{VHC} est le taux de mortalité lié au VHC) ; λ est l'incidence.

TABLE 4.2 – Paramètres annuels dans le modèle compartimental à deux états

Paramètre	Valeur	Référence
β : proportion de nouveaux usagers de drogues	2%	ANRS-Coquelicot [67, 68]
γ : Taux de séroréversion* VHC	0.001	Le Page, 2013 [85]
μ_1 : Taux de mortalité toutes causes chez les personnes anti-VHC négatives	0.7%	Smit, 2008 [129]
μ_2 : Taux de mortalité toutes causes chez les personnes anti-VHC positives	1.3%	Smit, 2008 [129]

* défini par l'absence d'anticorps anti-VHC chez un individu connu auparavant anti-VHC positif.

Le modèle de transmission du VHC s'exprime alors sous forme d'équations différentielles par le système suivant :

$$\begin{cases} \dfrac{dN(a,t)}{d(a,t)} = -\lambda(a,t)N(a,t) - \mu_1 N(a,t) + \beta N(a,t) + \gamma P(a,t) \\ \\ \dfrac{dP(a,t)}{d(a,t)} = \lambda(a,t)N(a,t) - \mu_2 P(a,t) - \gamma P(a,t) \end{cases} \tag{4.6}$$

Les paramètres introduits dans le système d'équation sont présentés dans le Tableau 4.2 et la Figure 4.1.

L'incidence de l'infection à VHC, $\lambda(a,t)$, est donc exprimée à partir de la dérivée de la fonction de prévalence. En effet, en dérivant le système 4.6 selon l'âge a et le temps t, on obtient les dérivées partielles suivantes :

$$\begin{cases} \dfrac{\partial}{\partial a}N(a,t) + \dfrac{\partial}{\partial t}N(a,t) = -\lambda(a,t)N(a,t) - \mu_1 N(a,t) + \beta N(a,t) + \gamma P(a,t) \\ \\ \dfrac{\partial}{\partial a}P(a,t) + \dfrac{\partial}{\partial t}P(a,t) = \lambda(a,t)N(a,t) - (\mu_2 + \gamma)P(a,t) \end{cases} \tag{4.7}$$

Par définition, $N(a,t) + P(a,t) = 1$. Nous en déduisons l'incidence à partir de l'équation 4.7 par :

$$\lambda(a,t) = \left(\frac{\partial}{\partial a}P(a,t) + \frac{\partial}{\partial t}P(a,t) + (\beta - \mu_1) - (\beta - \mu_1 - \gamma)P(a,t) \right) / (1 - P(a,t)) \tag{4.8}$$

où $P(a,t)$ représente la prévalence estimée dans la section 4.2.2 (équations 4.4 ou 4.5). Nous pouvons donc remplacer $P(a,t)$ par $P(Y = 1|a,t)$.

Ainsi, avec le lien *logit*, la dérivée de la prévalence par âge a, à un temps t donné, peut être déduite de l'équation 4.4 :

$$
\begin{aligned}
& \frac{\partial}{\partial a} P(Y = 1 \mid a, t) \\
= \; & \frac{\partial(\exp(\alpha + \eta(a) + ct) \times (1 - \exp(\alpha + \eta(a) + ct)) - \exp(\alpha + \eta(a) + ct) \times \partial(1 - \exp(\alpha + \eta(a) + ct))}{[1 + \exp(\alpha + \eta(a) + ct)]^2} \\
= \; & \frac{\eta^{'}(a) \exp(\alpha + \eta(a) + ct) \times (1 - \exp(\alpha + \eta(a) + ct)) - \exp(\alpha + \eta(a) + ct) \times [-\eta^{'}(a) \exp(\alpha + \eta(a) + ct)]}{[1 + \exp(\alpha + \eta(a) + ct)]^2} \\
= \; & \frac{\eta^{'}(a) \exp(\alpha + \eta(a) + ct)}{[1 + \exp(\alpha + \eta(a) + ct)]^2}
\end{aligned} \tag{4.9}
$$

où $\eta^{'}(a)$ est la dérivée première du polynôme fractionnaire $\eta(a)$.

A partir des équations 4.8 et 4.9, on obtient l'incidence en fonction de l'âge, $\lambda(a)$, pour un temps t donné :

$$
\lambda(a|t) = \frac{\eta^{'}(a) \times p(a|t) \times (1 - p(a|t)) + (\beta - \mu_1) - (\beta - \mu_1 - \gamma) \times p(a|t)}{1 - p(a|t)} \tag{4.10}
$$

où $p(a, t) = P(Y = 1|a, t)$ est la prévalence estimée pour l'âge a calculée à un temps t donné.

En utilisant un lien *cloglog*, la dérivée de la prévalence par âge a, à un temps t donné, peut être déduite de l'équation 4.5 :

$$
\begin{aligned}
& \frac{\partial}{\partial a} P(Y = 1|a, t) \\
= \; & -\frac{\partial}{\partial a}(-\exp(\alpha + \eta(a) + ct)) \times \exp[-\exp(\alpha + \eta(a) + ct)] \\
= \; & \frac{\partial}{\partial a}(\alpha + \eta(a) + ct) \times \exp(\alpha + \eta(a) + ct) \times \exp[-\exp(\alpha + \eta(a) + ct)] \\
= \; & \eta^{'}(a) \times \exp(\alpha + \eta(a) + ct) \times \exp[-\exp(\alpha + \eta(a) + ct)]
\end{aligned} \tag{4.11}
$$

A un temps t donné, l'expression de $\lambda(a)$, à partir des équations 4.8 et 4.11, est alors donnée par :

$$
\lambda(a|t) = \frac{-\eta^{'}(a) \log(1 - p(a|t)) \times (1 - p(a|t)) + (\beta - \mu_1) - (\beta - \mu_1 - \gamma) \times p(a|t)}{1 - p(a|t)} \tag{4.12}
$$

Estimation de l'incidence globale

En utilisant les équations 4.10 ou 4.12 selon le choix de la fonction de lien, et la proportion estimée d'UD par âge dans les 2 enquêtes, nous avons estimé l'incidence de l'infection à VHC en 2004 et en 2011 chez les 18-55 ans.

La proportion d'UD d'âge a au temps t, notée $q(a|t)$, a été estimée en utilisant l'estimateur d'Horvitz-Thompson,

$$\hat{q}(a|t) = \frac{\sum_i^n w_i x_i(a,t)}{\sum_i^n w_i}$$

où w_i est le poids de sondage de l'individu i, $x_i(a,t) = 1$ si l'individu est d'âge a au temps t, 0 sinon et n est la taille de l'échantillon [62].

Pour une enquête donnée, l'incidence peut donc être définie par la moyenne arithmétique pondérée des incidences par âge :

$$\lambda(t) = \sum_a \hat{q}(a|t)\lambda(a|t).$$

Estimation de la variance

Les variances des estimations ont été estimées par bootstrap. Pour cela, nous avons généré 2000 échantillons à partir de la combinaison des deux enquêtes :

1. Les individus ont été dupliqués selon leur poids de sondage
2. Un nombre aléatoire a été attribué à chaque individu
3. Les individus ont été ordonnés suivant l'année d'enquête et leur nombre attribué à l'étape précédente
4. Les n premiers individus ont été sélectionnés pour constituer un échantillon aléatoire ($n = 813$ pour 2004 et $n = 1242$ pour 2011).

Nous avons donc généré 2000 combinaisons d'échantillons de 2004 et 2011. Pour chaque combinaison, la prévalence a été calculée en tenant compte du plan de sondage. Pour estimer l'incidence, chacun des paramètres β, γ et μ_1 présentés précédemment, a d'abord été tiré aléatoirement

dans une distribution (Tableau 4.3 et Figure 4.2) pour chaque combinaison. Puis, l'incidence a été calculée à partir l'équation mathématique modélisant la relation entre la prévalence et l'incidence.

Les bornes des intervalles de confiance à 95% ont été obtenus avec les 2.5 ième et 97.5 ième percentiles de la distribution des prévalences et des incidences estimées pour les 2000 échantillons générés.

TABLE 4.3 – Distribution des paramètres

Paramètre	**Valeur moyenne**	**Distribution**
β	0.02	Distribution Uniforme $\sim U(0.01, 0.03)$
γ	0.001	Distribution binomiale $\sim B(2063, 0.001)/2063$
μ_1	0.007	Distribution de Poisson $\sim P(7)$ divisé par 1000

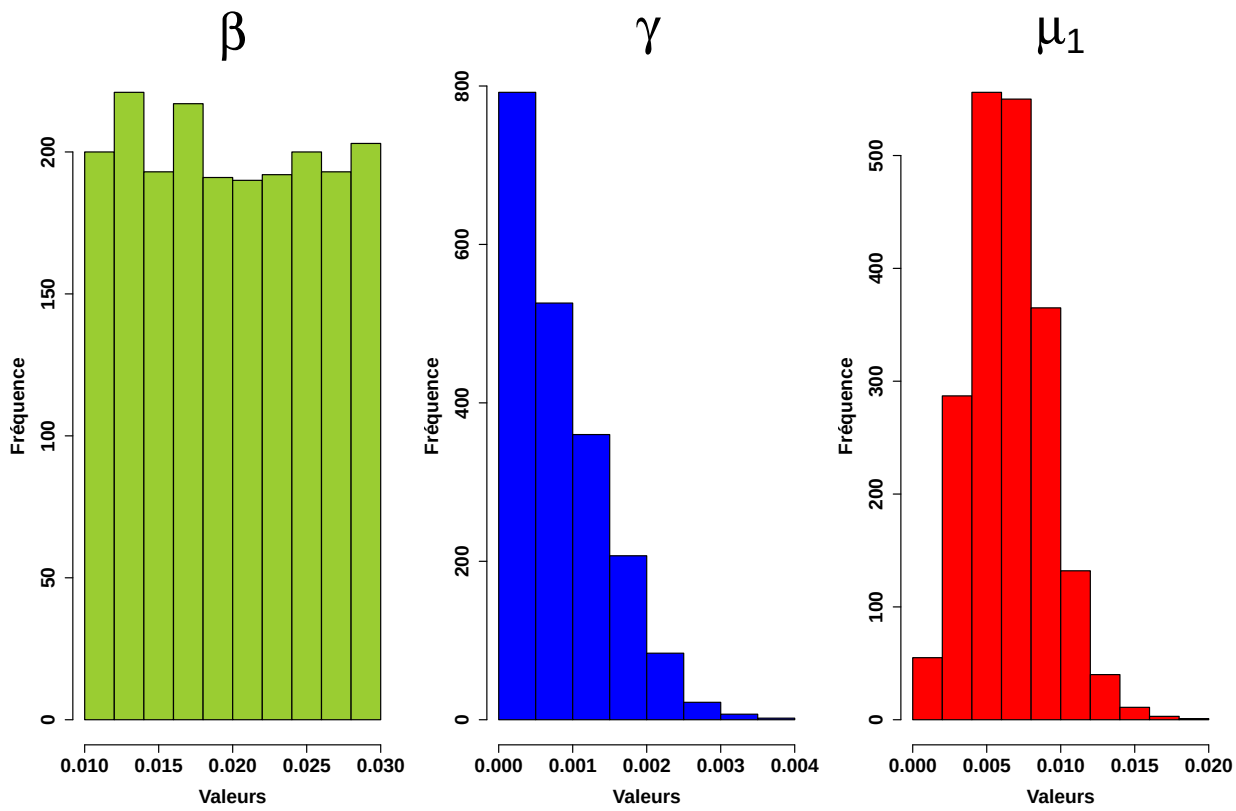


FIGURE 4.2 – Distribution des paramètres. β est la proportion de nouveaux usagers de drogues, γ est le taux de séroréversion du VHC et μ_1 est le taux de mortalité.

Les résultats de cette approche sont présentés dans la section suivante.

4.3 Approche par modèle mathématique : estimation de l'incidence de l'infection par le virus de l'hépatite C à partir des deux enquêtes ANRS-Coquelicot 2004 et 2011

La fusion des deux enquêtes a permis d'inclure 813 UD en 2004 et 1242 UD en 2011 (Tableau 4.4). Ces effectifs sont différents de ceux présentés dans les articles déjà publiés [67, 69] car les UD enquêtés auprès des cabinets de médecins généralistes en 2004 et ceux enquêtés dans les deux départements (Seine-et-Marne et Seine-Saint-Denis) en 2011 ont été exclus. De plus, le faible nombre d'UD retenus en 2004 s'explique par une proportion plus importante de DBS inexploitables par manque de matériel (i.e. une quantité de sang séché recueillie trop faible pour être analysée).

TABLE 4.4 – Statistiques descriptives des participants à l'enquête ANRS-Coquelicot, 2004 et 2011, France.

Participants	**Année**					
	2004 (N = 813)			**2011 (N = 1242)**		
	Sans poids de sondage (%)	**Avec poids de sondage (%)**	**95%IC**	**Sans poids de sondage (%)**	**Avec poids de sondage (%)**	**95%IC**
Age						
18-19 ans	0.9	0.4	0.1 - 1.1	0.9	0.6	0.3 - 1.4
20-25 ans	11.6	6.9	4.7 - 10.2	10.3	7.6	5.9 - 9.9
26-35 ans	43.7	48.5	39.9 - 57.2	25.9	26.1	22.6 - 29.9
36-45 ans	39.0	39.7	32.3 - 47.6	40.7	41.9	38.1 - 45.8
46-55 ans	4.7	4.2	2.6 - 6.5	19.9	21.3	17.5 - 25.7
56 ans et plus	0.2	0.3	0.00 - 1.9	2.3	2.5	1.6 - 4.0
Hommes	77.0	72.2	64.3 - 80.0	77.9	79.5	76.0 - 82.9
Rapportant l'usage de drogues par injection	73.3	72.0	63.4 - 80.7	63.9	65.7	61.7 - 69.7
Rapportant l'usage de drogues par injection dans le dernier mois précédant l'enquête	39.0	44.1	33.7 - 54.6	32.2	36.1	30.2 - 42.1
Consommateur de crack dans le dernier mois précédant l'enquête	24.6	41.0	30.3 - 51.6	27.8	34.4	29.7 - 39.1
Prevalence du VIH	10.1	10.8	5.4 - 16.2	8.0	9.4	6.8 - 12.0
Prévalence du VHC (approche biologique)	54.1	58.9	50.4 - 67.4	39.1	43.4	39.0 - 47.9
Prévalence du VHC (approche par mélange de lois)	52.5	58.2	49.7 - 66.7	39.2	43.2	38.9 - 47.7

La majorité des participants étaient des hommes (environ 77%), âgés de 26 à 45 ans (83% en 2004 et 67% en 2011). A partir des deux enquêtes, nous avons estimé que la plupart des UD reportaient un usage de drogues par injection (72.0%, $IC95\%, 63.4 - 80.7$ en 2004 et 65.7%, $IC95\%, 61.7 - 69.7$ en 2011).

La prévalence du VIH parmi les UD a été estimée à 10.8% ($IC95\%, 5.4 - 16.2$) en 2004 et 9.4% ($IC95\%, 6.8 - 12.0$) en 2011 (Tableau 4.4).

A partir du classement anti-VHC positif/négatif par la méthode biologique (section 2.4.1), nous avons estimé la prévalence du VHC parmi les UD à 58.9% ($IC95\%, 50.4?67.4$) en 2004 et à 43.4% ($IC95\%, 39.0 - 47.9$) en 2011. A partir du classement anti-VHC positif/négatif par mélange de lois (section 2.4.2), nous avons estimé la prévalence du VHC parmi les UD à 58.2% ($IC95\%, 49.7 - 66.8$) en 2004 et à 43.2% ($IC95\%$, $38.8 - 47.7$) en 2011.

Prévalence du virus de l'hépatite C par âge et par année

Différents modèles de régression ont été construits pour estimer la prévalence du VHC en fonction de l'âge et du temps, en ajustant ensuite sur le fait de s'injecter des drogues (oui/non), d'être séropositif au VIH ou encore de consommer du crack dans le dernier mois précédant l'enquête (oui/non). (Tableau 4.5). L'objectif était de fournir des estimations dans chaque sous-population. Pour l'ensemble des régressions, le modèle avec un lien *logit* a été retenu selon le critère AIC le plus faible dans la modélisation de la prévalence.

Pour chaque modèle de régression (excepté le modèle 2), l'âge et le temps étaient associés à l'infection au VHC ($p < 0.05$).

TABLE 4.5 – Modèles de régression logistique pour estimer la prévalence anti-VHC chez les usagers de drogues en France, 2004 et 2011, ANRS-Coquelicot

Variable	**Polynôme fractionnaire** $\eta(a)$	**Coefficient (écart-type)**	**p-Value**	**Intervalle de confiance à 95%**	**Odds ratio** (IC 95%)
Modèle 1					
Age	$(age/10)^{-1} - 0.27$	-18.15 (2.59)	< 0.001	[-23.25 , -13.06]	
	$(age/10)^3 - 49.14$	-0.008 (0.004)	0.023	[-0.016 , -0.001]	
Temps (ref : 2004)		-0.15 (0.03)	< 0.001	[-0.21 , -0.09]	0.86 (0.81 - 0.92)
Constante		0.81 (0.20)	< 0.001	[0.42 , 1.20]	
Modèle 2 (UDI actifs*)					
Age	$(age/10)^{-2} - 0.08$	-4.05 (22.53)	0.857	[-48.27 , 40.18]	
	$(age/10)^{-2} ln(age/10) - 0.10$	-44.66 (36.46)	0.221	[-116.24 , 26.91]	
Temps (ref : 2004)		-0.10 (0.06)	0.091	[-0.21 , -0.02]	0.91 (0.81 - 1.02)
Constante		1.52 (0.38)	< 0.001	[0.78 , 2.27]	
Modèle 3					
Age	$ln(age/10) - 1.30$	17.49 (4.12)	< 0.001	[9.38 , 25.60]	
	$ln(age/10)^2 - 1.68$	-5.22 (1.54)	0.001	[-8.28 , -2.20]	
Temps (ref : 2004)		-0.16 (0.04)	< 0.001	[-0.23 , -0.09]	0.85 (0.79 - 0.92)
Injection (ref : non)		2.88 (0.29)	< 0.001	[2.30 , 3.45]	17.73 (10.02 -31.36)
Constante		-1.22 (0.33)	< 0.001	[-1.87 , -0.58]	
Modèle 4					
Age	$(age/10^{-1} - 0.27$	-18.14 (2.57)	< 0.001	[-23.18 , -13.09]	
	$(age/10)^3 - 49.14$	-0.008 (0.004)	0.031	[-0.015 , -0.000]	
Temps (ref : 2004)		-0.15 (0.03)	< 0.001	[-0.21 , -0.09]	0.86 (0.81 - 0.91)
Crack (ref : non)		0.26 (0.19)	0.159	[-0.10 , 0.63]	1.30 (0.90 - 1.89)
Constante		0.71 (0.18)	< 0.001	[0.35 , 1.06]	
Modèle 5					
Age	$(age/10^{-1} - 0.27$	-17.60 (2.59)	< 0.001	[-22.68 , -12.52]	
	$(age/10^3 - 50.14$	-0.009 (0.004)	0.010	[-0.016 , -0.002]	
Temps (ref : 2004)		-0.16 (0.04)	< 0.001	[-0.24 , -0.09]	0.85 (0.79 - 0.91)
VIH (ref : séronégatif)		1.68 (0.31)	< 0.001	[1.07 , 2.29]	5.35 (2.90 - 9.86)
Constante		0.82 (0.26)	0.002	[0.31 , 1.32]	

Par exemple, pour le modèle 1 : $logit(P(Y = 1|a, t) = -18.15[(a/10)^{-1} - 0.27] - 0.01[(a/10)^3 - 49.14] - 0.15t + 0.81$,
pour l'âge a au temps t.
*UDI actifs : Usagers de drogues injecteurs actifs

Pour les Figures 4.3, 4.4, 4.5 et 4.7, les graphiques de gauche représentent les prévalences selon l'âge pour les deux enquêtes 2004 et 2011, estimées en tenant compte du plan de sondage, à partir des modèles de régression (les courbes) et estimées ponctuellement (les points).

Chez les UD, la prévalence a augmenté avec l'âge de façon monotone jusqu'à un plateau autour de 50 ans. Selon l'année d'enquête, une décroissance marquée de la prévalence par âge a été également observée (Figure 4.3).

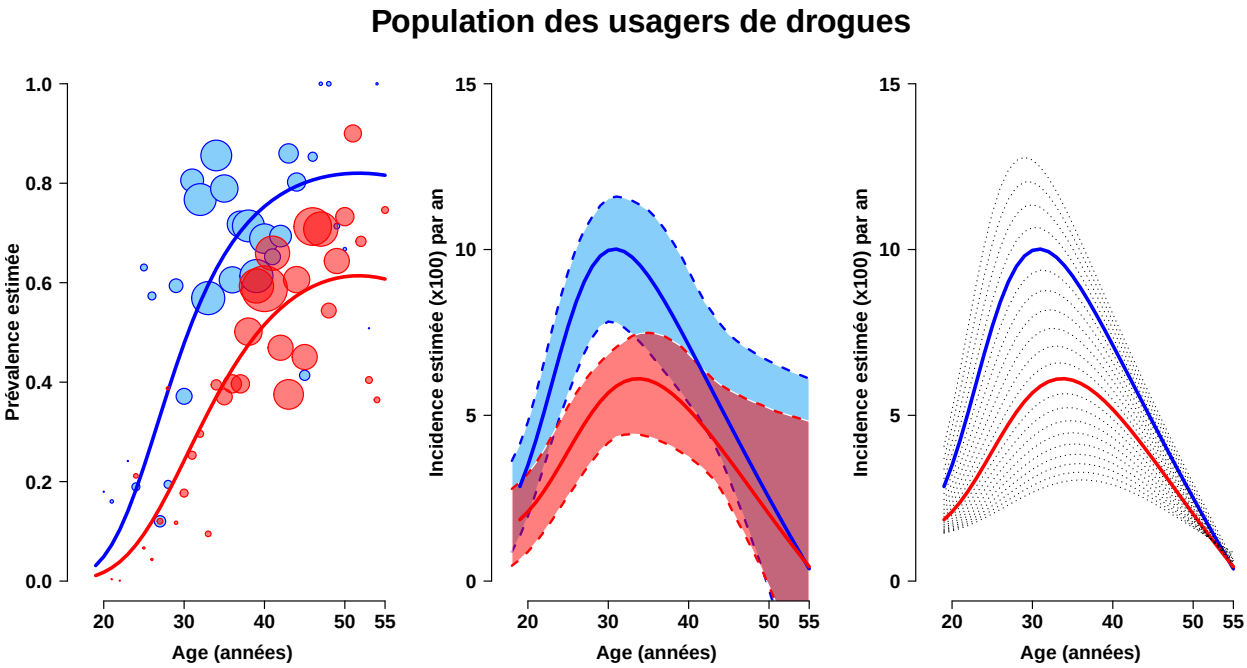


FIGURE 4.3 – (A gauche) Les courbes représentent la prévalence du VHC en fonction de l'âge, estimée à partir de modèles de régression logistique parmi les usagers de drogues en 2004 (courbe bleue) et en 2011 (courbe rouge). Les cercles représentent les prévalences estimées par âge. La taille des cercles est proportionnelle au nombre de personnes enquêtées en 2004 (cercles bleus) et en 2011 (cercles rouges).
(Au milieu) Les courbes représentent les incidences du VHC estimées en fonction de l'âge parmi les usagers de drogues en 2004 (courbe bleue) et en 2011 (courbe rouge) encadrées de leurs intervalles de confiance (courbes en pointillé).
(A droite) Estimations des incidences du VHC en fonction de l'âge parmi les usagers de drogues en 2000-2020. Les courbes sont obtenues à partir du modèle de régression, en 2004 (courbe bleue), en 2011 (courbe rouge) et les autres années (courbes en pointillé).

Parmi les UDI, la prévalence par âge et au cours du temps était plus élevée par rapport à la prévalence estimée chez les UD ne reportant aucun usage de drogues par injection au cours de leur vie ($OR = 17.7, IC95\%, 10.0 - 31.4$, Tableau 4.5, modèle 3 et Figure 4.4).

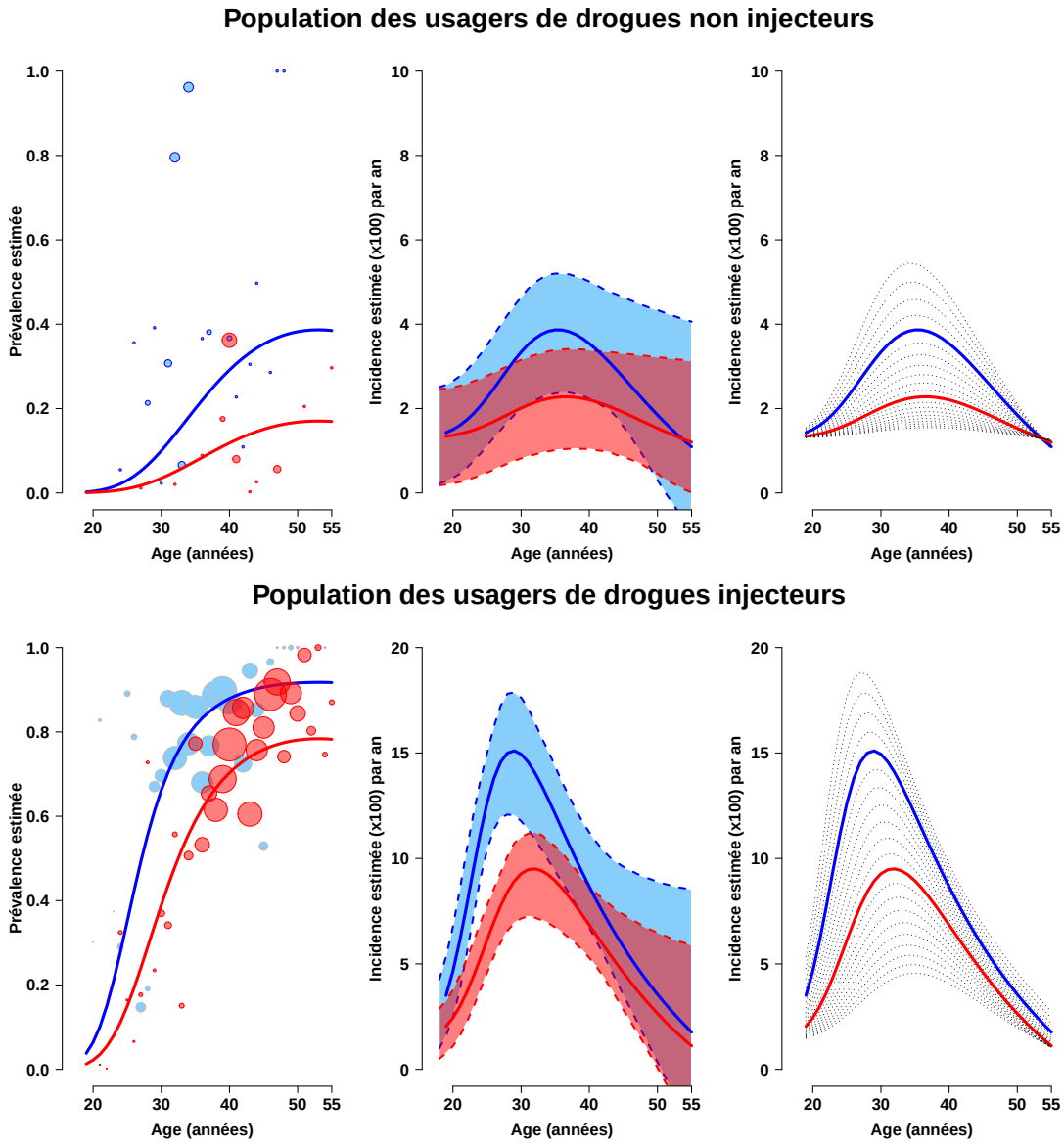


FIGURE 4.4 – (A gauche) Les courbes représentent la prévalence du VHC en fonction de l'âge, estimée à partir de modèles de régression logistique en 2004 (courbe bleue) et en 2011 (courbe rouge) parmi les usagers de drogues non injecteurs (en haut) et les usagers de drogues injecteurs (en bas). Les cercles représentent les prévalences estimées par âge. La taille des cercles est proportionnelle au nombre de personnes enquêtées en 2004 (cercles bleus) et en 2011 (cercles rouges).
(Au milieu) Les courbes représentent les incidences du VHC estimées en fonction de l'âge en 2004 (courbe bleue) et en 2011 (courbe rouge) encadrées de leurs intervalles de confiance (courbes en pointillé) parmi les usagers de drogues non injecteurs (en haut) et les usagers de drogues injecteurs (en bas).
(A droite) Estimations des incidences du VHC en fonction de l'âge en 2000-2020 parmi les usagers de drogues non injecteurs (en haut) et les usagers de drogues injecteurs (en bas). Les courbes sont obtenues à partir du modèle de régression, en 2004 (courbe bleue), en 2011 (courbe rouge) et les autres années (courbes en pointillé).

La plupart des UD étaient des UDI (68%, $IC95\%, 64.7 - 72.7$).

Parmi les UDI actifs, la prévalence a augmenté rapidement avec l'âge jusqu'à un plateau autour de 40 ans puis s'est stabilisée au delà de 80% (Figure 4.5). Les résultats du modèle de régression n'étaient pas significativement différents selon l'année d'enquête ($OR = 0.9, IC95\%, 0.8-1.0$, Tableau 4.5, modèle 2). L'allure des courbes le confirme. Aucune association significative avec l'âge et le temps n'a été trouvée (Tableau 4.5, modèle 2).

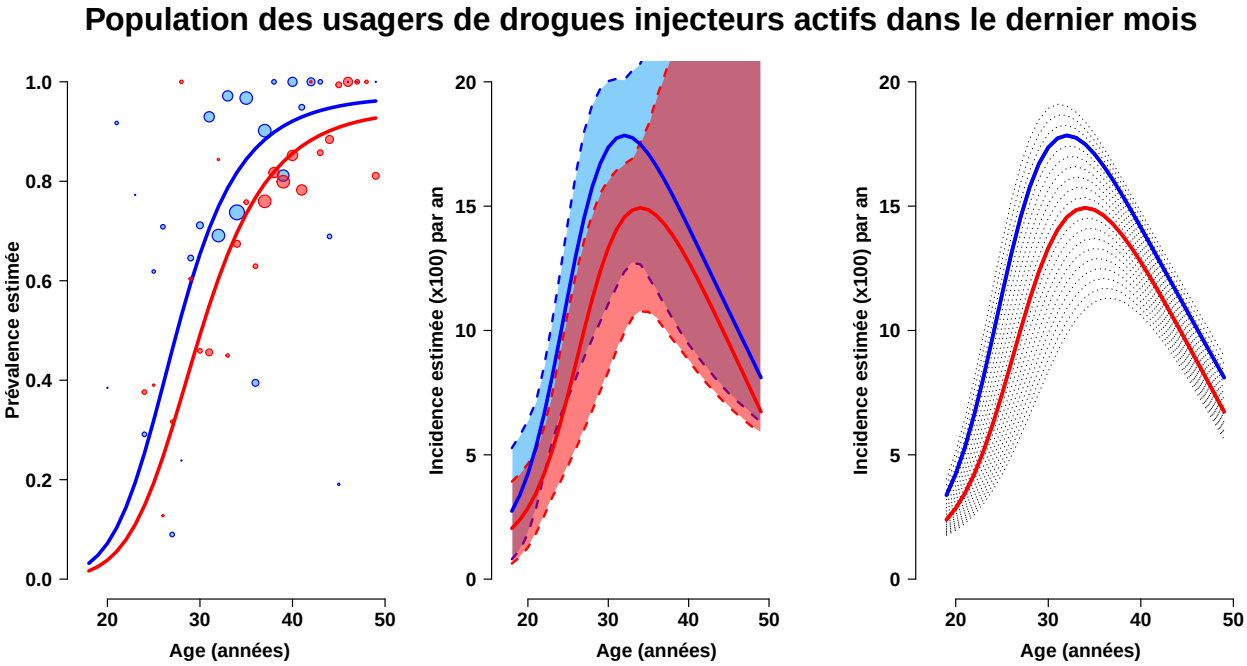


FIGURE 4.5 – (A gauche) Les courbes représentent la prévalence du VHC en fonction de l'âge, estimée à partir de modèles de régression logistique parmi les usagers de drogues injecteurs actifs en 2004 (courbe bleue) et en 2011 (courbe rouge). Les cercles représentent les prévalences estimées par âge. La taille des cercles est proportionnelle au nombre de personnes enquêtées en 2004 (cercles bleus) et en 2011 (cercles rouges).
(Au milieu) Les courbes représentent les incidences du VHC estimées en fonction de l'âge parmi les usagers de drogues injecteurs actifs en 2004 (courbe bleue) et en 2011 (courbe rouge) encadrées de leurs intervalles de confiance (courbes en pointillé).
(A droite) Estimations des incidences du VHC en fonction de l'âge parmi les usagers de drogues injecteurs actifs en 2000-2020. Les courbes sont obtenues à partir du modèle de régression, en 2004 (courbe bleue), en 2011 (courbe rouge) et les autres années (courbes en pointillé).

La Figure 4.6 représente la prévalence par âge, stratifiée sur la consommation de crack dans le dernier mois précédant l'enquête (oui/non), à partir des modèles de régression (les courbes) et estimées ponctuellement (les points). La plupart des UD n'étaient pas des consommateurs de

crack dans le mois précédant l'enquête (64.8%, $IC95\%, 58.3 - 71.3$). Entre 2004 et 2011, une diminution de la prévalence en fonction de l'âge a été observée mais aucune association n'était significative (Tableau 4.5, modèle 4).

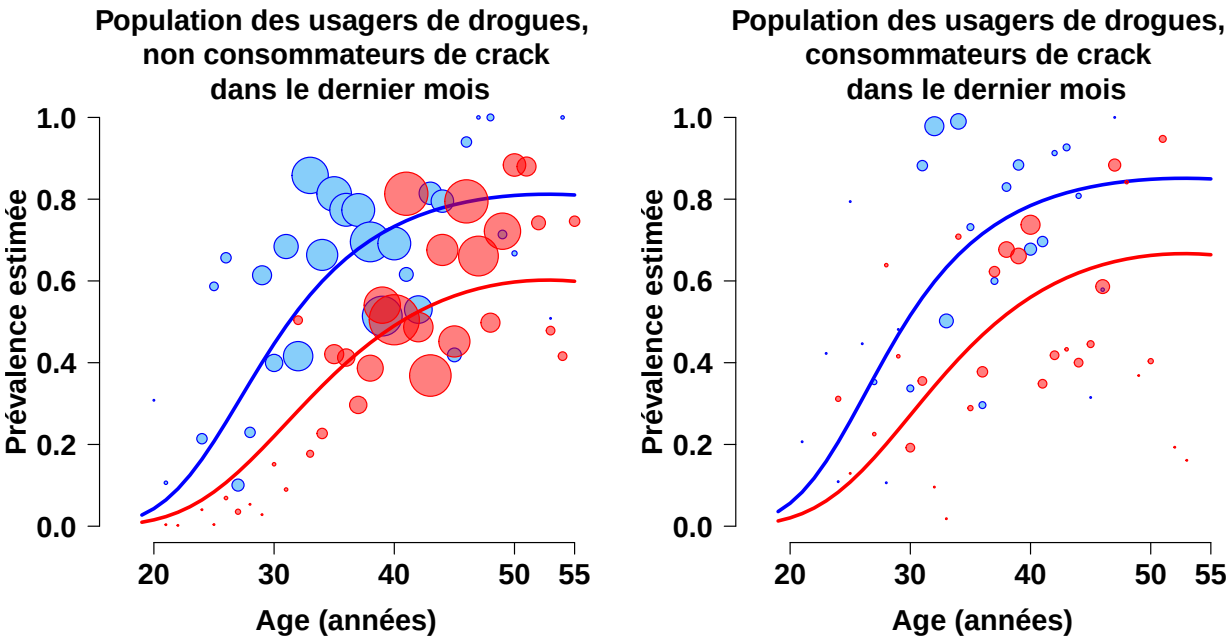


FIGURE 4.6 – (A gauche) Les courbes représentent la prévalence du VHC en fonction de l'âge, estimée à partir de modèles de régression logistique en 2004 (courbe bleue) et en 2011 (courbe rouge) parmi les usagers de drogues non consommateurs de crack dans le dernier mois (à gauche) et les usagers de drogues consommateurs de crack dans le dernier mois (à droite). Les cercles représentent les prévalences estimées par âge. La taille des cercles est proportionnelle au nombre de personnes enquêtées en 2004 (cercles bleus) et en 2011 (cercles rouges).

Enfin, la Figure 4.7 représente la prévalence par âge, stratifiée sur le statut VIH. La prévalence en fonction de l'âge et du temps était plus élevée chez les UD séropositifs au VIH, comparée à celle estimée chez les UD séronégatifs au VIH (OR=5.3, $IC95\%, 2.9 - 9.9$, Tableau 4.5, modèle 5), avec une prévalence chez les séropositifs au VIH supérieure à 80%.

Incidence de l'infection par le virus de l'hépatite C par âge et par année

Pour les Figures 4.3, 4.4, 4.5 et 4.7, les graphiques du milieu représentent les incidences selon l'âge pour les deux enquêtes 2004 et 2011, à savoir les taux de nouvelles personnes anti-VHC positives, encadrées de leurs intervalles de confiance et exprimées pour 100 personnes-années.

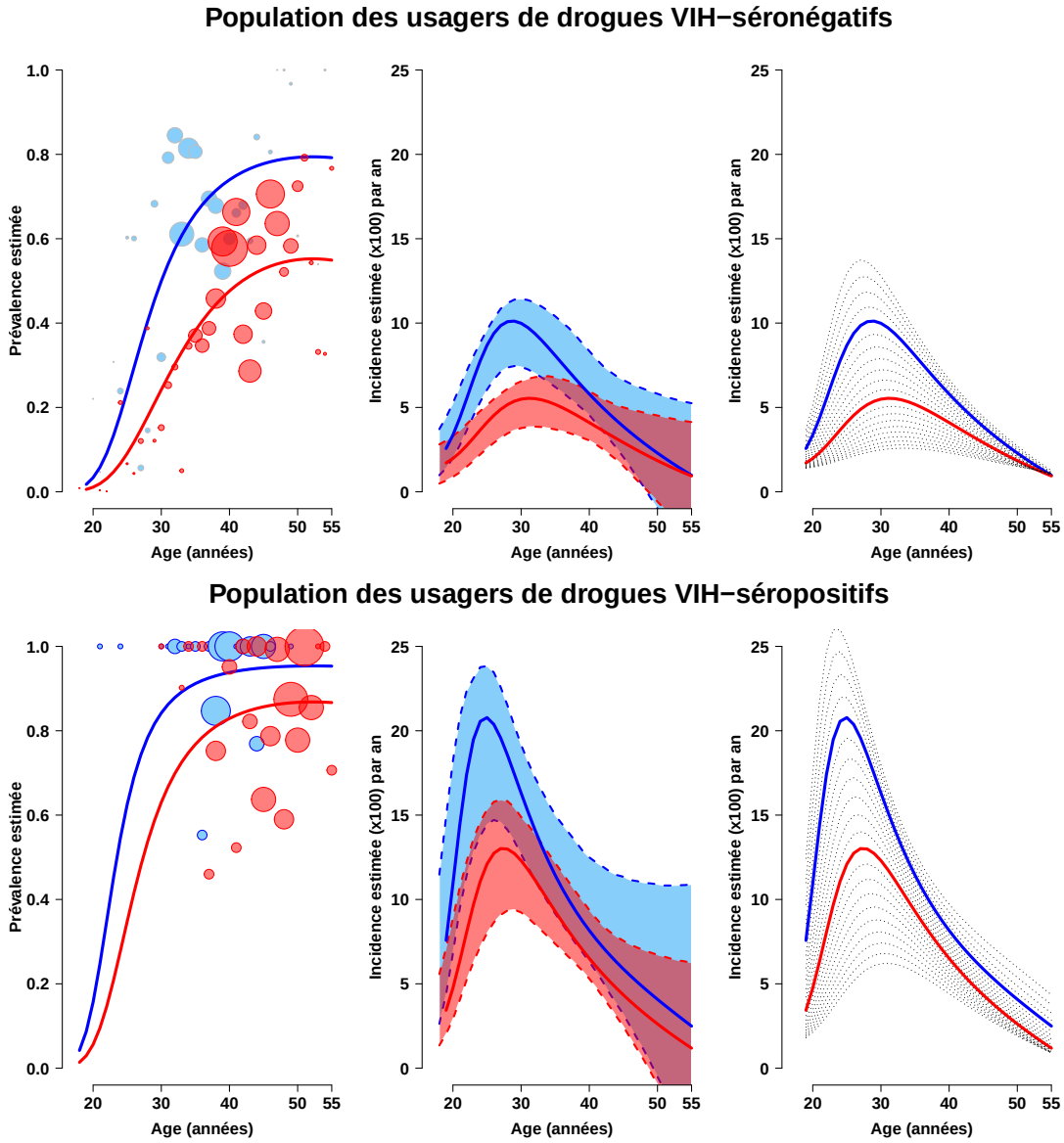


FIGURE 4.7 – (A gauche) Les courbes représentent la prévalence du VHC en fonction de l'âge, estimée à partir de modèles de régression logistique en 2004 (courbe bleue) et en 2011 (courbe rouge) parmi les usagers de drogues séronégatifs au VIH (en haut) et les usagers de drogues séropositifs au VIH (en bas). Les cercles représentent les prévalences estimées par âge. La taille des cercles est proportionnelle au nombre de personnes enquêtées en 2004 (cercles bleus) et en 2011 (cercles rouges).
(Au milieu) Les courbes représentent les incidences du VHC estimées en fonction de l'âge en 2004 (courbe bleue) et en 2011 (courbe rouge) encadrées de leurs intervalles de confiance (courbes en pointillé) parmi les usagers de drogues séronégatifs au VIH (en haut) et les usagers de drogues séropositifs au VIH (en bas).
(A droite) Estimations des incidences du VHC en fonction de l'âge en 2000-2020 parmi les usagers de drogues séronégatifs au VIH (en haut) et les usagers de drogues séropositifs au VIH (en bas). Les courbes sont obtenues à partir du modèle de régression, en 2004 (courbe bleue), en 2011 (courbe rouge) et les autres années (courbes en pointillé).

Les graphiques de droite représentent les incidences selon l'âge entre 2000 et 2020.

L'incidence a baissé au cours du temps. Elle a augmenté jusqu'à un âge donné (jusqu'à 34 ans en 2011 parmi l'ensemble des UD) avant de diminuer (Figures 4.3, 4.4, 4.5 et 4.7).

Parmi les UD, l'incidence la plus élevée a été estimée à 10.0/100 ($IC95\%, 7.8 - 11.6$) chez les personnes âgées de 31 ans en 2004 et à 6.1/100 ($IC95\%, 4.4 - 7.5$) chez les personnes âgées de 34 ans en 2011 (Figure 4.3, milieu).

Comparés aux UDI, l'incidence de l'infection à VHC chez les non UDI a diminué rapidement au cours du temps (Figure 4.4). Parmi les non UDI, l'incidence la plus élevée a été estimée à 3.9/100 ($IC95\%, 2.4 - 5.2$) chez ceux âgés de 35 ans en 2004 et à 2.3/100 ($IC95\%, 1.0 - 3.4$) chez ceux âgés de 37 ans en 2011. L'incidence de l'infection à VHC était toujours plus élevée chez les UDI comparés aux non UDI, la plus élevée chez ceux âgés de 29 ans, 15.1/100 ($IC95\%, 12.0 - 17.9$) en 2004 et chez ceux âgés de 32 ans, 9.5/100 ($IC95\%, 7.2 - 11.2$) en 2011 (Figure 4.4, milieu).

Les estimations obtenues dans la population des UDI actifs ont montré que la dynamique de l'infection VHC était relativement similaire entre 2004 et 2011, comme attendue d'après le modèle de régression (Figure 4.5 et Tableau 4.5, modèle 2).

Parmi les UD séropositifs au VIH, l'incidence par âge du VHC a augmenté jusqu'à un âge donné (25 ans en 2004 et 27 ans en 2011) avant de diminuer par la suite (Figure 4.7).

Incidence de l'infection par le virus de l'hépatite C en 2004 et en 2011 parmi les 18-55 ans

Nous avons estimé l'incidence de l'infection à VHC parmi les 18-55 ans compte tenu du faible nombre de participants plus âgés. Chez l'ensemble des UD, l'incidence de l'infection à VHC a baissé, passant de 7.9/100 ($IC95\%, 6.4 - 4.9$) en 2004 à 4.4/100 ($IC95\%, 3.3 - 5.9$) en 2011 (Tableau 4.6). Pour chaque sous-population étudiée, nous avons observé que l'incidence en 2011 était plus faible qu'en 2004. Toutefois, parmi les UDI actifs, l'incidence de l'infection à VHC était deux fois plus élevée par rapport aux autres UD, avec une incidence décroissant de 15.4/100 ($IC95\%, 11.9 - 19.3$) en 2004 à 11.2/100 ($IC95\%, 9.0 - 19.0$) en 2011.

TABLE 4.6 – Estimation de l'incidence de l'infection par le virus de l'hépatite C (pour 100 personnes-années) parmi les usagers de drogues, 18-55 ans, France, 2004 et 2011.

Populations	**Année 2004**			**Année 2011**		
	Taille d' échantillon	**Incidence pour 100 pa.**	**IC95%**	**Taille d' échantillon**	**Incidence pour 100 pa.**	**IC95%**
UD	811	7.9	6.4 - 9.4	1209	4.4	3.3 - 5.9
NUDI	216	3.1	1.9 - 4.5	434	2.0	0.9 - 3.2
UDI	594	10.8	9.0 - 12.8	775	6.1	5.0 - 8.0
UDI actifs (1 mois)	232	15.4	11.9 - 19.3	252	11.2	9.0 - 19.0
UD VIH-séronégatifs	753	7.4	5.8 - 8.9	1111	3.9	2.8 - 5.4
UD VIH-séropositifs	58	9.1	7.4 - 13.3	98	4.6	2.4 - 7.8

UD : usagers de drogues ; NUDI : usagers de drogues non injecteurs ; UDI : usagers de drogues injecteurs
pa. : personnes-années
Un UD n'a pas rapporté s'il était ou non un UDI. Les UDI actifs sont un sous-groupe des UDI.

A partir de la recherche de l'ARN du VHC réalisée en 2011, nous pouvions également utiliser l'approche biologique pour estimer l'incidence de l'infection à VHC chez les UD. Nous présentons cette seconde approche dans la section suivante.

4.4 Approche biologique : estimation de l'incidence de l'infection par le virus de l'hépatite C à partir de l'enquête ANRS-Coquelicot 2011

Cette approche a nécessité 3 étapes présentées dans les sections suivantes :

1. Estimation de la période fenêtre sur DBS,
2. Estimation du nombre de personnes ARN du VHC positives (ARN-VHC+) parmi les personnes anti-VHC négatives,
3. Estimation de l'incidence.

4.4.1 Mesure de la période fenêtre sur DBS

Les périodes fenêtres proposées dans la littérature ont été calculées sur des supports de prélèvements standards (sérum ou plasma) [20, 47, 113, 135]. En France, la période fenêtre utilisée est de 60 jours [13, 135]. La présence des anticorps anti-VHC est détectable au bout de 66 jours ($IC95\% : 38 - 94$) [13].

Nous nous sommes interrogés sur la modification de la période fenêtre avec l'utilisation d'un support de prélèvement DBS. Pour répondre à cette question, nous avons évalué dans un premier temps la perte de sensibilité des tests biologiques pour la détection de l'ARN du VHC et des anticorps anti-VHC avec l'utilisation du DBS. En, effet, une perte de sensibilité pourrait avoir un impact sur le délai de détection des deux marqueurs et donc sur le nombre de jours de la période fenêtre.

Cet élément ne remet pas en cause la fiabilité des tests biologiques réalisés sur DBS, discutée dans la section 1.3.3, mais uniquement la mesure de la période fenêtre. Par définition de la période fenêtre, la détection de l'ARN correspond au côté gauche de la fenêtre et la détection des anticorps correspond au côté droit (Figure 1.2).

Délai de détection de l'ARN du VHC (à gauche de la période fenêtre)

La différence moyenne du taux d'ARN du VHC détecté dans le sérum par rapport au DBS est de 1.60 log_{10} UI/mL $\pm$ 0.3 ou de 1.75 log_{10} UI/mL $\pm$ 0.3 selon la trousse commerciale testée et les recommandations du fabricant [131]. D'un autre côté, Tuaillon *et al.* ont montré une différence de charge virale (*i.e.* la quantification de l'ARN) du VHC de 2.27 log_{10} UI/mL $\pm$ 0.47 [148]. Nous avons alors fait l'hypothèse que la perte de sensibilité entre le sérum et le DBS était de l'ordre de 2 log_{10} UI/mL.

En partant de cette hypothèse et à partir de la modélisation de l'évolution de la charge virale du VHC, sous l'hypothèse d'un temps de doublement constant de 17 heures pendant la phase de montée de la charge virale [135], nous avons établi que la perte de sensibilité entre les deux supports (sérum *vs* DBS) correspond à un délai de 4 jours (Figure 4.8).

En France, la recherche de l'ARN du VHC doit être réalisée par une méthode sensible ayant un seuil de détection de l'ordre de 10-15 UI/mL, soit $\sim 1.04 - 1.16$ log_{10} UI/mL dans les conditions standards [8]. Ce seuil de détection apparaît généralement au septième jour (Figure 4.8). Sur DBS, le seuil de détection est donc estimé à $\sim 3.04 - 3.16$ log_{10} UI/mL soit $\sim 1100 - 1525$ UI/mL (correspondant à $\sim 2971 - 4118$ copies/mL) et l'ARN est détectable au onzième jour (voir Figure 4.8). Ceci nous conduit à un délai de $11 - 7 = 4$ jours pour la détection de l'ARN à partir d'un DBS par rapport au sérum.

