



Les méthodes et indicateurs en épidémiologie

Et quels usages en faire en sciences sociales

Kévin JEAN, Laboratoire MESuRS, Cnam Paris
kevin.jean@lecnam.net



Plan

1. Quelques repères historiques
2. Épidémiologie descriptive
3. Épidémiologie analytique
4. Quelques études en épidémiologie sociale
5. L'apport des autres disciplines quanti à l'épidémiologie

La figure emblématique de John Snow

- Londres, 1854
 - Epidémie de choléra
 - Croyance répandue : le choléra se transmet par l'air
 - Suspicion de John Snow : le choléra se transmet par l'eau
- ⇒ Cartographie de Londres



Carte de John Snow :

emplacements des pompes et cas de choléra



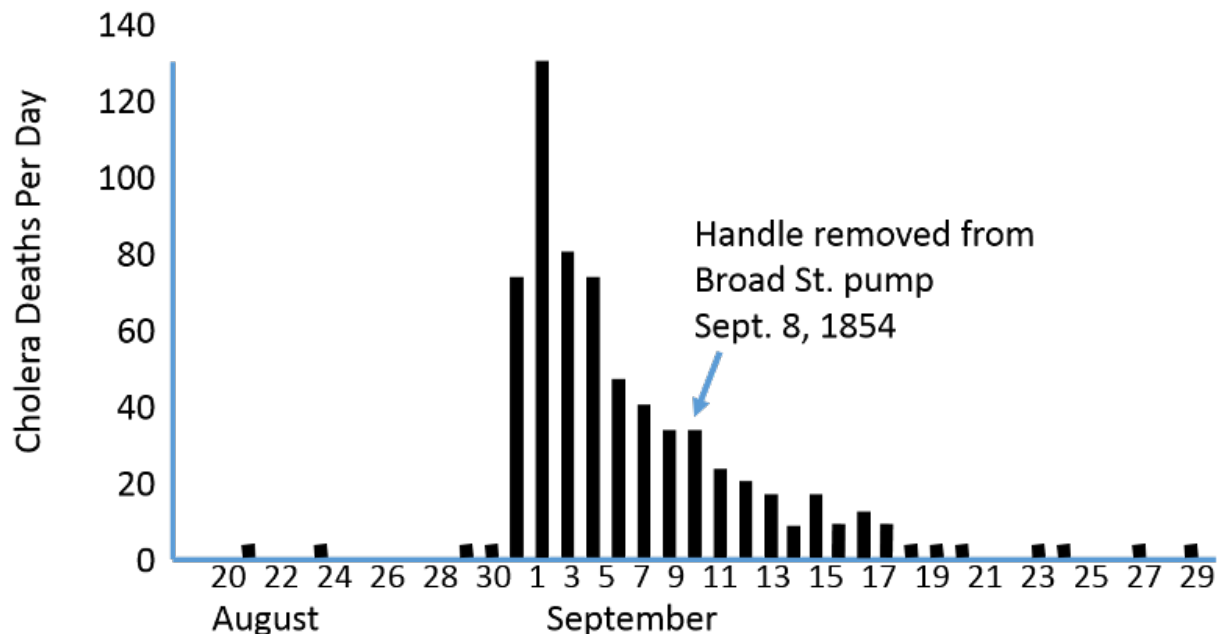
Données chiffrées

Water Supply	No. of Houses	Deaths from Cholera	Deaths per 10,000 houses
Southwark and Vauxhall Co.	40,046	1,263	315
Lambeth Co.	26,107	98	38
Other districts	256,423	1,422	56

*Southwark s'alimente en aval de Londres, Lambeth en amont
⇒ Snow suspecte l'eau souillée d'être à l'origine de l'épidémie*

Actions de John Snow et conséquences sur l'épidémie

- 7 septembre 1854 : Snow expose ses suspicions au « conseil de quartier »
- 8 septembre 1854 : pompe de Broad Street fermée





Définition

Epidémiologie (définition CDC)

Étude de la dynamique des *phénomènes de santé* dans les populations, dans le but de mettre en évidence les *facteurs* qui les déterminent et leur *rôle*, et de mettre en œuvre les *mesures de correction* appropriées

- **Décrire, comprendre, agir**

Définition

Épidémiologie sociale:

La branche de l'épidémiologie qui s'intéresse à la façon dont les **structures et les interactions sociales** conduisent à différentes expositions, et donc **différents états de santé** au sein des individus composant une population.