Délai de détection des anticorps anti-VHC (à droite de la période fenêtre)

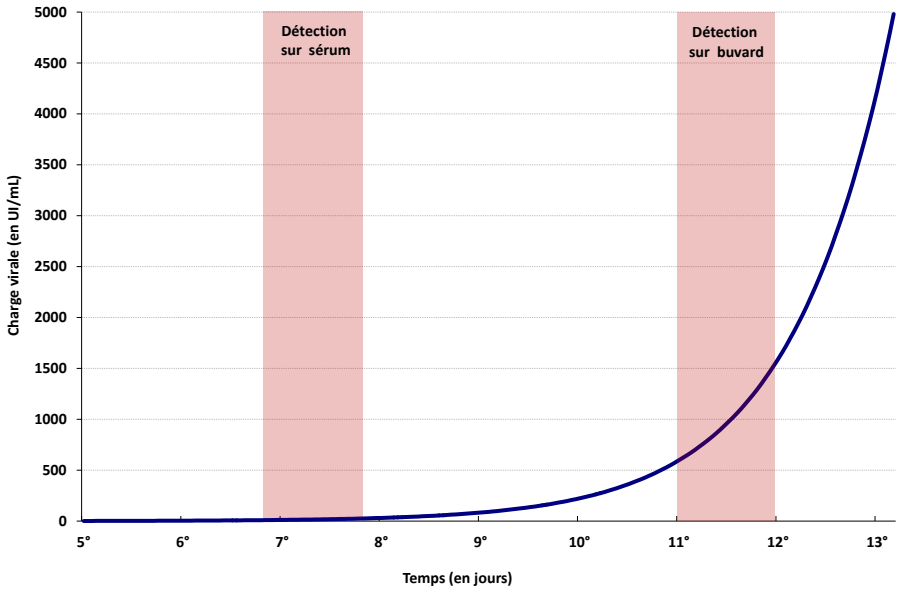


FIGURE 4.8 – Evolution de la charge virale du VHC sous l'hypothèse d'un temps de doublement constant de 17 heures pendant la phase de montée de la charge virale - modèle de Busch [135].

L'estimation du délai de détection des anticorps anti-VHC avec l'utilisation d'un test biologique sur DBS a été réalisée à l'aide de 14 panels d'échantillons de sérum, connus pour être anti-VHC positifs. Aucune différence significative n'a été observée entre le sérum et le DBS lors de la détection anti-VHC dans le sérum prélevé sur les séroconversions (test du log-rank, comparaison de deux courbes de survie). Toutefois, la baisse régulière des taux d'anticorps chez les patients guéris de leur infection par le VHC a conduit à une plus faible sensibilité pour détecter les anticorps anti-VHC sur le DBS à partir de ces patients. Par conséquent, la valeur de seuil pour les échantillons DBS a été abaissée à une valeur de 0,64 basée sur des données expérimentales, par rapport à la valeur unité lorsqu'on utilise le sérum ou le plasma, sans modifier le délai de détection des anti-VHC.

Ces différents éléments nous ont permis de fournir une estimation de la période fenêtre sur DBS de $60 - 4 = 56$ jours (Figure 4.9).

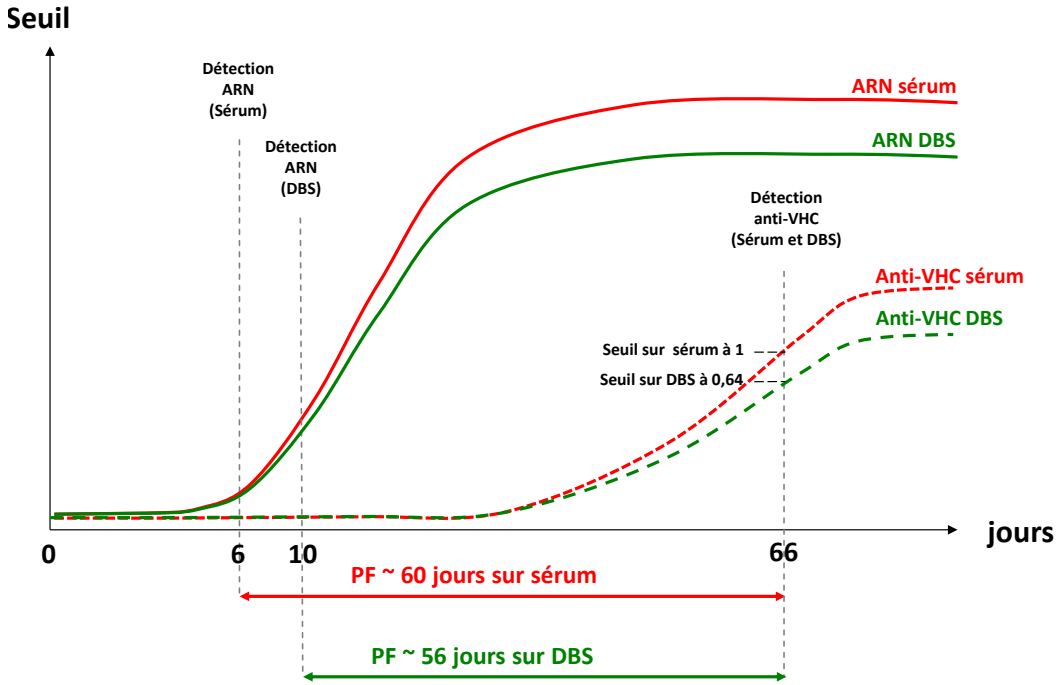


FIGURE 4.9 – Estimation de la période fenêtre selon le support de prélèvement (sérum ou DBS).

4.4.2 Estimation du nombre de personnes ARN du VHC positives parmi les personnes anti-VHC négatives

A partir de la recherche des anticorps anti-VHC et de l'abaissement du seuil de détection à 0.64, nous avons pu classer les individus en anti-VHC positif ou anti-VHC négatif. Il s'agit de l'approche classique des seuils, évoquée à la section 1.3.4 pour identifier les individus séropositifs et séronégatifs. Ainsi, 1568 UD ont été interrogés dont 37 questionnaires avec un volet biologique manquant. Sur 1531 participants à l'enquête avec un questionnaire complet, 92% avaient fourni des échantillons biologiques exploitables dont 562 ont été classés anti-VHC positifs et 856 anti-VHC négatifs. Parmi les 856 individus classés anti-VHC négatifs, 3 n'ont pas pu être testés pour la recherche de l'ARN du VHC par manque de matériel biologique (taille des gouttes de sang insuffisante).

Quatorze individus ont été classés en infection récente (i.e ARN-VHC+ et anti-VHC négatif) selon les critères suivants : (1) une charge virale comprise entre 1.6 et 2.5 log_{10} UI/mL et (2) un repassage pour la recherche de l'ARN dont le résultat était codé "OK", "QI" (quantité insuffisante) ou "génotypé". Neuf d'entre eux ont pu être génotypés. Aucun n'avait d'antécédent

de VHC ou n'avait jamais été testé. A partir de cet algorithme, 31 individus parmi les 853 anti-VHC négatifs n'ont pas pu être classés, dont 16 avec de l'ARN non répétable (matériel insuffisant) ; ils ont été considérés comme ARN du VHC négatifs.

En 2011, la proportion estimée du nombre d'UD en infection récente en tenant compte du plan de sondage était de 1.93% ($IC95\%, 0.55 - 3.32$). Chez les UD non injecteurs, elle a été estimée à 0.92% ($IC95\%, 0.00 - 1.90$) ; chez les UDI, elle a été estimée à 3.32% ($IC95\%, 0.35 - 6.29$) et chez les UDI actifs, elle a été estimée à 9.62% ($IC95\%, 1.14 - 17.80$).

4.4.3 Estimation de l'incidence

A partir des sections 4.4.1 et 4.4.2, et de l'équation 4.1, nous avons estimé l'incidence de l'infection à VHC parmi les UD anti-VHC négatifs (Tableau 4.7).

TABLE 4.7 – Estimation de l'incidence de l'infection par le virus de l'hépatite C (pour 100 personnes-années) parmi les usagers de drogues anti-VHC négatifs (approche biologique), France, 2011.

Populations	**Taille d'échantillon**	**Nombre de cas ANR+**	**Incidence pour 100 pa.** (PF = 56 jours)	**[IC95%]**	**Incidence pour 100 pa.** (PF = 46 jours)	**Incidence pour 100 pa.** (PF = 66 jours)
UD	853	14	12.6	[6.1 - 25.6]	15.3	10.7
NUDI	486	4	6.0	[2.1 - 17.1]	7.3	5.1
UDI	367	10	21.6	[8.7 - 52.0]	26.3	18.4
UDI actifs (1 mois)	117	9	62.7	[26.0 - 139.6]	76.3	53.2

UD : usagers de drogues ; NUDI : usagers de drogues non injecteurs ; UDI : usagers de drogues injecteurs
pa. : personnes-années ; PF : période fenêtre.

Chez les UD, l'incidence de l'infection à VHC a été estimée à 12.6/100 personnes-années avec l'hypothèse d'une période fenêtre de 56 jours. Elle était deux fois moins élevée chez les UD non injecteurs, 6.0/100 personnes-années. Chez les UDI, l'incidence de l'infection à VHC a été estimée à 21.6/100 personnes-années. Par contre, chez les UDI actifs, l'incidence estimée était très élevée, 62.7/100 personnes-années, presque 3 fois plus élevée que chez les UDI. Nous présentons également une analyse de sensibilité en faisant varier la période fenêtre de ± 10 jours [113] (Tableau 4.7, les deux dernières colonnes).

4.5 Discussion

Nous avons estimé l'incidence de l'infection à VHC chez les UD en France à partir de deux méthodes. Jusque là, la seule donnée publiée en France concernait une cohorte d'UDI suivie en 2000-2001 dans la région Nord-Est de la France (en excluant Paris et sa banlieue) avec une incidence estimée à 9/100 personnes-années [92].

A partir de la recherche des deux marqueurs biologiques du VHC, les UD ont été classés en infection récente ou non (Tableau 1.1). Puis, parmi les UD anti-VHC négatifs, ceux classés en équivoque ont été retestés pour la recherche de l'ARN du VHC et reclassés suivant un algorithme établi par avis d'experts (Section 2.4.1). Ici, une période fenêtre sur DBS de 56 jours a été estimée à partir de la période fenêtre sur sérum/plasma utilisée en France, d'une durée de 66 jours ($IC95\% : 38 - 94$) selon Barrera *et al.* [13]. Toutefois, plusieurs périodes fenêtres pour des supports biologiques sérum/plasma sont utilisées dans la littérature : 56 jours ($IC95\% : 45 - 68$) selon Glynn *et al.* [47], 51 jours ($IC95\% : 46 - 56$) selon Page-Shafer *et al.* [113] et 58 jours ($IC95\% : 46 - 76$) selon Brant *et al.* [19]. Ces deux termes ont une influence directe sur le calcul des estimations. Ainsi, l'incidence estimée chez les UD variait de 11 à 15 pour 100 personnes-années pour une période fenêtre passant de 66 à 46 jours. Chez les UDI, l'incidence de l'infection à VHC est estimée entre 18 et 26 pour 100 personnes-années pour une période fenêtre passant de 66 à 46 jours. Certains auteurs ont pointé la nécessité de disposer d'un large échantillon de personnes compte tenu de la courte durée de la période fenêtre par rapport à l'histoire naturelle du VHC à moins que l'incidence ne soit très élevée [30]. En effet, le faible effectif d'UDI actifs nous pousse à considérer les résultats d'estimations avec précaution. Page-Shaffer *et al.* ont notamment montré que l'estimation de l'incidence basée sur la méthode biologique (à partir de la recherche des deux marqueurs biologiques issus de données transversales) était deux fois plus élevée que l'incidence estimée à partir de cohortes (basée sur la recherche seule des anticorps anti-VHC), (53.7% *vs* 22.5%) [113]. Ces derniers ont également souligné que le délai de détection de l'ARN, variable d'un individu à l'autre, et l'histoire naturelle de l'infection durant la phase aiguë pouvaient être le résultat des différences entre les méthodes [113].

A partir de l'approche par modèle mathématique, nos estimations semblent cohérentes en se comparant à d'autres pays européens. En effet, en Angleterre, Pays de Galles et Irlande du Nord, l'incidence de l'infection à VHC chez les UDI a été estimée à 4-12 infections pour 100

personnes-années en 2011 [30] et entre 6 et 18 pour 100 personnes-années en 2013 [2], en utilisant un test d'avidité des anticorps anti-VHC. En Écosse, l'incidence de l'infection à VHC parmi les UDI a été estimée à 10 infections pour 100 personnes-années en 2013-2014 [2]. En Australie, l'incidence de l'infection à VHC parmi les UDI a diminué au cours du temps, de 10-15 pour 100 personnes-années en 2004 à 4/100 ($IC95\%, 1.3 - 12.3$) personnes-années en 2009 [65].

Notre étude montre que l'incidence globale est 2 fois plus élevée chez les UDI actifs par rapport aux UDI en 2011, ce qui explique pourquoi la prévalence n'a pas fluctué de manière significative entre les deux enquêtes chez les UDI actifs. Comparées aux autres pays Européens comme les Pays-Bas [31] ou la Suisse [154], l'incidence chez les UDI actifs demeure élevée en France.

Nos résultats montrent également que la consommation de crack dans le dernier mois précédant l'enquête est probablement un mauvais indicateur de prise de risque au cours de la vie par la consommation de crack. Ce n'est pas le crack lui-même qui expose les personnes au risque de transmission du VHC mais les lésions buccales causées par l'utilisation et le partage de pipes à crack en verre, facilement cassables et conduisant bien la chaleur. L'utilisation de pipes à crack en verre entraîne régulièrement des brûlures, des lésions ulcérées et des coupures sur les lèvres et dans la cavité buccale. De petites quantités de sang peuvent ainsi constituer un risque d'infection lorsque les utilisateurs partagent leurs pipes en verre lors de la consommation de crack.

Par ailleurs, les mesures de réduction des risques peuvent contribuer au déclin rapide et devraient être prises en compte dans la modélisation de la prévalence et de l'incidence au cours du temps. En effet, au niveau international, les mesures de réduction des risques ont été améliorées ces trois dernières décennies incluant les programmes d'échange d'aiguilles/seringues, l'accès aux traitements de substitution aux opiacés (méthadone et buprénorphine), le dépistage et le traitement thérapeutique [31, 65, 101, 102]. En France, l'accès aux traitements de substitution aux opiacés s'est amélioré mais des dispositifs innovants tels que les salles de consommation à moindre risque pourraient être envisagés [68]. Dans l'enquête ANRS-Coquelicot, la question suivante a été posée : "*au cours du dernier mois, vous est-il arrivé au moins une fois d'utiliser la même eau de javel qu'une autre personne pour nettoyer votre seringue ?*". Mais le pourcentage de données manquantes était élevé (70%).

Nos estimations basées sur des modèles souffrent néanmoins de limites potentielles également

présentes dans d'autres études, dont certaines sont énumérées par Cullen et al [30]. Premièrement, parmi les UD non injecteurs, l'incidence de l'infection à VHC a été estimée à 2/100 personnes-années en 2011. Cette estimation pourrait être le reflet d'un biais de classification en raison d'une sous-déclaration du mode de consommation (l'usage de drogues par injection). Deuxièmement, il aurait été intéressant d'inclure des variables supplémentaires - telles que la migration - dans les équations de régression modélisant la prévalence, si ces données avaient été disponibles. L'estimation de l'incidence pourrait également être grandement améliorée. En effet, la transmission de l'infection par le VHC a été simplifiée par un modèle compartimental simple à deux états. Plusieurs chercheurs ont développé des modèles compartimentaux à plus de 2 états, stratifiés sur plusieurs facteurs : le statut VIH, le fait d'être un UDI (oui/non), le fait d'être sous traitement (oui/non), ou le fait d'être guéri (i.e. ceux testés anti-VHC positifs et ARN du VHC négatif) (oui/non) [22, 54, 101]. Cependant, les résultats biologiques pour la recherche de l'ARN permettant d'identifier les individus guéris n'étaient pas disponibles en 2004.

Il aurait été également utile d'introduire des paramètres dépendant de l'âge (ou groupe d'âges), du genre et/ou du temps comme le taux de mortalité ou la proportion de nouveaux usagers de drogues. Mais nous n'avions pas ces données précises pour la population des UD français [27, 163]. Malgré ces limites, nous pensons que cette approche combinant un modèle de régression et un modèle compartimental est une méthode alternative pour estimer l'incidence à partir d'enquêtes épidémiologiques transversales en l'absence de cohorte. Mettre en place une troisième enquête parmi les UD pourrait être envisagé afin d'évaluer si le déclin de l'incidence de l'infection à VHC se poursuit depuis 2011. Malgré une augmentation potentielle des comportements à risque, un tel déclin peut être attendu compte tenu, d'une part, de la baisse de la prévalence au cours du temps, et d'autre part, des développements récents dans les mesures de réduction de risque et les nouvelles approches thérapeutiques.

Dans les chapitres précédents, nous avons estimé la prévalence et l'incidence de l'infection àu VHC chez les UD fréquentant des centres spécifiques dans 5 grandes villes françaises. Mais, qu'en est-il des personnes qui ne fréquentent aucun de ces lieux ? On pourrait penser qu'elles ont des profils différents, comme déjà discuté plus haut en introduction du manuscrit. Le prochain chapitre présente une technique d'enquête permettant de les atteindre et les outils statistiques permettant de mesurer la dynamique du VHC dans une population "cachée".

Chapitre 5

Échantillonnage conduit par les répondants - Enquête auprès d'individus ne fréquentant aucun lieu d'enquête

En l'absence de base de sondage d'individus appartenant à une population difficile d'accès, les personnes enquêtées sont vues dans les lieux spécialisés qui leur sont dédiés. Cependant, des personnes ne fréquentent pas ou très peu ces lieux et de fait ne peuvent pas être approchées lors des enquêtes.

S'agissant de l'étude de la population d'UD en France, nous avons présenté au chapitre 2 l'enquête ANRS-Coquelicot réalisée auprès d'UD fréquentant des centres spécialisés en 2004 puis en 2011. Cette enquête fait partie du projet de recherche sur les UD, intitulé "*Enquête multicentrique, multi-sites sur les fréquences et les déterminants des pratiques à risque de transmission du VIH, VHC et VHB chez les usagers de drogues - Enquête ANRS Coquelicot*" et coordonné par Marie Jauffret-Roustide. Ce projet a également pour objectif d'améliorer la "représentativité" de la diversité des UD afin de couvrir l'ensemble de la population des UD en France en accédant à des populations complémentaires, plus particulièrement :

- les UD non francophones (originaires des pays de l'Europe de l'Est),
- les UD incarcérés et

- la population invisible ou "cachée" qui ne fréquente aucun centre.

Les trois sous-populations citées plus haut ont des profils et des pratiques peu connus, certaines ont sans doute moins facilement accès au matériel de réductions des risques et bénéficient moins des conseils de prévention et de soins.

Lors du travail de terrain en 2011, il est apparu que certaines structures accueillaient un nombre important d'UD russophones. Beaucoup d'entre eux n'ont pas pu être inclus dans l'enquête de 2011, notamment en raison de la barrière de la langue. Une enquête auprès des russophones a donc été réalisée en 2013 à Paris selon le même schéma d'échantillonnage (TLS) que celles réalisées auprès des francophones en 2004 et 2011 si ce n'est une traduction du questionnaire en russe et le recrutement d'un enquêteur-interprète parlant le russe.

Nous ne détaillerons pas ici les deux autres enquêtes menées auprès des UD incarcérés et auprès des populations cachées pour lesquelles les plans de sondage sont en cours de construction.

Nous nous intéressons dans ce chapitre à l'échantillonnage auprès de la population des UD qui ne fréquente aucune structure. Une méthode d'échantillonnage originale est proposée pour atteindre cette population et connait un succès certain depuis une vingtaine d'années : l'échantillonnage conduit par les répondants, plus connu sous le terme anglais *respondent-driven sampling* (RDS). Cette technique d'enquête a été retenue pour répondre aux objectifs du projet cité plus haut concernant la population "cachée" des UD en France.

L'objectif de ce chapitre est donc de présenter le principe de l'échantillonnage RDS, de citer les principaux estimateurs, de les comparer et d'avancer quelques considérations pratiques pour la population des UD.

Cette partie a fait l'objet d'un papier publié dans la revue *Methodological Inovations* en 2016.

5.1 Contexte méthodologique - Historique

Cette technique d'enquête en réseau par chaînage a été introduite en 1997 par Heckathorn [56] et est utilisée principalement pour étudier les populations difficiles d'accès et obtenir sous certaines conditions des estimateurs sans biais dans ces populations. Entre 2003 et 2007, elle avait déjà été utilisée dans plus de 120 études à travers 20 pays différents, correspondant à plus de 32000 personnes recrutées [71]. A eux seuls, les Centers for Disease Control and Prevention (CDC) américains ont réalisé une enquête en 2005 dans 20 villes aux États-Unis incluant plus

de 13000 usagers de drogues [23]. En 2013 et depuis le milieu des années 1990, White et al. ont comptabilisé 460 études RDS de 69 pays incluant les États-Unis (151 études), la Chine (70) et l'Inde (32) [162]. Cet engouement des chercheurs pour cette méthode était la conséquence de leurs nombreuses années de frustration lorsqu'ils n'avaient à leur disposition que des méthodes présentant des inconvénients majeurs connus pour étudier ces populations. On peut citer notamment l'échantillonnage en réseau [15], en boule de neige [50], ciblé [157], en marche aléatoire [78] ou adaptatif en grappes [142].

Introduit en 1965 par Birnbaum et Sisken [15], l'*échantillonnage en réseau* (ou multiple) est un échantillonnage pour lequel des unités d'intérêt sont sélectionnées selon un plan de sondage classique puis ces unités sont interrogées pour déterminer l'ensemble des unités ayant une certaine caractéristique dans leur "voisinage" (e.g. les membres de la famille d'un individu, une liste de patients pour un centre de soins).

En 1961, Goodman [50] a proposé un *échantillonnage en boule de neige* où un échantillon aléatoire d'individus est construit dans un premier temps, puis chaque individu appartenant à l'échantillon nomme k individus appartenant à la population d'intérêt. Les individus nommés sont interrogés puis nomment à leur tour k individus et, la procédure est ensuite répétée jusqu'à former s étapes.

En 1989, Watters et Biernacki [157] ont développé une autre technique pour atteindre une population difficile d'accès, l'*échantillonnage ciblé* qui se compose de deux étapes : définition d'aires géographiques dans lesquelles interroger des individus puis utilisation d'un plan de sondage pour sélectionner les individus (sondage par quotas, échantillonnage en boule de neige). Parallèlement, Klovdahl [78] a défini l'*échantillonnage en marche aléatoire* où un répondant indique le nom de toutes ses relations appartenant à la population d'intérêt, un seul individu est tiré au sort par l'enquêteur puis l'individu sélectionné indique à son tour le nom de toutes ses relations appartenant à la population d'intérêt, etc.

L'*échantillonnage adaptatif en grappes* fait son apparition en 1990 [142]. Il s'agit tout d'abord de construire l'échantillon initial selon un plan de sondage classique. Si la variable d'intérêt d'un répondant de l'échantillon initial prend une certaine valeur, tous les "voisins" de cet individu sont inclus dans l'échantillon (foyers voisins pour un ménage, voisins d'un malade enquêté). Le processus est répété jusqu'à ce que l'on ne puisse plus ajouter de nouvel individu.

D'un point de vue pratique, lorsque l'on s'intéresse à des pratiques illégales comme l'usage

de drogues, il est inconcevable de localiser ou nommer des UD. De plus, d'un point de vue statistique, ces méthodes nécessitent généralement de disposer d'une base de sondage pour construire l'échantillon initial et ne reposent pas toujours sur une sélection aléatoire des individus ce qui peut entraîner un biais dans les estimations d'indicateurs épidémiologiques et de leurs variances.

Afin de s'affranchir de ces limites, la méthode RDS a donc été développée et appliquée à différentes populations.

5.2 Principe

Son principe est de construire un échantillon d'individus par les individus eux-mêmes, d'où son nom.

Tout d'abord, des individus de la population d'intérêt sont recherchés et recrutés. Ces individus sont appelés des **racines** et constituent la vague zéro. Il est important que ces racines aient un réseau social de taille suffisamment grande pour faciliter le recrutement d'unités d'intérêt (leurs pairs). Il est aussi préférable qu'elles ne se ressemblent pas en termes de caractéristiques socio-démographiques pour assurer une diversité dans l'échantillon. Ces racines vont ensuite recruter elles-mêmes, dans leur réseau social, des individus appartenant à la population d'intérêt. Ces derniers forment alors la première **vague** et vont à leur tour recruter d'autres individus appartenant à la population d'intérêt pour constituer la seconde vague. Le processus, illustré en Figure 5.1, se répète ainsi jusqu'à l'obtention d'une taille d'échantillon déterminée à l'avance.

Les individus ayant reçu un coupon se rendent dans un local dédié à l'enquête et répondent à un questionnaire pouvant être complété par des examens de santé et/ou des prélèvements biologiques. A l'issue de l'entretien, un ou plusieurs **coupons** en papier leur sont transmis. Chaque individu remet ensuite les coupons dont il dispose à des individus de son réseau social et appartenant à la population d'intérêt. Sur ces coupons sont renseignés l'adresse du local (parfois un numéro de téléphone), le nom de l'étude et un numéro d'identifiant. Ce numéro d'identifiant permet d'identifier qui a recruté qui, pour reconstruire le chaînage de recrutements. Puis, les individus recrutés par leurs pairs se rendent dans le local dédié, remettent leur coupon, répondent au questionnaire et reçoivent à leur tour des coupons à distribuer. Disposer d'un coupon fait donc partie des critères d'inclusion à l'enquête.

Une fois le questionnaire et le prélèvement biologique effectués, une compensation financière

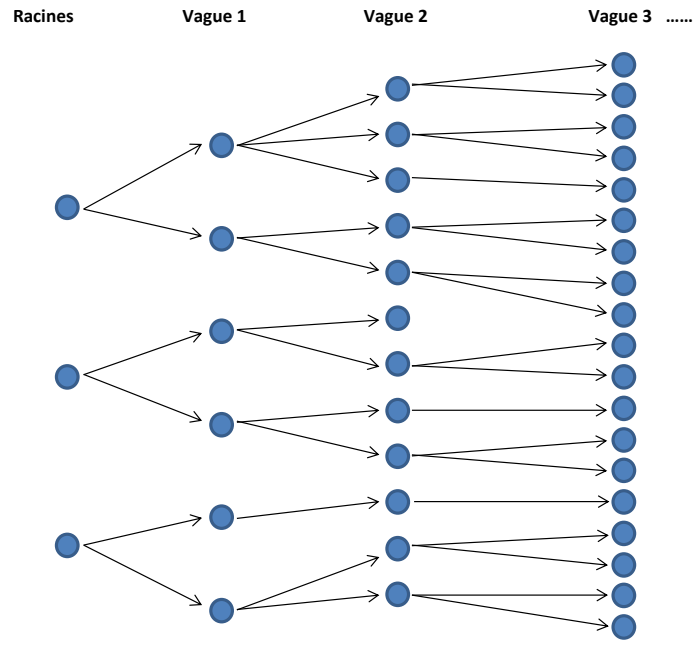


FIGURE 5.1 – Représentation des trois premières vagues de recrutement par RDS. Chaque cercle représente un individu et chaque flèche représente la distribution d'un coupon d'un individu à l'autre.

est généralement offerte aux répondants et aux recruteurs. Cette compensation incite les individus à participer et permet de réduire le biais dû au volontariat (les plus coopératifs). Toutefois, cela peut introduire un autre biais dans les estimations, par exemple, si les personnes les plus démunies participent davantage à l'enquête et si le niveau de vie est lié à la variable d'intérêt.

5.3 Estimation de la prévalence d'une maladie

Nous nous intéressons ici à la prévalence d'une maladie (proportion d'infectés). Soit une population composée de deux groupes d'individus (groupe A : individus infectés et groupe B : individus non infectés). Dans ce type d'échantillonnage, on peut s'attendre à ce que les propriétés de l'estimateur de la proportion d'infectés soient difficiles à déterminer. En effet, elles vont dépendre à la fois des caractéristiques du réseau social (incontrôlées par l'enquêteur) et des choix (contrôlés) dans l'échantillonnage, en termes de nombre de coupons, de vagues, etc.
Plusieurs estimateurs ont été proposés pour estimer une proportion dont :

- l'estimateur classique de Salganik-Heckathorn [57], appelé RDS-I ou estimateur *SH*

- l'estimateur de Volz-Heckathorn [152], appelé RDS-II ou estimateur *VH*

- l'estimateur de Gile [44], appelé RDS-SS (*successive sampling*).

Soit s un échantillon conduit par les répondants avec s_A (respectivement s_B) l'échantillon d'individus appartenant au groupe A (respectivement au groupe B). Soit p_A (respectivement p_B) la proportion d'individus infectés (respectivement la proportion d'individus non infectés). A partir de ces notations, les 3 estimateurs de la proportion p_A sont présentés dans le Tableau 5.1. De façon analogue, une estimation de la proportion p_B peut être calculée.

TABLE 5.1 – Estimateurs de la proportion d'individus infectés $\hat{P}_A$ à partir d'un échantillon RDS composé d'individus infectés (A) et non infectés (B).

Estimateurs	Expression mathématique	Paramètres
RDS-I	$\hat{P}_A^{SH} = \frac{\widehat{C^{BA}}}{\widehat{C^{BA}} + \widehat{C^{AB}} \frac{\hat{D}_A}{\hat{D}_B}}$	C^{AB} : proportion de B recrutés par A ; C^{BA} : proportion de A recrutés par B ; $\hat{D}_A$: estimateur du degré* moyen du groupe A et $\hat{D}_B$: estimateur du degré moyen du groupe B
RDS-II	$\hat{P}_A^{VH} = \frac{\sum_{i \in s_A} 1/\tilde{d}_i}{\sum_{i \in s} 1/\tilde{d}_i}$	$\tilde{d}_i$: degré* rapporté par l'individu i. Ici, $\hat{P}_A^{VH}$ est un estimateur de p_A asymptotiquement sans biais si π_i, la vraie probabilité d'inclusion, est proportionnelle à $\tilde{d}_i$ pour tout individu i.
RDS-SS	$\hat{P}_A^{SS} = \frac{\sum_{i \in s_A} 1/\tilde{\pi}_i}{\sum_{i \in s} 1/\tilde{\pi}_i}$	$\tilde{\pi}_i$: probabilité d'inclusion de l'individu i estimée à partir d'une procédure d'échantillonnage successif.

* Le degré d'un individu correspond au nombre de ses relations/contacts dans la population d'intérêt.

Il a été montré par simulations, en comparant les erreurs quadratiques des deux estimateurs les plus populaires, RDS-I et RDS-II, que la performance de RDS-II était presque toujours supérieure à celle de RDS-I [45]. L'estimateur RDS-II, dont les performances sont présentées dans la section 5.5, est le plus utilisé. Il est implémenté dans RDSAT (*Respondent-Driven Sampling Analysis Tool*), un programme gratuit dédié à l'analyse de données d'enquête par RDS. Sa variance est estimée par bootstrap [123].
RDS-II est asymptotiquement sans biais sous les hypothèses suivantes :

1. Le sondage est avec remise
2. La fraction de sondage est petite
3. Chaque individu ne recrute qu'un seul individu (*i.e.* nombre de coupons=1)
4. Les répondants indiquent précisément leur degré dans le réseau
5. Le recrutement est aléatoire
6. Les liens sont réciproques (réseau non-orienté)
7. Les individus de la population sont tous connectés

Le processus de recrutement peut donc être vu comme une chaîne de Markov irréductible de premier ordre et à espace fini d'états [56] :

- le recrutement d'un individu dépend uniquement de son recruteur et non des recruteurs précédents,
- tous les individus sont connectés, donc toute personne de la population d'intérêt peut être recrutée quel que soit son réseau social,
- la population d'intérêt est de taille finie,
- le processus est stationnaire : on définit ainsi la notion d'équilibre quand les caractéristiques des individus enquêtés demeurent constantes au fur et à mesure des vagues de recrutement et se rapprochent donc des caractéristiques de la population d'intérêt. Le processus converge alors vers un mélange équilibré d'individus recrutés et le recrutement des individus ne dépend plus des caractéristiques des racines initiales.

En 2011, Gile a proposé un nouvel estimateur appelé RDS-SS qui permet de s'affranchir de l'hypothèse 1 (sondage avec remise) [44]. Il est basé sur un échantillonnage successif (successive sampling, d'où son nom RDS-SS) équivalent à un tirage sans remise à probabilité proportionnelle au degré. RDS-SS diffère donc de RDS-II par le calcul des probabilités d'inclusion (Tableau 5.1).

Ces trois estimateurs sont désormais implémentés dans les packages RDS et RDS Analyst du logiciel R.

5.4 Caractéristiques d'un échantillon RDS

RDS est largement utilisée pour des études auprès de populations "cachées" mais, en réalité, les hypothèses assurant des estimateurs sans biais ne sont pas toujours vérifiées et vérifiables.

Population connectée : L'hypothèse que la population d'intérêt forme un seul réseau social n'est pas toujours réaliste. Par exemple, si l'on s'intéresse à la population des UD, certains UD forment un réseau social en fonction des drogues qu'ils consomment ou de l'endroit où ils vivent [87]. Dans ce cas, la probabilité que les personnes dans une sous-population recrutent des personnes dans une autre sous-population peut être considérablement réduite. Ainsi, l'utilisation d'un échantillonnage RDS dans chaque sous-population est recommandé.

Tirage avec remise : Cette hypothèse n'est pas appliquée en pratique. L'échantillonnage ne s'effectue jamais avec remise pour des raisons pratiques (pas d'intérêt) et statistiques (perte de puissance). Un individu est interrogé une seule fois.

Nombre de coupons. Afin d'éviter que la chaîne de recrutement ne s'arrête trop tôt en cas de non-participation de l'un des individus qui reçoit un coupon, plusieurs coupons (2 ou 3) sont souvent utilisés en pratique. Goel et Salganik ont montré que le recrutement multiple augmentait la variance de l'estimateur RDS-II [48].

Degré : Pour un individu donné, on nomme *degré*, la taille de son réseau social, c'est-à-dire le nombre de ses relations dans la population d'intérêt. Par exemple, si on s'intéresse à des UD et si un UD connaît n UD dans son réseau social, on dit que son degré est égal à n. Dans la vraie vie, il est difficile pour un individu de signaler précisément le nombre de ses relations dans la population d'intérêt. C'est pourquoi en pratique, on lui demandera : "combien de personnes pouvez-vous recruter ?".

Liens réciproques : Les liens sociaux ne sont pas toujours réciproques. Dans un réseau d'UD, la relation entre un usager et un fournisseur n'est pas forcément la même dans les deux sens. Il est aussi possible qu'un UD donne un coupon à un inconnu rencontré dans un lieu public pour recevoir une compensation financière. C'est pourquoi, on demande au recruté la nature de sa relation avec le recruteur. S'ils sont étrangers, alors le recruté est exclu de l'échantillon.

Sélection des racines : Sélectionner des racines en fonction de leur degré et de leurs caractéristiques n'est pas nécessaire pour atteindre une loi stationnaire si le nombre de vagues de recrutement est assez grand.

Homophilie : Lorsqu'un individu a tendance à recruter des personnes qui lui ressemblent, notamment en ce qui concerne la variable d'intérêt, on parle d'*homophilie*, même si plusieurs définitions existent [29, 149]. Ainsi, si la variable d'intérêt est le statut sérologique, l'homophilie sera forte si les personnes infectées (respectivement non infectées) ne recrutent que des personnes infectées (respectivement non infectées).

Recrutement aléatoire. Dans les faits, cela est impossible à vérifier. Le fait que le recrutement ne soit pas aléatoire engendre un biais dit d'homophilie si le recrutement est corrélé à la variable d'intérêt [57].

Un des atouts majeurs de RDS, contrairement à l'échantillonnage boule de neige, est qu'il préserve l'anonymat des répondants, compte tenu par exemple du caractère illicite de certaines pratiques dans les populations ciblées (utilisation de drogues, recours à la prostitution).

RDS est un processus stochastique complexe et plus précisément un processus de branchement sans remise sur un graphe arbitraire de relations sociales qui commence avec un échantillon de convenance (les racines) [45]. De ce fait, plusieurs publications récentes ont évalué et comparé le comportement des 3 estimateurs en cas de violation d'une ou plusieurs de ces hypothèses par des études de simulations et non analytiquement en raison de la complexité du processus [45, 46, 90, 110, 146].

5.5 Performances de l'estimateur RDS-II

Gile et Handcock [45] ont évalué les performances de RDS-II, en particulier en fonction de la sélection des racines (sélection de racines infectées ou non infectées, sélection aléatoire, etc.), du comportement des répondants (*i.e.* si l'homophilie est faible ou forte), de la fraction de sondage et du nombre de degrés, dépendant ou non du statut infecté/non-infecté. Les auteurs ont montré l'existence d'un biais dépendant des racines et du niveau d'homophilie. La prévalence réelle (connue car simulée) était sous-estimée quand les racines étaient non infectées et cette sous-estimation était plus importante quand l'homophilie était forte. En effet, en cas de forte homophilie, les racines non infectées tendaient à recruter des individus non infectés qui continuaient à recruter des individus non infectés et ainsi de suite. En conséquence, l'échantillon construit était essentiellement composé d'individus non infectés, entraînant une sous-estimation

de la prévalence. Le biais tendait vers zéro quand les racines étaient sélectionnées aléatoirement, indépendamment de l'homophilie. Quand les racines étaient infectées, la prévalence était sur-estimée, et plus l'homophilie augmentait, plus la sur-estimation était importante.

Gile et Handcock ont également montré que le biais de l'estimateur dépendait du ratio entre le nombre moyen de degrés des individus infectés et le nombre moyen de degrés des individus non infectés ainsi que de la fraction de sondage. Le biais augmentait quand la fraction de sondage et le ratio augmentaient. Ainsi, quand des individus infectés avaient un nombre de degrés plus important que les individus non infectés, la prévalence tendait à être sous-estimée.

Une autre étude a examiné à partir de simulations, la violation de chacune des hypothèses [90]. Les principales conclusions de cette étude indiquaient un biais important quand le réseau social était unidirectionnel (*i.e.* quand un individu connaissait un autre individu mais pas l'inverse) ou quand les répondants choisissaient de recruter des individus qui avaient des caractéristiques corrélées à la variable d'intérêt (*e.g.* malade (oui/non)). D'un autre côté, les auteurs de l'étude ont conclu que l'estimateur était robuste en cas de sondage sans remise, d'un taux de participation faible, de quelques erreurs sur les degrés ou la sélection des racines. Ces conclusions sont différentes de celles avancées par Gile et Handcock concernant les hypothèses 1 et 5.

Une récente étude, toujours basée sur des simulations, a montré un biais important quand les degrés déclarés étaient erronés [110]. Les auteurs ont démontré que l'obtention de degrés corrects pour des individus ayant de petits degrés était particulièrement importante car ces individus avaient des poids de sondage élevés et avaient probablement moins de chance d'être infectés. Il était donc essentiel de pouvoir recueillir des degrés exacts à l'aide de questions précises.

Les propriétés de l'estimateur RDS-II peuvent aussi dépendre de l'effet-plan (*design effect* en anglais). En utilisant des données réelles, Goel et Salganik [49] ont montré que l'effet-plan pouvait être assez important. Ils ont évalué un effet-plan compris entre 5.7 et 58.3 avec une médiane à 11. Cela montre que la variance de l'estimateur est élevée dans des enquêtes épidémiologiques aux plans de sondage "classiques". Cette variance augmentait quand le nombre de coupons augmentait et quand l'homophilie augmentait [90]. Récemment, même si l'effet-plan variait selon les pays et les populations étudiées, les chercheurs ont recommandé un effet-plan entre 2 et 4

pour estimer la taille d'échantillon dans les études RDS [72, 160].

Les études de simulations montrent que le biais et la variance de l'estimateur RDS-II dépend d'un ensemble d'hypothèses plus ou moins contrôlables par la personne en charge de l'enquête. En réalité, on peut donc s'attendre à ce que ces hypothèses ne soient pas vérifiées. La littérature montre que des hypothèses (hypothèses 3 à 6) sont fréquemment transgressées. Nous pouvons citer par exemple l'utilisation de plusieurs coupons [71, 97], la difficulté des répondants à renseigner précisément leur degré [99], un recrutement non aléatoire [41, 88, 155] ou une population non connectée [6, 64, 94, 114].

5.6 Considérations pratiques et recommandations pour la population des usagers de drogues

Les hypothèses mentionnées plus haut doivent être vérifiées dans chaque étude, mais pour des études dans des populations d'UDI, elles doivent également dépendre d'un certain nombre de préoccupations pratiques.

Premièrement, comment les chercheurs peuvent-ils considérer que la structure sociale de la population étudiée est conforme à une structure entièrement en réseau ? Plus précisément, existe-t-il ou non des sous-populations au sein de la population, ce qui aurait des effets importants sur le recrutement ? Il pourrait exister des sous-populations particulières au sein de la population d'intérêt qui permettent de réduire considérablement la probabilité que les individus dans une sous-population puissent recruter des individus dans une autre sous-population : par exemple, des personnes qui injectent des médicaments différents, ou des membres de différentes populations ethniques, ou des injecteurs de drogues qui vivent dans des zones géographiques différentes dans la même ville [87]. La recherche qualitative et/ou ethnographique peut souvent être utilisée pour identifier des sous-populations potentielles au sein de la population globale d'UDI où il serait très peu probable qu'un membre d'un groupe recrute un membre de l'autre groupe. Si cela semble être le cas, alors il est peut-être préférable de considérer deux études dans deux populations d'UDI différentes. Cela requiert bien sûr une taille d'échantillon pour chacune des sous-populations suffisamment grande pour les analyses statistiques, ce qui peut augmenter considérablement le coût de l'étude. Cependant, il n'existe aucune règle de décision pour l'utilisation de données qualitatives pour prendre la décision de mener des études RDS séparées pour différents groupes de populations au sein de la population d'intérêt.

Deuxièmement, la taille de l'échantillon est-elle être assez grande pour atteindre l'*équilibre* ? Pour rappel, l'équilibre se produit dans un échantillon RDS lorsque les caractéristiques importantes de l'échantillon (le genre, le statut sérologique, l'âge, l'origine ethnique, l'injection de drogues, etc.) restent constantes au cours des vagues successives de recrutement des personnes. L'équilibre est donc une indication que le recrutement des individus est indépendant des caractéristiques des racines. Il est généralement nécessaire d'avoir des tailles d'échantillons de plusieurs centaines d'individus pour atteindre l'équilibre. Ne pas parvenir à l'équilibre crée une forte suspicion que l'échantillon produira des estimations biaisées.

Troisièmement, l'étude doit avoir la capacité de traiter un grand nombre de personnes à la fois. Un des atouts de RDS est que l'on peut généralement recruter un grand nombre d'individus dans un court laps de temps. En effet, le recrutement dans l'échantillonnage RDS est une progression géométrique, car le nombre d'individus potentiels ayant des coupons et qui répondent aux critères d'admissibilité à l'étude s'accroît rapidement. Si chaque personne recrute en moyenne deux personnes supplémentaires qui désirent participer à l'étude, le nombre de personnes voulant participer doublera à chaque vague de recrutement (Figure 5.1). Travailler avec un grand nombre de personnes nécessite alors l'ordonnancement des nominations de recherche. Cela ne peut être fait que si les coupons sont remis à chaque individu (le coupon est valable pour seulement un moment précis à une date précise) ou en demandant aux individus de venir au local de l'enquête pour planifier un rendez-vous. Comme la population des UD est généralement peu fiable pour respecter un rendez-vous précis, la planification précise des rendez-vous signifie que certains usagers ne participeront pas parce qu'ils ne se présenteront pas au local de l'enquête à l'heure prévue. Il peut donc y avoir une différence importante entre les individus qui respectent ou non les rendez-vous ce qui engendre une autre source de biais possible dans l'étude. Avoir une certaine flexibilité pour enquêter des personnes, même si elles ne se présentent pas au bon moment permet de réduire ce biais, mais devient très laborieux quand le personnel de l'étude envisage de traiter un grand nombre de personnes.

En outre, dans certaines enquêtes, afin d'éviter et de contrôler les doublons possibles, une combinaison de mesures biométriques de chaque répondant peut être utilisé (par exemple la

longueur de chaque avant-bras) [58,150] ou d'autres identifiants spécifiques (*e.g.* le nom de jeune fille de la mère ou la date de naissance de l'individu enquêté) [121].

5.7 Discussion

Cette section ne prétend pas répondre à toutes les questions pratiques concernant l'étude de populations d'usagers de drogues par une approche RDS. Elle tente d'élargir le débat au-delà de l'hypothèse théorique à certaines des questions pratiques qui peuvent être toutes aussi importantes dans la conduite et l'interprétation d'une étude RDS.

RDS est largement utilisée pour des études auprès de populations "cachées" même si toutes les hypothèses ne sont toujours pas vérifiées. C'est la raison pour laquelle, depuis peu, la communauté scientifique interpelle à nouveau et insiste sur deux axes importants dans l'utilisation de ce type d'échantillonnage, à savoir :

- l'aspect méthodologique visant à poursuivre l'étude des propriétés des estimateurs existants et à améliorer les estimateurs et leurs variances, et
- l'aspect pratique visant à vérifier les hypothèses lorsqu'une enquête est menée, à discuter des résultats obtenus en cas violation de ces hypothèses et à proposer quelques recommandations.

Des articles très récents montrent en effet que cette recherche se poursuit, avec une prise de conscience que RDS doit être évalué sous différents angles [36, 79, 88, 104, 105, 122, 124, 165].

Cependant, certaines questions restent sans réponse pour ceux qui souhaitent mettre en place un échantillonnage RDS. En ce qui concerne le choix de l'estimateur, la question se pose de savoir s'il faut continuer à utiliser l'estimateur RDS-II et sa variance bootstrap, ou si l'estimateur RDS-SS, le plus récent, doit être utilisé sachant qu'il est désormais implémenté dans le logiciel R. D'un point de vue plus pratique, cela soulève d'autres questions pour l'enquêteur : les conditions sont-elles respectées (*i.e.* les hypothèses sont-elles vraies ?) pour pouvoir utiliser cette méthode d'échantillonnage ? Des études pilotes sont-elles nécessaires pour déterminer les caractéristiques du réseau social ? Ce type d'enquête est-il à exclure dans certains cas ?

Salganik a souligné la nécessité de rendre disponibles les données issues d'enquêtes RDS afin d'enrichir la recherche sur l'évaluation et le développement de RDS à partir de données

réelles [124]. White *et al.* [162] ont proposé un ensemble de critères à renseigner pour des études RDS à partir du guide "STROBE-RDS" (Reporting of Observational Studies in Epidemiology for respondent-driven sampling studies) développé pour des enquêtes transversales [153].

Au final, des outils diagnostics et des recommandations pratiques ont été proposés pendant et après le recueil des données pour améliorer l'échantillonnage RDS et l'inférence qui en découle [46]. Cela semble d'autant plus important puisque RDS s'est étendu à des domaines autres que l'étude de populations difficiles à atteindre, en particulier dans les enquêtes téléphoniques [86], dans les enquêtes par internet [126, 134] ou même dans le recrutement de participants pour évaluer l'efficacité des mesures de prévention pour le VIH dans les essais cliniques [130].

Pour recruter rapidement des individus difficiles à atteindre, RDS est une méthode simple et coût-efficace. Payer des personnes pour recruter d'autres personnes serait généralement plus coût-efficace que payer l'équipe d'enquêteurs pour recruter des personnes ayant un large réseau social. Toutefois, comme attendu par les chercheurs, RDS ne peut pas être appliqué dans de nombreuses situations. C'est le cas notamment lorsqu'il y a une forte homophilie et lorsqu'on n'arrive pas à atteindre l'équilibre au fur et à mesure de la construction de l'échantillon. En effet, il est possible qu'il n'y ait pas un réseau social unique dans la population d'intérêt (*i.e.* notion de "*petit monde*") [77] mais plutôt une population fragmentée en plusieurs sous-populations (plusieurs réseaux) nécessitant d'être considérées séparément. Il est donc possible que cette population ne soit pas socialement connectée. Par exemple, les travailleurs du sexe qui passent par internet pour rencontrer leurs clients ne sont pas suffisamment inter-connectés pour maintenir des chaînes de recrutement. Les hommes ayant des relations sexuelles avec d'autres hommes qui se rencontrent dans des lieux anonymes (*e.g.* dans un parc) ne sont pas nécessairement inter-connectés. Pour les UDI, l'une des clés dans le succès de RDS est de recruter des racines ayant un large réseau social et avec lesquelles les gens sont confiants. Cette notion de confiance est cruciale pour améliorer le recrutement des UDI avec l'échantillonnage RDS.

Dans des situations où les hypothèses sous-jacentes dans la théorie RDS ne tiennent pas ou lorsque le recrutement ne permet pas d'atteindre un large nombre d'individus, les données peuvent être analysées comme un échantillon de convenance tout en sachant qu'il n'est pas "représentatif" de la population d'intérêt.

Le plus important, si RDS ne peut être utilisé pour une certaine population d'intérêt est sans doute de fournir des indications importantes pour les actions de promotion de la santé pour cette population. Ces deux éléments vont de pair, à savoir, si RDS ne fonctionne pas, il en est certainement de même pour les campagnes et autres mesures de prévention pour la santé.

Pour conclure, si les hypothèses théoriques peuvent être satisfaites et les problèmes pratiques minimisés, l'échantillonnage RDS peut se définir, dans un cadre scientifique, comme la méthode d'échantillonnage la plus sophistiquée et la plus coût-efficace pour l'étude de populations ne fréquentant aucun lieu d'enquête en recrutant rapidement un grand nombre de personnes.

Discussion et perspectives

Afin de pouvoir mettre en oeuvre des messages de prévention, il est nécessaire de disposer de données épidémiologiques fiables auprès de populations spécifiques d'accès difficile, en situation de précarité ou mal connues. Dans ce contexte, notre réflexion s'est portée sur le développement d'outils statistiques dans le champ des sondages et leur utilisation pratique pour atteindre ces populations et fournir des indicateurs épidémiologiques corrects.
Nous nous sommes particulièrement intéressés à la population des UD vis-à-vis de l'hépatite C, une maladie évoluant à bas bruit durant de longues années avant la survenue de complications graves (cirrhose, cancer primitif du foie). L'usage de drogues par injection est connu comme la principale source de contamination par le VHC dans les pays industrialisés. En France, jusqu'en 2012, la seule estimation de prévalence du VHC publiée et calculée à partir d'un échantillon aléatoire d'UD avec un recueil de prélèvements biologiques, provenait de l'enquête ANRS-Coquelicot menée en 2004 dans 5 villes (Lille, Strasbourg, Paris, Bordeaux et Marseille) avec une prévalence des anticorps anti-VHC estimée à 60% [67]. La seule estimation d'incidence de l'infection à VHC concernait une cohorte d'UDI suivie en 2000-2001 dans la région Nord-Est de la France (en excluant Paris et sa banlieue) avec une incidence estimée à 9/100 personnes-années [92]. Dans un objectif de réduction de la diffusion de cette maladie, il était donc nécessaire de disposer d'estimations plus récentes concernant cette population.

Nous présentons dans ce chapitre une synthèse globale de notre travail (les résultats ayant déjà été discutés dans les différents chapitres) et les perspectives de recherches potentielles.

Synthèse des résultats

A partir des données de deux enquêtes ANRS-Coquelicot réalisées en France en 2004 et 2011, nous avons estimé la prévalence et l'incidence de l'infection à VHC par âge et en fonction du temps chez les UD fréquentant des centres dédiés (CSAPA, CAARUD, structures d'hébergement). Actuellement, ce sont les seules enquêtes nationales réalisées auprès des UD avec un recueil de prélèvements biologiques.
En général, les enquêtes par sondage passent par la construction d'un échantillon issu de la population d'intérêt. L'expérience montre que cette construction est d'autant plus délicate lorsque cette population est difficile d'accès. C'est le cas des UD, notamment en raison de leurs pratiques illicites.
Par définition, cette population d'UD est caractérisée par sa partie visible (UD accessibles par les lieux d'enquêtes) et cachée (UD qui ne fréquentent aucun lieu répertorié). Nous avons donc considéré deux types d'échantillonnage permettant d'atteindre ces deux parties de notre population d'intérêt, à savoir respectivement l'échantillonnage lieux-moments (TLS) et l'échantillonnage conduit par les répondants (RDS).

Nous avons tout d'abord formalisé le TLS dans le cadre théorique du sondage indirect, puis proposé un estimateur pour un total et une proportion en s'inspirant de la méthode généralisée du partage des poids initialement développée pour la statistique publique [83]. Cette technique d'enquête est généralement privilégiée pour des enquêtes réalisées auprès de populations accessibles par des lieux qu'elles fréquentent (des centres dédiés, des structures d'hébergement, des lieux de convivialité, la rue, etc.). L'estimateur basé sur le plan (*design-based*) que nous proposons ici intègre la fréquentation des lieux d'enquêtes et permet de fournir des estimations sans biais de prévalences. Nous avons pu démontrer la robustesse de cet estimateur en générant différents types de populations (prévalence variant de 1% à 90%, distribution des fréquentations plus ou moins hétérogènes, fréquentations des lieux variables selon la maladie étudiée) et cela, même en cas d'erreurs sur les fréquentations déclarées par les individus interrogés, en évaluant principalement l'impact de la fréquentation sur le biais de l'estimateur. A l'issue de ces travaux, nous recommandons fortement l'utilisation de l'estimateur basé sur le plan par rapport à un estimateur ignorant la fréquentation des lieux, lorsqu'un biais peut être lié à cette fréquentation.

De manière générale, nous préconisons l'utilisation de ce type d'échantillonnage associant l'estimateur basé sur le plan présenté ici pour estimer la prévalence d'une maladie dans des études auprès de populations accessibles par les lieux d'enquêtes. Ces travaux trouvent d'ores et déjà une application directe via l'étude actuelle du VIH auprès des hommes ayant des relations sexuelles avec d'autres hommes et fréquentant des lieux de convivialité (enquête Prevagay 2015).

En l'absence d'un test biologique permettant de distinguer les infections récentes des infections anciennes en 2004 et en l'absence de cohorte au niveau national au moment de ce travail de thèse, l'incidence de l'infection à VHC a été estimée à partir de ces deux enquêtes épidémiologiques transversales. Cette approche consistait à combiner un modèle compartimental à deux états (séropositif *versus* séronégatif) modélisant la transmission du VHC chez les UD, et un modèle de régression pour estimer la prévalence du VHC par âge en fonction du temps en incluant la variable âge en continu. Ce modèle mathématique intégrait également les plans de sondages mis en place lors des enquêtes épidémiologiques. Plus précisément, nous disposions des résultats des tests biologiques réalisés pour la recherche des anticorps anti-VHC sur les prélèvements biologiques recueillis dans les deux enquêtes.
Pour chaque enquête, nous avons d'abord modélisé la distribution des résultats quantitatifs des tests biologiques par un mélange de lois afin de classer les UD comme anti-VHC positif ou anti-VHC négatif. Nous avons pu déduire la prévalence globale pour chaque année d'enquête sans faire d'hypothèse sur la valeur du seuil de détection des anticorps anti-VHC. Généralement, cette valeur de seuil est fournie avec la trousse commerciale des tests biologiques pour une utilisation dans des conditions standards (en termes de conditionnement, stockage, support de prélèvement, etc.). Les estimations de prévalences globales présentées dans cette thèse sont différentes de celles déjà publiées car elles ont été calculées pour des "sous-populations" (exclusion des UD ayant consulté un médecin généraliste prescripteur de traitements de substitution aux opiacés en 2004 et exclusion de deux départements en 2011) issues des enquêtes initiales. Nous avons ensuite estimé les prévalences du VHC par âge et en fonction du temps en utilisant des modèles de régression logistique ajustés sur d'autres caractéristiques (statut VIH, injecteur de drogues (oui/non), consommateur de crack (oui/non)), la variable âge étant modélisée par des polynômes fractionnaires. Enfin, nous avons estimé l'incidence de l'infection à VHC par âge et en fonction du temps à partir de la relation formulée entre la prévalence et l'incidence, chez

les UD fréquentant des centres dédiés.

Malgré les limites inhérentes à cette approche, elle reste une alternative satisfaisante pour estimer l'incidence à partir d'enquêtes épidémiologiques transversales en l'absence de cohorte ou de tests biologiques permettant de détecter les infections récentes.
Ainsi, à partir des données de ces deux enquêtes, nous avons pu évaluer le niveau d'infection par le VHC chez les UD fréquentant des dispositifs spécialisés en termes de prévalence et d'incidence. En 2011, nous avons estimé la prévalence du VHC chez les UD à 43.7%, $IC95\% = [39.5\%, 47.9\%]$ et l'incidence de l'infection à VHC a été estimée à 4.4 ($IC95\%, 3.3 - 5.9$) pour 100 personnes-années. Chez les UDI actifs, l'incidence de l'infection à VHC a été estimée en 2011 à 11.2, ($IC95\%, 9.0 - 19.0$) pour 100 personnes-années.

A cela, nous pouvons noter une réflexion toujours active de la communauté scientifique sur l'utilisation de prélèvements alternatifs (le DBS, par exemple) ou le développement de nouveaux tests biologiques (par exemple, le test d'avidité du VHC) dont les avancées impactent les aspects logistiques du volet biologique de ces enquêtes. L'évolution de ces techniques biologiques contribuent à améliorer les conditions d'enquêtes séroépidémiologiques et donc potentiellement l'accès à certaines populations. Nous pouvons notamment citer l'utilisation du DBS en remplacement de la ponction veineuse auprès d'UD ou auprès de populations interrogées dans la rue, voire à domicile.

Perspectives de recherche

Plusieurs perspectives de recherche ont émergé de ce travail de thèse.

Tout d'abord, nous avons comparé l'estimateur basé sur le plan à un estimateur ignorant la fréquentation des individus puis, évalué l'influence de la fréquentation sur le biais de l'estimateur. Dans une future recherche, nous pourrions aussi étudier d'autres sources de biais (type de prestations, durée de la prestation, etc.). Toutefois, même si elles existent, nous rappelons que l'estimateur basé sur le plan de sondage reste à privilégier lorsqu'un biais peut être lié à la fréquentation. Pour cela, cette fréquentation doit être la plus précise possible. Il est alors

nécessaire d'intégrer des questions liées à la fréquentation dans le questionnaire à remplir lors de l'entrevue mais aussi de bien réfléchir à la formulation de ces questions :

- Combien de fois ? Réponse ouverte (1, 2, 3, ...,10 fois) ou catégorielle (moins de 5 fois, entre 5 et 15 fois, plus de 15 fois).
- Sur quelle période ? les dernières 24 heures, une semaine.
- en listant l'ensemble des lieux et/ou des moments ?
- etc.

Le recueil de ces informations dépend à la fois de la population d'intérêt (UD, HSH, personnes sans domicile, migrants) et des lieux à enquêter (centres spécifiques, consultations médicales, lieux festifs, la rue). Des études spécifiques selon les populations étudiées pourraient être envisagées pour élaborer un guide des questions précises liées à la fréquentation à intégrer dans les questionnaires.

Puis, nous avons évoqué la perspective d'un travail pour comparer l'estimateur basé sur un plan de sondage à un estimateur basé sur un modèle comme celui développé par Gustafson [53], à partir de données simulées, en discutant les avantages et les inconvénients de ces deux estimateurs.

Les réflexions et conclusions que nous apportons dans le cadre de cette thèse pourraient également s'appliquer aux enquêtes téléphoniques. En effet, l'absence de base de sondage des numéros filaires et/ou mobiles et le rattachement d'un individu à un ou plusieurs numéros de téléphone par analogie aux fréquentations associent l'utilisation de l'échantillonnage TLS d'une part et l'estimateur basé sur le plan proposé ici d'autre part.

De plus, concernant la population d'UD qui ne fréquente aucun lieu d'enquête, il apparaît intéressant de disposer de données en France afin de pouvoir, d'une part comparer les profils des UD appartenant à la population cachée avec ceux ayant un "rapport problématique aux drogues" et se rendant dans des centres, et d'autre part évaluer l'utilisation de l'échantillonnage RDS sur les aspects théoriques et pratiques.

Par ailleurs, l'estimation de l'incidence a nécessité la combinaison des deux enquêtes répétées disponibles. Une troisième enquête permettrait d'améliorer le modèle mathématique liant la

prévalence à l'incidence.

Lorsqu'une enquête est répétée dans le temps, c'est-à-dire lorsque l'on dispose d'estimations ponctuelles de prévalences sur plusieurs années, une question qui se pose naturellement est de déterminer s'il existe une tendance à la hausse ou à la baisse de la prévalence, en fonction de nombreuses variables, notamment socio-démographiques. La réponse à cette question n'est pas simple car l'estimation de la tendance à partir de plusieurs enquêtes transversales nécessite de prendre en compte les plans de sondages, la taille des enquêtes et la composition de la population qui peuvent ne pas être identiques d'une enquête à l'autre. Peu de papiers abordent le problème de la combinaison d'enquêtes pour estimer correctement une tendance temporelle de la prévalence ou d'autres paramètres à partir de plusieurs enquêtes. Une perspective serait de proposer une méthode adéquate pour combiner des enquêtes en tenant compte des plans de sondage, de la taille des enquêtes et de la composition de la population selon ce que l'on souhaite estimer (une moyenne, un total, une tendance, une force d'association).

Conclusion

Les travaux présentés ici ont permis de proposer des méthodes statistiques permettant d'estimer correctement des prévalences et des incidences à partir d'enquêtes épidémiologiques transversales répétées dans le temps, auprès de populations difficile d'accès, composées d'individus bénéficiant de services et ayant des comportements hétérogènes concernant la fréquentation de ces services. Nous avons en particulier montré l'importance de la prise en compte de la fréquentation des lieux pour mieux comprendre les populations difficiles d'accès vis-à-vis de certaines maladies. Si la population des UD demeure un réservoir de transmission du VHC, une diminution significative de l'incidence de cette maladie devrait se poursuivre compte tenu de la baisse de la prévalence, des mesures de réduction des risques et des avancées thérapeutiques malgré une potentielle augmentation des comportements à risque des UD.

Table des figures

Liste des tableaux

Annexes

Annexe 1 : Protocole de prélèvement

INSTITUT
DE VEILLE SANITAIRE

Département des Maladies Infectieuses
Unité VIH/sida-IST-VHC-Hépatite B chronique

PROTOCOLE DU PRELEVEMENT

Le prélèvement est réalisé par auto-piqûre du bout d'un doigt, avec une micro-lancette à lame rétractable comme celle utilisée plusieurs fois par jour par les diabétiques pour le contrôle de leur glycémie.
Le sang sera déposé sur du papier buvard sur les 6 emplacements prédéterminés.

La piqûre se fera sur la main gauche pour les droitiers, ou la main droite pour les gauchers. Le prélèvement s'effectue au niveau de la dernière phalange des 3 derniers doigts ; la piqûre se fera sur le côté et non directement sur la pulpe (cf illustration)

1) Le matériel nécessaire au prélèvement est réuni sur un plateau déposé sur la table :
Une micro-lancette à lame rétractable et à usage unique, 2 compresses, un pansement autocollant ou un flacon de spray hémostatique, le buvard sur lequel l'étiquette avec le numéro d'anonymat sera déjà collée, un flacon de solution hydro-alcoolique.
La boite de déchets (containeur plastique) et la boite de séchage pour le buvard sont à proximité.

2) Mettre une dose de solution hydro-alcoolique au creux de la main, étaler le produit puis frotter les 2 mains en insistant sur les bouts de doigts, jusqu'au séchage complet de la solution.

3) Retirer la languette de sécurité de la micro-lancette.
Avec l'autre main, maintenir fermement la micro-lancette sur le côté du doigt. Appuyer sur le bouton déclencheur. Retirer la micro-lancette du doigt et la déposer dans la boite à déchets.

4) Déposer 6 gouttes de sang sur le buvard aux emplacements indiqués, en respectant la quantité. Si il n'y a pas assez de sang, masser et/ou presser doucement le bout du doigt pour l'obtention d'une goutte suffisante. Il peut éventuellement être nécessaire de repiquer (recommencer à 3 en changeant de matériel et d'emplacement)

Insérer le buvard dans une fente du couvercle de la boite de séchage en carton.

5) Après le recueil du sang,
- appliquer une compresse sur le point de ponction et appuyer avec le pouce de la même main. Quand le questionnaire est rempli, retirer la compresse et le jeter (boite à déchets) ; mettre le pansement autocollant
ou
- vaporiser du spray hémostatique sur le point de ponction.

6) Quand le questionnaire est terminé, et dans un délai d'au moins 15 minutes, l'enquêteur actionne la trappe de la boite afin que le buvard glisse au fond de la boite.

Annexe 2 : Fiche prévisite et fiche visite

Supports enquête scientifique

Fiche de prévisite

Coquelicot 2011

Étude sur la santé des usagers de drogues et leurs besoins en matière de réduction des risques

Nom de l'enquêteur : .....

Date de la prévisite : |__|__| |__|__| |__|__|__|__|
JOUR MOIS ANNÉE

Heure du rendez-vous : |__|__| h |__|__|

Nom du service visité : .....

Adresse complète du service : .....

Nom du responsable du service : .....

Numéro de téléphone du responsable du service : |__|__|__|__|__|__|__|__|__|__|

Adresse mail du responsable du service : .....@.....

Nom du référent Coquelicot (si différent du responsable) : .....

Adresse mail du référent Coquelicot : .....@.....

Compte-rendu de l'entretien avec la personne responsable

Effectif total (tous profils confondus)[1] : |__|__|__|

Nombre (ou proportion) de mineurs : |__|__|__|

Nombre (ou proportion) de non-francophones : |__|__|__|

Effectif total de personnes éligibles[2] : |__|__|__|

Lieu prévu pour les entretiens : .....

.....

.....

Ce lieu répond-il aux exigences de confidentialité ? ☐ Oui ☐ Non

Moment et lieu du contact avec les personnes : .....

.....

.....

.....

Description de la méthode de tirage aléatoire : .....

.....

.....

Inserm
Institut national de la santé et de la recherche médicale

InVS
INSTITUT DE VEILLE SANITAIRE

[1] Il s'agit du nombre de personnes différentes présentes par jour en moyenne dans le service.

[2] Il s'agit du nombre de personnes différentes ayant déjà sniffé et/ou injecté au moins un produit, au moins une fois au cours de leur vie, majeures et francophones, présentes par jour en moyenne dans le service. Par ailleurs, les consommateurs exclusifs de crack/free base par voie fumée sont également éligibles.

INSTITUT DE VEILLE SANITAIRE, 12 rue du Val d'Osne 94415 Saint-Maurice Cedex France - Tél. : 33 (0)1 41 79 67 00 - Fax : 33 (0)1 41 79 67 67
www.invs.sante.fr - **Réalisé et imprimé** par BRI - 93000 Bobigny - France - Tirage : 180 exemplaires

Supports enquête scientifique

Fiche de visite

Coquelicot 2011

Étude sur la santé des usagers de drogues et leurs besoins en matière de réduction des risques

Nom de l'enquêteur : .....

Nom du service : .....

Date de la visite : |__|__| |__|__| |__|__|__|__|
JOUR MOIS ANNÉE

Heure d'arrivée : |__|__| h |__|__| Heure de départ : |__|__| h |__|__|

Effectif total (tous profils confondus)[1] : |__|__|__|

Nombre (ou proportion) de mineurs : |__|__|__|

Nombre (ou proportion) de non-francophones : |__|__|__|

Effectif total de personnes éligibles[2] : |__|__|__|

Organisation des contacts : .....

.....

.....

Description de la méthode de tirage aléatoire : .....

.....

.....

Lieu de passation du questionnaire : .....

.....

.....

Ce lieu répond-il aux aux conditions de confidentialité ? ☐ Oui ☐ Non

Autres informations : .....

.....

.....

.....

Inserm
Institut national de la santé et de la recherche médicale

InVS
INSTITUT DE VEILLE SANITAIRE

[1] Il s'agit du nombre total de personnes différentes dans le service sur le temps d'intervention.

[2] Il s'agit du nombre de personnes différentes ayant déjà sniffé et/ou injecté au moins un produit, au moins une fois au cours de leur vie, majeures et francophones, présentes dans le service sur le temps d'intervention. Par ailleurs, les consommateurs exclusifs de crack/free base par voie fumée sont également éligibles.

INSTITUT DE VEILLE SANITAIRE, 12 rue du Val d'Osne 94415 Saint-Maurice Cedex France - Tél. : 33 (0)1 41 79 67 00 - Fax : 33 (0)1 41 79 67 67
www.invs.sante.fr - **Réalisé et imprimé** par BRI - 93000 Bobigny - France - Tirage : 1 500 exemplaires

Annexe 3 : Article publié dans la revue *Biostatistics*

Biostatistics (2015), **16**, 3, *pp.* 565–579
doi:10.1093/biostatistics/kxu061
Advance Access publication on January 18, 2015

Design-based inference in time-location sampling

LUCIE LEON*

French Institute for Public Health Surveillance, Saint-Maurice 94415, France
l.leon@invs.sante.fr

MARIE JAUFFRET-ROUSTIDE

French Institute for Public Health Surveillance, Saint-Maurice 94415, France and Cermes3, Inserm U988/UMR CNRS 8211/Ehess/Paris Descartes University, Paris

YANN LE STRAT

French Institute for Public Health Surveillance, Saint-Maurice 94415, France

Summary

Time-location sampling (TLS), also called time-space sampling or venue-based sampling is a sampling technique widely used in populations at high risk of infectious diseases. The principle is to reach individuals in places and at times where they gather. For example, men who have sex with men meet in gay venues at certain times of the day, and homeless people or drug users come together to take advantage of services provided to them (accommodation, care, meals). The statistical analysis of data coming from TLS surveys has been comprehensively discussed in the literature. Two issues of particular importance are the inclusion or not of sampling weights and how to deal with the frequency of venue attendance (FVA) of individuals during the course of the survey. The objective of this article is to present TLS in the context of sampling theory, to calculate sampling weights and to propose design-based inference taking into account the FVA. The properties of an estimator ignoring the FVA and of the design-based estimator are assessed and contrasted both through a simulation study and using real data from a recent cross-sectional survey conducted in France among drug users. We show that the estimators of prevalence or a total can be strongly biased if the FVA is ignored, while the design-based estimator taking FVA into account is unbiased even when declarative errors occur in the FVA.

Keywords: Hard-to-reach populations; Indirect sampling; Inference; Sampling weights; Time-location sampling; Venue-based sampling.

1. Introduction

Studying populations at high risk of infectious diseases is crucial to implement adequate prevention messages and interventions to reduce transmission. Drug users, men who have sex with men (MSM), sex workers, homeless people, and certain immigrants are examples of vulnerable populations particularly

*To whom correspondence should be addressed.

exposed to Hepatitis B and C, HIV, sexually transmitted infections, and other diseases. However, performing an epidemiological survey in these populations is difficult in many countries because of the illicit nature of certain practices, such as the use of drugs or prostitution. Specifically adapted sampling techniques have been developed over the past decades to survey such hard-to-reach populations (Sudman *and others*, 1988; Spreen, 1992; Thompson and Frank, 2000; Semaan *and others*, 2002; Magnani *and others*, 2005; Kalton, 1993; Tourangeau *and others*, 2014). One of these techniques, time-location sampling (TLS), also called time-space sampling or venue-based sampling, is widely used (Muhib *and others*, 2001; Stueve *and others*, 2001; Magnani *and others*, 2005), especially for surveys among MSM (Parsons *and others*, 2008; Paquette and De Wit, 2010; Paz-Bailey *and others*, 2014).

Pioneered in public health by the Centers for Disease Control and Prevention in the Young Men's Survey, TLS is a method for reaching individuals in places and at times where they congregate rather than where they live (MacKellar *and others*, 1996; Valleroy *and others*, 2000). Drug users are surveyed in specialized centers where they receive services (e.g. needle exchange, medical examinations, accommodation) (Jauffret-Roustide *and others*, 2009; Sutton *and others*, 2012). Homeless people are surveyed in support centers offering accommodation, care or free meals, or are surveyed in street locations (Chew *and others*, 2013). MSM are surveyed in gay venues (e.g. bars, clubs, saunas, etc.) (Wejnert *and others*, 2013).

Some authors have considered issues to ensure the validity of TLS in producing unbiased estimates in terms of proportions of individuals covered by surveys, the duration of the sampling period, the eligibility and the range (in terms of number of visits) of the venues, and the "representativeness" of the resulting sample (Stueve *and others*, 2001; Pollack *and others*, 2005; MacKellar *and others*, 2007; Parsons *and others*, 2008). The heterogeneity of the frequencies of venue attendance (FVA), also referred to as multiplicity, has often been highlighted and remains a major point of debate with respect to the efficacy and accuracy of TLS. Some individuals visit only one venue during the course of a survey while others visit dozens of venues in different places at different times. Literature has shown that among MSM these frequencies are heterogeneous from one individual to another, depending on several individual characteristics (Gustafson *and others*, 2013).

The first objective of this paper is to present TLS in the context of statistical sampling theory, which to our knowledge, has never yet been described. Although, some authors have introduced TLS as a "multi-step" procedure (Stueve *and others*, 2001; Pollack *and others*, 2005), it has only recently been presented as a two-stage or three-stage sampling design (Karon and Wejnert, 2012). Some authors still consider TLS a non-random sampling technique (Meyer and Wilson, 2009) and others have raised the question about whether it is necessary to weight or not to weight in TLS (Jenness *and others*, 2011; Xia and Torian, 2013; Risser and Montealegre, 2014). Our second objective is to investigate this latter point by introducing sampling weights which incorporate the FVA in a design-based estimator as an alternative to a recently proposed model-assisted estimator (Gustafson *and others*, 2013). Our estimator uses the indirect sampling framework and the generalized weight share method (GWSM) (Lavallée, 1995, 2007). The properties of an estimator ignoring the FVA and of the design-based estimator which takes FVA into account were assessed both by a simulation study and by using data from a national cross-sectional survey carried out in France among drug users in 2011. In addition, we explored the behavior of the alternative design-based estimator when errors occur in the FVA.

2. TIME-LOCATION SAMPLING

We focus on a population of individuals, named B, attending centers (locations) at certain times. We consider that these centers are open at various times during the survey period. For simplicity but without loss of generality, we consider that the opening time unit for the centers is a half-day. The following is also valid for other populations, irrespective of the type of center, the number of centers, and the time unit.

2.1 *Sampling design and sampling weights*

TLS can be viewed as a three-stage sampling design as illustrated in Figure 1. At the first stage, n centers are randomly drawn from a sampling frame of N centers indexed by l ($l = 1, \ldots, N$). At the second stage, for each center l ($l = 1, \ldots, n$) drawn at the first stage (named primary sampling unit (PSU)), we build a sampling frame of the N_l opening half-days during the survey period, indexed by k ($k = 1, \ldots, N_l$). We draw at random n_l half-days from the N_l half-days for each center l (named the secondary sampling units (SSUs)). We then establish a schedule representing each randomly drawn center and each randomly drawn half-day for the survey. To illustrate this, Figure 2 depicts an opening time schedule for 5 centers during a 4-week survey period. Finally, at the third stage, one or more investigators visit the centers according to the opening time schedule. For each center l and for each half-day k drawn, they randomly select n_{kl} among N_{kl} eligible individuals who attend these centers. Individuals represent the tertiary sampling units (TSUs).

At the first stage, either a simple random sampling without replacement (SRSWR) or an unequal random sampling without replacement is used. For the latter, the inclusion probability of a center is proportional to an available quantitative auxiliary variable, e.g. the average daily number of individuals attending the center. At the second and third stages, SRSWR is widely used. In most cases, the investigator does not have any list of individuals when arriving at a center. Systematic sampling is then often chosen as follows: the investigator randomly draws a person who arrives at the center, then selects the other individuals according to their ranking order of arrival using a sampling fraction defined *a priori*. Sometimes, a stratified sampling can also be employed, e.g. individuals are stratified by sex, age groups, nationalities, or any other characteristics of interest. The random selections of the sampling units at each stage (centers, half-days, individuals) aim to reduce selection biases.

To make inference in the population from the random sample, a sampling weight is assigned to each surveyed individual. The (first-order) inclusion probability for a unit is equal to the probability that this unit belongs to the sample. A sampling weight defined as the inverse of an inclusion probability can be expressed as the product of the sampling weights calculated at each stage of the design. We introduce the notation of the inclusion probabilities in Table 1 (column 2), under the assumption that an SRSWR is used at each stage. At the first stage, the sampling weight of a center l is $w_l = 1/\pi_l$. At the second stage, the sampling weight of a half-day k for the center l is $w_{k|l} = 1/\pi_{k|l}$. At the third stage, the sampling weight of an individual i surveyed in the center l during the half-day k is $w_{i|kl} = 1/\pi_{i|kl}$. The final inclusion probability of an individual i is $\pi_i = \pi_l \times \pi_{k|l} \times \pi_{i|kl}$ and his/her final sampling weight is:

$$w_i = w_l \times w_{k|l} \times w_{i|kl}. \tag{2.1}$$

2.2 *The Horvitz–Thompson estimator*

Very often, the main objective of cross-sectional surveys including time-location surveys is to estimate parameters of interest such as a total (e.g. population size, number of infected individuals), a proportion (e.g. proportion of infected individuals, called prevalence), or a mean (e.g. the mean value of a biomarker). For each individual i in the population B, let us consider a binary variable of interest y_i corresponding to his/her serological status for the disease of interest: y_i equals 1 if i is infected and 0 if not.

The Horvitz–Thompson estimator (Horvitz and Thompson, 1952) of the total number of infected individuals in the population $T = \sum_{i \in B} y_i$ is:

$$\hat{T} = \sum_{i \in s^B} w_i y_i, \tag{2.2}$$

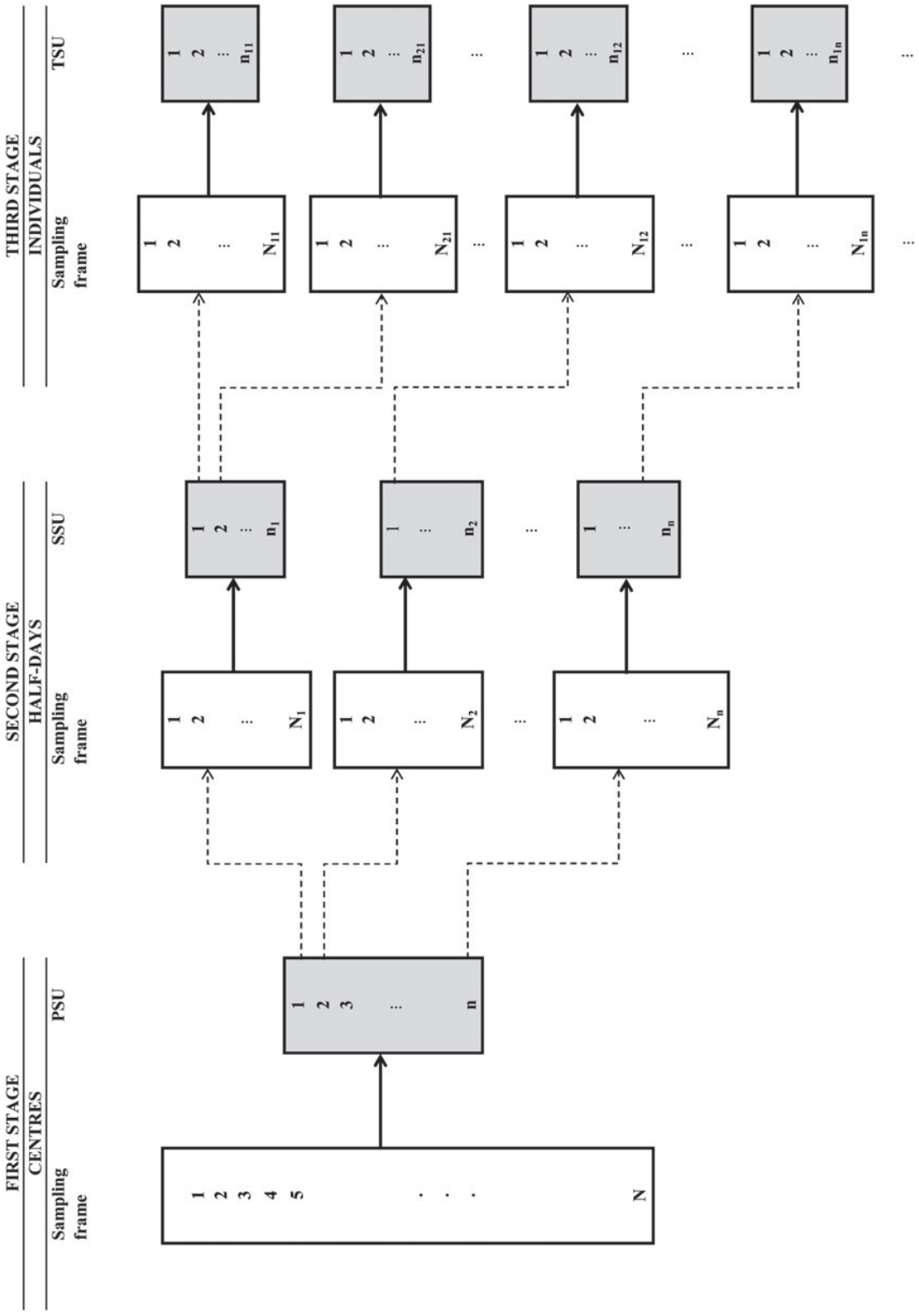


Fig. 1. Three-stage sampling design. Bold arrows represent the drawings and dashed arrows represent the sampling frames built for the units drawn at the first and second stages. PSU, primary sampling unit; SSU, secondary sampling unit; TSU, tertiary sampling unit.

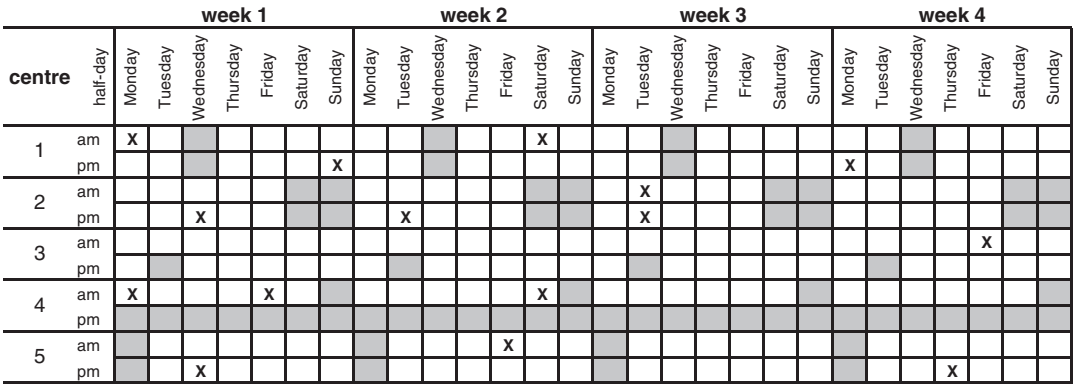


Fig. 2. Schedule for 5 randomly drawn centers in a 4-week time-location survey. Centers are visited during the randomly drawn half-days (cross squares) among opening half-days (white squares). Gray squares represent the closing half-days.

Table 1. *Inclusion probabilities and totals expressions under the SRSWR assumption used at each stage of a three-stage sampling design*

Stage	First-order	Second-order†	Δ quantities	Totals
1	$\pi_l = \frac{n}{N}$	$\pi_{ll'} = \frac{n}{N}\left(\frac{n-1}{N-1}\right)$	$\Delta_{ll'} = \pi_{ll'} - \pi_l\pi_{l'}$	$T = \sum_{l=1}^{N} t_l$
2	$\pi_{k\vert l} = \frac{n_l}{N_l}$	$\pi_{kk'\vert l} = \frac{n_l}{N_l}\left(\frac{n_l-1}{N_l-1}\right)$	$\Delta_{kk'\vert l} = \pi_{kk'\vert l} - \pi_{k\vert l}\pi_{k'\vert l}$	$t_l = \sum_{k=1}^{N_l} t_{k\vert l}$
3	$\pi_{i\vert kl} = \frac{n_{kl}}{N_{kl}}$	$\pi_{ii'\vert kl} = \frac{n_{kl}}{N_{kl}}\left(\frac{n_{kl}-1}{N_{kl}-1}\right)$	$\Delta_{ii'\vert kl} = \pi_{ii'\vert kl} - \pi_{i\vert kl}\pi_{i'\vert kl}$	$t_{k\vert l} = \sum_{i=1}^{N_{kl}} y_i$

†$\pi_{ll} = \pi_l$; $\pi_{kk\vert l} = \pi_{k\vert l}$; $\pi_{ii\vert kl} = \pi_{i\vert kl}$.

where s^B is the sample drawn from the population B using TLS described above. The population size N^B, which is unknown in most cases, in particular for hard-to-reach individuals, is estimated by $\hat{N}^B = \sum_{i\in s^B} w_i$. The prevalence $P = T/N^B$ is estimated by:

$$\hat{P} = \frac{\hat{T}}{\hat{N}^B}. \tag{2.3}$$

The variance of a total estimator (Särndal *and others*, 2003), with respect to the sampling design, is estimated using the second-order inclusion probabilities (which constitute the joint inclusion probability of 2 distinct units) and other notations introduced to simplify the following formula (see Table 1, columns 3–5):

$$\widehat{Var}(\hat{T}) = \sum_{l=1}^{n}\sum_{l'=1}^{n} \Delta_{ll'} \frac{\hat{t}_l}{\pi_l}\frac{\hat{t}_{l'}}{\pi_{l'}} + \sum_{l=1}^{n} \frac{\widehat{Var}(\hat{t}_l)}{\pi_l} + \sum_{l=1}^{n}\frac{1}{\pi_l}\sum_{k=1}^{n_l}\frac{\widehat{Var}(\hat{t}_{k\vert l})}{\pi_{k\vert l}} \tag{2.4}$$

where l and l' denote 2 distinct centers, k and k' denote 2 distinct half-days, i and i' denote 2 distinct individuals and where

$$\hat{t}_l = \sum_{k=1}^{n_l} \frac{\hat{t}_{k|l}}{\pi_{k|l}}, \quad \hat{t}_{k|l} = \sum_{i=1}^{n_{kl}} \frac{y_i}{\pi_{i|kl}},$$

$$\widehat{\mathrm{Var}}(\hat{t}_l) = \sum_{k=1}^{n_l}\sum_{k'=1}^{n_l} \Delta_{kk'|l} \frac{\hat{t}_{k|l}}{\pi_{k|l}} \frac{\hat{t}_{k'|l}}{\pi_{k'|l}} \quad \text{and} \quad \widehat{\mathrm{Var}}(\hat{t}_{k|l}) = \sum_{i=1}^{n_{kl}}\sum_{i'=1}^{n_{kl}} \Delta_{ii'|kl} \frac{y_i}{\pi_{i|kl}} \frac{y_{i'}}{\pi_{i'|kl}}.$$

$\widehat{\mathrm{Var}}(\hat{N}^B)$ is calculated in a similar way by using (2.4) and assuming that $y_i = 1$ for any $i \in B$.

The estimated variance of the estimated prevalence is:

$$\widehat{\mathrm{Var}}(\hat{P}) = \widehat{\mathrm{Var}}\left(\frac{\hat{T}}{\hat{N}^B}\right) = \frac{1}{\hat{N}^{B2}}\{\widehat{\mathrm{Var}}(\hat{T}) - 2\hat{P}\,\widehat{\mathrm{Cov}}(\hat{T}, \hat{N}^B) + \hat{P}^2\widehat{\mathrm{Var}}(\hat{N}^B)\}, \tag{2.5}$$

where

$$\widehat{\mathrm{Cov}}(\hat{T}, \hat{N}^B) = \sum_{l=1}^{n}\sum_{l'=1}^{n} \Delta_{ll'} \frac{\hat{t}_l}{\pi_l} \frac{\hat{N}_l}{\pi_{l'}} + \sum_{l=1}^{n} \frac{\widehat{\mathrm{Cov}}(\hat{t}_l, \hat{N}_l)}{\pi_l} + \sum_{l=1}^{n} \frac{1}{\pi_l} \sum_{k=1}^{n_l} \frac{\widehat{\mathrm{Cov}}(\hat{t}_{k|l}, \hat{N}_{k|l})}{\pi_{k|l}} \tag{2.6}$$

with

$$\hat{N}_l = \sum_{k=1}^{n_l} \frac{\hat{N}_{k|l}}{\pi_{k|l}}, \quad \hat{N}_{k|l} = \sum_{i=1}^{n_{kl}} \frac{1}{\pi_{i|kl}}, \quad \widehat{\mathrm{Cov}}(\hat{t}_l, \hat{N}_l) = \sum_{k=1}^{n_l}\sum_{k'=1}^{n_l} \Delta_{kk'|l} \frac{\hat{t}_{k|l}}{\pi_{k|l}} \frac{\hat{N}_{k'|l}}{\pi_{k'|l}}$$

and

$$\widehat{\mathrm{Cov}}(\hat{t}_{k|l}, \hat{N}_{k|l}) = \sum_{i=1}^{n_{kl}}\sum_{i'=1}^{n_{kl}} \Delta_{ii'|kl} \frac{y_i}{\pi_{i|kl}} \frac{1}{\pi_{i'|kl}}.$$

Note that if the second-order inclusion probabilities are easy to calculate when using SRSWR, their calculation is more complicated and sometimes intractable with other samplings and depend on the sampling algorithms used (Tillé, 2006). With more complex sampling designs, variances may be estimated using jackknife or bootstrap procedures (Särndal *and others*, 2003).

The Horvitz–Thompson estimator is unbiased for any sampling design if $\pi_i > 0$ for all $i \in B$ and of course if the inclusion probabilities are correctly calculated. For a population whose individuals attend several centers delivering services, the calculation of inclusion probabilities is more challenging than that for a population whose individuals are more static in time and space and who can be selected only once at most. In a time-location survey, the inclusion probability of an individual depends on his/her FVA.

In order to collect this information on FVA, we ask the respondents a set of questions to discover which venues they attend. One of the questions may be for example "how often did you go to any of the following venues during the previous 5 days?". Other more detailed questions may be asked according to the type of center (Gustafson *and others*, 2013). Then, the number of centers attended by each individual can be taken into account in a new estimator. As an alternative to the Horvitz–Thompson estimators ((2.2) and (2.3)) which can be biased when FVA is heterogeneous, we propose an unbiased design-based estimator incorporating the FVA. As this estimator is developed within the framework of indirect sampling, we firstly introduce the indirect sampling and then develop the new estimator.

3. Indirect sampling

Let us consider a population A containing N^A units indexed by j ($j = 1, \ldots, N^A$) and the population of interest B in which we want to estimate a function of interest (proportion, total) that contains N^B units indexed by i ($i = 1, \ldots, N^B$). A link between these 2 populations A and B is defined as the correspondence between any unit $j \in A$ with any unit $i \in B$ which allows switching back and forth between A and B. Indirect sampling designates a sampling for which: (1) a sample of units $j \in A$ named s^A is randomly drawn to access the units $i \in B$ and, (2) the units $i \in B$ are linked to, at least, one unit $j \in A$ (Deville and Lavallée, 2006; Lavallée, 2007). The correspondence between the 2 populations can be represented by a link matrix L of size $N^A \times N^B$. Each element $l_{ji} \geqslant 0$ defines the link between $i \in B$ and $j \in A$ and, if there is no link between the units, this quantity is 0. To illustrate indirect sampling, let consider a population A of 5 units and a population B of 3 units represented in Figure 1 in Section S.1 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org). The link matrix is:

$$L = \begin{pmatrix} l_{11} & 0 & 0 \\ 0 & l_{22} & 0 \\ l_{31} & 0 & 0 \\ 0 & 0 & l_{43} \\ 0 & 0 & l_{53} \end{pmatrix}.$$

A link l_{ji} is (1) bijective if a unit $i \in B$ has a one-to-one link with a unit $j \in A$, (2) injective if a unit $i \in B$ has at most one link with a unit $j \in A$, or (3) surjective if a unit $i \in B$ has at least one link with a unit $j \in A$.

Therefore, TLS can be viewed as a three-stage indirect sampling design where, at the third stage, population A is the population of services, population B is the population of individuals who receive these services in the centers and where the FVA of individuals is equal to the sum of the links between A and B, as illustrated in Section S.2 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org). The population of services offered by the centers exists but a list associated with this population is not available, except in special cases (e.g. accommodation). The estimator introduced in the following section is theoretically based on services received by the individuals surveyed. We will see in Section 5 how this estimator is calculated in practice even when we do not know which specific services individuals benefit from, and have only information about the centers they visited and how often.

In the framework of indirect sampling, we can use the GWSM, first described by Lavallée (1995), to provide a relevant sampling weight for each individual interviewed. A new sampling weight is assigned to each individual $i \in s^B$, basically defined as a weighted arithmetic mean of the sampling weights of the population of services involving the links between i and the services he/she received.

4. The alternative design-based estimator

The sampling weight of a service $j \in s^A$ is $w_j = w_i$ as defined in (2.1). If unit $j \in s^A$ is linked to unit $i \in s^B$, $l_{ji} \geqslant 0$ and if these 2 units are not related to each other, $l_{ji} = 0$. Note that some authors have highlighted the importance of the choices of link values that influence the precision of the estimates issued from indirect sampling, even though, in most applications, the values of l_{ji} for the linked units are equal to one (Lavallée and Caron, 2001; Deville and Lavallée, 2006). Then, for each unit $i \in s^B$, we can calculate the total number of links $L_i^B = \sum_{j \in A} l_{ji}$.

Finally, the final sampling weight incorporating the FVA for each unit $i \in s^B$ is defined as:

$$\tilde{w}_i = \frac{1}{L_i^B} \sum_{j \in s^A} l_{ji} w_i. \tag{4.1}$$

The alternative design-based estimators for the totals T and N^B and the prevalence P are, respectively:

$$\hat{T}_G = \sum_{i \in s^B} \tilde{w}_i y_i, \tag{4.2}$$

$$\hat{N}_G^B = \sum_{i \in s^B} \tilde{w}_i, \tag{4.3}$$

$$\hat{P}_G = \frac{\hat{T}_G}{\hat{N}_G^B}. \tag{4.4}$$

It has been demonstrated that these estimators are unbiased (Lavallée, 2007). Their respective variances are estimated using the same expressions proposed in (2.4)–(2.6) with

$$\widehat{\mathrm{Var}}(\hat{t}_{k|l}) = \sum_{j=1}^{n_{kl}} \sum_{j'=1}^{n_{kl}} \Delta_{ii'|kl} \frac{z_j}{\pi_{i|kl}} \frac{z_{j'}}{\pi_{i'|kl}},$$

$$\widehat{\mathrm{Cov}}(\hat{t}_{k|l}, \hat{N}_{k|l}) = \sum_{j=1}^{n_{kl}} \sum_{j'=1}^{n_{kl}} \Delta_{ii'|kl} \frac{z_j}{\pi_{i|kl}} \frac{1}{\pi_{i'|kl}}, \quad \text{and} \quad z_j = \sum_{i=1}^{n_i} \frac{l_{ji}}{L_i^B} y_i.$$

In real life, it is often pointless to ask participants what specific services they used or even their total number of visits over a survey period. First, individuals may be hesitant to answer this question. They go to centers for particular reasons and do not see why is it of any interest to spend time trying to remember what they did in the past, especially after a potentially long interview. This question can also be viewed by respondents as a check on illicit practices. Second, participants may find it difficult to answer such questions accurately due to forgetfulness or confusion as regards center identification. This is more marked in precarious populations or when there is a large number of centers (e.g. in a big city). Finally, the practical conditions of the interview rarely allow the collect of such detailed information, for example, when administering a questionnaire in the street or in a squat.

For all these reasons, researchers ask few questions about FVA over a short past period. This point is developed in greater detail in the next section.

However, it is important to note that we do not need detailed information about the services individuals benefit from to calculate the design-based estimator. Indeed, individuals are generally randomly drawn when they arrive at a center, irrespective of whether or not they are going to benefit from one or more services. Their inclusion probabilities therefore do not depend on the number of services they receive but on their number of visits to centers. Accordingly, we simply need to count the number of their visits at different centers.

Now, we will illustrate the properties of both the established Horvitz–Thompson and the alternative design-based estimators first using a cross-sectional survey (French ANRS-Coquelicot survey) conducted in 2011 and then using a comprehensive simulation study.

5. French ANRS-Coquelicot survey

5.1 *Design*

The French ANRS-Coquelicot survey was conducted in 2004 (Jauffret-Roustide *and others*, 2009) and in 2011 (Jauffret-Roustide *and others*, 2013) among drug users residing in metropolitan cities in France, to estimate the prevalence of hepatitis C virus (HCV) infection (based on serum testing), to assess the frequencies of at-risk practices and to follow the dynamics of the epidemic. In each city, we performed a comprehensive inventory of all centers providing services to drug users as follows: accommodation services including residential centers, hotel rooms, "sleep-in" centers (French social service accommodation centers), drug treatment centers including those providing methadone maintenance and psychotherapy, low threshold services including needle exchange programs and outreach work teams. We then constructed a sampling frame by each half-day that centers were open.

A two-stage TLS was used. All listed centers participated in the survey. At the first stage, we selected half-days in all centers using an SRSWR. At the second stage, at each center/half-day visit, drug users were selected using systematic random sampling (except for residential centers where all users were included in the survey). Participants were included if they provided written consent to be interviewed and to provide a self-obtained finger-prick blood samples in the form of a dried blood spot for HCV testing. Inclusion criteria for the survey were: $>$ 18 years of age, injected or snorted drugs "at least once during one's life", spoke French and agreed to participate in the survey by providing written, informed consent. The study questionnaire lasted approximately 45 min and was administered by professional interviewers with no ties to the recruitment centers. Interviewers had been trained for hard-to-reach populations and especially for interviews with drug users, in order to minimize social desirability bias associated with drug consumption and at-risk practices. We included 1568 drug users in the ANRS-Coquelicot study in 2011.

5.2 *Data collection regarding FVA*

As mentioned above, collecting an accurate list of services which each participant benefits from or the visits he/she makes over the whole survey period is unrealistic. Most researchers focus on asking few questions regarding FVA with some restrictions: the frequency of attendance is collected as a discrete variable (with some categories), over a short past period and sometimes using a limited number of centers (Karon and Wejnert, 2012; Gustafson *and others*, 2013).

In our survey, we asked 2 questions about FVA: (1) Yesterday and in the previous 3 days, did you attend one or more centers? If so, where and how many times? (2) Including the center where we are now, what other center or centers have you already attended today or do you intend on attending today? The FVA distribution in this survey is represented in Figure 4 in Section S.4.1 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org).