Kaufman et Oakes, 2017



Épidémiologie descriptive vs. analytique

Descriptive

- Étudie les **variations** dans la fréquence des maladies
 - Où?
 - Quand?
 - Qui?

Analytique

- Étudie le lien causal entre une maladie et une exposition
- Vise à quantifier cette relation



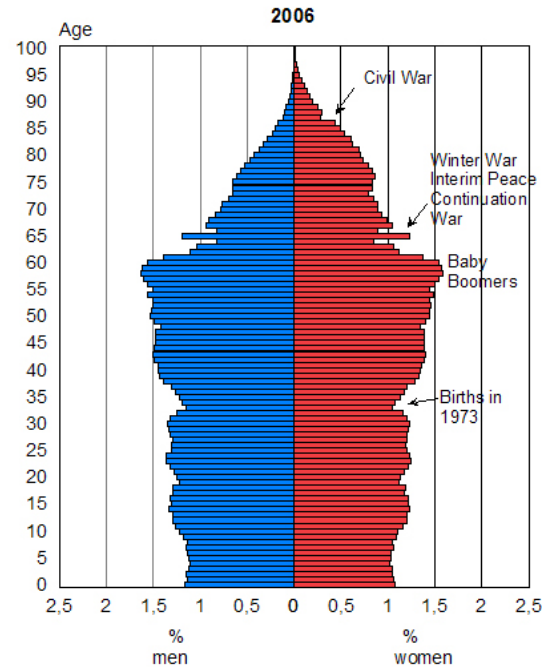
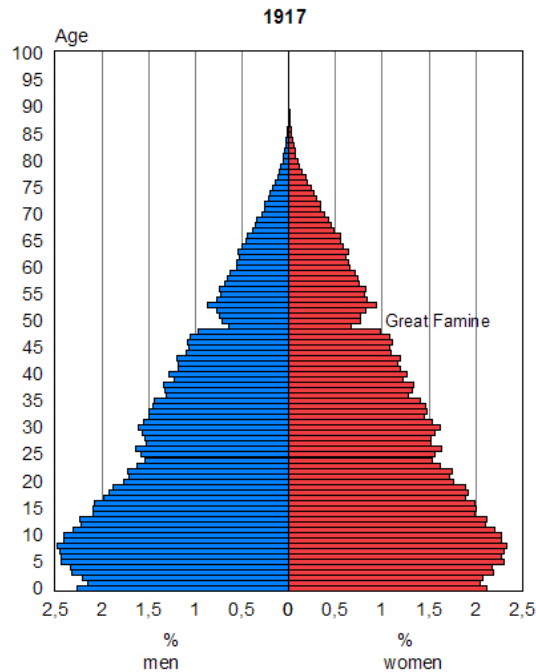
Indicateurs épidémiologiques

- **Rapports** (ratios) : relation entre deux quantités
 - Exemple : hommes/femmes *Sans unité, <, > ou = à 1*
- **Proportions** : un rapport dont le dénominateur inclut le numérateur
 - Exemple : garçons / totalité des naissances *Sans unité, toujours ≤ 1
Peut être exprimé en %*
- **Taux** : vitesse de transfert d'un état à un autre
 - Exemple : nombre de décès / (population totale x nb unités de temps) *Sans unité, toujours ≤ 1
Peut être exprimé en %, ‰*