5.3 *Results*

In 2011, the survey was conducted over 11 weeks (May–July) in 121 centers and 1568 drug users were interviewed. The Horvitz–Thompson estimate of the population size was $\hat{T} = 48\,147$ (95% confidence interval [43 741; 52 553]) individuals while our design-based estimate was $\hat{T}_G = 43\,710$ (95% confidence interval [39 667; 47 753]). The Horvitz–Thompson estimate of the HCV prevalence among drug users was $\hat{P} = 43.4\%$, (95% confidence interval [39.3%; 47.6%]) while our design-based estimate was $\hat{P}_G = 43.7\%$, (95% confidence interval [39.5%; 47.9%]). The 2 estimates are close, probably because of the low variance of the number of links declared by drug users. This low observed variance may be due to measurement. We assumed that FVA did not vary over the 11-week survey period. We therefore decided to only collect data on FVA for the previous 5 days. Furthermore, drug users may be reluctant to answer to these questions

Table 2. *Parameters used to generate erroneous links*

Error	$L_i^{B,error}$	k
1	$L_i^B \times (k+1)$	$k \sim \mathcal{U}(-0.5, 0.5)$
2	$L_i^B + k$	$k \in [-(L_i^B - 1); L_i^B]$
3	$L_i^B \times (k+1)$	$k \sim \mathcal{U}(-0.5, 0)$

for the reasons described in Section 4. Errors in the declared FVA may occur, leading to a possible underestimation of variance.

To investigate the impact of these errors on the estimates in greater detail, we conducted a simulation that we present in the following section.

6. SIMULATION STUDY

6.1 *Simulation process*

We generated several populations of individuals attending centers to benefit from one or more services during a fixed period. These simulated populations have prevalences ranging from 1% to 90% and the number of links (e.g. the FVA) depends or not on the serological status of each individual. Then, we generated 10 000 samples from each population. For each sample generated: $\hat{N}^B$, $\hat{T}$, $\hat{P}$, $\hat{N}_G^B$, $\hat{T}_G$, $\hat{P}_G$ were calculated. To explore the properties of our design-based estimator when errors occur in the FVA, we generated 3 kinds of errors, presented in Table 2. More details on the simulation process are given in Section S.3 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org).

6.2 *Results*

The simulation study shows that, for any scenario, the design-based estimator is unbiased irrespective of the prevalence (Figure 3, and Figures in Section S.6 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org)). On the contrary, the Horvitz–Thompson is biased for several scenarios, particularly for scenarios 13–16 where the FVA depends on serological status, as illustrated in Figure 3 for estimated prevalences and in Tables of Section S.5 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org) for estimated population sizes.

Figure 4 presents the relative bias for all the estimated prevalences. For scenarios 13–16, the estimated prevalences from the Horvitz–Thompson estimator, despite being unbiased such as those from our design-based estimator, are 1.05–2.22 times higher than the true prevalence.

Coverage probabilities of the estimated prevalences ranged from 87% to 100% using the alternative design-based estimator and from 0% (scenarios 13–16) to 95% using the Horvitz–Thompson estimator (values represented by circles in Section S.7 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org)).

When errors occurred in the declared FVA, we observed a low bias (scenarios 13–16) or sometimes an absence of bias (scenarios 1–12) in the estimations of prevalence using the alternative design-based estimator (Figures available in Section S.8 of supplementary material available at *Biostatistics* online (http://biostatistics.oxfordjournals.org)) irrespective of the true prevalence in the population. We expected that the observed bias would increase due to the kinds of errors introduced in the FVA and presented in Table 2 (link error 1 $\leqslant$ link error 3 $\leqslant$ link error 2) as illustrated in Section S.9 of supplementary material

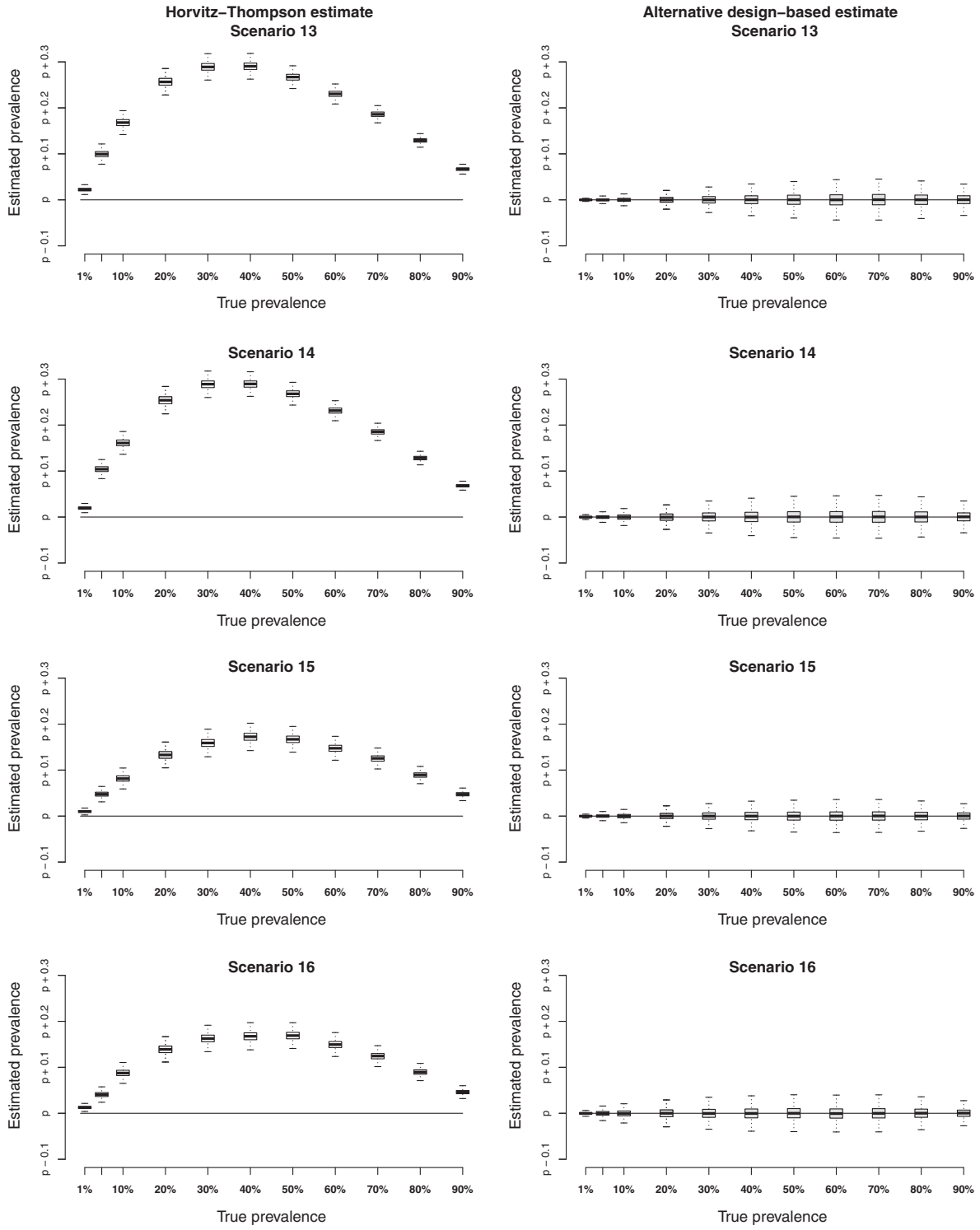


Fig. 3. Boxplots of estimated prevalences from the Horvitz–Thompson estimator (left) and from the alternative design-based estimator (right) for the scenarios 13–16. On each graph, the straight line represents the true prevalence in the simulated population for each scenario.

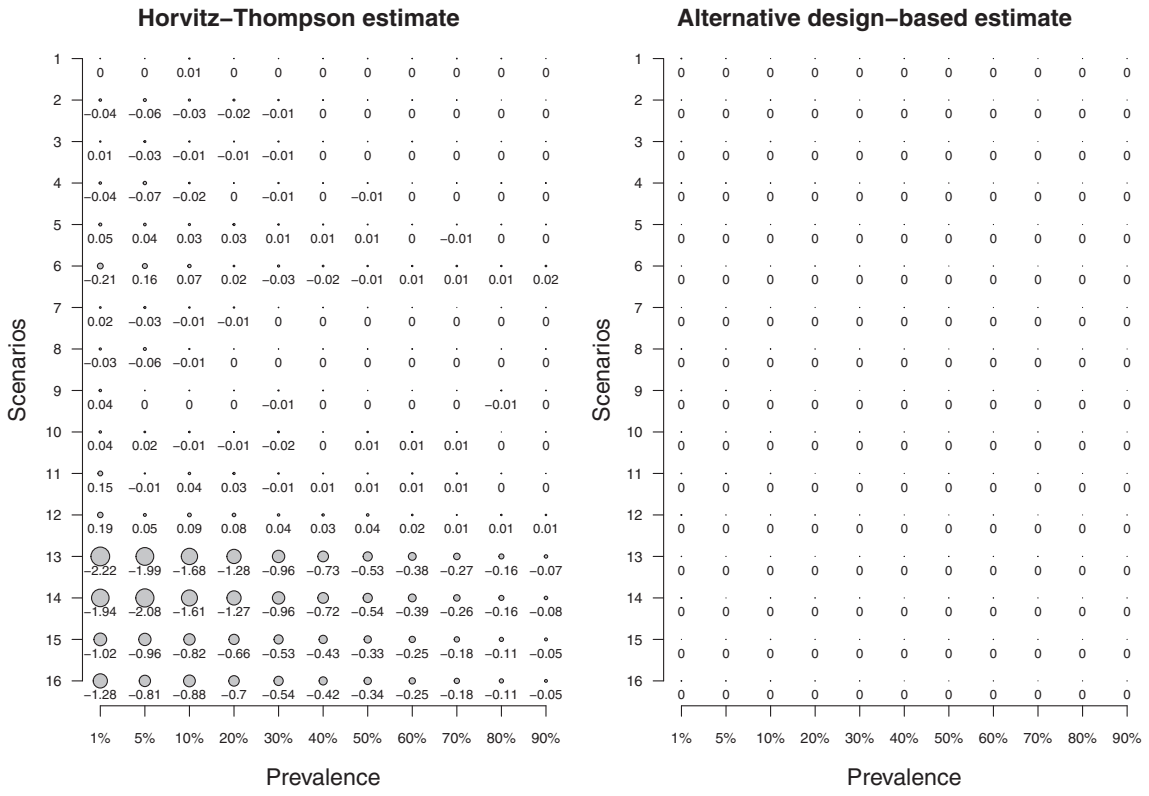


Fig. 4. Relative bias represented by the circles according to the scenarios and the different prevalences for the Horvitz–Thompson (left) and the alternative design-based (right) estimates.

available at *Biostatistics* online (http://biostatistics.oxfordjournals.org) when we estimated the number of infected individuals using both the Horvitz–Thompson and alternative design-based estimators.

7. Discussion

We presented and implemented TLS as a multi-stage indirect sampling design and proposed a design-based estimator using the GWSM to provide accurate estimations for a total or a proportion when the population of interest in a survey is hard-to-reach and frequents specific venues. This design-based estimator takes into account the FVA of individuals, which was sometimes heterogeneous.

In the Coquelicot survey, the design-based estimator we proposed was adjusted for visits and showed results similar to those found using established Horvitz–Thompson estimator, due to the low variance in the FVA declared by participants. We did not carefully investigate why this observed variance was low but can put forward several explanations. First of all, the variance in the studied population of drug users was low. This is not the most likely explanation however as the participants had heterogeneous characteristics, particularly in terms of drug usage/consumption and therefore we expected them to have heterogeneous FVA. If this assumption is true, there is no benefit to using our estimator instead of the Horvitz–Thompson estimator in order to estimate proportions. However, the benefit is positive and real when estimating a total which must include the FVA.

A second most likely assumption is that the true variance is higher than that observed in the sample due to the difficulty in accurately collecting FVA. Indeed, participants with a great number of visits are not interested in spending time trying to recollect all their visits. The consequence is underestimated variance.

There is probably no perfect way to collect accurate information on FVA. It depends on the population studied and the surveyed locations. Future specific studies are needed to propose guideline questions to include related to FVA in questionnaires used in time-location surveys.

In the simulation study, we proposed different scenarios to cover several hard-to-reach populations with several prevalence values and with several FVA depending or not on the serological status. We concluded that collecting data on FVA during a face-to-face interview is crucial to modify the sampling weights in order to build an unbiased estimator. Even if errors occur in the FVA, the bias is reduced. Instead, ignoring FVA leads to severe bias and a weak coverage probability, in particular when FVA depends on serological status.

Our simulation mainly focused on the impact of FVA on the estimator bias. We did not investigate how other sources of bias could play a role in the robustness of the alternative design-based estimator. However, even if other sources of bias exist, our alternative estimator should always outperform the established Horvitz–Thompson estimator when FVA bias exists.

Furthermore, it could be interesting in a future extension of this study to compare our design-based estimator with the model-based estimator developed by Gustafson *and others* (2013) which focuses on simulated data, and to discuss the pros and the cons of these 2 estimators. From the results of the present study, we can already state that the use of indirect sampling coupled with the GWSM could solve several of the problems encountered in phone surveys when multiple communication with the same person because of both landline and mobile telephoning, must be taken into account.

8. Software

The ANRS-Coquelicot surveys were analyzed using STATA 12.1. The simulation study was implemented using the R software package (R version 3.0.2).

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors thank the reviewer and the associate editors for their insight and helpful comments.
Conflict of Interest: None declared.

References

Chew, Ng. R. A., Muth, S. Q. and Auerswald, C. L. (2013). Impact of social network characteristics on shelter use among street youth in san francisco. *Journal of Adolescent Health* **53**, 381–386.

Deville, J. C. and Lavallée, P. (2006). Indirect sampling: the foundations of the Generalised Weight Share Method. *Survey Methodology* **32**, 165–176.

Gustafson, P., Gilbert, M., Xia, M., Michelow, W., Robert, W., Trussler, T., McGuire, M., Paquette, D., Moore, D. M. and Gustafson, R. (2013). Impact of statistical adjustment for frequency of venue attendance in a venue-based survey of men who have sex with men. *American Journal of Epidemiology* **177**(10), 1157–1164.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

JAUFFRET-ROUSTIDE, M., LE STRAT, Y., COUTURIER, E., THIERRY, D., RONDY, M., QUAGLIA, M., RAZAFANDRATSIMA, N., EMMANUELLI, J., GUIBERT, G., BARIN, F. *and others*. (2009). A national cross-sectional study among drug-users in France: epidemiology of HCV and highlight on practical and statistical aspects of the design. *BMC Infectious Diseases* **9**, 113.

JAUFFRET-ROUSTIDE, M., PILLONEL, J., WEILL-BARILLET, L., LÉON, L., LE STRAT, Y., BRUNET, S., BENOIT, T., CHAUVIN, C., LEBRETON, M., BARIN, F. *and others*. (2013). Estimation de la séroprévalence du VIH et de l'hépatite C chez les usagers de drogues en France - premiers résultats de l'enquête ANRS-COQUELICOT 2011. *Bulletin Epidémiologique Hebdomadaire* **39–40**, 504–509.

JENNESS, S. M., NEAIGUS, A., MURRILL, C. S., GELPI-ACOSTA, C., WENDEL, T. AND HAGAN, H. (2011). Recruitment-adjusted estimates of HIV prevalence and risk among men who have sex with men: effects of weighting venue-based sampling data. *Public Health Reports* **126**, 635–642.

KALTON, G. (1993). Sampling considerations in research on HIV risk and illness. *Methodological Issues in AIDS Behavioral Research*. New York: Plenum Press.

KARON, J. M. AND WEJNERT, C. (2012). Statistical methods for the analysis of time-location sampling data. *Journal of Urban Health* **89**, 565–586.

LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology* **21**, 25–32.

LAVALLÉE, P. (2007). *Indirect Sampling*. New York: Springer.

LAVALLÉE, P AND CARON, P. (2001). Estimation using the Generalised Weight Share Method: the case of record linkage. *Survey Methodology* **27**, 155–169.

MACKELLAR, D. A., GALLAGHER, K. M., FINLAYSON, T., SANCHEZ, T., LANSKY, A. AND SULLIVAN, P. S. (2007). Surveillance of HIV risk and prevention behaviors of men who have sex with men—a national application of venue-based, time-space sampling. *Public Health Reports* **122**(Suppl. 1), 39–47.

MACKELLAR, D., VALLEROY, L., KARON, J., LEMP, G. AND JANSEN, R. (1996). The young men's survey: methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports* **111**, 138–144.

MAGNANI, R., SABIN, K., SAIDEL, T. AND HECKATHORN, D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* **19** (Suppl. 2), S67–S72.

MEYER, I. H. AND WILSON, P. A. (2009). Sampling lesbian, gay, and bisexual populations. *Journal of Counseling Psychology* **56**, 23–31.

MUHIB, F. B., LIN, L. S., STUEVE, A., MILLER, R. L., FORD, W. L., JOHNSON, W. D. AND SMITH, P. J. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports* **116** (Suppl. 1), 216–222.

PAQUETTE, D. AND DE WIT, J. (2010). Sampling methods used in developed countries for behavioural surveillance among men who have sex with men. *AIDS and Behavior* **14**, 1252–1264.

PARSONS, J. T., GROV, C. AND KELLY, B. C. (2008). Comparing the effectiveness of two forms of time-space sampling to identify club drug-using young adults. *Journal of Drug Issues* **38**, 1061–1082.

PAZ-BAILEY, G., PHAM, H., OSTER, A. M., LANSKY, A., BINGHAM, T., WIEGAND, R. E., DINENNO, E., SKARBINSKI, J. AND HEFFELFINGER, J. D. (2014). Engagement in HIV care among HIV-positive men who have sex with men from 21 cities in the United States. *AIDS and Behavior* **18** (Suppl. 3), 348–358.

POLLACK, L. M., OSMOND, D. H., PAUL, J. P. AND CATANIA, J. A. (2005). Evaluation of the center for disease control and prevention's HIV behavioral surveillance of men who have sex with men: sampling issues. *Sexually Transmitted Diseases* **32**, 581–589.

RISSER, J. M. AND MONTEALEGRE, J. R. (2014). Comparison of surveillance sample demographics over two cycles of the National HIV Behavioral Surveillance Project, houston, texas. *AIDS and Behavior* **18** (Suppl. 3), 382–390.

Särndal, C. E., Swensson, B. and J. Wretman, J. (2003). *Model Assisted Survey Sampling*. New York: Springer.

Semaan, S., Lauby, J. and Leibman, J. (2002). Street and network sampling in evaluation studies of HIV risk-reduction interventions. *AIDS Reviews* **4**, 213–223.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Sociological Methodology* **36**(1), 34–58.

Stueve, A., O'Donnell, L. N., Duran, R., Doval, A. S. and Blome, J. (2001). Time-space sampling in minority communities: results with young latino men who have sex with men. *American Journal of Public Health* **91**, 922–926.

Sudman, S., Sirken, M. G. and Cowan, C. D. (1988). Sampling rare and elusive populations. *Science* **240**, 991–996.

Sutton, A. J., McDonald, S. A., Palmateer, N., Taylor, A. and Hutchinson, S. J. (2012). Estimating the variability in the risk of infection for hepatitis C in the Glasgow injecting drug user population. *Epidemiology and Infection* **140**, 2190–2198.

Thompson, S. K. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology* **26**, 87–98.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M. and Bates, N. (2014). *Hard-to-Survey Populations*. Cambridge: Cambridge University Press.

Valleroy, L. A., MacKellar, D. A., Karon, J. M., Rosen, D. H., McFarland, W., Shedan, D. A., Stoyanoff, S. R., LaLota, M., Celentano, D. D., Koblin, B. A. *and others*. (2000). HIV prevalence and associated risks in young men who have sex with men. *Journal of the American Medical Association* **284**, 198–204.

Wejnert, C., Le, B., Rose, C. E., Oster, A. M., Smith, A. J., Zhu, J. and Paz-Bailey, G. for the NHBS study group. (2013). HIV infection and awareness among men who have sex with men-20 cities, United States, 2008 and 2011. *Plos One* **8**(10), 1–9.

Xia, Q. and Torian, L. V. (2013). To weight or not to weight in time-location sampling: why not do both? *AIDS and Behavior* **17**, 3120–3123.

[*Received July 8, 2014; revised December 15, 2014; accepted for publication December 17, 2014*]

Biostatistics (2014), **0**, 0, *pp.* 1–23
doi:10.1093/biostatistics/2014˙Biostatistics˙SupplMat˙R2

Design-based inference in time-location sampling : Supplementary Materials

LUCIE LEON*, MARIE JAUFFRET-ROUSTIDE, YANN LE STRAT
French Institute for Public Health Surveillance, France
l.leon@invs.sante.fr

S.1. Example of links

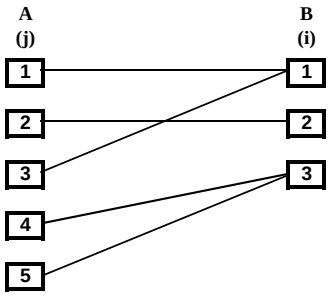


Fig. 1. Population A of 5 units and population B of 3 units with their links.

S.2. Three-stage indirect sampling design

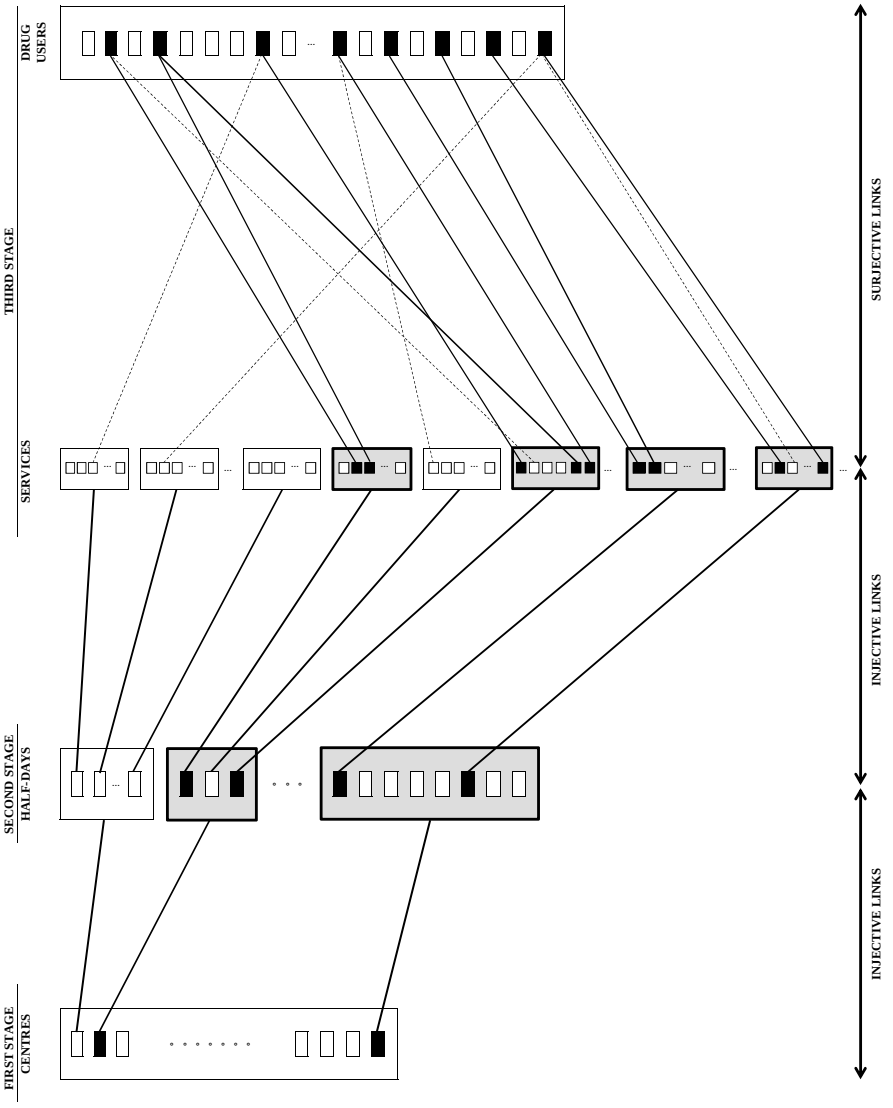


Fig. 2. Three-stage indirect sampling design. First stage: PSU (black squares) and not drawn centres (white squares). Second stage: SSU (black squares) and not drawn half-days (white squares) among centres drawn (gray rectangles). Third stage: TSU (black squares) and not drawn services (white squares) among half-days drawn (gray rectangles). Bold lines represent known links and dashed lines represent links declared by drug users.

S.3 Simulation process

We generated N centres offering services during a fixed survey period. We assumed that each centre l ($l = 1, \ldots, N$) was open for N_l half-days during the survey period and offered N_{kl} services during the half-day k ($k = 1, \ldots, N_l$). A total of $N_A = \sum_{l=1}^{N} \sum_{k=1}^{N_l} N_{kl}$ services was thus proposed by the N centres during the survey period. We noted N_B the size of the population B of individuals who may receive one or more services. First we built the sampling frames of services and individuals (steps 1-3). Secondly, we drew samples of centres, half-days and services (steps 4-6). Finally we obtain the sample of individuals (steps 7):

- **Step 1**. We built the sampling frame of services by generating a matrix with N_A rows, representing all the services and three columns identifying - for each service - the centre, the half-day and the service itself.

- **Step 2**. We generated a population of N^B individuals. For each individual i ($i = 1, \ldots, N_B$), we generated at random his/her total number of links over the survey period, L_i^B, following a negative binomial distribution of mean μ and variance $\theta\mu$ with dispersion parameter $\theta \geqslant 1$. We chose 12 sets of values for μ and θ (called scenarios 1 to 12) introduced in Table 1. Note that for the ANRS-Coquelicot survey, the FVA follows a negative binomial distribution with $(\mu, \theta) = (1, 1)$. We also generated the serological status of each individual to generate his/her number of links according to his/her status. For that, we generated the number of links of seropositive (respectively seronegative) individuals following a negative binomial distribution of mean μ_1 (respectively μ_2) and variance $\theta_1\mu_1$ (respectively $\theta_2\mu_2$). We chose 4 sets of values for these 4 parameters (called scenarios 13 to 16, Table 1). For the ANRS-Coquelicot survey, $(\mu_1, \mu_2, \theta_1, \theta_2) = (1, 1, 1, 1)$. For each scenario, we generated the serological status in order to obtain 11 prevalences ranging from 1% to 90% ($P = 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$). The distribution of links are represented in Section *S*.4.2 (Figures 4 to 9) for each scenario.

- **Step 3**. We associated as many services as the total number of links generated for each individual. Then, to the matrix built in step one we merged three other columns identifying for each service, the individual associated to the service itself, his/her total number of links and his/her serological status, with the constraint that an individual cannot be linked to more than one service in each half-day.

- **Step 4**. We generated a sample of centres by drawing n random numbers between 1 and N using SRSWR.

- **Step 5**. For each centre l drawn, we generated a sample of half-days by drawing n_l random numbers between 1 and N_l using SRSWR.

- **Step 6**. For each centre l and for each half-day k drawns, we generated a sample of services by drawing n_{kl} random numbers between 1 and N_{kl} using SRSWR.

- **Step 7**. A sample of n_B individuals ($n_B \leqslant n_A$) was obtained using the individuals linked to the sampled services.

Then, we generated 10000 samples of size $n_A = 2000$ for each of the 16 scenarios and for each of the 11 prevalence values. For each sample generated: $\hat{N}^B$, $\hat{T}$, $\hat{P}$, $\hat{N}_G^B$, $\hat{T}_G$, $\hat{P}_G$ were calculated. To explore the properties of our design-based estimator when errors occur in the FVA, we generated

three kinds of errors, presented in Table 2 of the main manuscript. Indeed, in real life, it is often difficult for individuals to remember all FVA particularly if they are multiple and varied. Furthermore interviewed individuals are sometimes vulnerable or consume substances which do not facilitate memorization.

Table 1. Parameters associated to the 16 scenarios

Scenario	μ	θ	μ_1	θ_1	μ_2	θ_2
1	3	1	–	–	–	–
2	3	3	–	–	–	–
3	3	5	–	–	–	–
4	3	10	–	–	–	–
5	3	20	–	–	–	–
6	3	50	–	–	–	–
7	5	1	–	–	–	–
8	5	3	–	–	–	–
9	5	5	–	–	–	–
10	5	10	–	–	–	–
11	5	20	–	–	–	–
12	5	50	–	–	–	–
13	–	–	3	1	10	1
14	–	–	3	1	10	10
15	–	–	5	1	10	1
16	–	–	5	1	10	10

S.4.1 Distribution of number of links - ANRS-Coquelicot study

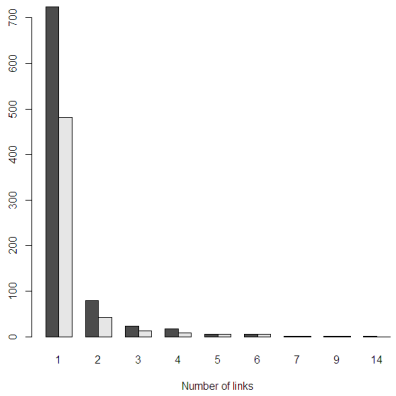


Fig. 3. Frequency distribution of the number of links observed in the Coquelicot survey. The dark bars represent the uninfected individuals and the clear rectangles represent the infected individuals

S.4.2 Distributions of number of links - Simulation study

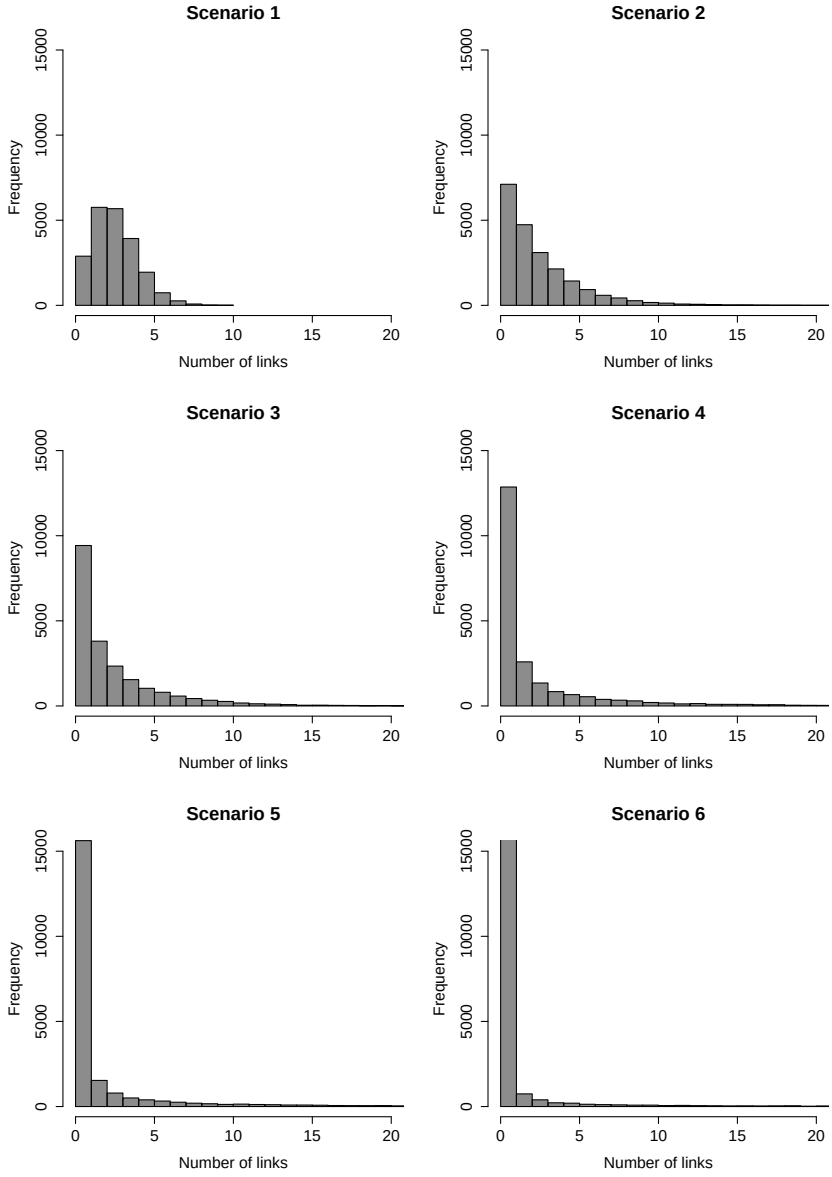


Fig. 4. Frequency distribution of the number of links generated from a negative binomial distribution of mean $\mu = 3$ and variance $\theta\mu$ with $\theta = 1, 3, 5, 10, 20, 50$ corresponding to the scenarios 1 to 6.

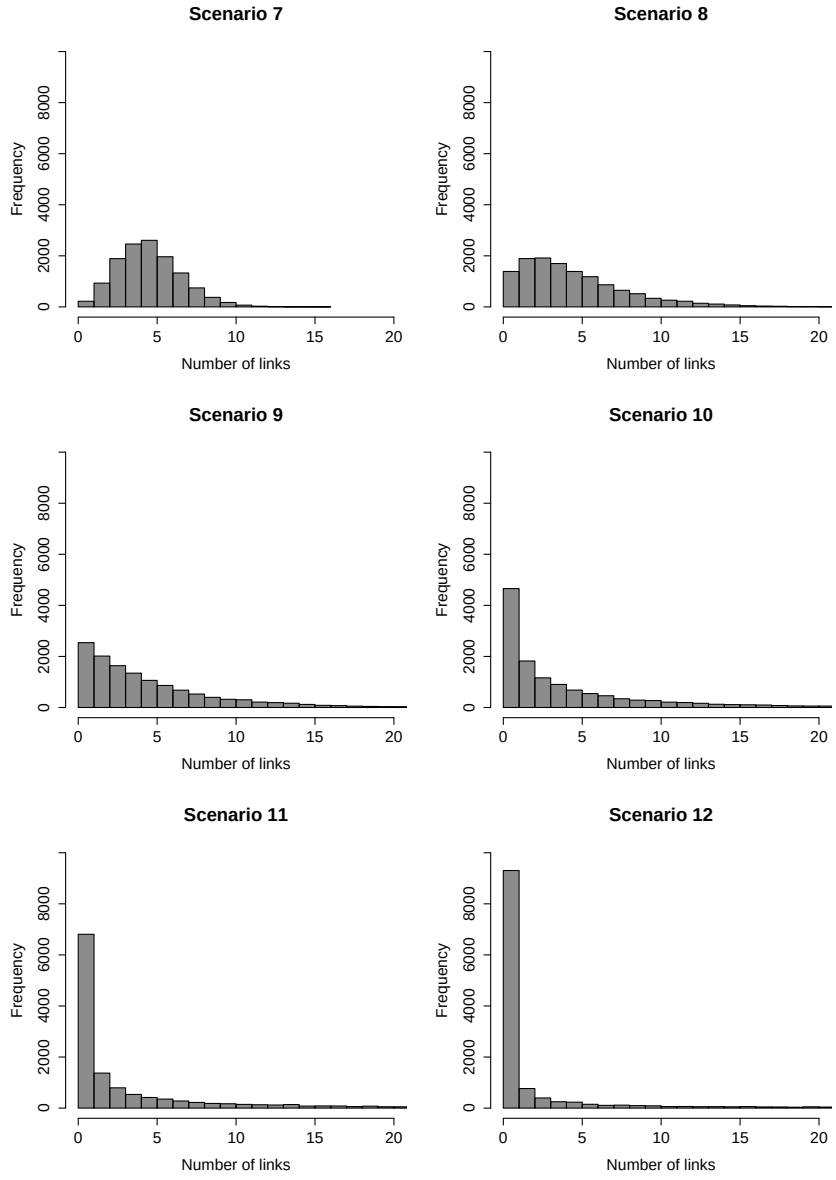


Fig. 5. Frequency distribution of the number of links generated from a negative binomial distribution of mean $\mu = 5$ and variance $\theta\mu$ with $\theta = 1, 3, 5, 10, 20, 50$ corresponding to the scenarios 7 to 12.

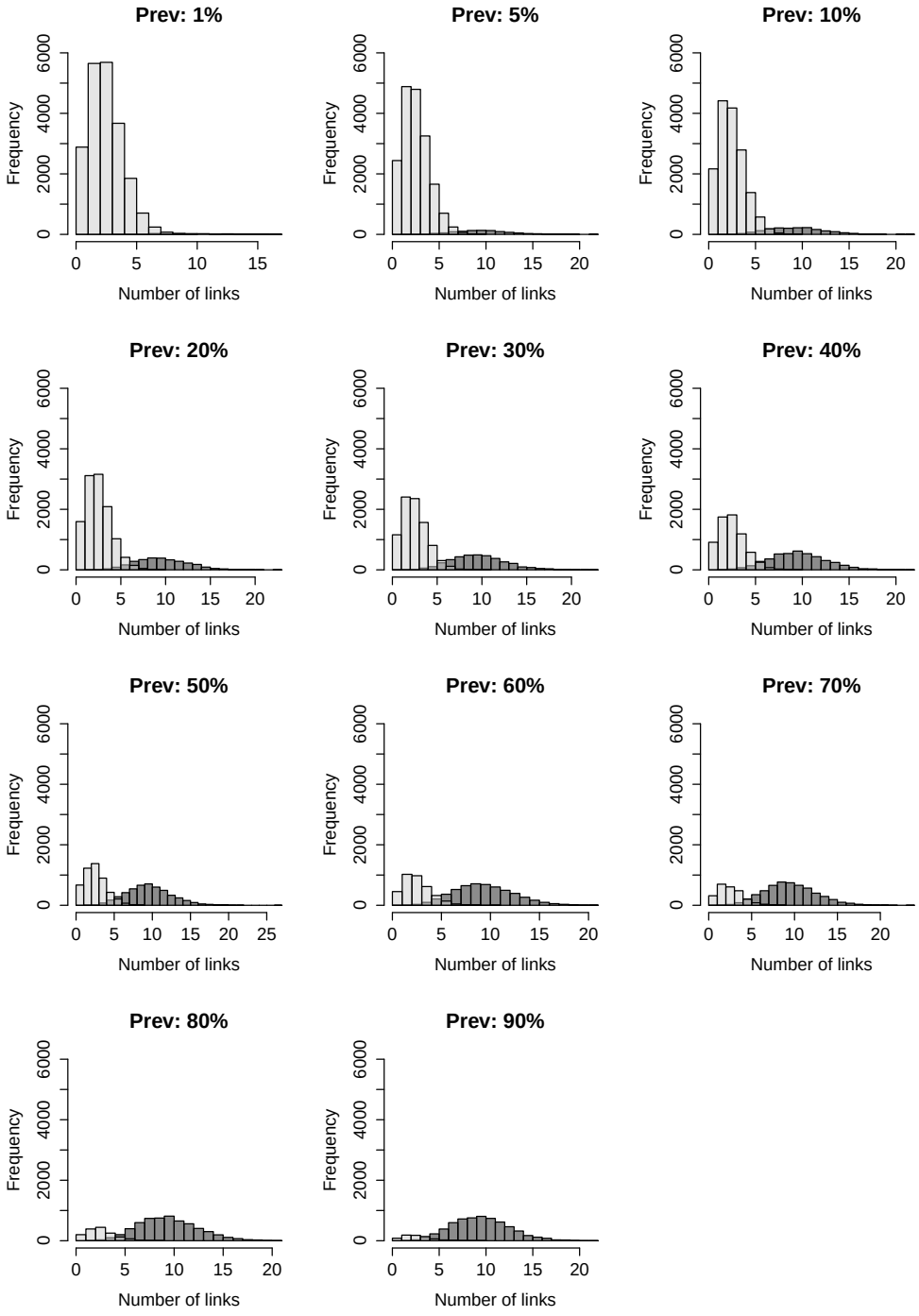


Fig. 6. Scenario 13: Frequency distribution of the number of links generated from two negative binomial distributions: the first distribution of mean $\mu_1 = 3$ and variance μ_1 for the non-infected individuals (clear bars) and the second one of mean $\mu_2 = 10$ and variance μ_2 for the infected individuals (dark bars). Prevalence varies from 1% to 90%.

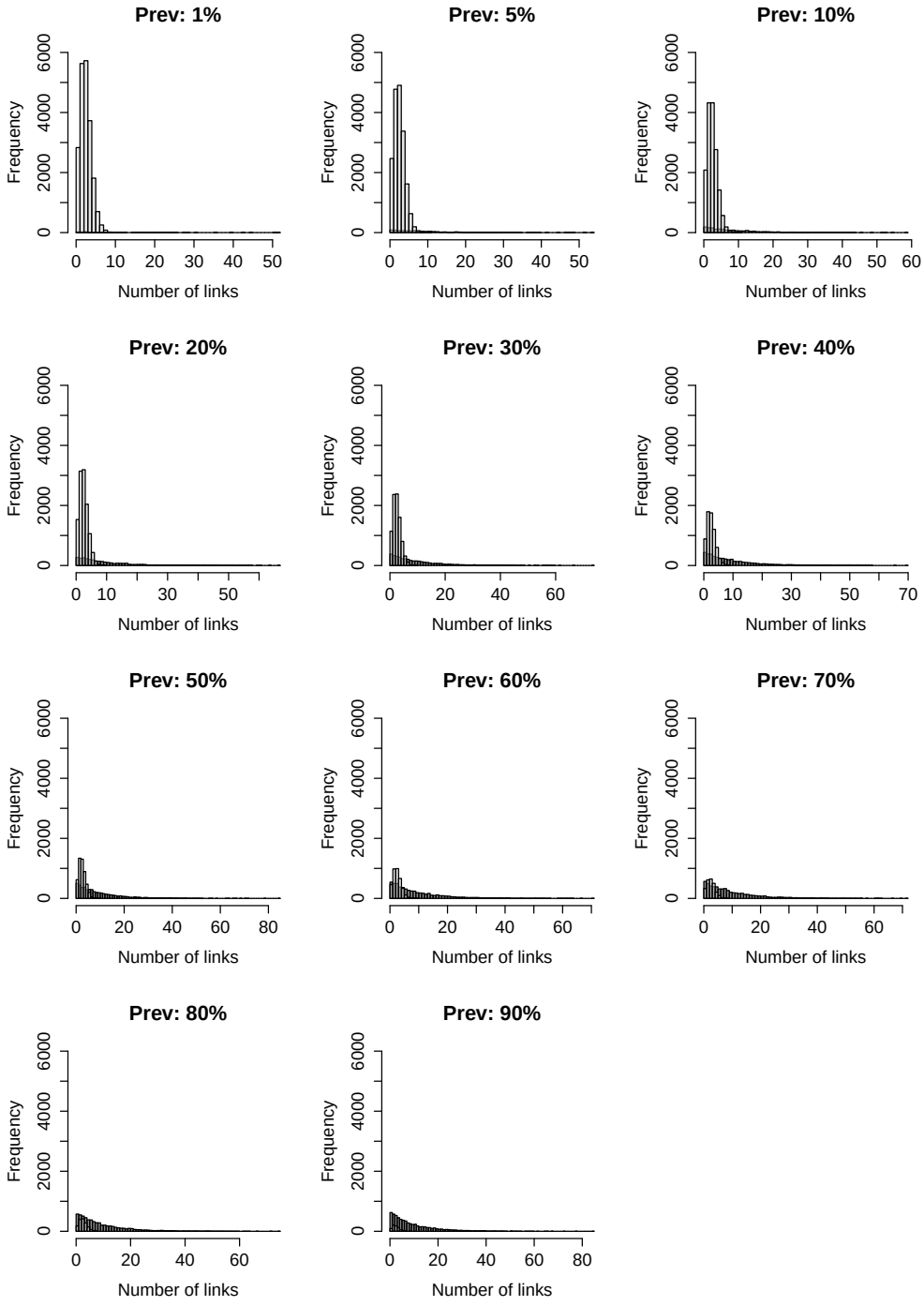


Fig. 7. Scenario 14: Frequency distribution of the number of links generated from two negative binomial distributions: the first distribution of mean $\mu_1 = 3$ and variance μ_1 for the non-infected individuals (clear bars) and the second one of mean $\mu_2 = 10$ and variance $\theta_2\mu_2$ with $\theta_2 = 10$ for the infected individuals (dark bars). Prevalence varies from 1% to 90%.

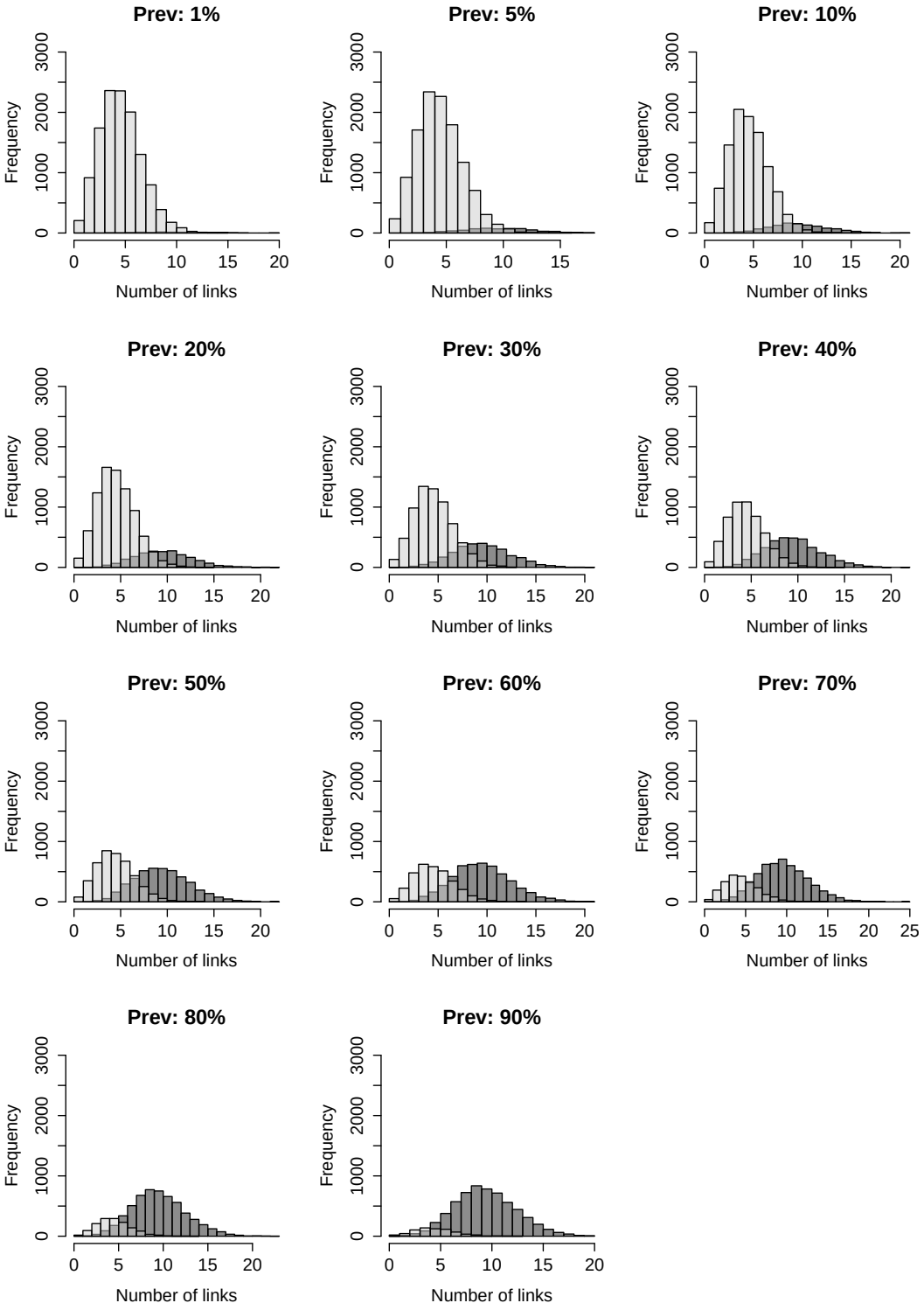


Fig. 8. Scenario 15: Frequency distribution of the number of links generated from two negative binomial distributions: the first distribution of mean $\mu_1 = 5$ and variance μ_1 for the non-infected individuals (clear bars) and the second one of mean $\mu_2 = 10$ and variance μ_2 for the infected individuals (dark bars). Prevalence varies from 1% to 90%.

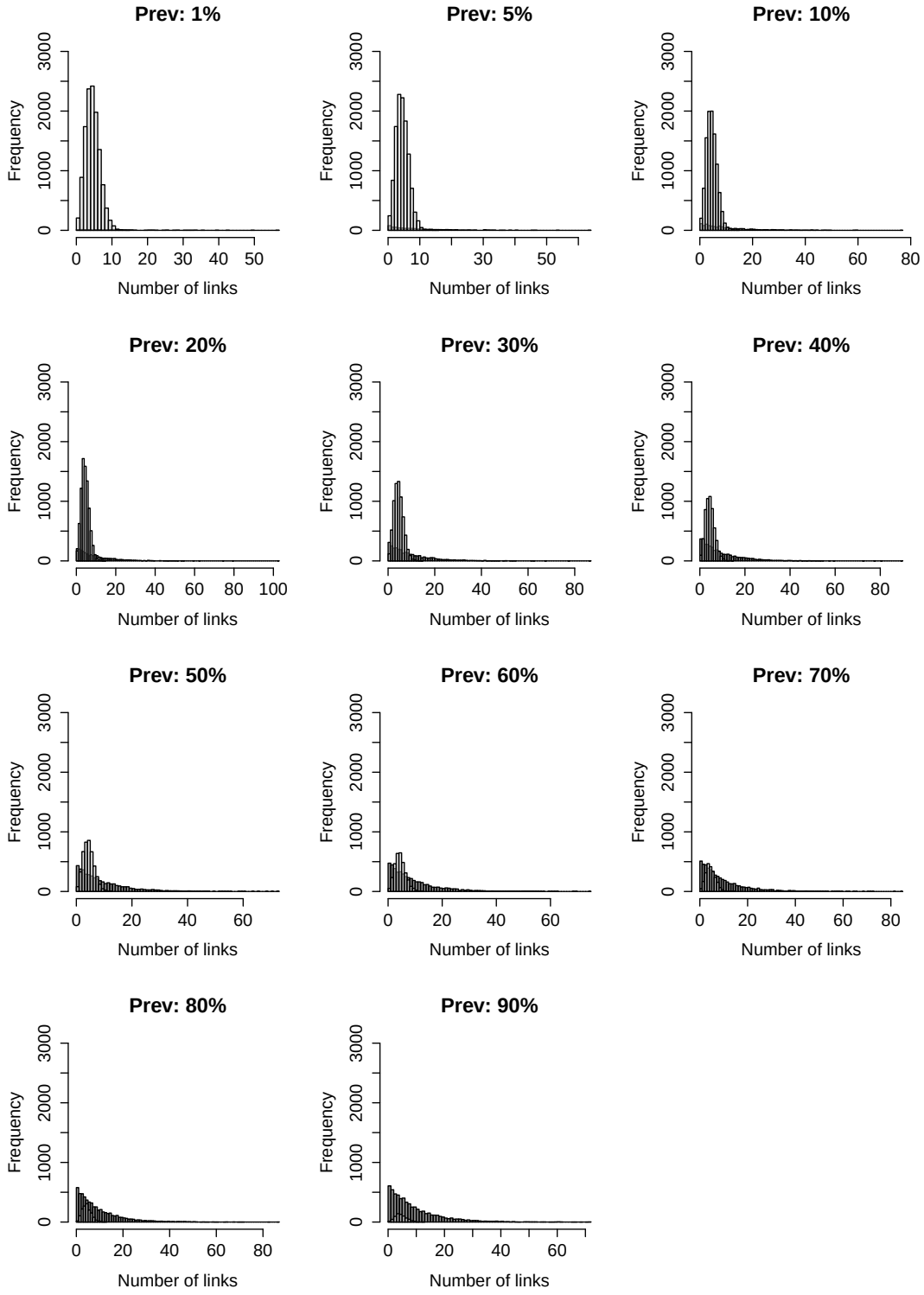


Fig. 9. Scenario 16: Frequency distribution of the number of links generated from two negative binomial distributions: the first distribution of mean $\mu_1 = 5$ and variance μ_1 for the non-infected individuals (clear bars) and the second one of mean $\mu_2 = 10$ and variance $\theta_2\mu_2$ with $\theta_2 = 10$ for the infected individuals (dark bars). Prevalence varies from 1% to 90%.

S.5. Population size data

Table 2. True population size for the scenarios 1-16.

Scenarios	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%
1-6	21300	21300	21300	21300	21300	21300	21300	21300	21300	21300	21300
7-12	12800	12800	12800	12800	12800	12800	12800	12800	12800	12800	12800
13-14	21000	19000	17500	14500	12500	11000	9800	9000	8100	7500	6900
15-16	12500	12300	11500	10600	9800	9200	8600	8000	7500	7100	6800

Table 3. Estimated population size from the Horvitz-Thompson estimator and its 95%CI according to each scenario and each prevalence value.

Scenario	1%	5%	10%	20%
1	21297 [20812; 21802]	21301 [20794; 21810]	21297 [20798; 21809]	21304 [20810; 21805]
2	21301 [20580; 22047]	21295 [20557; 22053]	21300 [20573; 22052]	21294 [20555; 22043]
3	21308 [20469; 22183]	21313 [20467; 22172]	21301 [20485; 22146]	21307 [20464; 22157]
4	21297 [20305; 22284]	21306 [20324; 22273]	21288 [20309; 22277]	21295 [20307; 22289]
5	21295 [20227; 22373]	21299 [20218; 22398]	21298 [20223; 22394]	21302 [20183; 22380]
6	21299 [20111; 22451]	21308 [20167; 22459]	21291 [20114; 22467]	21302 [20153; 22479]
7	12804 [12550; 13071]	12799 [12543; 13067]	12798 [12540; 13060]	12802 [12545; 13073]
8	12800 [12367; 13251]	12802 [12368; 13250]	12803 [12367; 13254]	12800 [12369; 13255]
9	12804 [12252; 13371]	12797 [12246; 13357]	12799 [12234; 13370]	12796 [12256; 13350]
10	12803 [12123; 13515]	12797 [12095; 13500]	12800 [12103; 13528]	12806 [12114; 13519]
11	12803 [11987; 13637]	12801 [11979; 13637]	12798 [11986; 13643]	12796 [11980; 13655]
12	12801 [11880; 13753]	12806 [11884; 13744]	12807 [11859; 13761]	12803 [11866; 13735]
13	21001 [20493; 21520]	19003 [18489; 19533]	17500 [16972; 18041]	14503 [14020; 15009]
14	21000 [20491; 21529]	19003 [18469; 19536]	17497 [16955; 18049]	14498 [13960; 15039]
15	12499 [12245; 12754]	12302 [12038; 12584]	11500 [11243; 11761]	10601 [10350; 10859]
16	12502 [12242; 12764]	12301 [12022; 12595]	11499 [11202; 11796]	10603 [10288; 10937]

Scenario	30%	40%	50%	60%
1	21303 [20801; 21814]	21300 [20816; 21818]	21297 [20786; 21799]	21302 [20796; 21819]
2	21304 [20566; 22069]	21293 [20552; 22041]	21301 [20551; 22052]	21298 [20583; 22042]
3	21299 [20483; 22140]	21302 [20467; 22145]	21298 [20466; 22123]	21301 [20450; 22133]
4	21305 [20320; 22295]	21297 [20323; 22273]	21295 [20333; 22274]	21302 [20333; 22293]
5	21302 [20223; 22380]	21298 [20219; 22362]	21309 [20211; 22367]	21301 [20234; 22384]
6	21309 [20132; 22463]	21310 [20160; 22503]	21293 [20108; 22454]	21303 [20125; 22464]
7	12801 [12541; 13070]	12799 [12536; 13064]	12799 [12535; 13063]	12799 [12543; 13066]
8	12800 [12361; 13247]	12799 [12363; 13249]	12800 [12368; 13240]	12799 [12361; 13246]
9	12797 [12229; 13367]	12799 [12250; 13365]	12803 [12257; 13365]	12804 [12254; 13378]
10	12805 [12120; 13507]	12798 [12106; 13510]	12805 [12115; 13519]	12797 [12098; 13518]
11	12802 [11994; 13632]	12808 [11987; 13657]	12805 [11986; 13655]	12804 [11984; 13640]
12	12794 [11853; 13747]	12811 [11905; 13749]	12802 [11887; 13753]	12808 [11905; 13744]
13	12501 [12054; 12960]	11002 [10591; 11426]	9798 [9443; 10168]	9001 [8687; 9333]
14	12499 [11969; 13027]	11004 [10516; 11491]	9797 [9341; 10279]	8998 [8564; 9451]
15	9800 [9561; 10052]	9201 [8969; 9444]	8600 [8387; 8821]	8000 [7809; 8196]
16	9799 [9473; 10138]	9199 [8863; 9537]	8598 [8262; 8951]	7998 [7664; 8350]

Scenario	70%	80%	90%
1	21297 [20805; 21806]	21300 [20801; 21818]	21299 [20793; 21804]
2	21304 [20573; 22069]	21299 [20564; 22052]	21301 [20570; 22037]
3	21305 [20482; 22156]	21300 [20483; 22140]	21299 [20486; 22136]
4	21299 [20319; 22330]	21304 [20348; 22278]	21305 [20341; 22280]
5	21303 [20227; 22407]	21302 [20224; 22363]	21299 [20210; 22357]
6	21306 [20151; 22468]	21304 [20128; 22492]	21304 [20144; 22479]
7	12800 [12544; 13062]	12799 [12547; 13063]	12800 [12542; 13064]
8	12800 [12359; 13252]	12801 [12366; 13242]	12802 [12357; 13250]
9	12795 [12243; 13376]	12801 [12251; 13368]	12805 [12252; 13374]
10	12797 [12113; 13519]	12804 [12106; 13514]	12797 [12121; 13496]
11	12811 [11991; 13644]	12807 [11990; 13631]	12799 [11981; 13628]
12	12795 [11866; 13755]	12795 [11855; 13729]	12803 [11848; 13761]
13	8099 [7825; 8384]	7501 [7278; 7727]	6900 [6736; 7080]
14	8101 [7686; 8528]	7500 [7121; 7894]	6899 [6538; 7275]
15	7500 [7325; 7680]	7101 [6955; 7251]	6800 [6676; 6929]
16	7502 [7160; 7858]	7101 [6751; 7455]	6799 [6461; 7153]

S.6. Estimated prevalences

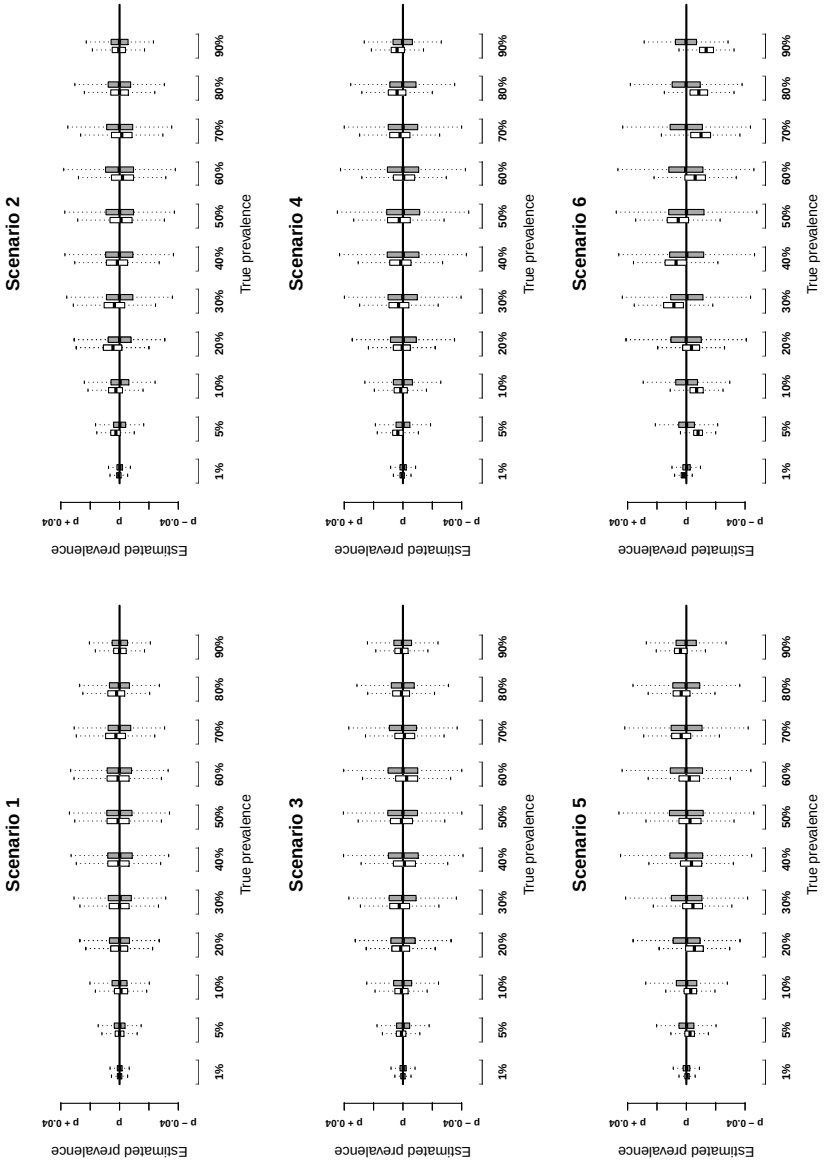


Fig. 10. White boxplots represent the Horvitz-Thompson estimated prevalences and gray boxplots represent the alternative design-based estimated prevalences for the scenarios 1 to 6. On each graph, the straight line represents the true prevalence.

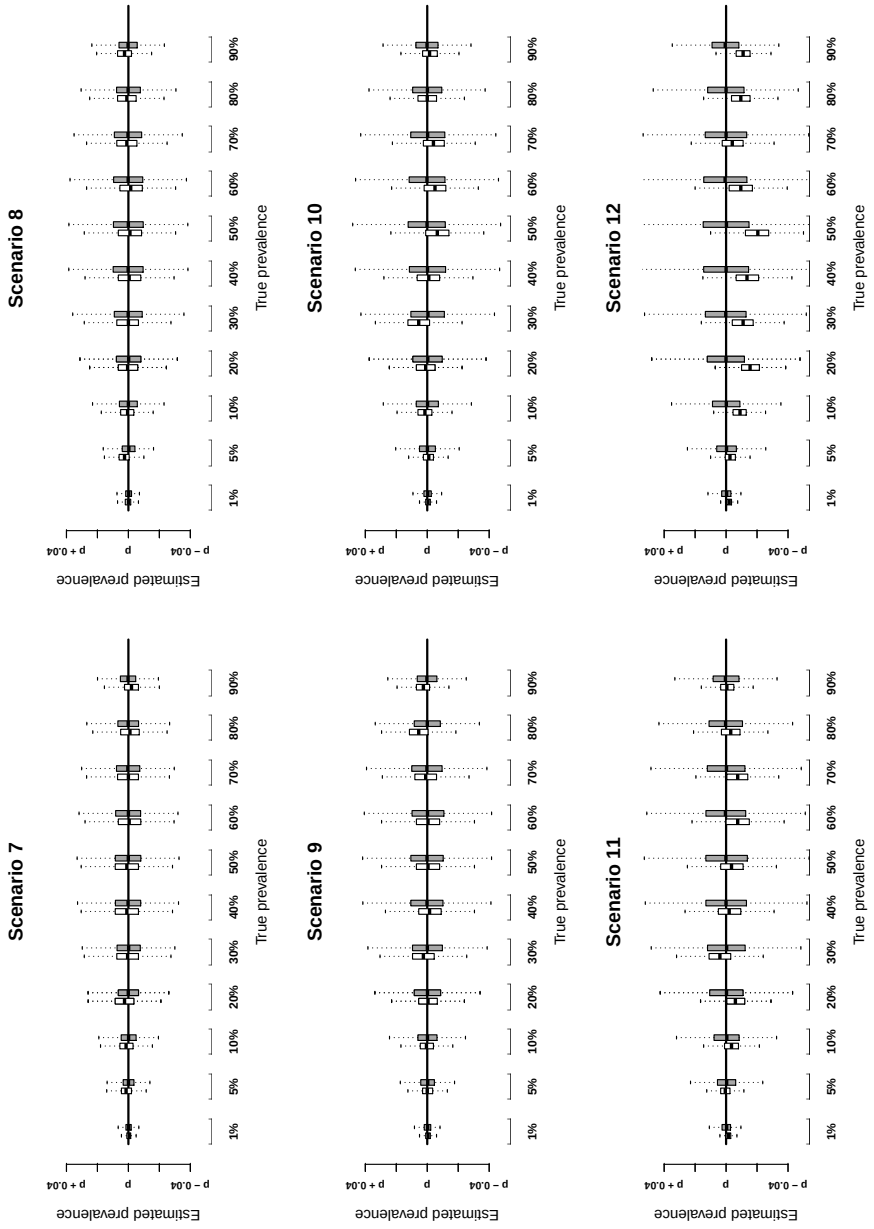


Fig. 11. White boxplots represent the Horvitz-Thompson estimated prevalences and gray boxplots represent the alternative design-based estimated prevalences for the scenarios 7 to 12. On each graph, the straight line represents the true prevalence.

S.7. Coverage probability

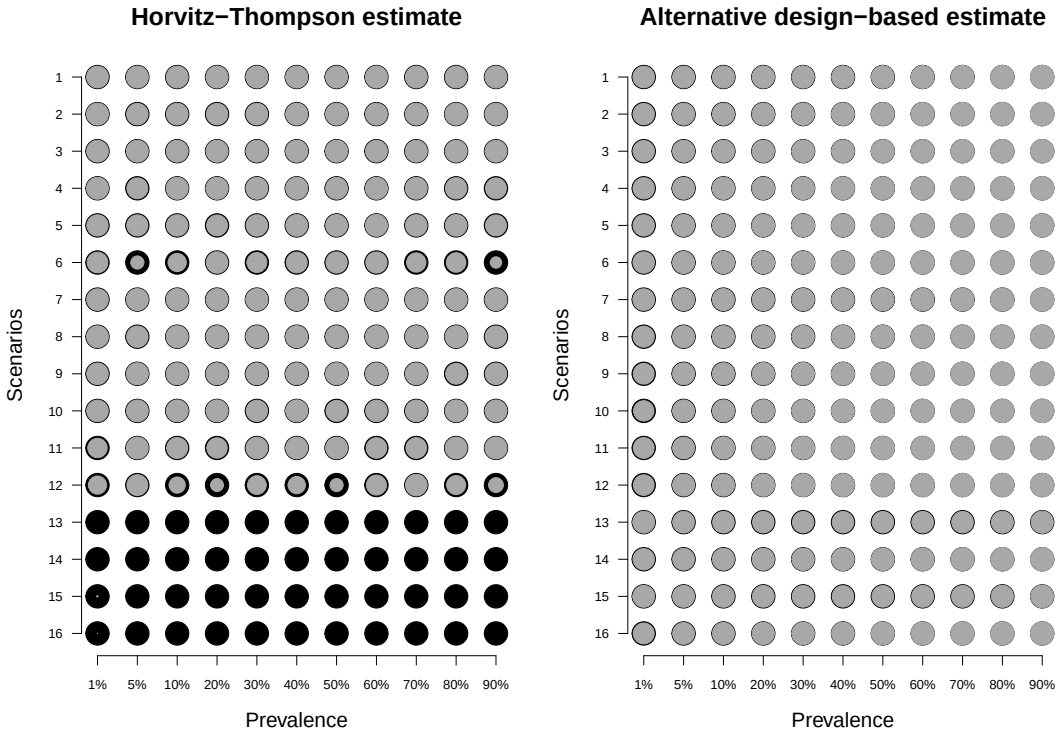


Fig. 12. Coverage probability: a full black circle indicates a zero coverage probability and an full gray circle indicates a one coverage probability for the Horvitz-Thompson (left) and the alternative design-based (right) estimates.

S.8. Estimated prevalences when errors occur in the FVA

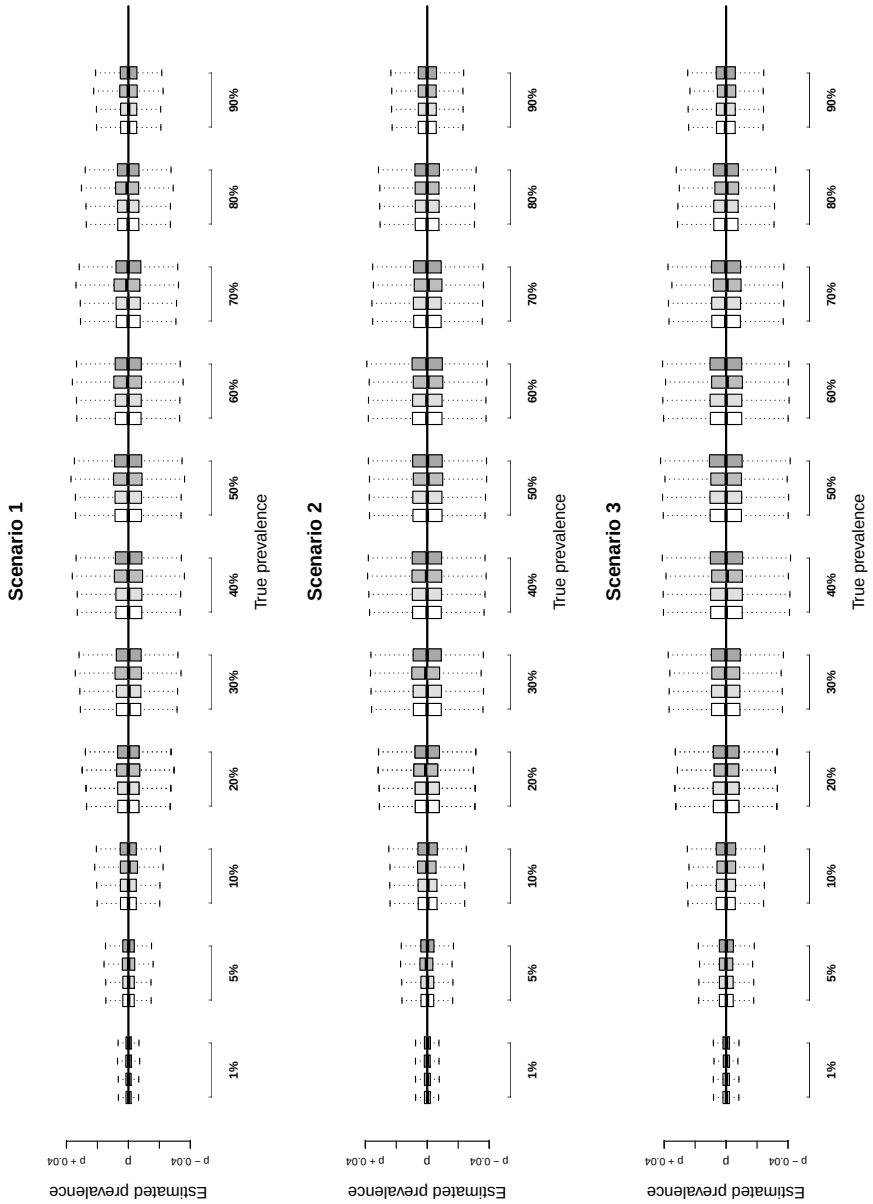


Fig. 13. All boxplots represent estimated prevalences from the alternative design-based estimator for the scenarios 1-3: no link error (white) and the three links errors (gray). On each graph, the straight line represents the true prevalence.

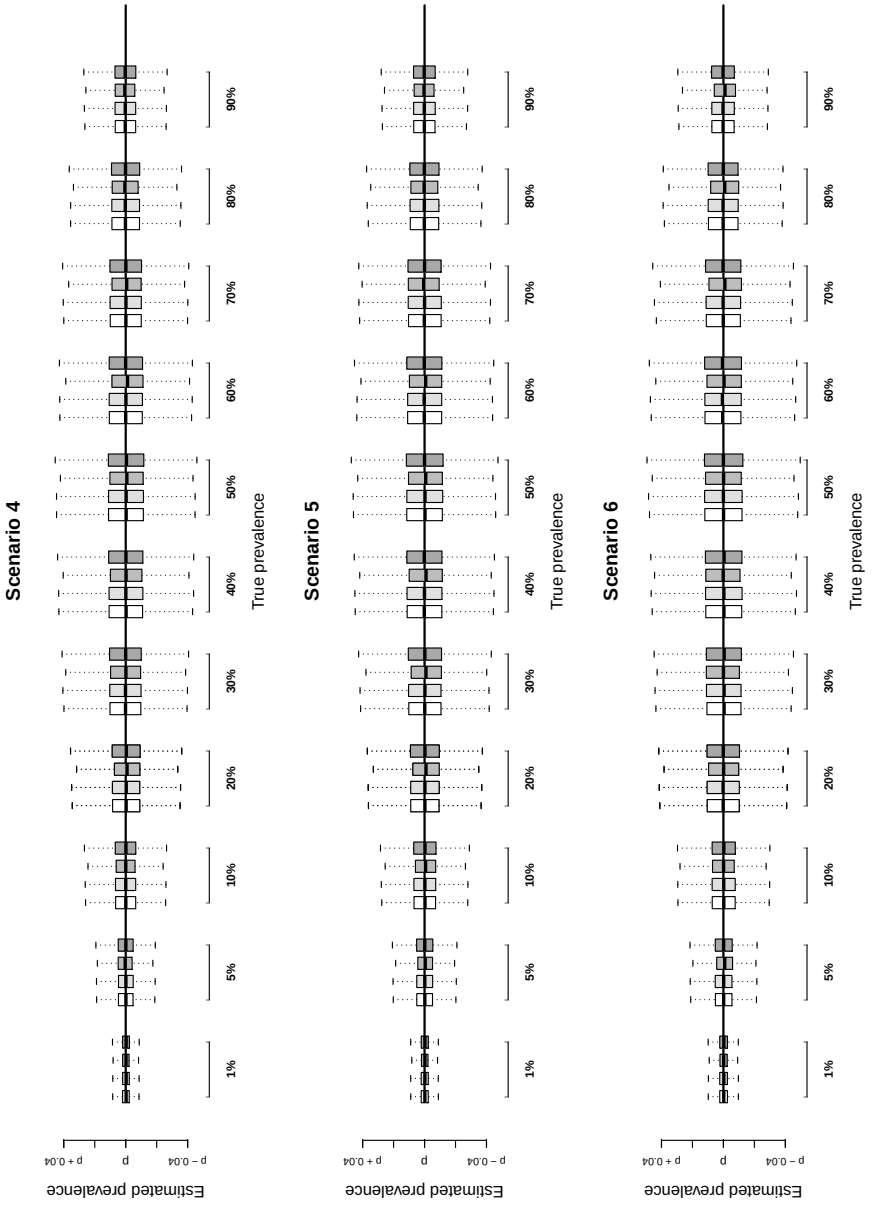


Fig. 14. All boxplots represent estimated prevalences from the alternative design-based estimator for the scenarios 4-6: no link error (white) and the three link errors (gray). On each graph, the straight line represents the true prevalence.

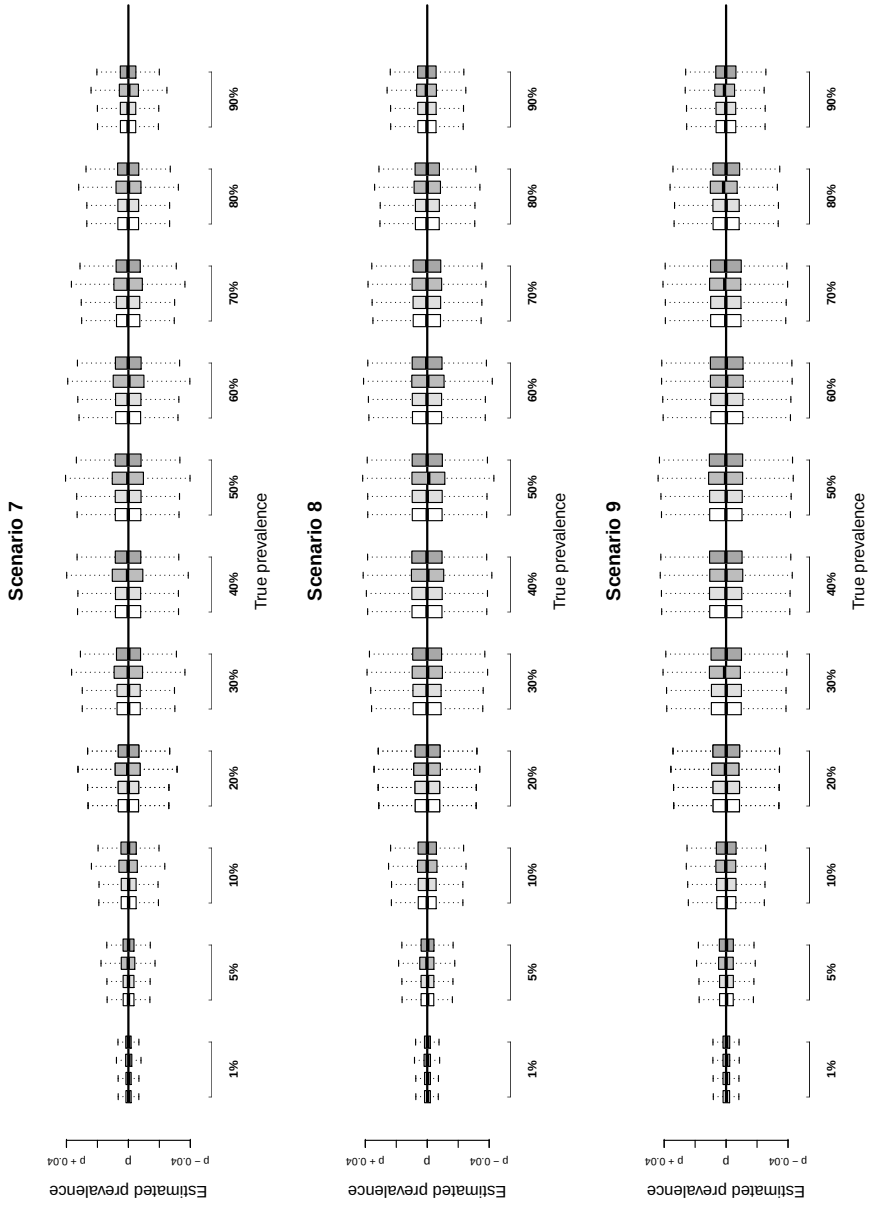


Fig. 15. All boxplots represent estimated prevalences from the alternative design-based estimator for the scenarios 7-9: no link error (white) and the three link errors (gray). On each graph, the straight line represents the true prevalence.

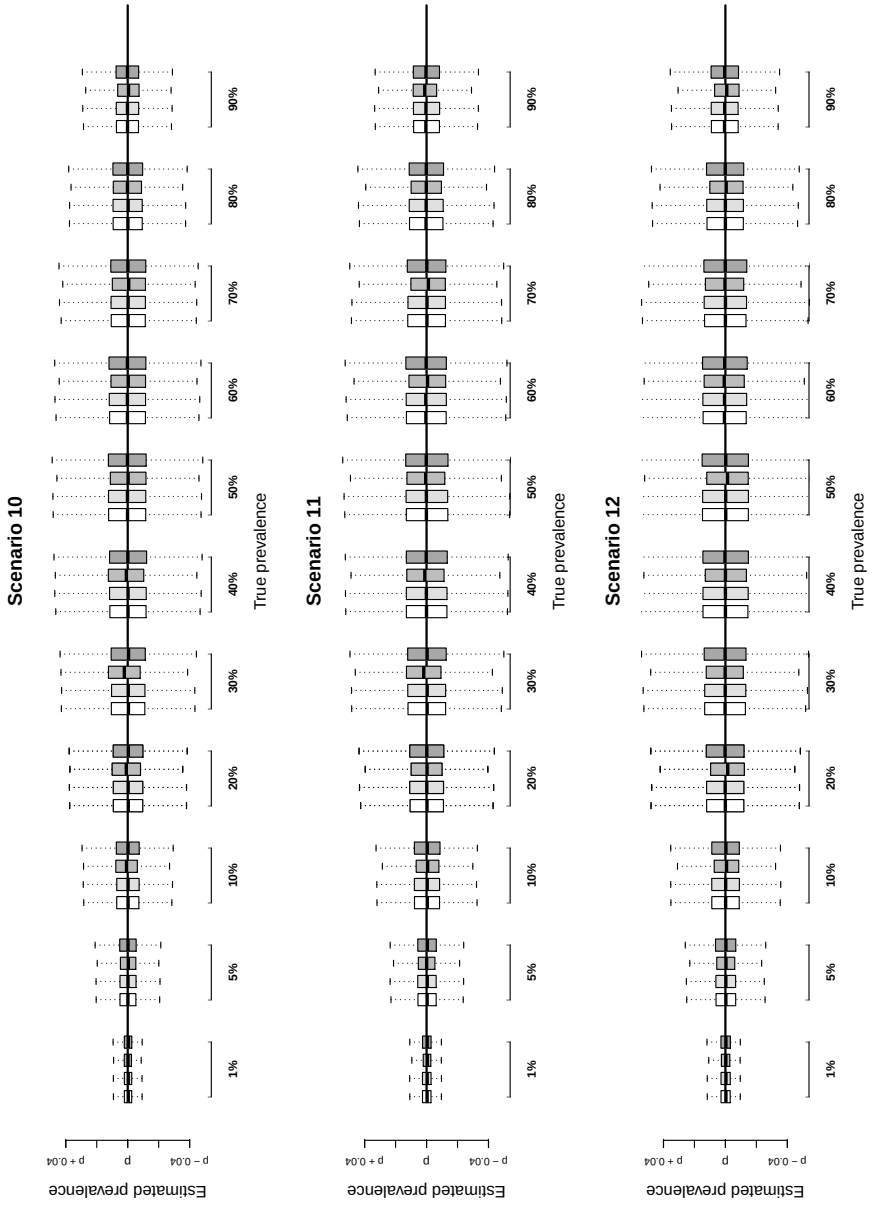


Fig. 16. All boxplots represent estimated prevalences from the alternative design-based estimator for the scenarios 10-12: no link error (white) and the three link errors (gray). On each graph, the straight line represents the true prevalence.

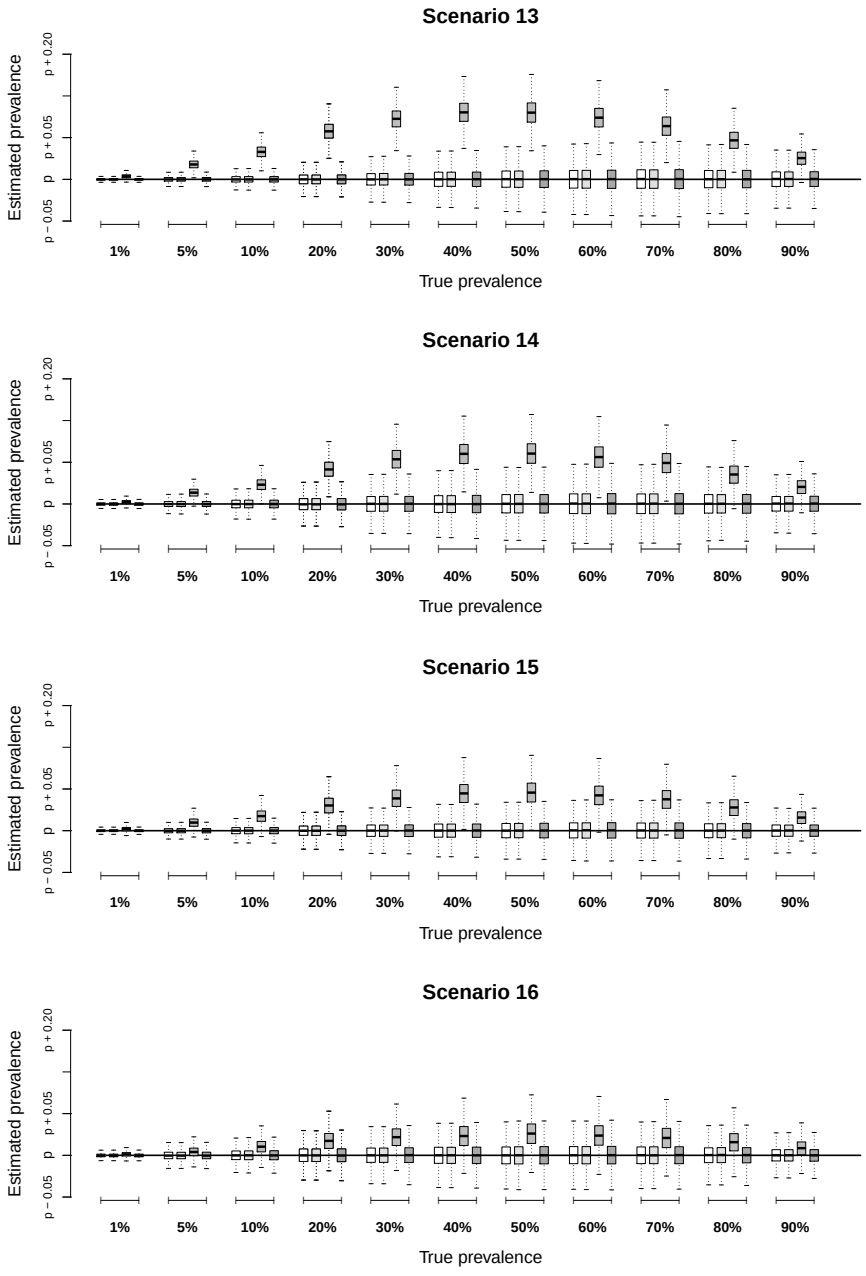


Fig. 17. All boxplots represent estimated prevalences from the alternative design-based estimator for the scenarios 13-16: no link error (white) and the three link errors (gray). On each graph, the straight line represents the true prevalence.

S.9. Estimated number of infected individuals

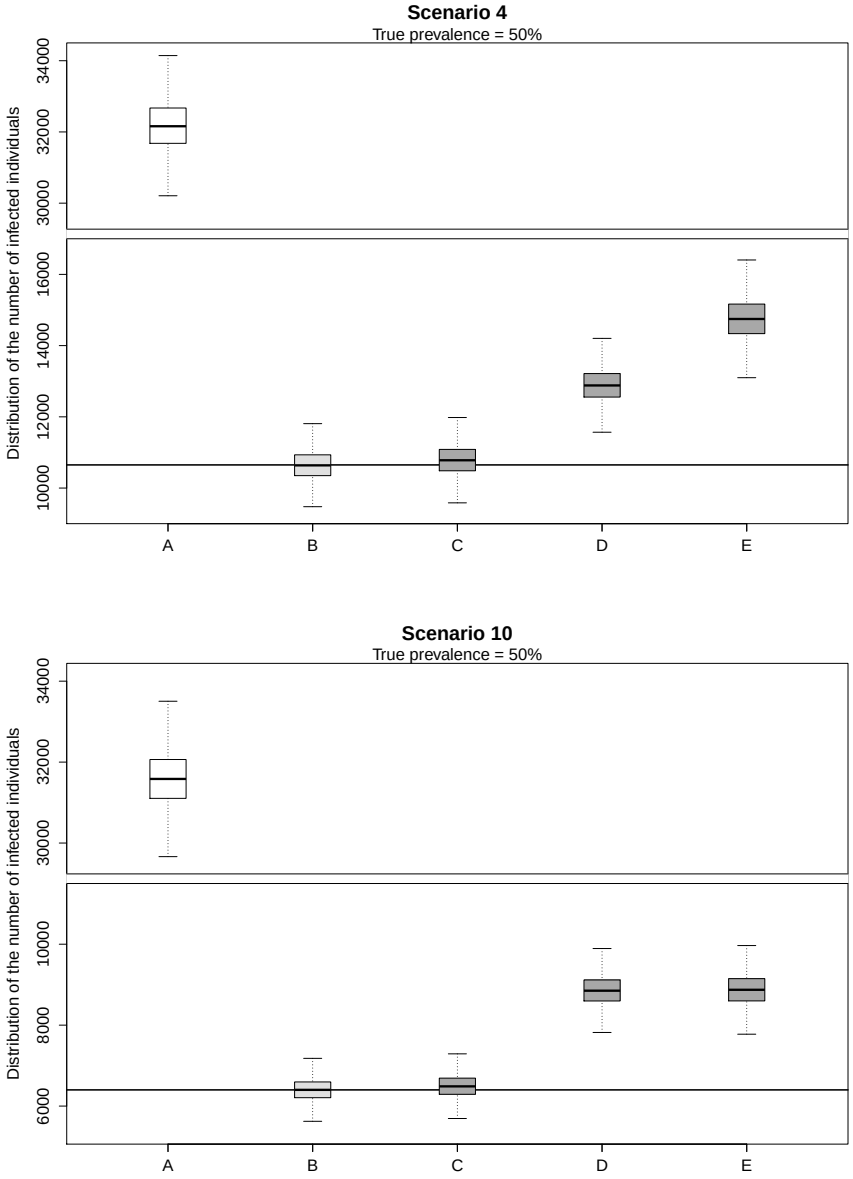


Fig. 18. All boxplots represent the estimated numbers of infected individuals for the scenarios 4 and 10: from the Horvitz-Thompson estimator (A) and from the alternative design-based estimator (B-E), with respectively no link error , link errors 1, link errors 2, link errors 3. On each graph, the straight line represents the true number of infected individuals.

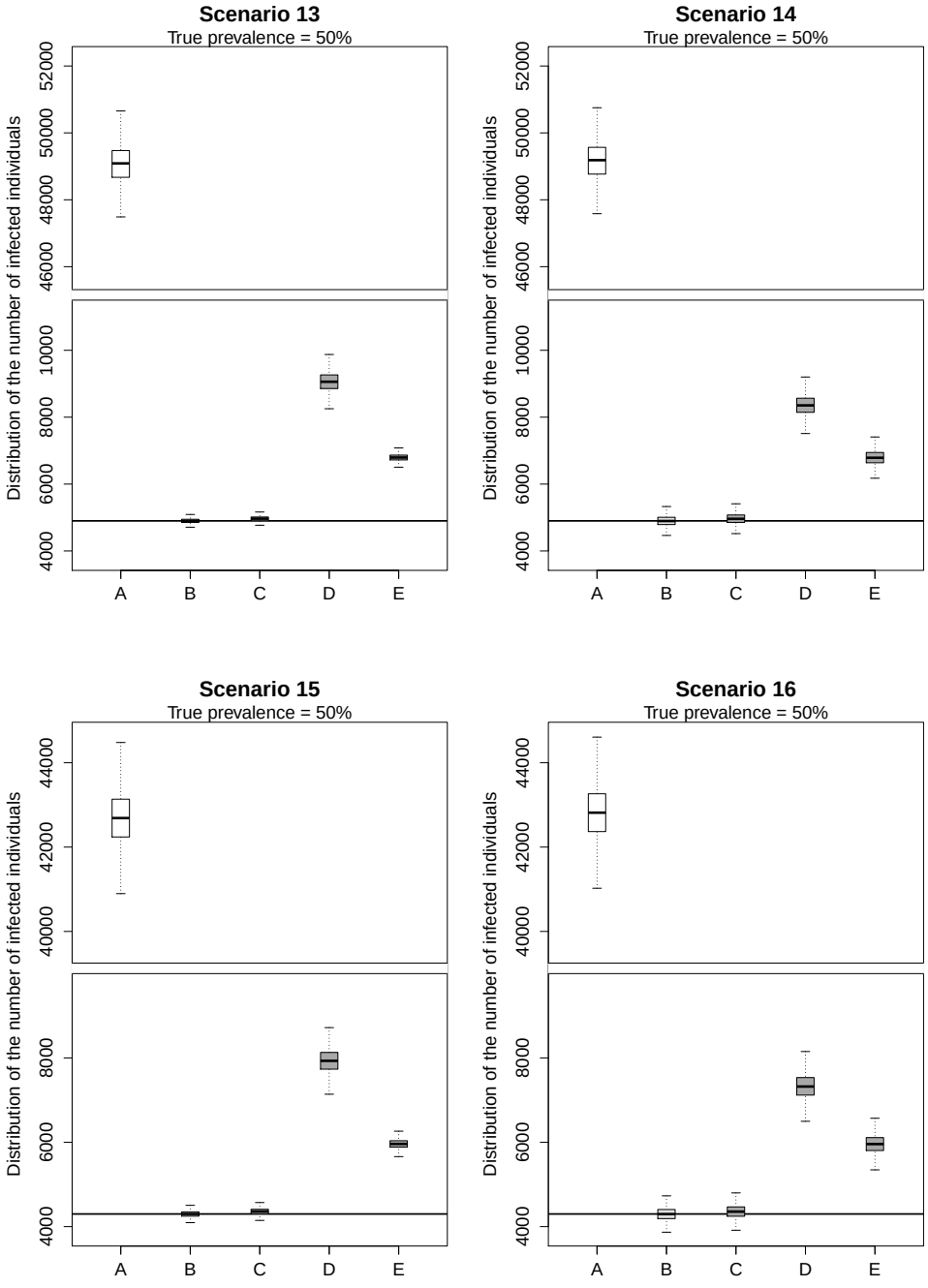


Fig. 19. All boxplots represent estimated numbers of infected individuals for the scenarios 13-16 : from the Horvitz-Thompson estimator (A) and from the alternative design-based estimator (B-E), with respectively no link error , link errors 1, link errors 2, link errors 3. On each graph, the straight line represents the true number of infected individuals.

Annexe 4 : Article publié dans la revue *Epidemiology and Infection*

Epidemiol. Infect., Page 1 of 13. © Cambridge University Press 2016
doi:10.1017/S0950268816002934

Age- and time-dependent prevalence and incidence of hepatitis C virus infection in drug users in France, 2004–2011: model-based estimation from two national cross-sectional serosurveys

L. LEON[1]*, S. KASEREKA[1], F. BARIN[2], C. LARSEN[1], L. WEILL-BARILLET[1], X. PASCAL[1], S. CHEVALIEZ[3], J. PILLONEL[1], M. JAUFFRET-ROUSTIDE[1,4] AND Y. LE STRAT[1]

[1] *Santé publique France, French National Public Health Agency, Saint-Maurice, France*
[2] *Centre National de Référence du VIH & INSERM UMR966, Centre Hospitalier Universitaire & Université François-Rabelais, Tours, France*
[3] *Department of Virology, Hôpital Henri Mondor, Centre National de Référence des Hépatites B, C et Delta, Créteil, France*
[4] *Cermes3 (Inserm U988/CNRS UMR 8211/EHESS/Paris Descartes University), Paris, France*

Received 30 March 2016; Final revision 5 October 2016; Accepted 9 November 2016

SUMMARY

Hepatitis C virus (HCV) infection is a public health issue worldwide. Injecting drug use remains the major mode of transmission in developed countries. Monitoring the HCV transmission dynamic over time is crucial, especially to assess the effect of harm reduction measures in drug users (DU). Our objective was to estimate the prevalence and incidence of HCV infection in DU in France using data from a repeated cross-sectional survey conducted in 2004 and 2011. Age- and time-dependent HCV prevalence was estimated through logistic regression models adjusted for HIV serostatus or injecting practices. HCV incidence was estimated from a mathematical model linking prevalence and incidence. HCV prevalence decreased from 58·2% [95% confidence interval (CI) 49·7–66·8] in 2004 to 43·2% (95% CI 38·8–47·7) in 2011. HCV incidence decreased from 7·9/100 person-years (95% CI 6·4–9·4) in 2004 to 4·4/100 person-years (95% CI 3·3–5·9) in 2011. HCV prevalence and incidence were significantly associated with age, calendar time, HIV serostatus and injecting practices. In 2011, the highest estimated incidence was in active injecting DU (11·2/100 person-years). Given the forthcoming objective of generalizing access to new direct antiviral agents for HCV infection, our results contribute to decision-making and policy development regarding treatment scale-up and disease prevention in the DU population.

Key words: Drug users, hepatitis C virus, incidence, mixture model, prevalence.

INTRODUCTION

Hepatitis C virus (HCV) infection is a public-health issue worldwide and injecting drug use is still the major mode of HCV transmission, especially through the sharing of injecting equipment [1, 2]. Although public health prevention measures have been introduced in a large number of high-income countries (syringe-exchange programmes, opioid substitution treatments, consumption rooms and, to a lesser extent, treatment for prevention), the level of HCV transmission in drug users (DU) is still a public health issue as current harm reduction intervention strategies on HCV transmission have had mixed success [2–4]. A marked decrease in incidence has been observed

* Author for correspondence: Ms. L. Léon, Santé publique France, French National Public Health Agency, 12 rue du Val d'Osne, 94415 Saint-Maurice Cedex France.
(Email: Lucie.Leon@santepubliquefrance.fr)

in some countries [2, 5]. One example is the city of Amsterdam which has seen a marked decrease tending towards zero, partly thanks to harm reduction measures combined with changes over time in the type of drugs used and the consumption patterns of injecting drug users (IDU) [5]. Despite these positive developments, HCV incidence remains high in IDU [4] and consequently regular estimation of HCV prevalence and incidence in the DU population is crucial to assess the impact of harm reduction measures.

The most suitable way to estimate incidence is to conduct a prospective cohort where high-risk individuals are followed up over time and are tested for anti-HCV seroconversion [6–9]. Such a cohort would need to follow a large number of hard-to-reach subjects for a long time. This is difficult, expensive and time consuming.

Alternative approaches exist, for example implementing sentinel surveillance and using one or repeated cross-sectional surveys. Essential to all these alternatives is the collection of blood samples [plasma, serum, or dried blood spots (DBS)] for anti-HCV antibodies and/or HCV RNA testing.

There are several different ways cross-sectional surveys can be used. The first is to estimate HCV incidence from retrospective cohorts built from cross-sectional surveys when only HCV antibodies are collected [2]. Another is the use of mathematical models [3, 10–13]. In general, supplementary data are needed to build these models, including disease-related mortality, annual number of new DU and clinical and behavioural data (e.g. lifetime history of injecting drug use). Yet another cross-sectional approach involves HCV RNA testing of anti-HCV-negative samples. Using the proportion of new infected persons (i.e. HCV RNA positive in anti-HCV-negative persons) and the window period (i.e. the mean number of days during which HCV RNA is detectable before HCV antibodies develop) [10–13], HCV incidence is estimated using a simple formula [9, 14]. Finally, another way to estimate incidence is to apply an anti-HCV avidity-testing algorithm to identify samples compatible with recent primary infection [15, 16].

In France, no cohort studying HCV infection in DU at the national level is currently in place. Although an anti-HCV IgG avidity assay to identify recent HCV infection has recently been developed in France it has not yet been applied in practice [17, 18].