Exemple d'un taux en santé

- Comparer des taux bruts masque les différences d'âge



Standardisation directe ou indirecte

Figure 1: Concept of direct standardization

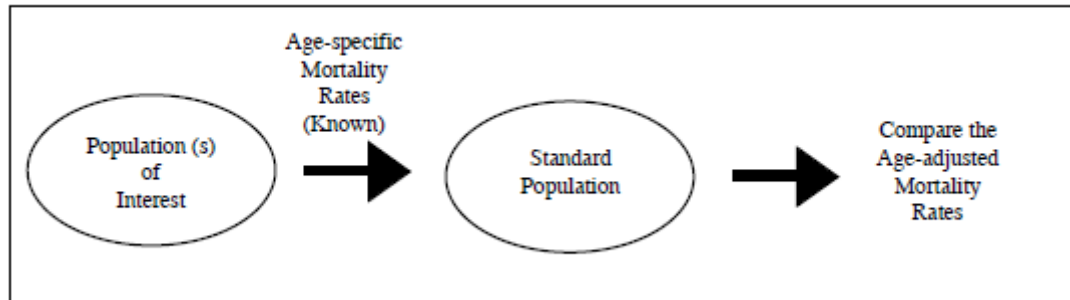
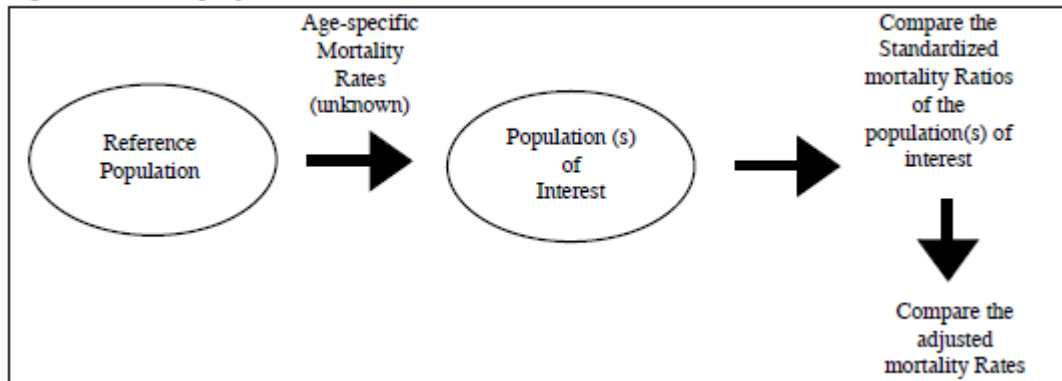


Figure 3: Concept of indirect standardization



Exemple de standardisation indirecte

14 décès par cirrhose observée chez une population professionnelle d'hommes médecins sur une année (Norvège, 2010)

	Population de référence: Norvège (hommes)	Population d'étude: médecins (hommes)	
Age	Taux de décès annuel par cirrhose (x10,000)	Nb	Nb décès attendus
25-34	0.13	12860	
35-44	0.47	11510	
45-54	1.62	10330	
55-64	2.71	7790	
Total			

Exemple d'application

14 décès par cirrhose observée chez une population professionnelle d'hommes médecins sur une année (Norvège, 2010)

	Population de référence: Norvège (hommes)	Population d'étude: médecins (hommes)	
Age	Taux de décès annuel par cirrhose (x1,000)	Nb	Nb décès attendus
25-34	0.13	12860	0.17
35-44	0.47	11510	0.54
45-54	1.62	10330	1.67
55-64	2.71	7790	2.11
Total			4.49

$$\text{SMR} = O/A = 14 / 4.49 = 3.11$$



Au-delà de la mortalité

Il peut être pertinent de s'intéresser à des indicateurs de santé:

- réversibles ou transitoires
- non lié à la mortalité mais tout de même pertinent dans une perspective de santé