The aims of this paper were to estimate age- and time-dependent prevalence and incidence of HCV infection in DU in France from 2004 to 2011 using two national cross-sectional surveys (ANRS-Coquelicot studies) conducted in 2004 and 2011, based on blood testing [19, 20]. We built a mathematical model based on the relationship between prevalence and incidence. HCV antibodies were used as biological markers to estimate both prevalence and incidence, as HCV RNA was not available in the first survey.

METHODS

Data sources

The French ANRS-Coquelicot survey is a repeated cross-sectional serosurvey conducted in 2004 [19] and 2011 [20] in DU recruited in five French metropolitan cities (Lille, Strasbourg, Paris, Bordeaux, Marseille). In the 2011 survey two additional administrative departments (Seine-Saint-Denis and Seine-et-Marne, which are suburbs of Paris) were also included. The surveys' objectives were to estimate the prevalence of anti-HIV and anti-HCV antibodies, to assess at-risk practices associated with HCV transmission and to evaluate the dynamics of the HIV and HCV epidemics in this population.

For each survey and in each city (and department in 2011), time-location sampling was used (described in a previous paper [21]). Briefly, a comprehensive inventory was built of all the centres providing services to DU (including high- and low-threshold services). We then constructed a sampling frame based on half-day opening times of centres. All listed centres participated in the survey. Half-days were randomly drawn in all centres using simple random sampling without replacement. At each centre/half-day visit, DU were selected using systematic random sampling, except for residential centres where all users were included in the survey.

Information on participants' socio-demographic situation, health status, access to HCV screening, knowledge of HCV transmission modes, drug use, and at-risk practices was collected.

In order to have similar populations, we excluded DU interviewed in general practitioners' offices in 2004, as these locations were not used in the 2011 survey. We also excluded those interviewed in the two administrative departments only in 2011 for the same reason.

Studied variables

We focused on certain variables known to be associated with HCV infection: age, HIV serostatus and injecting practices [injected drugs at least once during lifetime (yes/no), injected drugs in the month before

http://dx.doi.org/10.1017/S0950268816002934

the study interview (yes/no)], and crack use (yes/no) [19, 20]. Crack use was defined as the consumption of crack (sniffing, snorting, injecting, smoking) in the month before the interview. Although injecting drug use and the sharing of syringes and injecting equipment remains the major mode of transmission, crack use is suspected to be a possible risk for HCV infection [19] as chipped or hot glass pipes can cause lesions in the mouth and hands, exposing users to infection. An IDU was defined as someone who reported injecting drug use at least once in her/his lifetime. An active injecting drug user (AIDU) was defined as someone who reported injecting drug use in the month before the study interview.

Laboratory data

Blood samples on blotting papers were collected during the interview by participants who agreed to provide self-obtained finger-prick blood samples on DBS for anti-HIV and HCV antibody testing. Six drops, corresponding to ~50 µl capillary whole blood, were spotted onto filter paper card (Whatman 903™, GE Healthcare Europe GmbH, Germany).

DBS samples in 2004 and in 2011 were screened using the same assays: HCV 3.0 Ortho ELISA and Ortho HIV1/2 Ab capture ELISA for HCV and HIV antibodies, respectively [19]. Positive anti-HIV samples (i.e. defining HIV positivity in a DU) were confirmed by serotyping and/or Western blot [22]. Details of the serological data analysis are provided in Supplementary Appendix 1.

Analyses

Statistical analyses were based on a five-step process:

Step 1. Anti-HCV data were modelled using a mixture of normal distributions to discriminate between HCV-seronegative and HCV-seropositive individuals.

Step 2. We deduced the global HCV prevalence from the classification obtained in step 1.

Step 3. We estimated age- and time-dependent HCV prevalence using regression models.

Step 4. We deduced age- and time-dependent HCV incidence from the prevalence estimated in step 3 using a model-based approach.

Step 5. We estimated the global HCV incidence for the year of each survey (i.e. 2004 and 2011), using the incidence estimated in step 4 and the estimated proportion of DU.

Mixture model

To avoid inconclusive classifications arising from the use of specified biological thresholds (e.g. a cut-off value provided by the manufacturer), the distribution of the quantitative results of antibody tests was modelled using an underlying mixture model – also called the direct method – rather than the usual threshold method [23]. We used a six-component and five-component mixture model on data from the 2004 and 2011 surveys, respectively, to identify persons who were seronegative or seropositive for HCV, according to reactivity level in the anti-HCV assay. Details of the model selection strategy are provided in Supplementary Appendix 2. Component densities were assumed to be normally distributed. We assumed that levels 1–3, corresponding to the lowest reactivity, represented the negative results of the anti-HCV test. Levels 4–6 (for 2004) and levels 4–5 (for 2011) were assumed to represent the positive results of the anti-HCV test.

Sampling weights

To produce estimates in the DU population, all the analyses took into account the sampling designs (sampling weights, stratifications, primary sampling units) of the two surveys.

As these surveys were based on time-location random sampling, DU attendance frequency in centres was incorporated into the sampling weights [21]. We appended the two datasets and adapted the sampling weights according to year of each survey, gender and age group (dichotomized into age >30 years or not) [24]. We decided to create this dichotomization as individuals aged <30 years in the 2004 sample were able to benefit from all the harm reduction measures available in France in 2004.

Estimation of age- and time-dependent prevalence

From our mixture model, each individual was classified as seropositive or seronegative. For each individual i, let us consider a binary variable of interest Y_i corresponding to the HCV classification ($Y_i = 1$ if i is seropositive and 0 if not). $P(Y = 1|a, t)$ is the probability of being seropositive at age a at calendar time t. A multivariate regression mel was used to estimate age- and time-dependent HCV prevalence, including age as a continuous covariate. The two most popular approaches to deal with a continuous covariate are to use splines or fractional polynomials. In the former, a generalized additive model is used. In

http://dx.doi.org/10.1017/S0950268816002934

the latter, a generalized regression model is performed. As we expected a simple shape reflecting a simple relationship between HCV infection and age, we chose this latter approach to build the multivariable model [25]. The generalized linear model can be expressed by:

$$g\big(E[Y|a,t]\big) = g(P(Y=1|a,t)) = \alpha + \eta(a) + ct, \quad (1)$$

where g is a link function, α is the intercept, c is the regression coefficient associated with time t and $\eta(a)$ a fractional polynomial function for age a. Fractional polynomials are an extension to classic polynomials, and are used for possible improvements in fit where the powers can be real values [26]. Different power transformation models are used instead of a straight line to estimate the relationship between the outcome variable and a continuous covariate, and to select the best-fitting model (i.e. the one with the highest likelihood value).

The fractional polynomial of degree m for the linear predictor, associated with age a, is defined as: $\eta_m\big(a, b, p_1, p_2, \ldots, p_m\big) = \sum_{j=0}^{m} b_j H_j(a)$, where m is an integer, b is a vector of regression coefficients, $p_1 \leqslant p_2 \ldots \leqslant p_m$ is a sequence of powers and $H_j(a)$ is a transformation function given by:

$$H_j(a) = \begin{cases} a^{p_j} & \text{if } p_j \neq p_{j-1} \\ H_{j-1}(a) \times \ln(a) & \text{if } p_j = p_{j-1} \\ & \text{with } p_0 = 0 \text{ and } H_0 = 1. \end{cases}$$

In our study, two link functions were tested: the complementary log-log link [$\log(-\log[1-x])$] and the logit link [$\log(x/[1-x])$], also used in previous studies [27]. The best model was selected using Akaike's Information Criterion (AIC). Using a logit link, age- and time-dependent prevalence from equation (1) is expressed as:

$$P(Y=1|a,t) = \frac{\exp\big(\alpha + \eta(a) + ct\big)}{1 + \exp\big(\alpha + \eta(a) + ct\big)}. \quad (2)$$

Using a complementary log-log link, $P(Y=1|a,t) = 1 - \exp(-\exp(\alpha + \eta(a) + ct))$. It is straightforward to include additional covariates to adjust for any specific characteristics of interest such as the HIV serostatus or at-risk behaviors (e.g. injecting practices, crack use). Five regression models were performed to estimate HCV prevalence as a function of age and time (models 1 and 2), on age, time and injecting practices (model 3), on age, time and crack use (model 4) and finally, on age, time and HIV serostatus (model 5). We considered these five different models instead of one global model in order to estimate prevalence and incidence in each sub-population.

Estimation of age-dependent incidence

$\lambda(a,t)$ is the age- and time-dependent incidence of HCV infection. $N(a,t)$ is the proportion of anti-HCV negative persons of age a at time t, and $P(a,t)$ the proportion of anti-HCV positive persons (i.e. the prevalence) of age a at time t. We assumed that HCV transmission could be represented by a two-state compartmental model [28], corresponding to the anti-HCV negative (N) and the anti-HCV positive (P) states, as illustrated in Figure 1, and expressed by the following differential equations:

$$\left.\begin{aligned} \frac{dN(a,t)}{d(a,t)} &= -\lambda(a,t)N(a,t) - \mu_1 N(a,t) + \beta N(a,t) \\ &\quad + \gamma P(a,t) \\ \frac{dP(a,t)}{d(a,t)} &= \lambda(a,t)N(a,t) - \mu_2 P(a,t) - \gamma P(a,t). \end{aligned}\right\} \quad (3)$$

Parameters presented in Figure 1 and introduced in equation (3) are shown in Table 1 [19, 29, 30].

Given that $N(a,t) + P(a,t) = 1$ for each age and time, we can express incidence from equation (3) by:

$$\begin{aligned} \lambda(a,t) = \Big(&\frac{\partial}{\partial a}P(a,t) + \frac{\partial}{\partial t}P(a,t) + (\beta - \mu_1) \\ &- (\beta - \mu_1 - \gamma)P(a,t)\Big) \\ &/(1 - P(a,t)), \end{aligned}$$

where $P(a,t)$ represents the prevalence estimated in the previous section. We can thus replace $P(a,t)$ by $P(Y=1|a,t)$ hereafter. With a logit link, age-dependent prevalence, for a given time t, can be derived from equation (2):

$$\begin{aligned} \frac{\partial P(a,t)}{\partial a} &= \frac{\partial P(Y=1|a,t)}{\partial a} \\ &= \frac{\eta'(a)\exp\big(\alpha + \eta(a) + ct\big)}{\big[1 + \exp\big(\alpha + \eta(a) + ct\big)\big]^2}, \end{aligned}$$

where $\eta'(a)$ is the first derivative of the fractional polynomial $\eta(a)$ with respect to age a. The estimated age-dependent incidence for a given time t is therefore:

$$\begin{aligned} \lambda(a|t) = \big(&\eta'(a)p(a|t)(1 - p(a|t)) + (\beta - \mu_1) \\ &- (\beta - \mu_1 - \gamma)p(a|t)\big) \\ &/(1 - p(a|t)), \end{aligned} \quad (4)$$

where $p(a|t) = P(Y=1|a,t)$ is the estimated prevalence for age a calculated at a given time t. Using a

http://dx.doi.org/10.1017/S0950268816002934

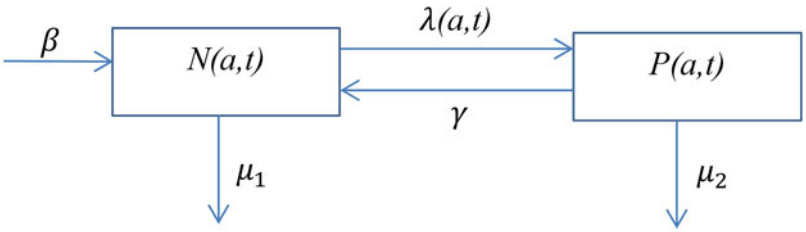


Fig. 1. Two-state compartmental model for HCV transmission. β is the proportion of new drug users; γ is the seroreversion (defined as the absence of HCV antibodies in a person previously known to be HCV positive) rate; μ_1 is the all-cause mortality rate in those without HCV infection; μ_2 ($=\mu_1+\mu_{HCV}$, μ_{HCV} is the HCV-related mortality rate) is the all-cause mortality rate in those with HCV infection and λ is the incidence rate.

Table 1. *Annual parameters in the two-state compartmental model*

Parameter	Parameter value	Reference
β: proportion of new drug users	2%	ANRS-Coquelicot data [19, 20]
γ: HCV seroreversion* rate	0·001	Le Page *et al.* [29]
μ_1: all-cause mortality rate in those without HCV infection	0·7%	Smit *et al.* [30]
μ_2: all-cause mortality rate in those with HCV infection	1·3%	Smit *et al.* [30]

* Defined as the absence of HCV antibodies in a person previously known to be anti-HCV positive.

complementary log-log link, the estimated age-dependent incidence at time t is given by:

$$\lambda(a|t) = \big(-\eta'(a)\log(1-p(a|t))(1-p(a|t)) + (\beta-\mu_1) - (\beta-\mu_1-\gamma)p(a|t)\big) /(1-p(a|t)).$$

Estimation of global incidence

Using the previous estimation of age-dependent incidence $\lambda(a|t)$ and the estimated proportion of DU by age, we calculated the global incidence of HCV infection for each time survey. The proportion of DU of age a at time t, noted $q(a|t)$, was estimated using the Horvitz–Thompson estimator $\hat{q}(a|t) = \sum_i^n w_i x_i(a,t) / \sum_i^n w_i$, where w_i is the sampling weight of the individual i, $x_i(a,t) = 1$ if the individual i is of age a at time t and 0 otherwise, and where n is the survey sample size [31].

For a given survey time, the global incidence can thus be expressed by the weighted arithmetic mean of the age-dependent incidences $\lambda(t) = \sum_a \hat{q}(a|t)\lambda(a|t)$.

A bootstrap method was used to estimate the variance of estimates, detailed in Supplementary Appendix 3. All analyses were performed using Stata v. 12.1 (StataCorp., USA) and the R 3.1.2 program (R Foundation for Statistical Computing, Austria).

RESULTS

Descriptive characteristics of DU

In the ANRS-Coquelicot surveys, 1462 and 1568 DU were included in 2004 and 2011, respectively, and blood samples were available in 79% and 92% of the participants [19, 20]. The final dataset combining these two surveys included 813 DU in 2004 and 1242 DU in 2011, after excluding individuals surveyed in general practitioners' offices only in 2004, and those surveyed in the two administrative departments only in 2011. In addition, DBS from 2004 were deemed invalid when there was insufficient material (i.e. DBS <6 mm in diameter). Table 2 presents descriptive statistics about the participants for both of the surveys after exclusion. The majority of the participants were men (~77%) aged 26–45 years (83% in 2004 and 67% in 2011).

From both surveys, we estimated that most DU reported injecting drug use [72·0% (95% confidence interval (CI) 63·4–80·7) in 2004 and 65·7% (95% CI 61·7–69·7) in 2011] and that <50% had used crack in the previous month (Table 2). HIV prevalence was estimated at 10·8% (95% CI 5·4–16·2) in 2004 and 9·4% (95% CI 6·8–12·0) in 2011.

Using the threshold method, HCV prevalence in DU was estimated at 58·9% (95% CI 50·4–67·4) in 2004 and 43·4% (95% CI 39·0–47·9) in 2011. Using the mixture model, HCV prevalence in DU was estimated at 58·2% (95% CI 49·7–66·8) in 2004 and 43·2% (95% CI 38·8–47·7) in 2011.

Estimated age- and time-dependent prevalence

The logit link provided the lowest AIC for all regressions when modelling prevalence. Table 3 presents the

http://dx.doi.org/10.1017/S0950268816002934

Table 2. *Descriptive statistics of participants, France, 2004 and 2011, ANRS-Coquelicot*

	2004 ($N = 813$)			2011 ($N = 1242$)		
Participants	Unweighted (proportion)	Weighted (proportion)	95% CI	Unweighted (proportion)	Weighted (proportion)	95% CI
Age, years						
18–19	0·9	0·4	0·1–1·1	0·9	0·6	0·3–1·4
20–25	11·6	6·9	4·7–10·2	10·3	7·6	5·9–9·9
26–35	43·7	48·5	39·9–57·2	25·9	26·1	22·6–29·9
36–45	39·0	39·7	32·3–47·6	40·7	41·9	38·1–45·8
46–55	4·7	4·2	2·6–6·5	19·9	21·3	17·5–25·7
⩾56	0·2	0·3	0·00–1·9	2·3	2·5	1·6–4·0
Men	77·0	72·2	64·3–80·0	77·9	79·5	76·0–82·9
Reporting injecting drug use	73·3	72·0	63·4–80·7	63·9	65·7	61·7–69·7
Reporting injecting drug use in the month previous to the study interview	39·0	44·1	33·7–54·6	32·2	36·1	30·2–42·1
Crack use in the previous month	24·6	41·0	30·3–51·6	27·8	34·4	29·7–39·1
HIV prevalence	10·1	10·8	5·4–16·2	8·0	9·4	6·8–12·0
HCV prevalence (threshold method)	54·1	58·9	50·4–67·4	39·1	43·4	39·0–47·9
HCV prevalence (direct method)	52·5	58·2	49·7–66·7	39·2	43·2	38·9–47·7

CI, Confidence interval.

results obtained from the different regression models. For each model (except model 2), age and time were significantly associated ($P < 0·05$) with HCV infection.

For all the Figures below (except Fig. 1), the left panel represents the prevalence according to age from the two surveys, estimated from the regression model (curves) and from the pointwise design-based prevalence estimates (circles).

In all DU, prevalence monotonically increased with age until reaching a plateau at age ~50 years. A marked decrease in prevalence was observed regardless of age between 2004 and 2011 (Fig. 2, left panel).

Age- and time-dependent HCV prevalence was higher in IDU than in those who did not report injecting drug use [odds ratio (OR) 17·7, 95% CI 10·0–31·4; Table 3; Fig. 3, left panel]. Most DU were IDU (68·9%, 95% CI 64·7–72·7).

In AIDU, prevalence sharply increased with age until it reached a plateau at age ~40 years, then stabilized at ~80% (Fig. 4, left panel). The results and the shape of the prevalence according to the year of the survey, were not significantly different (OR 0·9, 95% CI 0·8–1·0, Table 3). No significant association with age or with time was found (model 2, Table 3). No significant association with crack use was found (model 4, Table 3).

Estimated age- and time-dependent prevalence was higher in HIV-positive DU than in HIV-negative DU (OR 5·3, 95% CI 2·9–9·9, Table 3, model 5), with the global HCV prevalence exceeding 80% (Fig. 5, left panel).

Estimated age- and time-dependent incidence

In Figures 2–5, the middle panel represents the estimated HCV incidence expressed as the rate of new anti-HCV positive persons/100 person-years, according to age both in 2004 and 2011, with their confidence intervals. The right panel represents estimated incidence according to age for each year between 2000 and 2020.

Overall, incidence decreased over time. It increased until a given age (for example, in 2011, age 34 years in all DU) before decreasing thereafter (Figs 2–5).

In all DU, the highest incidence was estimated at 10·0/100 (95% CI 7·8–11·6) in those aged 31 years in 2004, and at 6·1/100 (95% CI 4·4–7·5) in those aged 34 years in 2011 (Fig. 2, middle panel).

HCV incidence in non-IDU decreased faster over time than in IDU (Fig. 3, middle and right panels). The highest incidence in non-IDU was estimated at 3·9/100 (95% CI 2·4–5·2) in those aged 35 years in 2004, and at 2·3/100 (95% CI 1·0–3·4) in those aged 37 years in 2011. HCV incidence was always higher in IDU than in their non-injecting counterparts. The highest HCV incidence was estimated at 15·1/100 (95% CI 12·0–17·9) in those aged 29 years in 2004, and at 9·5/100 (95% CI 7·2–11·2) in those aged 32 years in 2011 (Fig. 3, middle panel).

http://dx.doi.org/10.1017/S0950268816002934

Table 3. *Logistic regression models performed to estimate anti-HCV prevalence in drug users, France, 2004 and 2011, ANRS-Coquelicot*

Variable	Fractional polynomial transformation $\eta(a)$	Regression coefficient estimate (S.E.)	*P*	95% CI	OR
Model 1					
Age	$(age/10)^{-1} - 0{\cdot}27$	−18·15 (2·59)	<0·001	−23·25 to −13·06	
	$(age/10)^{3} - 49{\cdot}14$	−0·008 (0·004)	0·023	−0·016 to −0·001	
Time (ref.: 2004)		−0·15 (0·03)	<0·001	−0·21 to −0·09	0·86 (0·81–0·92)
Intercept		0·81 (0·20)	<0·001	0·42 to 1·20	
Model 2 (AIDU)					
Age	$(age/10)^{-2} - 0{\cdot}08$	−4·05 (22·53)	0·857	−48·27 to 40·18	
	$(ag\ (age/10)^{-2}\ \ln(age/10) - 0{\cdot}10$	−44·66 (36·46)	0·221	−116·245 to 26·91	
Time (ref.: 2004)		−0·10 (0·06)	0·091	−0·21 to −0·02	0·91 (0·81–1·02)
Intercept		1·52 (0·38)	<0·001	0·78 to 2·27	
Model 3					
Age	$\ln(age/10) - 1{\cdot}30$	17·49 (4·12)	<0·001	9·38 to 25·60	
	$\ln(age/10)^{2} - 1{\cdot}68$	−5·22 (1·54)	0·001	−8·28 to −2·20	
Time (ref.: 2004)		−0·16 (0·04)	<0·001	−0·23 to −0·09	0·85 (0·79–0·92)
Reporting injecting drug use (ref: no)		2·88 (0·29)	<0·001	2·30 to 3·45	17·73 (10·02–31·36)
Intercept		−1·22 (0·33)	<0·001	−1·87 to −0·58	
Model 4					
Age	$(age/10)^{-1} - 0{\cdot}27$	−18·14 (2·57)	<0·001	−23·18 to −13·09	
	$(age/10)^{3} - 49{\cdot}14$	−0·008 (0·004)	0·031	−0·015 to −0·000	
Time (ref.: 2004)		−0·15 (0·03)	<0·001	−0·21 to −0·09	0·86 (0·81–0·91)
Crack user (ref.: no)		0·26 (0·19)	0·159	−0·10 to 0·63	1·30 (0·90–1·89)
Intercept		0·71 (0·18)	<0·001	0·35 to 1·06	
Model 5					
Age	$(age/10)^{-1} - 0{\cdot}27$	−17·60 (2·59)	<0·001	−22·68 to −12·52	
	$(age/10)^{3} - 50{\cdot}14$	−0·009 (0·004)	0·010	−0·016 to −0·002	
Time (ref.: 2004)		−0·16 (0·04)	<0·001	−0·24 to −0·09	0·85 (0·79–0·91)
HIV (ref: HIV-negative)		1·68 (0·31)	<0·001	1·07 to 2·29	5·35 (2·90–9·86)
Intercept		0·82 (0·26)	0·002	0·31 to 1·32	

AIDU, Drug user reporting active injecting drug use; OR, odds ratio; CI, confidence interval.

For example, for model 1: $\text{logit}(P(Y = 1|a, t) = -18{\cdot}15\left[(a/10)^{-1} - 0{\cdot}27\right] - 0 \cdot 01\left[(a/10)^{3} - 49{\cdot}14\right] - 0{\cdot}15t + 0{\cdot}81$, for age a at time t.

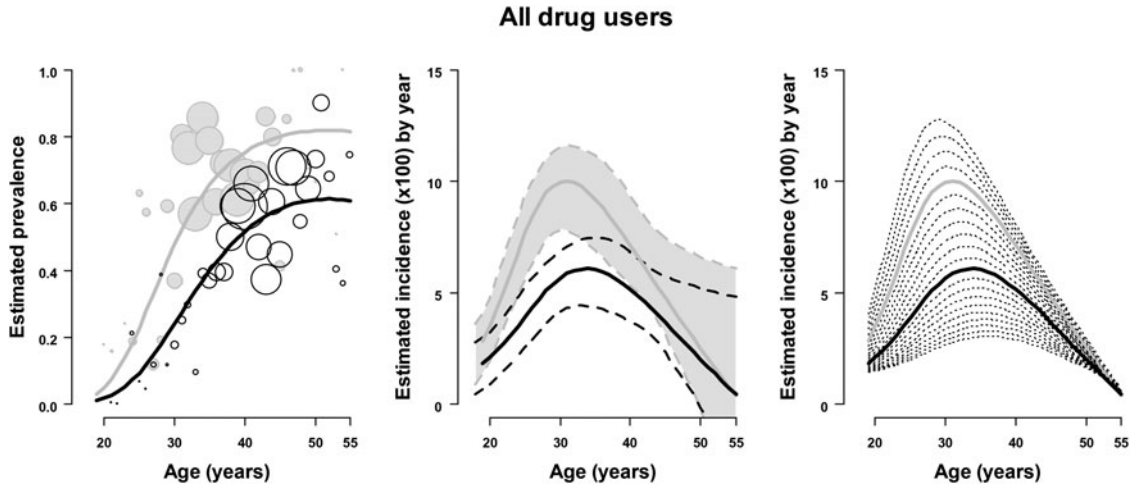


Fig. 2. *Left panel*: Curves represent the age-dependent HCV prevalence estimates from the logistic models in drug users in 2004 (grey) and 2011 (black). Circles represent the estimated prevalence by age. Their size is proportional to the number of persons in 2004 (solid grey circles) and 2011 (open circles). *Middle panel*: Curves represent the age-dependent HCV incidence estimates in drug users in 2004 (grey) and 2011 (black) with their confidence intervals (dashed curves). *Right panel*: Age-dependent HCV incidence estimates in drug users over 2000–2020. Curves were obtained from the model in 2004 (grey curves), 2011 (black curves) and the other years (dotted curves).

Estimates in AIDU showed that HCV dynamic infection was relatively similar between 2004 and 2011, as expected from the regression model (Fig. 4, model 2, Table 3).

In HIV-positive DU, HCV incidence increased until age 25 years in 2004 and until age 27 years in 2011 before decreasing thereafter (Fig. 5, middle and right panels). Even age-dependent HCV incidence consistently drifted towards older ages with time, the corresponding DU being younger in this subpopulation than the other subpopulations.

Estimated global incidence in those aged 18–55 years

We estimated a global incidence in those aged 18–55 years because of the very small number of older participants.

In all DU, HCV incidence was lower in 2011 (4·4/100, 95% CI 3·3–5·9) than in 2004 (7·9/100, 95% CI 6·4–9·4) (Table 4). For each DU subpopulation studied, we observed that HCV incidence was lower in 2011 than in 2004.

HCV incidence was found to be twice as high in AIDU as in other DU, decreasing from 15·4/100 (95% CI 11·9–19·3) in 2004 to 11·2/100 (95% CI 9·0–19·0) in 2011 (Table 4).

DISCUSSION

We estimated age- and time-dependent prevalence and incidence of HCV infection in DU in France by first modelling prevalence data from two repeated cross-sectional surveys and then building a model linking prevalence and incidence. We estimated that HCV prevalence in DU in France increased with age in 2004 and 2011, and decreased over time in all DU. HCV incidence was also dependent on age and declined from 11/100 person-years in 2004 to 6/100 person-years in 2011.

Prior to this work, the only published HCV incidence estimate available in France came from a 2000–2001 cohort of injecting DU in the northeast region of France (excluding Paris and its suburbs) and equalled 9/100 person-years [32].

The present study exhaustively recruited harm reduction facilities and care services for DU in five French metropolitan cities. We included both high- and low-threshold services, which enabled us to obtain a wide range of services serving different profiles within the DU population attending specialized services. Furthermore, the ANRS-Coquelicot survey is the only study that includes biological data to measure HCV prevalence.

Our estimates are consistent with those from some other European countries. In England, Wales and Northern Ireland, the incidence of HCV infection in IDU was estimated at 4–12 infections/100 person-years in 2011 [15] and between 6 and 18/100 person-years in 2013 [33], using an anti-HCV avidity testing method. In Scotland, HCV incidence in IDU was estimated at 10 infections/100 person-years in 2013–2014 [33]. In Australia, HCV incidence in IDU declined over

http://dx.doi.org/10.1017/S0950268816002934

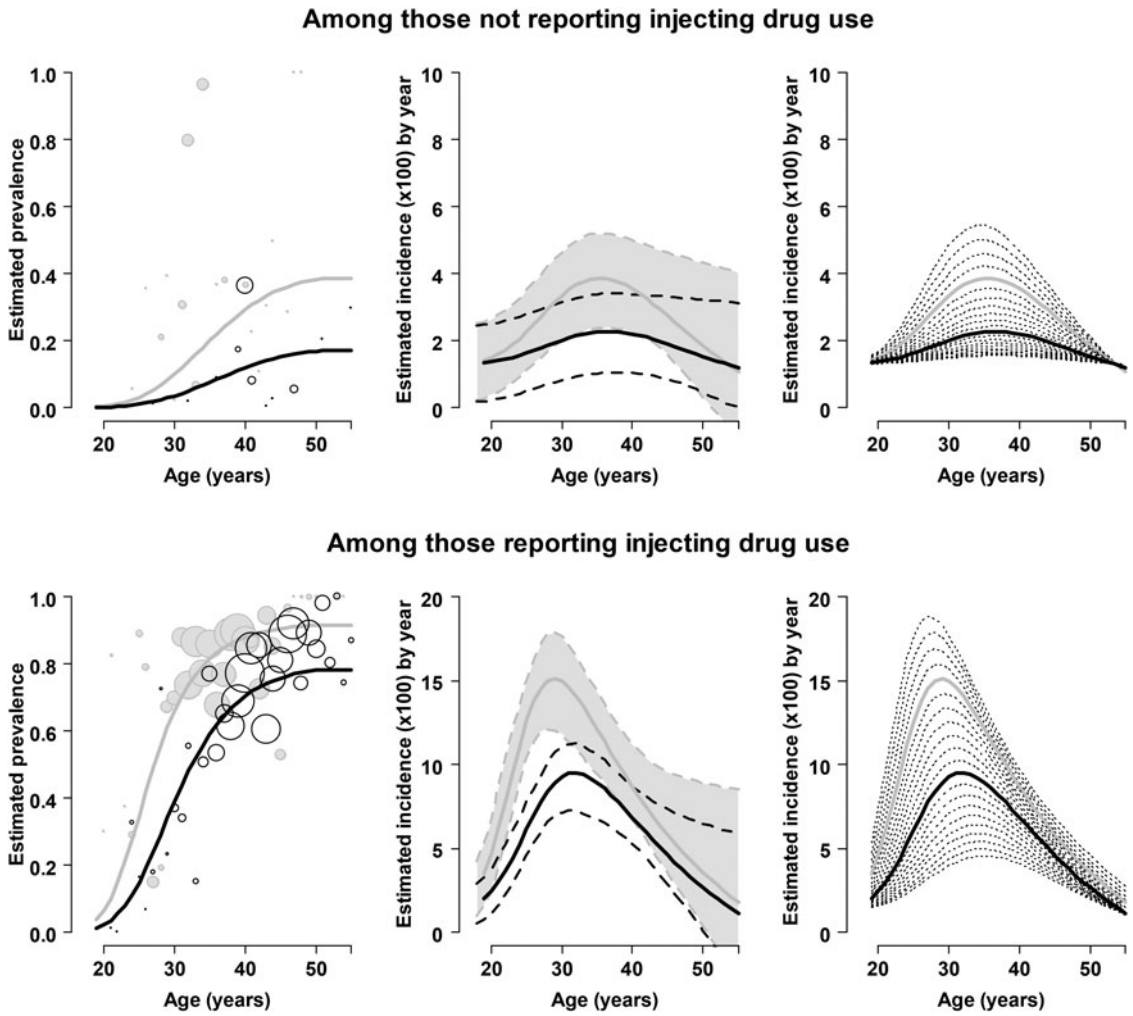


Fig. 3. *Left panels*: Curves represent the age-dependent HCV prevalence estimates from the logistic models in 2004 (grey) and 2011 (black) in those not reporting injecting drug use (*top panels*) and those reporting injecting drug use (*bottom panels*). Circles represent the estimated prevalence by age. Their size is proportional to the number of individuals in 2004 (solid grey circles) and 2011 (open circles). *Middle panels*: Curves represent the age-dependent HCV incidence estimates in 2004 (grey) and 2011 (black) with their confidence intervals (dashed curves) in those not reporting injecting drug use (*top panels*) and those reporting injecting drug use (*bottom panels*). *Right panels*: Age-dependent HCV incidence estimates over 2000–2020 in those not reporting injecting drug use (*top panels*) and those reporting injecting drug use (*bottom panels*). Curves were obtained from the model in 2004 (grey curves), 2011 (black curves) and the other years (dotted curves).

time, between 10 and 15/100 person-years in 2004 to 4/100 (95% CI 1·3–12·3) person-years in 2009 [2].

Our study shows that global HCV incidence was twice as high in AIDU as in IDU in 2011, which explains why the prevalence did not vary significantly between the two surveys in AIDU. Compared to other European countries such as The Netherlands [5] and Switzerland [34], HCV incidence in AIDU remains high in France. In parallel with this work, we estimated a HCV incidence of 49/100 person-years in 2011–2013 in AIDU in Paris and its suburbs, using a biological approach based on HCV RNA positives in negative anti-HCV tests, and an estimated window period of 56 days [35]. The two estimates (11% and 49%) are not directly comparable. The first came from five French cities throughout mainland France, each city having its own characteristics in terms of DU profiles, while the second focused on individuals surveyed in Paris and its suburbs [20]. Some authors have pointed out that in order to use this method effectively, a large sample size is needed due to the short window period unless incidence is very high [15]. Other authors have highlighted that variability in HCV RNA detection and the disease's natural history during early infection may also result in differences between methods [11].

http://dx.doi.org/10.1017/S0950268816002934

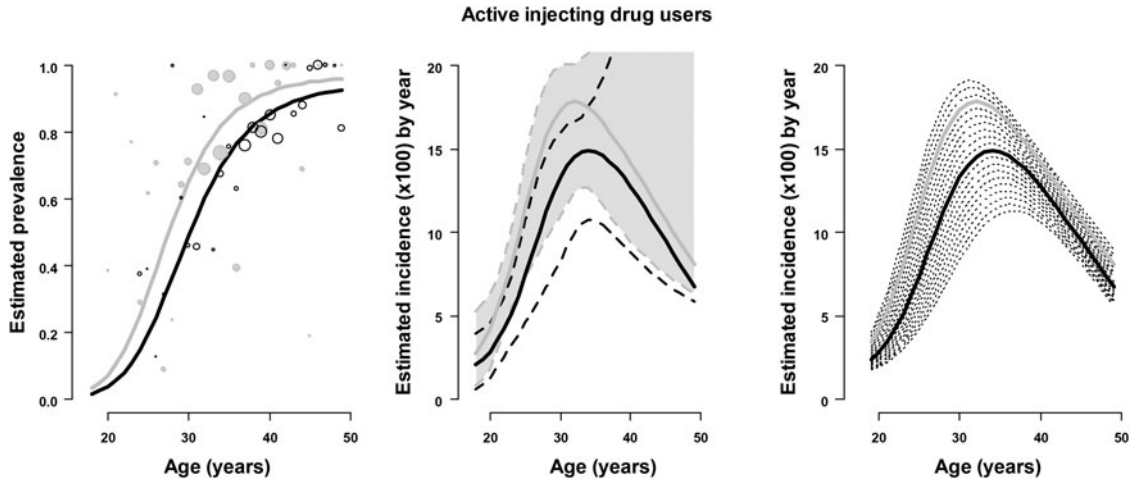


Fig. 4. *Left panel*: Curves represent the age-dependent HCV prevalence estimates from the logistic models in those reporting active injecting drug use in 2004 (grey) and 2011 (black). Circles represent the estimated prevalence by age. Their size is proportional to the number of individuals in 2004 (solid grey circles) and 2011 (open circles). *Middle panel*: Curves represent the age-dependent HCV incidence estimates in active injecting drug users in 2004 (grey) and 2011 (black) with their confidence intervals (dashed curves). *Right panel*: Age-dependent HCV incidence estimates in active injecting drug users over 2000–2020. Curves were obtained from the model in 2004 (grey curves), 2011 (black curves) and the other years (dotted curves).

Our results showed that crack use in the previous month is probably a poor proxy for lifetime risk from crack use. Indeed, it is not crack itself that exposes individuals to the risk of HCV transmission, but the oral lesions caused by chipped and very hot glass pipes. The use of glass pipes regularly results in burns and ulcerated lesions and cuts on lips and in oral cavities. Small amounts of blood may constitute a risk of infection when users share their glass pipes during crack consumption.

Harm reduction measures may contribute to a faster decline in incidence/prevalence and should be taken into account when modelling prevalence and incidence over time. At an international level, harm reduction measures have been greatly improved since the late 1980s [2, 5]. Today they include needle-and-syringe exchange programmes, access to opiate substitution treatment, HCV screening and HCV treatment [2, 3, 5, 36]. In France, access to opioid substitutive treatments has improved but some harm reduction measures, available in other countries, are still not available, such as supervised consumption rooms [37]. In the ANRS-Coquelicot surveys, the following question about needle-and-syringe cleaning (bleach) was included: '*Over the last month, did you at least once use the same water/bleach to clean your needle/syringe?*'. However, the percentage of missing data was high (70%). Furthermore, as the question '*Have you been cured of your HCV infection?*' was asked only in the 2011 survey, we were unable to consider HCV treatment for HCV-positive individuals in our modelling approach.

Our model-based estimates should also be interpreted with caution as they suffer from many potential limitations also present in other studies, some of which are listed by Cullen *et al.* [15]. First, in those not reporting injecting drug use, HCV incidence was estimated at 2/100 (95% CI 0·9–3·2) person-years in 2011. However, this estimation may be a reflection of misclassification bias due to the under-reporting of injecting drug use. Second, to model prevalence, it would have been interesting to include additional variables – such as migration – in the regression models, if that data had been available. Incidence estimation could be greatly improved as the transmission of HCV infection is driven by a basic two-state model. Third, we considered one homogeneous population of DU while other authors have considered two or more populations based on different IDU risk behaviours [3].

Many studies have developed compartmental models with more than two states, stratified by several factors including HIV serostatus, being an injector or not, being on HCV treatment or not, being cured (i.e. those who are anti-HCV positive and HCV-RNA negative) or not [3, 38, 39]. However biological results on HCV RNA to identify those individuals cured of the disease were not available in 2004. It would have been also useful to incorporate parameters depending

http://dx.doi.org/10.1017/S0950268816002934

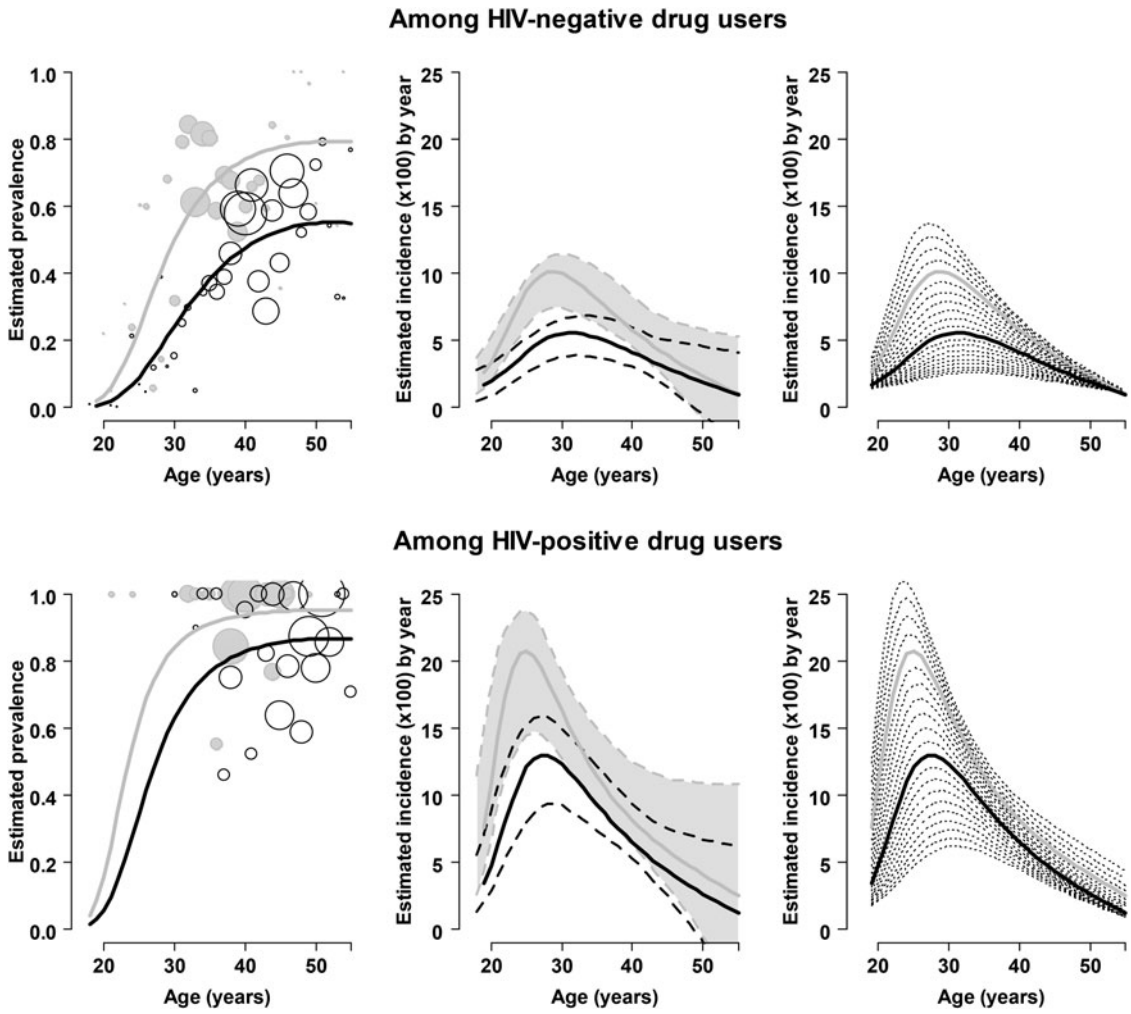


Fig. 5. *Left panels*: Curves represent the age-dependent HCV prevalence estimates from the logistic models in 2004 (grey) and 2011 (black) in HIV-negative drug users (*top panels*) and HIV-positive drug users (*bottom panels*). Circles represent the estimated prevalence by age. Their size is proportional to the number of individuals in 2004 (solid grey circles) and 2011 (open circles). *Middle panels*: Curves represent the age-dependent HCV incidence estimates in 2004 (grey) and 2011 (black) with their confidence intervals (dashed curves) in HIV-negative drug users (*top panels*) and HIV-positive drug users (*bottom panels*). *Right panels*: Age-dependent HCV incidence estimates over 2000–2020 in HIV-negative drug users (*top panels*) and HIV-positive drug users (*bottom panels*). Curves were obtained from the model in 2004 (grey curves), 2011 (black curves) and the other years (dotted curves).

on age (or age group), gender and/or time in the mortality rate, the proportion of new DU. However, precise data on the French DU population are not available [4, 40].

Despite these methodological limitations, we believe that our approach combining a regression model with a compartmental model is an alternative method to estimate incidence from cross-sectional data in the absence of cohort.

Implementing a third cross-sectional survey in DU should be considered, to evaluate whether the decline in HCV incidence has continued since 2011. Despite a potential increase of at-risk behaviours, such a decline is to be expected given recent developments in harm reduction measures and new therapeutic approaches. Since June 2016, all individuals at risk of HCV transmission in France, including IDU, have been eligible for HCV antiviral treatment with new direct-acting antivirals (DAA). Compared with anti-HCV regimens using pegylated interferon and ribavirin, DAA have a very high success rate, better tolerance, a shorter prescribed course and easier adherence. HCV transmission models have shown that even modest increases in successful treatment of HCV infection in persons who inject drugs can decrease prevalence and incidence [3]. Assessing DAA treatments' impact

http://dx.doi.org/10.1017/S0950268816002934

Table 4. *Estimation of HCV incidence (per 100 person-years) in drug users, 18–55 years age group, France, 2004 and 2011*

	2004			2011		
Participants	Sample size	Incidence per 100 py	95% CI	Sample size	Incidence per100 py	95% CI
All drug users	811	7·9	6·4–9·4	1209	4·4	3·3–5·9
Not reporting injecting drug use	216	3·1	1·9–4·5	434	2·0	0·9–3·2
Reporting injecting drug use	594	10·8	9·0–12·8	775	6·1	5·0–8·0
Reporting active injecting drug use (injected during month previous to study interview)	232	15·4	11·9–19·3	252	11·2	9·0–19·0
HIV-negative drug users	753	7·4	5·8–8·9	1111	3·9	2·8–5·4
HIV-positive drug users	58	9·1	7·4–13·3	98	4·6	2·4–7·8

py, Person-years; CI, confidence interval.
One individual did not report if he was an injecting drug user or not. Active injecting drug users are a subgroup of injecting drug users.

on prevalence and incidence in the French context is crucial. Furthermore, the declining trend of injecting drugs observed in most European countries, reflected in the two ANRS-Coquelicot surveys in France, could lead to a decline of prevalence and incidence of HCV.

SUPPLEMENTARY MATERIAL

For supplementary material accompanying this paper visit https://doi.org/10.1017/S0950268816002934.

DECLARATION OF INTEREST

None

REFERENCES

1. **Mathers BM, *et al.*** Mortality among people who inject drugs: a systematic review and meta-analysis. *Bulletin of the World Health Organization* 2013; **91**: 102–123.
2. **Iversen J, *et al.*** Reduction in HCV incidence among injection drug users attending needle and syringe programs in Australia: a linkage study. *American Journal of Public Health* 2013; **103**: 1436–1444.
3. **Martin NK, *et al.*** Hepatitis C virus treatment for prevention among people who inject drugs: Modeling treatment scale-up in the age of direct-acting antivirals. *Hepatology* 2013; **58**: 1598–1609.
4. **Wiessing L, *et al.*** Hepatitis C virus infection epidemiology among people who inject drugs in Europe: a systematic review of data for scaling up treatment and prevention. *PLoS ONE* 2014; **9**: e103345.
5. **De Vos AS, *et al.*** Decline in incidence of HIV and hepatitis C virus infection among injecting drug users in Amsterdam; evidence for harm reduction? *Addiction* 2013; **108**: 1070–1081.
6. **Craine N, *et al.*** Incidence of hepatitis C in drug injectors: the role of homelessness, opiate substitution treatment, equipment sharing, and community size. *Epidemiology and Infection* 2009; **137**: 1255–1265.
7. **Bravo MJ, *et al.*** HCV seroconversion among never-injecting heroin users at baseline: no predictors identified other than starting injection. *International Journal of Drug Policy* 2012; **23**: 415–419.
8. **Sun H-Y, *et al.*** Recent hepatitis C virus infections in HIV-infected patients in Taiwan: incidence and risk factors. *Journal of Clinical Microbiology* 2012; **50**: 781–787.
9. **Luciani F, *et al.*** A prospective study of hepatitis C incidence in Australian prisoners. *Addiction* 2014; **109**: 1695–1706.
10. **Balogun M, *et al.*** Prevalence and incidence of hepatitis C in injecting drug users attending genitourinary medicine clinics. *Epidemiology and Infection* 2009; **137**: 980–987.
11. **Page-Shafer K, *et al.*** Testing strategy to identify cases of acute hepatitis C virus (HCV) infection and to project HCV incidence rates. *Journal of Clinical Microbiology* 2008; **46**: 499–506.
12. **Hope V, *et al.*** Measuring the incidence, prevalence and genetic relatedness of hepatitis C infections among a community recruited sample of injecting drug users, using dried blood spots. *Journal of Viral Hepatitis* 2011; **18**: 262–270.
13. **Busch MP, Shafer KAP.** Acute-phase hepatitis C virus infection: implications for research, diagnosis, and treatment. *Clinical Infectious Diseases* 2005; **40**: 959–961.
14. **Brant L, *et al.*** Diagnosis of acute hepatitis C virus infection and estimated incidence in low-and high-risk English populations. *Journal of Viral Hepatitis* 2008; **15**: 871–877.
15. **Cullen K, *et al.*** Factors associated with recently acquired hepatitis C virus infection in people who inject

drugs in England, Wales and Northern Ireland: new findings from an unlinked anonymous monitoring survey. *Epidemiology and Infection* 2015; **143**: 1398–1407.
16. **Shepherd SJ, *et al.*** A hepatitis C avidity test for determining recent and past infections in both plasma and dried blood spots. *Journal of Clinical Virology* 2013; **57**: 29–35.
17. **Patel EU, *et al.*** Use of hepatitis C virus (HCV) immunoglobulin G antibody avidity as a biomarker to estimate the population-level incidence of HCV infection. *Journal of Infectious Diseases* 2016: jiw005.
18. **Gaudy-Graffin C, *et al.*** Use of an anti-hepatitis C virus (HCV) IgG avidity assay to identify recent HCV infection. *Journal of Clinical Microbiology* 2010; **48**: 3281–3287.
19. **Jauffret-Roustide M, *et al.*** A national cross-sectional study among drug-users in France: epidemiology of HCV and highlight on practical and statistical aspects of the design. *BMC Infectious Diseases* 2009; **9**: 1.
20. **Weill-Barillet L, *et al.*** Hepatitis C virus and HIV seroprevalences, sociodemographic characteristics, behaviors and access to syringes among drug users, a comparison of geographical areas in France, ANRS-Coquelicot 2011 survey. *Revue d'Épidémiologie et de Santé Publique* 2016.
21. **Leon L, Jauffret-Roustide M, Le Strat Y.** Design-based inference in time-location sampling. *Biostatistics* 2015; **16**: 565–579.
22. **Semaille C, *et al.*** Monitoring the dynamics of the HIV epidemic using assays for recent infection and serotyping among new HIV diagnoses: experience after 2 years in France. *Journal of Infectious Diseases* 2007; **196**: 377–383.
23. **Bollaerts K, *et al.*** Estimating the population prevalence and force of infection directly from antibody titres. *Statistical Modelling* 2012; **12**: 441–462.
24. **Korn EL, Graubard BI.** *Analysis of Health Surveys*: John Wiley & Sons, 2011.
25. **Binder H, Sauerbrei W, Royston P.** Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine* 2013; **32**: 2262–2277.
26. **Royston P, Ambler G, Sauerbrei W.** The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**: 964–974.
27. **Shkedy Z, *et al.*** Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine* 2006; **25**: 1577–1591.
28. **Cousien A, *et al.*** Dynamic modelling of hepatitis C virus transmission among people who inject drugs: a methodological review. *Journal of Viral Hepatitis* 2015; **22**: 213–229.
29. **Le Page A, Robertson P, Rawlinson W.** Discordant hepatitis C serological testing in Australia and the implications for organ transplant programs. *Journal of Clinical Virology* 2013; **57**: 19–23.
30. **Smit C, *et al.*** Risk of hepatitis-related mortality increased among hepatitis C virus/HIV-coinfected drug users compared with drug users infected only with hepatitis C virus: a 20-year prospective study. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2008; **47**: 221–225.
31. **Horvitz DG, Thompson DJ.** A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 1952; **47**: 663–685.
32. **Lucidarme D, *et al.*** Incidence and risk factors of HCV and HIV infections in a cohort of intravenous drug users in the North and East of France. *Epidemiology and Infection* 2004; **132**: 699–708.
33. **Anon.** Hepatitis C in the UK. Public Health England, 2014.
34. **Wandeler G, *et al.*** Hepatitis C virus infections in the Swiss HIV Cohort Study: a rapidly evolving epidemic. *Clinical Infectious Diseases* 2012: cis694.
35. **Jauffret-Roustide M, *et al.*** High biological-based HCV incidence and increasing frequency of high-risk practices among IDUs in Paris: What are the implications for harm reduction models? *First European Conference on Addictive Behaviours and Dependencies*, 2015.
36. **Martin NK, *et al.*** Cost-effectiveness of hepatitis C virus antiviral treatment for injection drug user populations. *Hepatology* 2012; **55**: 49–57.
37. **Jauffret-Roustide M, Pedrono G, Beltzer N.** Supervised consumption rooms: the French Paradox. *International Journal of Drug Policy* 2013; **24**: 628–630.
38. **Castro Sanchez AY, *et al.*** A mathematical model for HIV and hepatitis C co-infection and its assessment from a statistical perspective. *Epidemics* 2013; **5**: 56–66.
39. **Hagan H, *et al.*** Hepatitis C virus infection among HIV-positive men who have sex with men: protocol for a systematic review and meta-analysis. *Systematic reviews* 2014; **3**: 1.
40. **Cousien A, *et al.*** Hepatitis C treatment as prevention of viral transmission and liver-related morbidity in persons who inject drugs. *Hepatology* 2015.

http://dx.doi.org/10.1017/S0950268816002934

APPENDIX 1 – DETAILS ON LABORATORY DATA

Fingerstick whole blood was collected on dried blood spots (DBS). Six drops, corresponding to approximately 50 µL of capillary whole blood, were spotted onto filter paper card (Whatman 903™, GE Healthcare Europe GmbH, Freiburg, Germany), The filter paper was then placed onto a horizontal clean dry surface to air dry for at least 1 hour. Each dried DBS was then stored in an individual sealed plastic bag with a desiccant package at -80°C until analysis.

Screening for antibodies to HCV was performed by ELISA using the HCV 3.0 Ortho assay (Ortho-Clinical-Diagnostics, Raritan, NJ) with a sensitivity of 99.1% (95%CI 97.4 - 99.8) and a specificity of 98.2% (95%CI 94.9 - 99.6) [1]. The DBS were cut out with a punch to obtain a circle 6 mm in diameter, which was placed in 250 µl of 0.01 M sodium phosphate buffer containing 10% bovine serum albumin and 0.05% Tween 20, then incubated at room temperature for one hour in an ultrasonic cleaner. The eluted serum samples were directly used to fill the wells of ELISA microplates (200 µL per well). Subsequent steps were carried out in strict compliance with the manufacturer's recommendations.

APPENDIX 2 – DESCRIPTION AND CHOICE OF THE MIXTURE MODEL

We investigated whether the distribution of the quantitative results of anti-HCV antibody tests depended on HIV serostatus, gender, blotting paper quality and year of each survey. Distribution was only associated with year of each survey. We then decided to apply a mixture model for the distributions in 2004 and 2011.

In a *c*-component mixture model, $f_i(x)$ is the distribution for the ith component and p_i is the proportion of samples from the ith component. The overall density of results, F, is a mixture of the *c* component densities,

$$F(x) = \sum_{i}^{c} p_i f_i(x).$$

Each $f_i(x)$ is a normal distribution with mean μ_i and standard deviation σ_i.

The simplest model (also called a 2-component mixture model) includes two states, interpreted in this case as individuals with and individuals without HCV infection. Due to the poor fit of this simplest mixture model, we applied different mixture models with a varying number of components (from 2 to 6). The model with the lowest Bayesian information criterion was chosen and is represented in bold in Table A1.

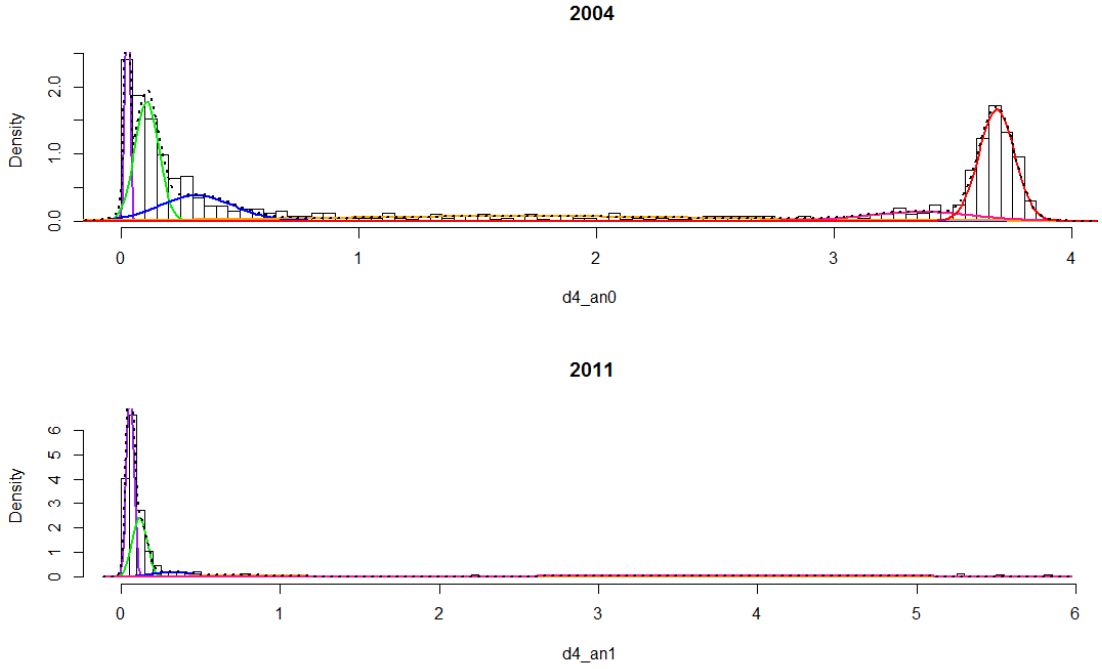
Table A1. Model selection procedure for mixture models, France, 2004 and 2011.

Number of normal components in the mixture	2004		2011	
	AIC	BIC	AIC	BIC
2	1855.17	1878.67	-223.83	-199.53
3	990.33	1072.93	-422.76	-383.88
4	879.9	931.61	-538.52	-485.07
5	800.35	866.16	-554.71	**-486.67**
6	753.35	**833.26**	-557.14	-474.52

* AIC: Akaike information criterion; BIC: Bayesian information criterion

For the year of each survey, the density functions of the components of the finite mixture models provided by the retained model are shown in Figure A1.

Figure A1. The density functions of the components of the finite mixture models, France, 2004 and 2011.



The best model was based on 6 and 5 component distributions (Table A2) in 2004 and 2011, respectively. Table A2 indicates the estimated parameters of the mixtures and our interpretation (anti-HCV status) for each component.

Table A2. Parameters of the final mixture models, France, 2004 and 2011.

ith component (Reactivity level)	Anti-HCV status	2004			2011		
		p_i	μ_i	σ_i	p_i	μ_i	σ_i
1	negative	0.09	0.03	0.01	0.44	0.06	0.02
2	negative	0.23	0.11	0.05	0.28	0.12	0.05
3	negative	0.15	0.32	0.16	0.06	0.32	0.13
4	positive	0.15	1.68	0.85	0.05	0.79	0.29
5	positive	0.07	3.36	0.22	0.16	3.86	1.32
6	positive	0.31	3.68	0.07	-	-	-

A maximum of 6 components was chosen in order to avoid over-fitting the number of mixture model components.

HCV classification obtained from both approaches (threshold and direct methods using the above mixture model) for comparison is presented in Table A3.

Table A3. HCV classification according to the threshold and direct methods, France, 2004 and 2011. ANRS-Coquelicot surveys

Method	Threshold method			
	2004		2011	
Direct method	**0**	**1**	**0**	**1**
0	373	13	747	8
1	0	427	9	478

0 represents the anti-HCV negative results; 1 represents the anti-HCV positive results

A discrepancy about 1.5% was observed between the two classification methods. The direct method allows us to identify anti-HCV antibodies without using the manufacturer's cut-off value [2, 3].

APPENDIX 3 – CONFIDENCE INTERVALS FOR PREVALENCE AND INCIDENCE

We generated 2000 samples from the 2004-2011 combined dataset, using a 6-step process.

- Step 1: Individuals were expanded according to their sampling weights
- Step 2: A random number was assigned to each individual
- Step 3: Individuals were ordered according to the year of each survey and these random numbers
- Step 4: The first *n* individuals were selected to constitute a simple random sample. (*n = 813* for 2004 and *n=1242* for 2011).
- Step 5: Prevalence and incidence were estimated by assuming that β followed a uniform distribution *U(0.01,0.03)*, γ followed a binomial distribution *B(size =2063, probability =0.001)* divided by 2063, and μ_1 followed a Poisson distribution *P(7)* divided by 1000

- Step 6: 95% confidence intervals were obtained with the 2.5th and 97.5th percentiles of the distribution.

Confidence bands for prevalence are presented in Figures A2, A3, A4 and A5 according to each sub-population.

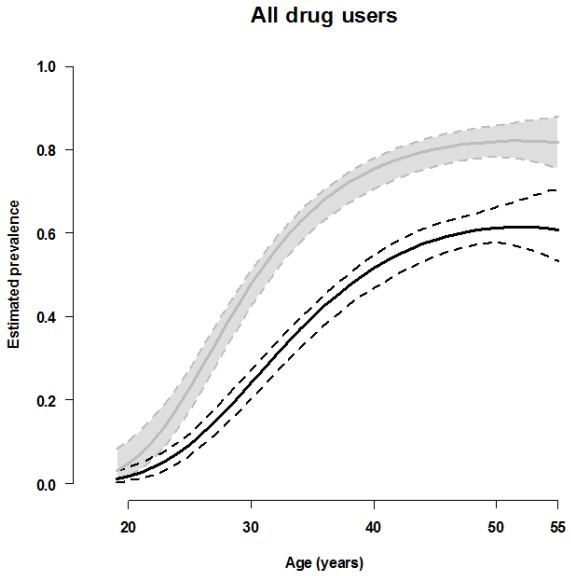


Figure A2. Curves represent the age-dependent HCV prevalence estimates from the logistic models among drug users in 2004 (gray) and 2011 (black) with their confidence intervals (dashed curves).

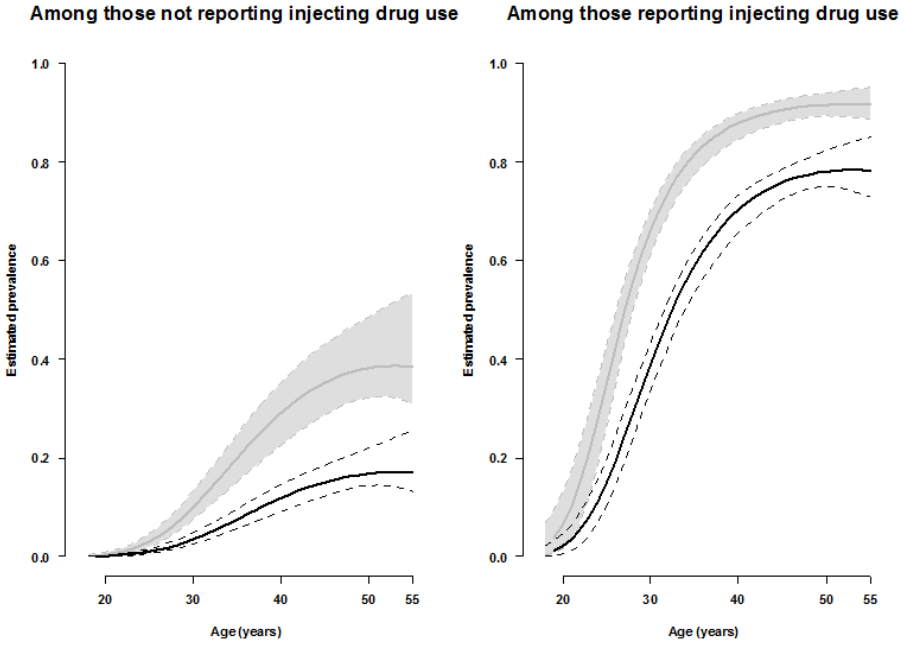


Figure A3. Curves represent the age-dependent HCV prevalence estimates from the logistic models in 2004 (gray) and 2011 (black) with their confidence intervals (dashed curves) among those not reporting injecting drug use (Left) and those reporting injecting drug use (Right).

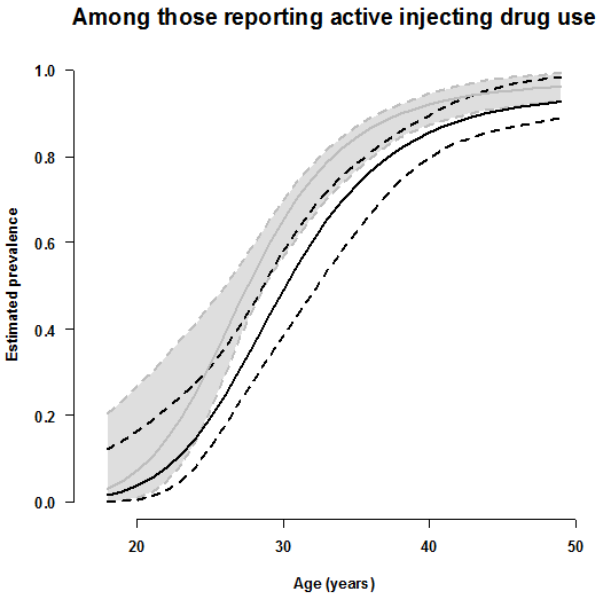


Figure A4. Curves represent the age-dependent HCV prevalence estimates from the logistic models among those reporting active injecting drug use in 2004 (gray) and 2011 (black) with their confidence intervals (dashed curves).

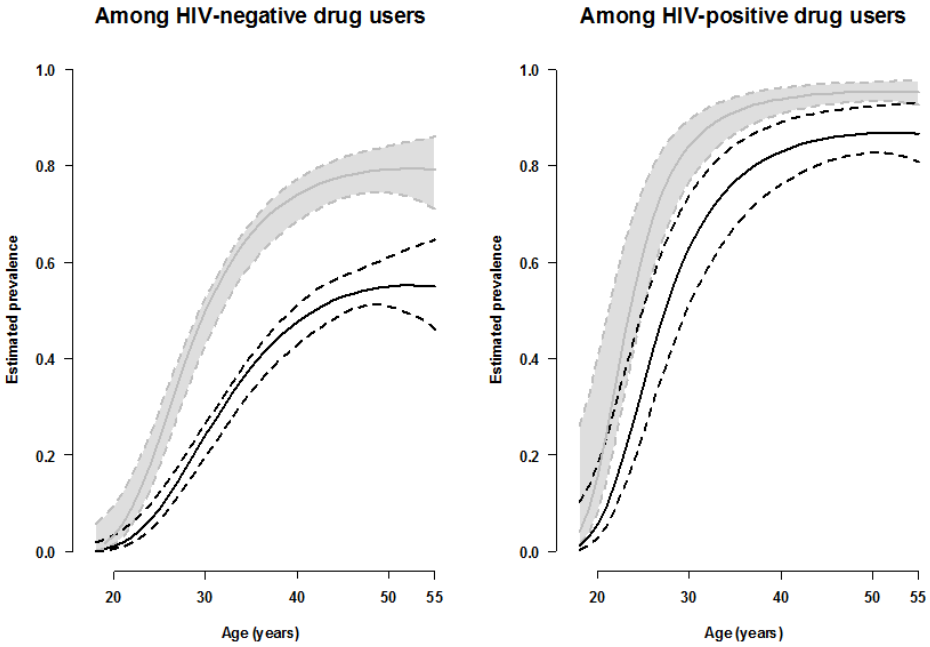


Figure A5. Curves represent the age-dependent HCV prevalence estimates from the logistic models in 2004 (gray) and 2011 (black) with their confidence intervals (dashed curves) among HIV-negative drug users (Left) and HIV-positive drug users (Right).

References

(1) **Soulier A, et al.** Dried blood spots: a tool to ensure broad access to hepatitis C screening, diagnosis and treatment monitoring. *Journal of Infectious Diseases* 2015: jiv423.
(2) **Bollaerts K, et al.** Estimating the population prevalence and force of infection directly from antibody titres. *Statistical Modelling* 2012; **12**(5): 441-462.
(3) **Kafatos G, et al.** Is it appropriate to use fixed assay cut-offs for estimating seroprevalence? *Epidemiology and Infection* 2016; **144**(04): 887-895.

Annexe 5 : Article publié dans la revue *Methodological Innovations*

MI

Original Article

Methodological Innovations
Volume 9 1–9
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2059799116672878
mio.sagepub.com
SAGE

Update on respondent-driven sampling: Theory and practical considerations for studies of persons who inject drugs

Lucie Léon[1], Don Des Jarlais[2], Marie Jauffret-Roustide[1,3] and Yann Le Strat[1]

Abstract

In the last 5 years, more than 600 articles using respondent-driven sampling has been published. This article aims to provide an overview of this sampling technique with an update on the key questions that remain when using respondent-driven sampling, with regard to its application and estimators. Respondent-driven sampling was developed by Heckathorn in 1997 and was based on the principle of individuals recruiting other individuals, who themselves were recruited in previous waves. When there is no sampling frame, respondent-driven sampling has demonstrated its ability to capture individuals belonging to "hidden" or "hard-to-reach" populations in numerous epidemiological surveys. People who use drugs, sex workers, or men who have sex with men are notable examples of specific populations studied using this technique, particularly by public agencies such as the Centers for Disease Control and Prevention in the United States. Respondent-driven sampling, like many others, is based on a set of assumptions that, when respected, can ensure an unbiased estimator. Based on a literature review, we will discuss, among other topics, the effect of violating these assumptions. A special focus is made on surveys of persons who inject drugs. Publications show two major thrusts—methodological and applied researches—for providing practical recommendations in conducting respondent-driven sampling studies. The reasons why respondent-driven sampling did not work for a given population of interest will usually provide important insights for designing health-promoting interventions for that population.

Keywords

Respondent-driven sampling, persons who inject drugs, bias, variance

Introduction

It is crucial to study populations that are at higher risk of contracting infectious diseases in order to implement interventions to prevent transmission of these diseases. People who use drugs, men who have sex with men, sex workers, and some immigrants are examples of populations that are more exposed and therefore vulnerable to HIV, hepatitis B and C, and sexually transmitted infections, in particular (Gile et al., 2015; Le et al., 2010).

However, it is difficult to conduct a sero-epidemiological survey within these populations because of the illicit nature of some practices, such as drug use or sex work (depending on the country's legislation). Moreover, these populations are often stigmatized, and the individuals who comprise them are hard to reach because their practices are hidden (hence the term "hidden population") and their living conditions make it difficult for interviewers to approach them on account of location (e.g. in the case of squatting) or sometimes for safety reasons.

[1]Santé Publique France, French National Public Health Agency, Saint-Maurice, France
[2]Baron Edmond de Rothschild Chemical Dependency Institute, Beth Israel Medical Center, New York, NY, USA
[3]Inserm U988, CNRS UMR 8211, EHESS, Paris Descartes University, Cermes3, Paris, France

Corresponding author:
Lucie Léon, Santé Publique France, French National Public Health Agency, F-94415 Saint-Maurice, France.
Email: lucie.leon@santepubliquefrance.fr

When no sampling frame exists and traditional survey techniques cannot be used, techniques specifically designed for hard-to-reach populations have been proposed and used (Magnani et al., 2005; Spreen, 1992). Examples of such sampling techniques are snowball (Goodman, 1961), network (Birnbaum and Sirken, 1965; Granovetter, 1976; Sudman et al., 1988), targeted (Watters and Biernacki, 1989), random walk (Klovdahl, 1989), adaptive cluster (Thompson and Seber, 1996), time-space (Mackellar et al., 1996; Muhib et al., 2001), and link-tracing (Chow and Thompson, 2003; Félix-Medina and Thompson, 2004; Thompson and Frank, 2000). Capture–recapture is another method to estimate a population size (Ruiz et al., 2016). However, from a statistical standpoint, some of these techniques are not based on a random selection of individuals. This can lead to a bias in the estimates of epidemiological indicators and their variances since some individuals in the population of interest may have a zero probability of being recruited. Other techniques use, in the first stage, lists of places in which individuals belonging to the population of interest can be interviewed, as in the case of time-location sampling (Karon and Wejnert, 2012). As a result, these sampling methods are not useful for surveying hard-to-reach populations.

To overcome these statistical and practical limitations, a new method, called respondent-driven sampling (RDS), was developed by the sociologist Douglas Heckathorn (1997) in the late 1990s. The objectives of this sampling were to build a sample of socially networked individuals belonging to a hidden or hard-to-reach population and produce an unbiased estimate of the functions of interest (e.g. prevalence or strength of association) in this population.

Starting in the early 2000s, this sampling method, seen as innovative particularly for dealing with selection bias, was used to a great extent to survey hard-to-reach populations. Between 2003 and 2007, it was already used in more than 120 studies in 20 different countries, representing more than 32,000 individuals recruited for the only behavioral and biological HIV studies (Malekinejad et al., 2008; Montealegre et al., 2013; White et al., 2015). In 2005, the American Centers for Disease Control and Prevention (CDC, 2009) alone surveyed more than 13,000 drug users in 20 US cities. RDS is also often used in low-income countries, in order to implement surveys on hard-to-reach populations, due to the low cost of this type of study and its ability of reaching many people within a very short time period. In 2010, a special issue in this journal (Vol 5, issue 2) was devoted to methods for hard-to-reach populations. In this issue, RDS was discussed (Johnston and Sabin, 2010) and compared with time-location sampling (Semann, 2010). In 2013 (since the mid-1990s), 460 studies from 69 countries used RDS (White et al., 2015).

Researchers' enthusiasm for this method was the consequence of their many years of frustration with having only the previous methods at their disposal, all of which were known to have major drawbacks. However, RDS, based on strong assumptions, began to be widely used before there was enough time to determine the validity of the method, assess the conditions for applying it, and ensure that the underlying assumptions were respected. Its use outpaced its methodological development, which made the results of some studies open to question.

This article aims to briefly describe the principle of RDS, to identify the main estimators used, in particular the RDS-II estimator that is unbiased under certain assumptions, and to describe the behavior of the RDS-II estimator when certain assumptions are violated. Finally, practical considerations for RDS studies applied to persons who inject drugs (PWID) are discussed.

Principle of RDS

The principle of this sampling is fairly simple. First, the study investigator looks for individuals (called *seeds*) who belong to the population of interest and know a sizable number of individuals in it. The investigator contacts these seeds and, in a location adapted to the survey, administers a questionnaire to them, possibly supplemented by medical examinations and/or biological sampling. When they leave, they are given one or more coupons that bear a unique identifier, the address of the survey premises, and the name of the study. Each seed is to give the coupons to people he or she knows or in some cases more precisely with people with whom they had sexual intercourse or had shared injecting equipment. Quite often, participants are paid to take part in the study and recruit others. Once the seeds have distributed the coupons to their peers, the latter go to the premises to complete the questionnaire and the medical examinations and to recruit others. These persons recruited by the seeds constitute Wave 1, as illustrated in Figure 1. When they have finished, the individuals recruited in Wave 1 are given coupons that they in turn give to others, who make up Wave 2 and also go to the survey location. This process is repeated until the pre-determined sample size is reached. The numbers on the coupons identify who recruited whom to allow researchers to reconstruct the recruitment chains.

RDS can be viewed as a technique for populations described by *small-world theory*, in which any individual in a given population may be indirectly associated, via his or her social network, with any other individual belonging to the same population through approximately six intermediaries (Killworth and Bernard, 1978). According to this theory, starting with a sampling method based on recruitment chains, any individual should be able to be included in the sample with a strictly positive probability.

For a given individual i, the true number of his or her relationships is called the *degree*, noted d_i, that is, the size of this individual's social network within the population of interest. Thus, if the focus is on drug users and if a user knows n users in his or her social network, he or she will be considered to have a degree equal to n. There is probably a difference between the true degree and the reported degree,

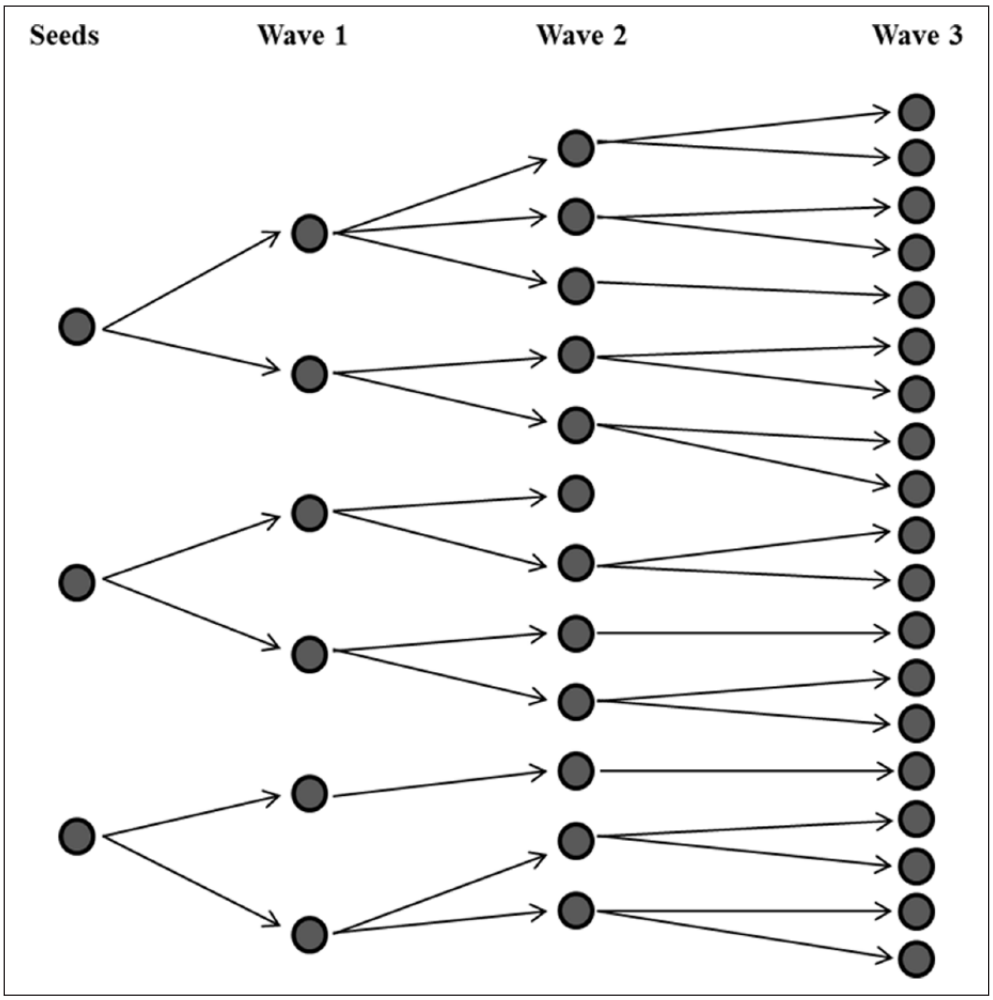


Figure 1. Simplified representation of the first three waves of RDS recruitments. Each circle represents an individual and each arrow represents the distribution of a coupon from an individual to another. For a given wave, individuals are not recruited at the same time and the social network of each individual is not represented.

noted $\widetilde{d}_i$. When an individual tends to recruit persons who resemble him or her, especially with regard to the variable of interest, this is classically defined as homophily, even if several definitions coexist (Crawford et al., 2015; Tyldum and Johnston, 2014). Homophily will be high (1.0) if the infected persons (or non-infected persons, as the case may be) recruit only other infected (or non-infected) persons.

Estimators

RDS is a complex stochastic process since in theory, it is a branching process without replacement, on an arbitrary graph of social relationships that begins with a seed convenience sample (Gile and Handcock, 2010). Thus, if we are interested in a population composed of two groups of individuals (e.g. group A: infected persons and group B: non-infected persons), we can expect that the properties of the estimator of the infected proportion, noted P_A, will not be easy to determine. Those properties will depend both on the characteristics of the social network, preferential recruitment (uncontrolled by the investigator), and on choices (controlled) in the sampling in terms of the number of coupons, the number of waves, and so on.

Several estimators have been proposed to estimate a proportion. Two of the most popular estimators are RDS-I, also called the classical estimator or the SH (Salganik–Heckathorn) estimator (Heckathorn, 2002), and RDS-II, also called the VH (Volz–Heckathorn) estimator (Volz and Heckathorn, 2008). We note s, the respondent-driven sample, and s_A and s_B, the sample of individuals belonging to group A and group B, respectively. We note n_A and n_B, the two sample sizes.

RDS-I (or SH) estimator

Following classical notations (Tomas and Gile, 2011), we note $\widehat{C_{AB}}$, the proportion of individuals recruited by

members of group A who are members of group B; $\widehat{C_{BA}}$, the proportion of individuals recruited by members of group B who are members of group A; and $\widehat{\widetilde{D}}_g$, an estimate of the mean degree in group g (g=A or B)

$$\widehat{\widetilde{D}}_g = \frac{n_g}{\sum_{i \in s_g} \left(\frac{1}{\tilde{d}_i} \right)}$$

The RDS-I estimator for P_A is given by

$$\hat{P}_A = \frac{\widehat{C_{BA}}}{\widehat{C_{BA}} + \widehat{C_{AB}} \left(\frac{\widehat{\widetilde{D}}_A}{\widehat{\widetilde{D}}_B} \right)}$$

RDS-II (or VH) estimator

The RDS-II estimator for P_A is given by

$$\hat{P}_A = \frac{\sum_{i \in s_A} \frac{1}{\tilde{d}_i}}{\sum_{i \in s} \frac{1}{\tilde{d}_i}}$$

Using simulations that compare the mean square errors of the two estimators, it was shown that the performance of RDS-II was almost always superior to that of RDS-I (Gile and Handcock, 2010).

The RDS-II estimator is currently the estimator used in RDSAT (respondent-driven sampling analysis tool), a free program for analyzing data from RDS surveys. Its variance is estimated using a bootstrap method (Salganik, 2006). It has been shown that the RDS-II estimator is asymptotically unbiased under the following assumptions (Volz and Heckathorn, 2008):

1. The sample is selected with replacement.
2. The sampling fraction is small.
3. Each individual recruits only one individual (number of coupons = 1).
4. The respondents state precisely what their degree in the network is.
5. Recruitment of each individual (including seeds) is random.
6. Relationships are reciprocal (undirected network).
7. Population consists of one connected component (every individual can be reached by a finite path from any other individual).

The estimator's lack of bias is therefore based on a priori assumptions, and it is legitimate to ask how this estimator will behave if one or more of those assumptions are violated. Since RDS is a complex process, the estimator's properties are studied through simulations and not through analytical developments. Several recent publications have measured the performance of RDS-II and shown that this estimator could be biased in some circumstances (see subsection "Performances of the RDS-II estimator").

In 2011, Gile (2011) proposed a new estimator called RDS-SS. It is based on successive sampling, equivalent to probability proportional to size without replacement sampling. RDS-SS iteratively estimates both the degree distribution and the inclusion probabilities. Gile shows that the RDS-SS estimator offers an interesting alternative to the RDS-II estimator in terms of bias related to the sampling fraction and the ratio of the average number of degrees between infected and non-infected persons. That being said, this estimator, like others, can induce biases; these are presented in a summary table in an article by Tomas and Gile (2011). Recently, these three RDS estimators have been implemented in the R (R Core Team, free statistical software) packages RDS (Handcock et al., 2009) and RDS Analyst (Handcock et al., 2013). However, the great majority of researchers used a publicly available software application (RDSAT) in which only RDS-II estimator is implemented. We will therefore come back to the latter's performance.

Performances of the RDS-II estimator

Gile and Handcock (2010) measured the performance of RDS-II, in particular the procedure for selecting seeds (non-infected, random and infected seeds), the behavior of respondents (whether homophily is weak or strong (individuals recruit more in their own group)), the sampling fraction, and the mean degree according to disease status (infected or non-infected). The authors show that a bias is induced by seed selection and the level of homophily. The real prevalence (simulated) is underestimated when the seeds are non-infected, and this underestimation is greater when homophily is strong. This is due to the fact that when homophily is strong, non-infected seeds tend to recruit non-infected individuals who will in turn recruit non-infected individuals. Ultimately, the sample will be essentially composed of non-infected persons, leading to an underestimation of prevalence. Bias is close to zero when seeds are selected randomly, regardless of the level of homophily. When the seeds are infected, prevalence will be overestimated, and the stronger the homophily, the greater this overestimation will be.

Gile and Handcock also showed that the bias of the estimator depends on the ratio of the mean number of degrees of infected persons to that of non-infected persons and the sampling fraction. Bias increases when both the sampling fraction and the ratio increase. Thus, when infected persons have a higher mean number of degrees than non-infected persons, prevalence is increasingly underestimated. Another study examined in-depth simulations to test the violation of each assumption (Lu et al., 2012). The main findings of that study indicate a major bias when the network is one-directional

(i.e. when one individual may know another individual but not vice versa) or when respondents choose to recruit individuals who have characteristics correlated with the variable of interest (such as disease status). On the other hand, the authors of that study consider that the estimator is robust with regard to sampling without replacement, low response rate, some degree-reporting errors, and the seed selection criterion. These conclusions are different from those of Gile and Handcock as regards sampling without replacement and seed selection.

A recent simulation-based study showed significant bias if degrees are inaccurately reported (Mills et al., 2014). The authors demonstrated that obtaining correct degrees for individuals reporting low degrees is of particular importance because these individuals have higher weights and are less likely to be infected. It is thus crucial to recover accurate degrees through some relevant questions which represent a real challenge.

The properties of the RDS-II estimator can also be considered in terms of the size of the design effect. Using real data, Goel and Salganik (2010) showed that the design effect could be sizable. Based on their data, they obtain a range for the design effect between 5.7 and 58.3 and a median of 11. In relation to epidemiological surveys using more traditional survey designs, this shows that the variance of the estimator is high. This variance increases when the number of coupons increases and homophily increases (Lu et al., 2012). Recently, even if design effects varied across countries and populations, researchers recommended a design effect between 2 and 4 in RDS studied to estimate the sample size (Johnston et al., 2013; Wejnert et al., 2012).

Violation of assumptions

The simulation studies show that the bias and variance of the RDS-II estimator depend on a set of assumptions (listed in subsection "Estimators") that cannot be controlled by the person in charge of the survey. It can therefore be expected that in reality, these assumptions will not be checked to potential biases. The literature shows that some of the assumptions (assumptions 3–6) are indeed often violated. Examples of this are use of more than one coupon (Johnston et al., 2008; Malekinejad et al., 2008), respondents have difficulty stating precisely what their degree is (Marsden, 2005), non-random recruitment (Frost et al., 2006; Liu et al., 2012; Wang et al., 2005), or not all relationships are two-way (Abramovitz et al., 2009; Iguchi et al., 2009; Ma et al., 2007; Paquette et al., 2011).

Practical considerations for RDS studies of PWID

RDS has probably been used for more studies of PWID than with all other "hard-to-reach" populations combined, and a fair amount of practical knowledge for conducting RDS studies of PWID populations has accumulated (Malekinejad et al., 2008; Rudolph et al., 2011). Whether the RDS assumptions noted above will hold in any specific study of PWID will depend upon a number of practical concerns as well as the underlying theory.

First, how can the researchers determine that the social structure of the population to be studied conforms to the RDS assumption of a fully networked structure? Specifically, whether there are no separations within the population that would have large effects on recruitment. There may be critical divisions within the local PWID population that would greatly reduce the likelihood that people in one subgroup might recruit people in another group. Examples would include groups that inject different drugs, or members of different racial/ethnic groups, or PWID who live in different geographic areas of the same city (Linton et al., 2015).

Qualitative/ethnographic research can often be used to identify potential subgroups within the overall local PWID population where it would be very unlikely that a member of one group would recruit a member of the other group. If this does appear to be the case, then it may be best to reformulate the research as two studies of two different PWID populations. This will, of course, require a sample size for each of the subgroups large enough for the needed statistical analyses. And it will greatly increase the cost of the study. There is no ironclad decision rule for using qualitative data for making the decision to conduct separate RDS studies for different groups within a total local PWID population.

Second, is the sample size large enough to reach "equilibrium"? Equilibrium occurs for an RDS sample, when the important characteristics of the sample (gender, HIV status, age, race/ethnicity, drugs injected, etc.) remain constant over following waves of subject recruitment. Equilibrium is an indication that subject recruitment is no longer determined by the characteristics of the initial seeds. It often requires sample sizes of several hundred to know that one has reached equilibrium. Failure to reach equilibrium creates a strong suspicion that the estimators are biased.

Third, the study needs to have the capability of handling large numbers of subjects at once. One of the virtues of RDS is that it will typically recruit many subjects within a short time period. RDS recruitment is geometric, that is, the number of potential subjects—person with coupons who meet the study eligibility criteria grows quickly. If each subject recruits an average of two additional persons who desire to participate in the study, then the number of persons wanting to participate will double with each recruitment wave. Working with large numbers of subjects then requires scheduling of the research appointments. This can be done as the coupons are given to each subject (the coupon is valid for only a specific time on a specific date) or by asking subjects to come to the research site to schedule an appointment. As people who use drugs are usually not very good at keeping precise appointments, precise scheduling means that some potential subjects will not participate because they did not

present at the research site at the scheduled time. There may be important difference between subjects who do and subjects who do not keep tight appointments, creating another source of possible bias in the study. Having some flexibility to take subjects even though they do not present at the proper time would reduce such bias, but becomes quite difficult when the study staff are trying to process large number of subjects.

Furthermore, to avoid and to control for potential duplication in some surveys, a combination of biometric measures of each respondent can be used (e.g. length of each forearm) (Heckathorn et al., 2002; Uuskula et al., 2011) or other specific identifiers (e.g. mother's maiden name, birth date) (Rudolph et al., 2011).

This section has not attempted to address all of the practical issues that frequently arise in RDS studies of PWID. Rather, it attempts to broaden the discussion beyond the theoretical assumption to some of the practical issues that may be equally important to conducting and interpreting an RDS study. In general, these issues arise from the success of RDS as a method for rapidly recruiting large numbers of subjects within a scientific framework that, if the theoretical assumptions can be met and the practical problems minimized, can combine the most sophisticated sampling method and the greatest cost-efficiency for studying hard-to-reach populations such as PWID.

Discussion

Recent publications show two major thrusts in research on RDS. The first thrust is methodological research. There is a need to continue studying the properties of existing estimators and to improve the estimators and their variances. The second thrust is applied research, which consists in verifying the assumptions when a survey is conducted and addressing the practical issues in conducting RDS studies. These assumptions have often been ignored in the past, making it difficult to discuss the accuracy of the results produced. Very recent articles show that this research is ongoing, with a realization that the method had to be evaluated from different angles (Dombrowski et al., 2013; Lansky et al., 2012; Liu et al., 2012; McCreesh et al., 2011, 2012; Rudolph et al., 2013; Salganik, 2012; Wylie and Jolly, 2013).

However, some questions remain unanswered for anyone wanting to carry out RDS. As regards estimators, the question arises as to whether to continue using the RDS-II estimator and its bootstrap variance or whether the RDS-SS estimator should be used instead, knowing that it has been recently implemented only in the statistical software R. From a more practical standpoint, this raises further questions for the survey investigator: Are the conditions right (i.e. the assumptions are true) to use this sampling method? Should preliminary studies be carried out to determine the characteristics of the social network? Should this type of survey be ruled out in some cases? Salganik (2012) calls for the data from RDSs to be made available so that the evaluation and development of this sampling method can be completed. White et al. (2012) have proposed a set of information to be reported in RDS studies adapted from the STROBE guidelines developed for cross-sectional studies (Von Elm et al., 2007). Finally, diagnostic tools and practical recommendations have been recently proposed to be applied before, during, and after data collection to improve RDS sampling and inference (Gile et al., 2015).

This seems all the more important since this method is being extended to areas other than the study of hard-to-reach populations, particularly its use in telephone surveys (Lee et al., 2011), in web surveys (Schonlau et al., 2014; Stein et al., 2014), or even in recruiting participants to assess the effectiveness of HIV-prevention measures in clinical trials (Solomon et al., 2013).

RDS can be used simply as a time-efficient method of recruiting "hard-to-reach" populations. Paying subjects to recruit other subjects will be usually more cost-effective than paying research staff to recruit subjects in populations with strong social networks. There will be many occasions, however, when RDS does not work as well as the researchers hoped. There may be very high positive homophilies and failure to reach equilibria in the data, suggesting that there is not a single network in the population of interest (a "small world") but rather the population is fragmented into two or more subpopulations that need to be considered separately. It is also possible that the population of interest is not socially networked. For example, commercial sex workers who use the Internet for attracting customers may not be sufficiently networked with each other to sustain recruitment chains. Men who have sex with men who meet in anonymous locations, such as parks or restrooms, also may not be able to recruit each other. For PWID, one of the key of success is to recruit seeds with large networks and with whom people are confident; the question of confidence is crucial for improving the recruitment of PWID with RDS.

In such situations, where either the underlying assumptions in RDS theory do not hold or RDS recruiting does not produce large numbers of subjects, the data may still be analyzed as a convenience sample, with the knowledge that the sample is not representative of the underlying population of interest. Most importantly, if RDS does not work for a given population of interest, the reasons why RDS did not work will usually provide important insights for designing health-promoting interventions for that population. Both RDS and many health-promoting interventions rely on positive peer relationships, and if RDS does not work, the interventions are also likely to not work.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Abramovitz D, Volz EM, Strathdee SA, et al. (2009) Using respondent-driven sampling in a hidden population at risk of HIV infection: Who do HIV-positive recruiters recruit? *Sexually Transmitted Diseases* 36(12): 750–756. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-73949137536&partnerID=40&md5=ad170ecf764cbd76b1ed109c405745b0

Birnbaum ZW and Sirken MG (1965) Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics* 2(11): 1–8.

Centers for Disease Control and Prevention (CDC) (2009) HIV-associated behaviors among injecting-drug users—23 cities, United States, May 2005-February 2006. *Morbidity and Mortality Weekly Report* 58: 329–332.

Chow M and Thompson S (2003) Estimation with link-tracing sampling designs a Bayesian approach. *Survey Methodology* 29(2): 197–205.

Crawford FW, Aronow PM, Zeng Li, et al. (2015) *Identification of homophily and preferential recruitment in respondent-driven sampling* (Unpublished Work). Available at: https://arxiv.org/abs/1511.05397

Dombrowski K, Khan B, Moses J, et al. (2013) Assessing respondent driven sampling for network studies in ethnographic contexts. *Advances in Anthropology* 3(1): 1–9.

Félix-Medina M and Thompson S (2004) Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics* 20(1): 19–38.

Frost SDW, Brouwer KC, Firestone Cruz MA, et al. (2006) Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities: Recruitment dynamics and impact on estimates of HIV and syphilis prevalence. *Journal of Urban Health* 83(7 Suppl.): i83–i97. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-33845654037&partnerID=40&md5=04f1fb2665808d750c313682b6d9d490

Gile KJ (2011) Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association* 106(493): 135–146. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-79954461374&partnerID=40&md5=beb8e610516c8027f776ee9dec7bc6d6

Gile KJ and Handcock MS (2010) Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology* 40(1): 285–327. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-77949340082&partnerID=40&md5=05636a2fae2eeb516971b0d27e0cf5ac

Gile KJ, Johnston L and Salganik M (2015) Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A* 178(1): 241–269.

Goel S and Salganik MJ (2010) Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences USA* 107(15): 6743–6747.

Goodman LA (1961) Snowball sampling. *Annals of Mathematical Statistics* 32(1): 148–170.

Granovetter M (1976) Network sampling: Some first steps. *American Journal of Sociology* 81: 1267–1303.

Handcock MS, Fellows IE and Gile KJ (2013) RDS analyst: Analysis of respondent-driven sampling data. Working paper.

Handcock MS, Gile KJ and Neely WW (2009) *RDS: R Functions for Respondent-Driven Sampling*. R Package version 0.10.

Heckathorn DD (1997) Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems* 44(2): 174–199. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-0003992773&partnerID=40&md5=f004da404fee33b6cde964af2bfbdcaa

Heckathorn DD (2002) Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 49(1): 11–34.

Heckathorn DD, Semann S, Broadhead RS, et al. (2002) Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18–25. *AIDS and Behavior* 6(1). Available at: http://www.respondentdrivensampling.org/reports/steering.pdf

Iguchi MY, Ober AJ, Berry SH, et al. (2009) Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: Sampling methods and implications. *Journal of Urban Health* 86(Suppl. 1): S5–S31. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-68349084634&partnerID=40&md5=ab7b971d61fec7ba7b3883847b1a2413

Johnston LG and Sabin K (2010) Sampling hard-to-reach populations with respondent driven sampling. *Methodological Innovations Online* 5(2): 38–48.

Johnston LG, Chen YH, Silva-Santisteban A, et al. (2013) An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS and Behavior* 17(6): 2202–2210. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84879028371&partnerID=40&md5=e373905562557b90aa2b5e704dcfd53f

Johnston LG, Malekinejad M, Kendall C, et al. (2008) Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: Field experiences in international settings. *AIDS and Behavior* 12(Suppl. 1): S131–S141. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-46149085847&partnerID=40&md5=f455fb9cdecde405f993194d07214e45

Karon JM and Wejnert C (2012) Statistical methods for the analysis of time-location sampling data. *Journal of Urban Health* 89(3): 565–586.

Killworth P and Bernard H (1978) The reversal small-world experiment. *Social Networks* 1: 159–192.

Klovdahl A (1989) Urban social networks: Some methodological problems and possibilities. In: Kochen M (ed.) *The Small World*. Norwood, MA: Ablex Publishing, pp. 176–210.

Lansky A, Drake A, Wejnert C, et al. (2012) Assessing the assumptions of respondent-driven sampling in the national HIV Behavioral Surveillance System among injecting drug users. *Open AIDS Journal* 6: 77–82.

Le VS, Le SY, Barin F, et al. (2010) Population-based HIV-1 incidence in France, 2003–08: A modelling analysis. *The Lancet Infectious Diseases* 10(10): 682–687.

Lee R, Ranaldi J, Cummings M, et al. (2011) Given the increasing bias in random digit dial sampling, could respondent-driven sampling be a practical alternative? *Annals of Epidemiology* 21(4): 272–279.

Linton SL, Cooper HL, Kelley ME, et al. (2015) HIV infection among people who inject drugs in the United States: Geographically explained variance across racial and ethnic groups. *American Journal of Public Health* 105(12): 2457–2465.

Liu H, Li J, Ha T, et al. (2012) Assessment of random recruitment assumption in respondent-driven sampling in egocentric network data. *Social Networks* 1(2): 13–21.

Lu X, Bengtsson L, Britton T, et al. (2012) The sensitivity of respondent-driven sampling. *Journal of the Royal Statistical Society Series A: Statistics in Society* 175(1): 191–216. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84855955571&partnerID=40&md5=8171bf9fa45db898315965e6f8dff954

Ma X, Zhang Q, He X, et al. (2007) Trends in prevalence of HIV, syphilis, hepatitis C, hepatitis B, and sexual risk behavior among men who have sex with men. Results of 3 consecutive respondent-driven sampling surveys in Beijing, 2004 through 2006. *Journal of Acquired Immune Deficiency Syndromes* 45(5): 581–587.

McCreesh N, Frost SDW, Seeley J, et al. (2012) Evaluation of respondent-driven sampling. *Epidemiology* 23(1): 138–147. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-81255204005&partnerID=40&md5=0bcbb318f0a1ed8bf3fe201ef885f3b9

McCreesh N, Johnston LG, Copas A, et al. (2011) Evaluation of the role of location and distance in recruitment in respondent-driven sampling. *International Journal of Health Geographics* 10: 56.

Mackellar D, Valleroy L, Karon J, et al. (1996) The Young Men's Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports* 111(Suppl. 1): 138–144.

Magnani R, Sabin K, Saidel T, et al. (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* 19 (Suppl. 2): S67–S72.

Malekinejad M, Johnston LG, Kendall C, et al. (2008) Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior* 12(4 Suppl): S105–S130.

Marsden P (2005) Recent developments in network measurement. In: Carrington P, Scott J and Wasserman S (eds) *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, pp. 8–30.

Mills HL, Johnson S, Hickman M, et al. (2014) Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug and Alcohol Dependence* 142: 120–126.

Montealegre JR, Johnston LG, Murrill C, et al. (2013) Respondent driven sampling for HIV biological and behavioral surveillance in Latin America and the Caribbean. *AIDS and Behavior* 17(7): 2313–2340.

Muhib FB, Lin LS, Stueve A, et al. (2001) A venue-based method for sampling hard-to-reach populations. *Public Health Reports* 116(Suppl. 1): 216–222.

Paquette DM, Bryant J and de WJ (2011) Use of respondent-driven sampling to enhance understanding of injecting networks: A study of people who inject drugs in Sydney, Australia. *International Journal of Drug Policy* 22(4): 267–273.

Rudolph AE, Crawford ND, Latkin C, et al. (2011) Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: Implications for research and public health practice. *Annals of Epidemiology* 21(4): 280–289.

Rudolph AE, Fuller CM and Latkin C (2013) The importance of measuring and accounting for potential biases in respondent-driven samples. *AIDS and Behavior* 17(6): 2244–2252.

Ruiz MS, O'Rourke A and Allen ST (2016) Using capture-recapture methods to estimate the population of people who inject drugs in Washington, DC. *AIDS and Behavior* 20: 363–368.

Salganik MJ (2006) Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health* 83(6 Suppl.): i98–i112.

Salganik MJ (2012) Commentary: Respondent-driven sampling in the real world. *Epidemiology* 23(1): 148–150. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-83655183647&partnerID=40&md5=6a865f6ae1f465b8389c378ecc8789eb

Schonlau M, Weidmeir B and Kapteyn A (2014) Recruiting an internet panel using respondent-driven sampling. *Journal of Official Statistics* 30(2): 291–310.

Semann S (2010) Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodological Innovations Online* 5(2): 60–75.

Solomon SS, Lucas GM, Celentano DD. (2013) Beyond surveillance: A role for respondent-driven sampling in implementation science. *American Journal of Epidemiology* 178(2): 260–267. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84880546484&partnerID=40&md5=22f8419fccc8b8fe2b47978f6bcc0242

Spreen M (1992) Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin of Sociological Methodology* 36: 34–58.

Stein ML, van Steenbergen JE, Chanyasanha C, et al. (2014) Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: A pilot study in Thailand. *PLoS ONE* 9(1): e85256.

Sudman S, Sirken MG and Cowan CD (1988) Sampling rare and elusive populations. *Science* 240(4855): 991–996.

Thompson S and Frank O (2000) Model-based estimation with link-tracing sampling designs. *Survey Methodology* 26(1): 87–98.

Thompson S and Seber G (1996) *Adaptative Sampling*. New York: John Wiley & Sons.

Tomas A and Gile KJ (2011) The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electronic Journal of Statistics* 5: 899–934. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-83655205915&partnerID=40&md5=14a36018c4b1d1b3f64b9801798560cf

Tyldum G and Johnston LJ (2014) *Applying Respondent Driven Sampling to Migrant Populations*. New York: Palgrave Macmillan.

Uuskula A, Des Jarlais DC, Kals M, et al. (2011) Expanded syringe exchange programs and reduced HIV infection among new injection drug users in Tallinn, Estonia. *BMC Public Health* 11: 517.

Volz E and Heckathorn DD (2008) Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics* 24(1): 79–97.

Von Elm E, Altman DG, Egger M, et al. (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Med* 4(10): e296.

Wang J, Carlson RG, Falck RS, et al. (2005) Respondent-driven sampling to recruit MDMA users: A methodological assessment. *Drug and Alcohol Dependence* 78(2): 147–157.

Watters J and Biernacki P (1989) Targeted sampling: Options for the study of hidden populations. *Social Problems* 36(4): 416–430.

Wejnert C, Pham H, Krishna N, et al. (2012) Estimating design effect and calculating sample size for Respondent-driven sampling studies of injection drug users in the United States. *AIDS and Behavior* 16(4): 797–806. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84863717003&partnerID=40&md5=1b3f117e7f19f8dd458d1e2bb2739dc5

White RG, Hakim AJ, Salganik MJ, et al. (2015) Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies: "STROBE-RDS" statement. *Journal of Clinical Epidemiology* 68(12): 1463–1471.

White RG, Lansky A, Goel S, et al. (2012) Respondent driven sampling–where we are and where should we be going? *Sexually Transmitted Infections* 88(6): 397–399.

Wylie JL and Jolly AM (2013) Understanding recruitment: Outcomes associated with alternate methods for seed selection in respondent driven sampling. *BMC Medical Research Methodology* 13: 93.

Author biographies

Lucie Léon is biostatistician at the Direction of Infectious Diseases of Santé publique France, the French National Public Health Agency in France. She has a Master of Science and is currently working on her doctoral thesis in Biostatistics/Epidemiology.

Don Des Jarlais, Ph.D. is Director of Research for the Baron Edmond de Rothschild Chemical Dependency Institute at Beth Israel Medical Center, a Senior Research Fellow with the National Development and Research Institutes, Inc. and a Guest Investigator at Rockefeller University in New York. As a leader in the fields of AIDS and injecting drug use, Dr. Des Jarlais has published extensively on these topics.

Marie Jauffret-Roustide is Ph.D sociologist at Santé publique France and at the French Institute of Health and Medical Research in France. She leads sero-epidemiological and sociological studies on drug use practices and social processes of at-risk practices among people who inject drugs and crack users.

Yann Le Strat is Ph.D. biostatistician at the Direction of Infectious Diseases of Santé publique France, the French National Public Health Agency in France.

Bibliographie

[1] Greffe hépatique. *Rapport médical et scientifique de l'Agence de la biomédecine* (2013).

[2] *Hepatitis C in the UK*. Public Health England, 2014.

[3] EASL Recommendations on treatment of hepatitis C 2015. *Journal of Hepatology accepted manuscript* (2015).

[4] Epidémie d'hépatite C chez les usagers de drogues : oser prendre de vraies mesures. *Médecins du Monde* (mai 2009), 10 pages.

[5] Aaron, S., McMahon, J. M., Milano, D., Torres, L., Clatts, M., Tortu, S., Mildvan, D., and Simm, M. Intranasal transmission of hepatitis C virus : virological and clinical evidence. *Clinical Infectious Diseases 47*, 7 (2008), 931–934.

[6] Abramovitz, D., Volz, E. M., Strathdee, S. A., Patterson, T. L., Vera, A., and Frost, S. D. W. Using respondent-driven sampling in a hidden population at risk of HIV infection : Who do HIV-positive recruiters recruit ? *Sexually Transmitted Diseases 36*, 12 (2009), 750–756.

[7] AFEF. Recommandations sur la prise en charge des hépatites virales C. *Avec le soutien de la Société de Pathologie Infectieuse de Langue Française* (juin 2015).

[8] ANRS, and AFEF. Prise en charge des personnes infectées par le virus de l'hépatite B ou de l'hépatite C. *Rapport de Recommandations, sous la direction du Pr. Daniel Dhumeaux* (2014).

[9] Ardilly, P. *Les Techniques de sondage*. Editions TECHNIP, Paris, 2006.

[10] Arnaud, A., Chosidow, O., Détrez, M. A., Bitar, D., Huber, F., Foulet, F., Le Strat, Y., and Vandentorren, S. Prevalences of scabies and pediculosis corporis among homeless people in the paris region : results from two randomized cross-sectional surveys (HYTPEAC study). *British Journal of Dermatology 174* (2016), 104–112.

[11] Backmund, M., Reimer, J., Meyer, K., Gerlach, J. T., and Zachoval, R. Hepatitis C virus infection and injection drug users : prevention, risk factors, and treatment. *Clinical Infectious Diseases 40*, Suppl 5 (2005), 330–335.

[12] Balogun, M. A., Murphy, N., Nunn, S., Grant, A., Andrews, N. J., Teo, C. G., Ramsay, M. E., and Parry, J. V. Prevalence and incidence of hepatitis C in injecting drug users attending genitourinary medicine clinics. *Epidemiology and Infection 137* (2009), 980–987.

[13] Barrera, J. M., Francis, B., Ercilla, G., Nelles, M., Achord, D., Darner, J., and Lee, S. R. Improved detection of anti-hcv in post transfusion hepatitis by a third-generation ELISA. *VoxSanguinis 64* (1995), 15–18.

[14] Bennett, S., Gunson, R. N., McAllister, G. E., Hutchinson, S. J., Goldberg, D. J., Cameron, S. O., and Carman, W. F. Detection of hepatitis C virus RNA in dried blood spots. *Journal Clinical Virology 54*, 2 (2012), 106–109.

[15] Birnbaum, Z. W., and Sirken, M. G. Design of sample surveys to estimate the prevalence of rare diseases : three unbiased estimates. *Vital Health Statistics 11*, 2 (1965), 1–8.

[16] Bollaerts, K., Aerts, M., Shkedy, Z., Faes, C., Van der Stede, Y., Beutels, P., and Hens, N. Estimating the population prevalence and force of infection directly from antibody titres. *Statistical Modelling 12*, 5 (2012), 441–462.

[17] Bouhnik, A., Préau, M., Schiltz, M., Lert, F., Abodia, Y., Spire, B., and the VESPA Study Group, . Unprotected sex in regular partnerships among homosexual men living with HIV : a comparison between sero-nonconcordant and seroconcordant couples (ANRS-EN12-VESPA Study). *AIDS 21*, Suppl 1 (2007), S43–S48.

[18] Bourdon, C. *Dépistage de l'infection par le virus de l'hépatite C : adaptation et évaluation d'un test sérologique combiné sur prélèvement de sang capillaire et prélèvement oral.* PhD thesis, Université Joseph FOURIER, Faculté de pharmacie de Grenoble, 2011.

[19] Brant, L., Ramsay, M. E., Balogun, M. A., Boxall, E., Hale, A., Hurrelle, M., Kaluba, L., Klapper, P., Lewis, D., Patel, B., Parry, J., and Irving, W. L. Diagnosis of acute hepatitis C virus infection and estimated incidence in low- and high-risk english populations. *Journal of Viral Hepatitis 15*, 12 (2008), 871–877.

[20] Busch, M. P., and Shafer, K. A. Acute-phase hepatitis C virus infection : implications for research, diagnosis, and treatment. *Clinical Infectious Diseases 40*, 7 (2005), 959–961.

[21] Cadet-Taïrou, A., Saïd, S., and Martinez, M. Profils et pratiques des usagers des CAARUD en 2012. *Tendances, Observatoire Français des Drogues et des Toxicomanes*, 98 (2015), 8 p.

[22] Castro Sanchez, A. Y., Aerts, M., Shkedy, Z., Vickerman, P., Faggiano, F., Salamina, G., and Hens, N. A mathematical model for HIV and hepatitis C co-infection and its assessment from a statistical perspective. *Epidemics 5*, 1 (2013), 56–66.

[23] CDC. HIV-associated behaviors among injecting-drug users–23 cities, United States, May 2005-February 2006. *Morbidity and Mortality Weekly Report 58* (2006), 329–332.

[24] Chevaliez, S., and Pawlotsky, J. Méthodes alternatives au prélèvement sanguin pour le diagnostic de l'infection par le virus de l'hépatite C. *Bulletin Épidémiologique Hebdomadaire 1* (2011), 1–5.

[25] Chow, M., and Thompson, S. Estimation with link-tracing sampling designs a Bayesian approach. *Survey Methodology 29* (2003), 197–205.

[26] Costes, J.-M., Vaissade, L., Colasante, E., Palle, C., Legleye, S., Janssen, E., Toufik, A., and Cadet-Taïrou, A. Prévalence de l'usage problématique de drogues en France - estimations 2006-. *Observatoire Français des Drogues et des Toxicomanies* (2009).

[27] Cousien, A., Tran, V. C., Deuffic-Burban, S., Jauffret-Roustide, M., Dhersin, J. S., and Yazdanpanah, Y. Dynamic modelling of hepatitis C virus transmission among people who inject drugs : a methodological review. *Journal of Viral Hepatitis 22*, 3 (2015), 213–229.

[28] Craine, N., Hickman, M., Parry, J. V., Smith, J., Walker, A. M., Russel, D., Nix, B., May, M., McDonald, T., and Lyons, M. Incidence of hepatitis C in drug injectors : the role of homelessness, opiate substitution treatment, equipment sharing, and community size. *Epidemiology and Infection 137*, 9 (2009), 1255–1265.

[29] Crawford, F. W., Aronow, P. M., Zeng, L., and Li, J. Identification of homophily and preferential recruitment in respondent-driven sampling. *Cornell Unisversity Library* (2015), 19 pages.

[30] Cullen, K. J. adn Hope, V. D., Croxford, S., Shute, J., Ncube, F., and Parry, J. V. Factors associated with recently acquired hepatitis C virus infection in people who inject drugs in England, Wales and Northern Ireland : new findings from an unlinked anonymous monitoring survey. *Epidemiology and Infection 143* (2015), 1398–1407.

[31] De Vos, A. S., van der Helm, J. J., Master, A., Prins, M., and E., K. M. Decline in incidence of HIV and hepatitis C virus infection among injecting drug users in Amsterdam ; evidence for harm reduction ? *Addiction 108*, 6 (2013), 1070–1081.

[32] Delile, J., Reiller, B., Foucher, J., de Ledinghen, V., and Gachie, J. Hépatite C chez les usagers de drogues. *Alcoologie et Addictologie 30*, 4 (2008), 385–394.

[33] Dény, P., and Roulot, D. *Le virus de l'hépatite C.* Elsevier, Paris, France, 2004.

[34] Deville, J., and Lavallée, P. Indirect sampling : the foundations of the Generalised Weight Share Method. *Survey Methodology 32* (2006), 165–176.

[35] Dokubo, E. K., Evans, J., Winkelman, V., Cyrus, S., Tobler, L. H., Asher, A., Briceno, A., and K., P. Comparison of hepatitis C virus RNA and antibody detection in dried blood spots and plasma specimens. *Journal of Clinical Virology 59*, 4 (2014), 223–227.

[36] Dombrowski, K., Khan, B., Moses, J., Channell, E., and Misshula, E. Assessing respondent driven sampling for network studies in ethnographic contexts. *Advances in Anthropology 3*, 1 (2013), 1–9.

[37] EMCDDA. Perspectives on drugs : hepatitis C treatment for injecting drug users. *European Monitoring Centre for Drugs and Drug Addiction* (2015).

[38] Esteban, J. I., Sauleda, S., and Quer, J. The changing epidemiology of hepatitis C virus infection in Europe. *Journal of Hepatology 48* (2008), 148–162.

[39] Etchevers, A., Bretin, P., Lecoffre, C., Bidondo, M. L., Le Strat, Y., Glorennec, P., and Le Tertre, A. Blood lead levels in young children in France, 2008-2009. *International Journal of Hygiene and Environmental Health 217*, 4-5 (2014), 528–537.

[40] Félix-Medina, M., and Thompson, S. Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics 20* (2004), 19–38.

[41] Frost, D. W., Brouwer, K. C., Firestone Cruz, M. A., Ramos, R., Ramos, M. E., Lozada, R. M., Magis-Rogriguez, M., and Strathdee, S. A. Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities : Recruitment dynamics and impact on estimates of HIV and syphilis prevalence. *Journal of Urban Health 83*, 7 (2006).

[42] Gaudy-Graffin, C., Lesage, G., Kousignian, I., Laperche, S., Girault, A., Dubois, F., Goudeau, A., and Barin, F. Use of an anti-hepatitis C virus (HCV) igG avidity assay to identify recent HCV infection. *Journal of Clinical Microbiology 48*, 9 (2010), 3281–3287.

[43] Gay, N. J., Vyse, A. J. Enquselassie, F., Nigatu, W., and Nokes, D. J. Improving sensitivity of oral fluid testing in IgG prevalence studies : application of mixture models to rubella antibody survey. *Epidemiology and Infection 130* (2003), 285–291.

[44] Gile, K. J. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association 106* (2011), 135–146.

[45] Gile, K. J., and Handcock, M. S. Respondent-driven sampling : an assessment of current methodology. *Sociological Methodology 40* (2010), 285–327.

[46] Gile, K. J., Johnston, L., and Salganik, M. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society - Series A 178* (2015), 241–269.

[47] Glynn, S. A., Wright, D. J., Kleinman, S. H., Hirschkorn, D., Tu, Y., Heldebrant, C., Smith, R., Giachetti, C. Gallarda, J., and Busch, M. P. Dynamics of viremia in early hepatitis C virus infection. *Transfusion 45*, 6 (2005), 994–1002.

[48] Goel, S., and Salganik, M. J. Respondent-driven sampling as Markov Chain Monte Carlo. *Statistics in Medicine 28*, 17 (2009), 2202–2229.

[49] Goel, S., and Salganik, M. J. Assessing respondent-driven-sampling. *Proceedings of the National Academy of Sciences 107*, 15 (2010), 6743–6747.

[50] Goodman, L. A. Snowball sampling. *Annals of Mathematical statistics 32*, 1 (1961), 148–170.

[51] Gower, E., Estes, C., Blach, S., Razavi-Shearer, K., and Razavi, H. Global epidemiology and genotype distribution of the hepatitis C virus infection. *Journal of Hepatology 61* (2014), S45–S57.

[52] Granovetter, M. Network sampling : some first steps. *American Journal of Sociology 81* (1976), 1267–1303.

[53] Gustafson, P., Gilbert, M., Xia, M., Michelow, W., Robert, W., Trussler, T., McGuire, M., Paquette, D., Moore, D. M., and Gustafson, R. Impact of statistical adjustment for frequency of venue attendance in a venue-based survey of men who have sex with men. *American Journal of Epidemiology 177*, 10 (2013), 1157–1164.

[54] Hagan, H., Neurer, J., Jordan, A. E., Des Jarlais, D. C., Wu, J., Dombrowski, K., Khan, B., Braithwaite, R., and Kessler, J. Hepatitis C virus infection among HIV-positive men who have sex with men : protocol for a systematic review and meta-analysis. *Systematic Reviews* (2014), 3–31.

[55] Hardelid, P., Williams, D., Dezateux, C., Tookey, P., Peckhal, C., Cubitt, W. D., and Cortina-Borja, M. Analysis of rubella antibody distribution from newborn dired blood spots using finite mixture models. *Epidemiology and Infection 136* (2008), 1698–1706.

[56] Heckathorn, D. D. Respondent-driven sampling : a new approach to the study of hidden populations. *Social Problems 44*, 2 (1997), 174–199.

[57] Heckathorn, D. D. Respondent-driven sampling II : deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems 49*, 1 (2002), 11–34.

[58] Heckathorn, D. D., Semaan, S., Broadhead, R. S., and Hughes, J. J. Extensions of respondent-driven sampling : A new approach to the study of injection drug users aged 18 –25. *AIDS and Behavior 6*, 1 (2002), 55–67.

[59] Hens, N., Aerts, M., Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P., and Beutels, P. Seventy-five years of estimating the force of infection from current status data. *Epidemiology and Infection 138*, 6 (2010), 802–812.

[60] Hesketh, T., Li, L., Ye, X., Wang, H., Jiang, M., and Tomkins, A. HIV and syphilis in migrant workers in eastern China. *Sexually Transmitted Infections 82*, 1 (2006), 11–14.

[61] Hope, V. D., Hickman, M., Ngui, S. L., Jones, S., Telfer, M., Bizarri, M., Ncube, F., and Parry, J. V. Measuring the incidence, prevalence and genetic relatedness of hepatitis C infections among a community recruited sample of injecting drug users, using dried blood spots. *Journal of Viral Hepatitis 18*, 4 (2011), 262–270.

[62] Horvitz, D., and Thompson, D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47* (1952), 663–685.

[63] Houghton, M. Discovery of the hepatitis c virus. *Liver International 29*, Suppl 1 (2009), 82–88.

[64] Iguchi, M. Y., Ober, A. J., Berry, S. H., Fain, T., Heckathorn, D. D., Gorbach, P. M., Heimer, R., Kozlov, A., Ouellet, L. J., Shoptaw, S., and Zule, W. A. Simultaneous recruitment of drug users and men who have sex with men in the United

States and Russia using respondent-driven sampling : Sampling methods and implications. *Journal of Urban Health 86*, 1 (2009), 83–97.

[65] Iversen, J., Wand, H., Topp, L., Kaldor, J., and Maher, L. Reduction in HCV incidence among injection drug users attending needle and syringe programs in Australia : a linkage study. *American Journal of Public Health 103*, 8 (2013), 1436–1444.

[66] Jauffret-Roustide, M., and Le Strat, Y. Conception, échantillonnage, organisation du terrain d'une enquête TLS et nouveaux développements. *Methodological Innovations Online 5 (2)* (2010), 26–37.

[67] Jauffret-Roustide, M., Le Strat, Y., Couturier, E., Thierry, D., Rondy, M., Quaglia, M., Razafandratsima, N., Emmanuelli, J., Guibert, G., Barin, F., and Desenclos, J. C. A national cross-sectional study among drug-users in France : epidemiology of HCV and highlight on practical and statistical aspects of the design. *BMC Infectious Diseases 9* (2009), 113.

[68] Jauffret-Roustide, M., Pedrono, G., and Beltzer, N. Supervised consumption rooms : the French paradox. *International Journal of Drug Policy 24*, 6 (2013), 628–630.

[69] Jauffret-Roustide, M., Pillonel, J., Weill-Barillet, L., Léon, L., Le Strat, Y., Brunet, S., Benoit, T., Chauvin, C., Lebreton, M., Barin, F., and Semaille, C. Estimation de la séroprévalence du VIH et de l'hépatite C chez les usagers de drogues en France - premiers résultats de l'enquête ANRS-COQUELICOT 2011. *Bulletin Épidémiologique Hebdomadaire 39-40* (2013), 504–509.

[70] Jenness, S. M., Neaigus, A., Murrill, C. S., Gelpi-Acosta, C., Wendel, T., and Hagan, H. Recruitment-adjusted estimates of HIV prevalence and risk among men who have sex with men : effects of weighting venue-based sampling data. *Public Health Reports 126* (2011), 635–642.

[71] Johnston, L., Malekinejad, M., C., K., Iuppa, I., and Rutherford, G. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance : field experiences in international settings. *AIDS and Behavior 12*, Suppl 4 (2008), 131–141.

[72] Johnston, L. G., Chen, Y.-H., Silva-Santisteban, A., and F., R. H. An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS and Behavior 17* (2013), 2202–2210.

[73] Judd, A., Parry, J., Hickman, M., T., M., Jordan, L., Lewis, K., Contreras, M., Dusheiko, G., Foster, G., Gill, N., Kemp, K., Main, J., Murray-Lyon, I., and Nelson, M. Evaluation of a modified commercial assay in detecting antibody to hepatitis C virus in oral fluids and dried blood spots. *Journal of Medical Virology 71* (2003), 49–55.

[74] Kafatos, G., Andrews, N. J., McConway, K. J., Maple, P. A. C., Brown, K., and Farrington, C. P. Is it appropriate to use fixed assay cut-offs for estimatig seroprevalence ? *Epidemiology and Infection* (2015), 1–9.

[75] Karon, J. M., and Wejnert, C. Statistical methods for the analysis of time-location sampling data. *Journal of Urban Health 89* (2012), 565–586.

[76] Kassak, K., Mahfoud, Z., Kreidieh, K., Shamra, S., Afifi, R., and Ramia, S. Hepatitis B virus and hepatitis C virus infections among female sex workers and men who have sex with men in Lebanon : prevalence, risk behaviour and immune status. *Sexual Health 8*, 2 (2011), 229–233.

[77] Killworth, P., and Bernard, H. The reversal small-world experiment. *Social Networks 1* (1978), 159–192.

[78] Klovdahl, A. S. Urban social networks : some methodological problems and possibilities. In *The small world*, M. Kochen ed. Ablex Publishing, Norwood, 1989, pp. 176–210.

[79] Lansky, A., Drake, A., Wejnert, C., Pham, H., Cribbin, M., and Heckathorn, D. D. Assessing the assumptions of respondent-driven sampling in the national HIV behavioral surveillance system among injecting drug users. *Open AIDS Journal 6* (2012), 77–82.

[80] Laporte, A., Douay, C., Détrez, M. A., Le Masson, V., Le Méner, E., and Chauvin, P. *La santé mentale et les addictions chez les personnes sans logement personnel d'Île-de-France, premiers résultats.* Instituts thématiques. Inserm, 2010.

[81] Lavallée, P. Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology 21* (1995), 25–32.

[82] Lavallée, P. *Le sondage indirect ou la méthode généralisée du partage des poids.* Editions de l'université de Bruxelles. Editions Ellipses, Bruxelles, 2002.

[83] Lavallée, P. *Indirect sampling.* Springer, New York, 2007.

[84] Lavallée, P., and Caron, P. Estimation using the Generalised Weight Share Method : the case of record linkage. *Survey Methodology 27* (2001), 155–169.

[85] Le Page, A. K., Robertson, P., and Rawlinson, W. D. Discordant hepatitis C serological testing in australia and the implications for organ transplant programs. *Journal of Clinical Virology 57*, 1 (2013), 19–23.

[86] Lee, R., Ranaldi, J., Cummings, M., Crucetti, J. B., Stratton, H., and McNutt, L. A. Given the increasing bias in random digit dial sampling, could respondent-driven sampling be a practical alternative ? *Annals of Epidemiology 21*, 4 (2011), 272–279.

[87] Linton, S. L., Cooper, H. L., Kelley, M. E., Karnes, C. C., Wolfe, M. E., Des Jarlais, D., Semaan, S., DiNenno, E., Finlayson, T., Sionean, C., Wejnert, C., Paz-Bailey, G., and National HIV Behavioral Surveillance study Group, . HIV infection among people who inject drugs in the United States : Geographically explained variance across racial and ethnic groups. *American Journal of Public Health 105*, 12 (2015), 2457–2465.

[88] Liu, H., Li, J., Ha, T., and Li, J. Assessment of random recruitment assumption in respondent-driven sampling in Egocentric Network Data. *Social Networks 1*, 2 (2012), 13–21.

[89] Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., and et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010 : a systematic analysis for the global burden of disease study 2010. *The Lancet 380* (2012), 2095–2128.

[90] Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B., Thorson, A., and Liljeros, F. The sensitivity of respondent-driven sampling. *Journal of the Royal Statistical Society - Series A 175*, Part 1 (2012), 191–216.

[91] Luciani, F., Bretaña, N. A., Teutsch, S. ans Amin, J., Topp, L., Dore, G. J., Maher, L., Dolan, K., Lloyd, A. R., and investigators, H. A prospective study of hepatitis C incidence in Australian prisoners. *Addiction 109*, 10 (2014), 1695–1706.

[92] Lucidarme, D., Bruandet, A., Ilef, D., Harbonnier, J., Jacob, C., Decoster, A., Delamare, C., Cyran, C., Van Hoenacker A., F., Frémaux D. and, Josse, P., Emmanuelli, J., Le Strat, Y., Desenclos, J., and Filoche, B. Incidence and risk factors of HCV and HIV infections in a cohort of intravenous drug users in the North and East of France. *Epidemiology and Infection 132*, 4 (2004), 699–708.

[93] Lucidarme, D., Duburque, C., Bulois, P., and Filoche, B. Evolution of HCV incidence in drug users in France. *Epidemiology and Infection 139* (2011), 1287–1295.

[94] Ma, X., Zhang, Q., He, X., Sun, W., Yue, H., Chen, S., Raymond, H. F., Li, Y., Xu, M., Du, H., and McFarland, W. Trends in prevalence of HIV, syphilis, hepatitis C, hepatitis B, and sexual risk behavior among men who have sex with men. results of 3 consecutive respondent-driven sampling surveys in Beijing, 2004 through 2006. *Journal of Acquired Immune Deficiency Syndromes 45*, 5 (2007), 581–587.

[95] MacKellar, D., Valleroy, L., Karon, J., Lemp, G., and Jansen, R. The young men's survey : methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports 111* (1996), 138–144.

[96] Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS 19*, Suppl 2 (2005), S67–S72.

[97] Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R., Rifkin, M. R., and Rutherford, G. W. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings : a systematic review. *AIDS and Behavior 12*, Suppl 4 (2005), S105–S130.

[98] Marcellin, P., Pequignot, F., Delarocque-Astagneau, E., Zarski, J., Ganna, N., Hillon, P., Antona, D., Bovet, M., Mecgain, M., Asselah, T., Desenclos, J., and Jougla, E. Mortality related to chronic hepatitis B and chronic hepatitis C in France : Evidence for the role of HIV coinfection and alcohol consumption. *Journal of Hepatology 48* (2008), 200–207.

[99] Marsden, P. Recent developments in network measurement. In *Models and Methods in social network analysis*, P. Carrington, J. Scott, and S. Wasserman ed. Ablex Publishing, New York, 2005, pp. 8–30.

[100] Martin, N. K., Hickman, M., Hutchinson, S. J., Goldberg, D. J., and Vickerman, P. Combination interventions to prevent HCV transmission among people who inject drugs : modeling the impact of antiviral treatment, needle and syringe programs, and opiate substitution therapy. *Clinical Infectious Diseases 57 Suppl 2* (2013), S39–S45.

[101] Martin, N. K., Vickerman, P., Grebely, J., Hellard, M., Hutchinson, S. J., Lima, V. D., Foster, G. R., Dillon, J. F., Goldberg, D. J., Dore, G., and Hickman, M. Hepatitis C virus treatment for prevention among people who inject drugs : modeling treatment scale-up in the age of direct-acting antivirals. *Hepatology 58*, 5 (2013), 1598–1609.

[102] Martin, N. K., Vickerman, P., Miners, A., Foster, G. R., Hutchinson, S. J., Goldberg D., J., and Hickman, M. Cost-effectiveness of hepatitis C virus antiviral treatment for injection drug user populations. *Hepatology 55*, 1 (2012), 49–57.

[103] Mathers, B., Degenhardt, L., Bucello, C., Lemon, J., Wiessing, L., and Hickman, M. Mortality among people who inject drugs : a systematic review and meta-analysis. *Bulletin World Health Organization 91* (2013), 102–123.

[104] McCreesh, N., Frost, S. D. W., Seeley, J., Katongole, J., Tarsh, M. N., Ndunguse, R., Jichi, F., Lunel, N. L., Maher, D., Johnston, L. G., Sonnenberg, P., Copas, A. J., Hayes, R. J., and White, R. G. Evaluation of respondent-driven sampling. *Epidemiology 23*, 1 (2012), 138–147.

[105] McCreesh, N., Johnston, L. G., Copas, A., Sonnenberg, P., Seeley, J., Hayes, R., Frost, S. D. W., and White, R. G. Evaluation of the role of location and distance in recruitment in respondent-driven sampling. *International Journal of Health Geographics 10* (2011), 56.

[106] Meffre, C., Le Strat, Y., Delarocque-Astagneau, E., Antona, D., and Desenclos, J. Prévalence des hépatites B et C en France en 2004. *Rapport Institut de Veille Sanitaire* (2007), 114 pages.

[107] Meffre, C., Le Strat, Y., Delarocque-Astagneau, E., Dubois, F., Antona, D., Lemasson, J., Warszawski, J., Steinmetz, J., Coste, D., Myer, J., Leiser, S., Giordanella, J., Gueguen, R., and Desenclos, J. Prevalence of hepatitis B and hepatitis C virus infections in France in 2004 : Social factors are important predictors after adjusting for known risk factors. *Journal of Medical Virology 82* (2010), 546–555.

[108] Messina, J. P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G. S., Pybus, O. G., and Barnes, E. Gobal distribution and prevalence of hepatitis C virus genotypes. *Hepatology 61*, 1 (2015), 77–87.

[109] Meyer, I. H., and Wilson, P. A. Sampling lesbian, gay, and bisexual populations. *Journal of Counseling Psychology 56* (2009), 23–31.

[110] Mills, H., Johnson, S., Hickman, M., Jones, N. S., and Colijn, C. Errors in reported degrees and respondent driven sampling : Implications for bias. *Drug and Alcohol Dependence 142* (2014), 120–126.

[111] Muhib, F. B., Lin, L. S., Stueve, A., Miller, R. L., Ford, W. L., Johnson, W. D., and Smith, P. J. A venue-based method for sampling hard-to-reach populations. *Public Health Reports 116*, Suppl 1 (2001), 216–222.

[112] Nelson, P. K., Mathers, B. M., Cowie, B., Hagan, H., Des Jarlais, D., Horyniak, D., and Degenhardt, L. Global epidemiology of hepatitis B and hepatitis C in people who inject drugs : results of systematic reviews. *The Lancet 378* (2011), 571–583.

[113] Page-Shafer, K., Pappalardo, B. L., Tobler, L. H., Phelps, B. H., Eldin, B. R., Moss, A. R., Wright, T. L., Wright, D. J., O'Brien, T. R., Caglioti, S., and Bush, M. P. Testing strategy to identify cases of acute hepatitis c virus (HCV) infection and to project HCV incidence rates. *Journal of Clinical Microbiology 46* (2008), 499–506.

[114] Paquette, D., Bryant, J., and De Wit, J. Use of respondent-driven sampling to enhance understanding of injecting networks : a study of people who inject drugs in Sydney, Australia. *International Journal of Drug Policy 22*, 4 (2011), 267–273.

[115] Paquette, D., and De Wit, J. Sampling methods used in developed countries for behavioural surveillance among men who have sex with men. *AIDS and Behavior 14* (2010), 1252–1264.

[116] Patel, E. U., Cox, A. L., Mehta, S. H., Boon, D., Mullis, C., Astemborski, J., Osburn, W. O., Quinn, J., Redd, A. D., Kirk, G. D., Thomas, D. L., Quinn, T. C., and Laeyendecker, O. Hepatitis C IgG antibody avidity as a biomarker to estimate population-level incidence. *Journal of Infectious Diseases* (2016).

[117] Pollack, L. M., Osmond, D. H., Paul, J. P., and Catania, J. A. Evaluation of the center for disease control and prevention's HIV behavioral surveillance of men who have sex with men : sampling issues. *Sexually Transmitted Diseases 32* (2005), 581–589.

[118] Risser, J. M., and Montealegre, J. R. Comparison of surveillance sample demographics over two cycles of the National HIV Behavioral Surveillance Project, Houston, Texas. *AIDS and Behavior 18*, Suppl 3 (2014), 382–390.

[119] Rota, M. C., Massari, M., Gabutti, G., Guido, M., De Donno, A., and Ciofi degli Atti, M. L. Measles serological survey in the Italian population : Interpretation of results using mixture model. *Vaccine 26* (2008), 4403–4409.

[120] Royston, P., Ambler, G., and Sauerbrei, W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology 28*, 5 (1999), 964–974.

[121] Rudolph, A. E., Crawford, N. D., Latkin, C., Heimer, R., Benjamin, E. O., Jones, K. C., and Fuller, C. M. Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies : Implications for research and public health practice. *Annals of Epidemiology 21*, 4 (2011), 280–289.

[122] Rudolph, A. E., Fuller, C. M., and Latkin, C. The importance of measuring and accounting for potential biases in respondent-driven samples. *AIDS and Behavior 17*, 6 (2013), 2244–2252.

[123] Salganik, M. J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health : Bulletin of the New York Academy of Medicine 83*, Suppl 6 (2006), 98–112.

[124] Salganik, M. J. Commentary : Respondent-driven sampling in the real world. *Epidemiology 23*, 1 (2012), 148–150.

[125] Särndal, C. E., Swensson, B., and J. Wretman, J. *Model assisted survey sampling*. Springer, New York, 2003.

[126] Schonlau M, Weidmeir B, K. A. Recruiting an internet panel using respondent-driven sampling. *Journal of Official Statistics 30*, 2 (2014), 291–310.

[127] Shepherd, S. J., Kean, J., Hutchinson, S. J., Cameron, S. O., Goldberg, D. J., Carman, W. F., Gunson, R. N., and Aitken, C. A hepatitis C avidity test for determining recent and past infections in both plasma and dried blood spots. *Journal of Clinical Virology 51*, 1 (2013), 29–35.

[128] Shkedy, Z.and Aerts, M., Molenberghs, G., Beutels, P., and Van Damme, P. Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine 25*, 9 (2006), 1577–1591.

[129] Smit, C., van den Berg, C., Geskus, R., Berkhout, B., Coutinho, R., and Prins, M. Risk of hepatitis-related mortality increased among hepatitis C virus/HIV-coinfected drug users compared with drug users infected only with hepatitis C virus : a 20-year prospective study. *Journal of Acquired Immune Deficiency Syndromes 47*, 2 (2008), 221–225.

[130] Solomon, S. S., Lucas, G. M., Celentano, D. D., Sifakis, F., and Mehta, S. H. Beyond surveillance : A role for respondent-driven sampling in implementation science. *American Journal of Epidemiology 178*, 2 (2013), 260–267.

[131] Soulier, A., Poiteau, L., Rosa, I., Hézode, C., Roudot-Thoraval, F., Pawlotsky, J.-M., and Chevaliez, S. Dried blood spots : a tool to ensure broad access to hepatitis C screening, diagnosis and treatment monitoring. *Journal of Infectious Diseases accepted manuscript* (2015).

[132] Spreen, M. Rare populations, hidden populations, and link-tracing designs : what and why ? *Sociological Methodology 36*, 1 (1992), 34–58.

[133] Stanaway, J. D., Flaxman, A. D., Naghavi, M., Fitzmaurice, C., Vos, T., Abubakar, I., Abu-Raddad, L. J., Assadi, R., Bhala, N., Benjamin Cowie, B., Forouzanfour, M. H., Groeger, J., Mohd Hanafi ah, K., Jacobsen, K. H., James, S. L., MacLachlan, J., Malekzadeh, R., Martin, N. K., Mokdad, A. A., Mokdad, A. H., Murray, C. J. L., Plass, D., Rana, S., Rein, D. B., Richardus, J. H., Sanabria, J., Saylan, M., Shahraz, S., So, S., Vlassov, V. V., Weiderpass, E., Wiersma, S. T., Younis, M., Yu, C., Zaki, M. E. S., and Cooke, G. S. The global burden of viral hepatitis from 1990 to 2013 : findings from the Global Burden of Disease Study 2013. *The Lancet* (2016).

[134] Stein, M. L., van Steenbergen, J. E., Chanyasanha, C., Tipayamongkholgul, M., Buskens, V., van der Heijden, P. G. M., Sabaiwan, W., Bengtsson, L., Xin, L., E., T. A., and Kretzschmar, M. E. E. Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens : a pilot study in Thailand. *Plos One 9*, 1 (2014), e85256.

[135] Stramer, S. L. *Blood safety in the new millennium*, american association of blood banks ed. 2001.

[136] Stueve, A., O'Donnell, L. N., Duran, R., Doval, A. S., and Blome, J. Time-space sampling in minority communities : results with young Latino men who have sex with men. *American Journal of Public Health 91* (2001), 922–926.

[137] Sudman, S., Sirken, M. G., and Cowan, C. D. Sampling rare and elusive populations. *Science 240* (1988), 991–996.

[138] Sun, H. Y., Chang, S. Y., Yang, Z. Y., Lu, C. L., Wu, H., Yeh, C. C., Liu, W. C., Hsieh, C. Y., Hung, C. C., and Chang, S. C. Recent hepatitis C virus infections in HIV-infected patients in Taiwan : incidence and risk factors. *Journal of Clinical Microbiology 50*, 3 (2012), 781–787.

[139] Sutton, A. J., Gay, N. J., Edmunds, W. J., Hope, V. D., Gill, O. N., and Hickman, M. Modelling the force of infection for hepatitis B and hepatitis C in injecting drug users in England and Wales. *BMC Infectious Diseases 6* (2006), 93.

[140] Sutton, A. J., McDonald, S. A., Palmateer, N., Taylor, A., and Hutchinson, S. J. Estimating the variability in the risk of infection for hepatitis C in the Glasgow injecting drug user population. *Epidemiology and Infection 140* (2012), 2190–2198.

[141] Thibault, V., Bara, J. L., Nefau, T., and Duplessy-Garson, C. Hepatitis C transmission in injection drug users : could swabs be the main culprit ? *Journal of Infectious Diseases 204*, 12 (2011), 1839–1842.

[142] Thompson, S., and Seber, G. *Adaptitive sampling.* John Wiley & sons, New York, 1996.

[143] Thompson, S. K., and Frank, O. Model-based estimation with link-tracing sampling designs. *Survey Methodology 26* (2000), 87–98.

[144] Tillé, Y. *Théorie des Sondages : échantillonnage et estimation en populations finies.* Dunod, Paris, 2001.

[145] Tillé, Y. *Sampling algorithms.* Springer, New York, 2006.

[146] Tomas, A., and Gile, K. J. The effect of differential recruitment non-response and non-recruitment on estimators for respondent-driven sampling. *Electronic Journal of Statistics, 5* (2011), 899–934.

[147] Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., and Bates, N. *Hard-to-Survey Populations.* Cambridge University Press, 2014.

[148] Tuaillon, E., Mondain, A.-M., Meroueh, F., Ottomani, L., Picot, M.-C., Nagot, N., Van de Perre, P., and Ducos, J. A hepatitis C avidity test for determining recent and past infections in both plasma and dried blood spots. *Hepatology 51*, 3 (2010), 752–758.

[149] Tyldum, G., and Johnston, L. J. *Applying respondent-driven-sampling to migrant populations*. Palgrave Macmillan, New York, 2014.

[150] Uusküla, A., Des Jarlais, D. C., Kals, M., Rüütel, K., Abel-Ollo, K., Talu, A., and Sobolev, I. Expanded syringe exchange programs and reduced HIV infection among new injection drug users in Tallinn, Estonia. *BMC Public Health 11* (2011), 517.

[151] Velter, A., Barin, F., Bouyssou, A., Guinard, J., Léon, L., Le Vu, S., Pillonel, J., Spire, B., and Semaille, C. HIV prevalence and sexual risk behaviors associated with awareness of HIV status among men who have sex with men in Paris, France. *AIDS and Behavior 17*, 4 (2013), 1266–1278.

[152] Volz, E., and Heckathorn, D. D. Probability based estimation theory for respondent-driven sampling. *Journal of Official Statistics 24*, 1 (2008), 79–97.

[153] von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., and Vandenbroucke, J. P. The strengthening the reporting of observational studies in epidemiology (STROBE) statement : guidelines for reporting observational studies. *PLoS Medicine 4*, 10 (2007), e296.

[154] Wandeler, G., Gsponer, T., Bregenzer, A., Günthard, H., Clerc, O., Calmy, A., Stöckle, M., Bernasconi, E., Furrer, H., Rauch, A., and Study, S. H. C. Hepatitis C virus infections in the swiss HIV cohort study : a rapidly evolving epidemic. *Clinical Infectious Diseases 55*, 10 (2012), 1408–1416.

[155] Wang, J., Carlson, R., Falck, R., Siegal, H., Rahman, A., and Li, L. Respondent-driven sampling to recruit MDMA users : a methodological assessment. *Drug and Alcohol Dependence 78*, 2 (2005), 147–157.

[156] Ward, J. The hidden epidemic of hepatitis C virus infection in the United States : Occult transmission and burden of disease. *IAS-USA - Topics in Antiviral Medicine 21*, 1 (2013), 15–19.

[157] Watters, J. K., and Biernacki, P. Targeted sampling : options for the study of hidden populations. *Social Problems 36*, 4 (1989), 416–430.

[158] Weill-Barillet, L., Pillonel, J., Semaille, C., Léon, L., Le Strat, Y., Pascal, X., Barin, F., and Jauffret-Roustide, M. Hepatitis C virus and HIV seroprevalences, sociodemographic characteristics, behaviors and access to syringes among drug users, a comparison of geographical areas in France, ANRS-Coquelicot 2011 survey. *Revue d'Épidémiologie et de Santé Publique* (2016).

[159] Wejnert, C., Le, B., Rose, C. E., Oster, A. M., Smith, A. J., Zhu, J., Paz-Bailey, G., and the NHBS study group. HIV infection and awareness among men who have sex with men 20 cities, United States, 2008 and 2011. *Plos One 8*, 10 (2013).

[160] Wejnert, C., Pham, H., Krishna, N., Le, B., and DiNenno, E. Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. *AIDS ansd Behavior 16* (2012), 797–806.

[161] Werb, D., Garfein, R., Kerr, T., Davidson, P., Roux, P., Jauffret-Roustide, M., Auriacombe, M., Small, W., and Strathdee, S. A. A socio-structural approach to preventing injection drug use initiation : rationale for the PRIMER study. *Harm Reduction Journal 13*, 1 (2016), 10 pages.

[162] White, R. G., Hakim, A. J., Salganik, M. J., Spiller, M. W., Johnston, L. G., Kerr, L., Kendall, C., Drake, A., Wilson, D., Orroth, K., Egger, M., and Hladik, W. Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies : "STROBE-RDS" statement. *Journal of Clinical Epidemiology 68*, 12 (2015), 1463–1471.

[163] Wiessing, L., Ferri, M., Grady, B., Kantzanou, M., Sperle, I., Cullen, K. J., EMCDDA, Hatzakis, A., Prins, M., Vickerman, P., Lazarus, J. V., Hope, V. D., and Matheï, C. Hepatitis C virus infection epidemiology among people who inject drugs in europe : A systematic review of data for scaling up treatment and prevention. *Plos One 9*, 7 (2014), 1–19.

[164] Wolter, K. M., Smith, P., and Blumberg, S. J. Statistical foundations of cell-phone surveys. *Survey Methodology 36* (2010), 203–215.

[165] Wylie, J. L., and Jolly, A. M. Understanding recruitment : outcomes associated with alternate methods for seed selection in respondent driven sampling. *BMC Medical Research Methodology* (2013), 13–93.

[166] Xia, Q., and Torian, L. V. To weight or not to weight in time-location sampling : why not do both ? *AIDS and Behavior 17* (2013), 3120–3123.

Titre : Estimation de prévalences et d'incidences à partir d'enquêtes épidémiologiques transversales répétées auprès de populations difficiles d'accès. Application au virus de l'hépatite C chez les usagers de drogues en France

Mots clefs : Populations difficiles d'accès, Usagers de drogues, Virus de l'hépatite C, Techniques de sondages, Prévalence, Incidence

Résumé : Le virus de l'hépatite C (VHC) est un problème majeur de santé publique dont les usagers de drogues (UD) constituent la principale source de contamination en France. Réaliser des enquêtes séro-épidémiologiques auprès de cette population pour suivre la dynamique du VHC s'avère difficile notamment en raison de leurs pratiques illicites. Cette population est en partie accessible par les lieux d'enquêtes et en partie "cachée" car ne fréquentant aucun lieu répertorié. Pour enquêter chaque partie, nous avons considéré l'échantillonnage lieux-moments (TLS) puis l'échantillonnage conduit par les répondants. Après avoir formalisé le TLS dans le cadre d'un sondage indirect, nous avons proposé un estimateur pour un total et une proportion, qui tient compte de la fréquentation multiple et hétérogène des lieux d'enquêtes. Nous recommandons cette méthode pour estimer la prévalence d'une maladie dans des études auprès de populations fréquentant des services, même en cas d'erreurs sur les fréquentations déclarées par les participants. L'enquête ANRS-Coquelicot réalisée en 2004 auprès des UD fréquentant des centres dédiés, puis répétée en 2011, a permis d'estimer la prévalence du VHC à 43,7%. A partir des deux enquêtes, nous avons ensuite estimé l'incidence de l'infection à VHC par âge et en fonction du temps en construisant un modèle mathématique reposant sur la formulation d'une relation entre la prévalence et l'incidence. Ce modèle consistait en la combinaison d'un modèle compartimental et d'un modèle de régression. L'incidence de l'infection à VHC a ainsi été estimée à 4,4/100 personnes-années en 2011. Cette approche est une alternative satisfaisante pour estimer l'incidence d'une maladie à partir d'enquêtes épidémiologiques transversales en l'absence de cohorte ou de tests biologiques permettant d'identifier les infections récentes. Compte tenu de la baisse de la prévalence, des mesures de réduction des risques et des avancées thérapeutiques, une diminution de l'incidence de l'infection à VHC devrait se poursuivre malgré une potentielle augmentation des comportements à risque des UD.

Title : Estimation of prevalences and incidences from repeated cross-sectional seroepidemiological surveys in hard-to-reach populations: application to hepatitis C among drug users in France

Keywords : Hard-to-reach populations, Drug users, Hepatitis C Virus, Design surveys, Prevalence, Incidence

Abstract : Hepatitis C virus (HCV) is a public-health issue that drug users (DU) remain the major source of contamination in France. Conducting seroepidemiological surveys among this population to assess the HCV dynamic is difficult particularly due to their illicit practices. This population can be accessible through survey locations or can be hidden (who does not visit any location). To survey each part, we presented time-location sampling (TLS) and respondent-driven sampling. We presented TLS in the context of an indirect sampling and proposed a design-based inference taking into account the frequency of venue attendance (FVA) to estimate a total or a proportion. We recommend this method for estimating the prevalence of a disease in surveys among hard-to-reach populations, even if errors occur in the FVA reported by the participants. The ANRS-Coquelicot survey carried out in 2004 among DU attending centres providing services to drug users, then repeated in 2011, allowed us to estimate the HCV prevalence at 43.7%. Using these two surveys, we estimated age- and time-dependent HCV incidence from a mathematical model linking prevalence and incidence. This model consisted in combining a compartmental model with a regression model. The HCV incidence was thus estimated at 4.4/100 person-years in 2011. This method is an alternative approach to estimate incidence of a disease from cross-sectional epidemiological data in the absence of cohort or biological tests to identify acute infections. The decline in HCV incidence is to be expected given decreasing prevalence, recent developments in harm reduction measures and new therapeutic approaches despite a potential increase of at-risk behaviors.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France