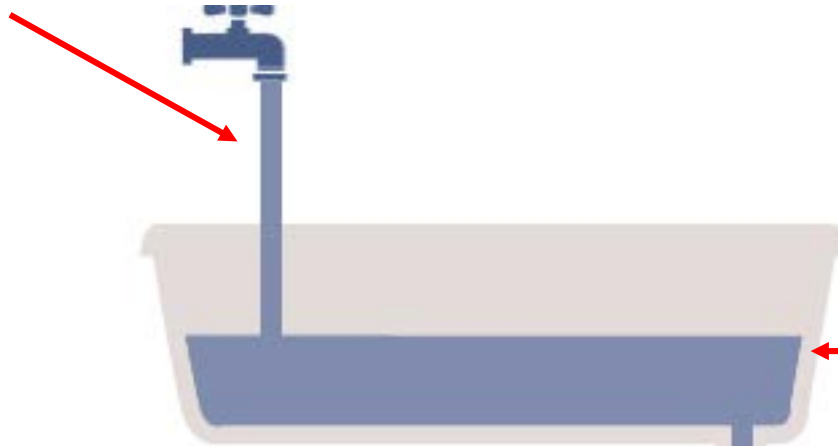
Pour cela l'épidémiologiste s'intéresse à:

- l'incidence: **nombre de nouveaux cas** d'un phénomène de santé qui adviennent au cours d'une **période de temps** donnée, dans une population à risque (**taux**)
- la prévalence: **proportion de personnes** d'une population à risque qui sont **affectées** par un phénomène de santé **à un instant donné** (**proportion**)



Incidence & Prévalence

Incidence = nouveaux cas



Prévalence = cas existants

Décès
Guérison
Emigration

Dramatic increase in HIV prevalence after scale-up of antiretroviral treatment

Jaffer Zaidi^a, Erofil Grapsa^a, Frank Tanser^a,
Marie-Louise Newell^{a,b} and Till Bärnighausen^{a,c}

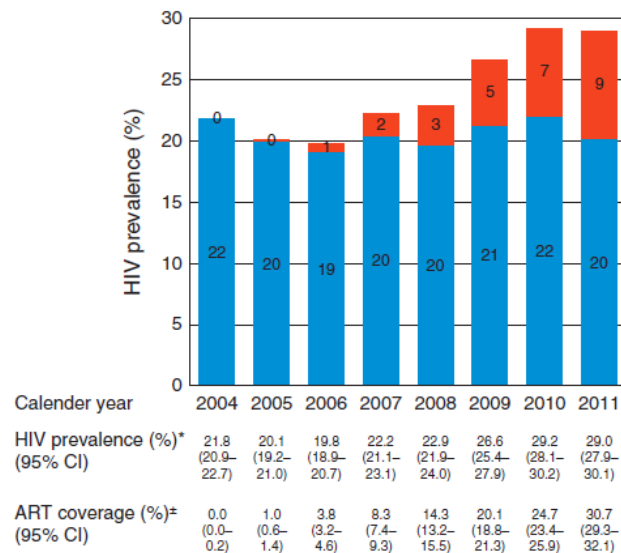


Fig. 1. HIV prevalence and ART coverage trends, 2004–2011. ART, antiretroviral treatment; CI, confidence interval. The figure is based on the analysis of complete cases. *HIV prevalence is expressed in percent of all adults aged 15–49 years. †ART coverage is expressed in percent of all HIV-infected adults aged 15–49 years. The total bars in the figure represent overall HIV prevalence. The red components of the bars represent the HIV-infected people who are receiving ART (in percent of the entire population aged 15–49 years); the blue components of the bars represent the HIV-infected people who are not receiving ART (in percent of the entire population aged 15–49 years).

Épidémiologie analytique

- Ou épidémiologie étiologique = recherche des causes
- S'intéresse généralement à une relation maladie-exposition
- Vise à quantifier une association
- Puis discute de l'éventuel caractère causal de cette association



Mesures du risque en épidémiologie

- Incidence (mais pas prévalence)

- Risque relatif (RR)

- Odds-ratio (OR)

- Excès de risque

- Fraction étiologique du risque

- Fraction de Risque attribuable

Mesure absolue

Mesure relative:

- Permet de comparer 2 sous-pop
- Exposés vs. Non exposés

Mesures d'impact:

- Combien de cas attribuables à une exposition
- Combien de cas évitables?

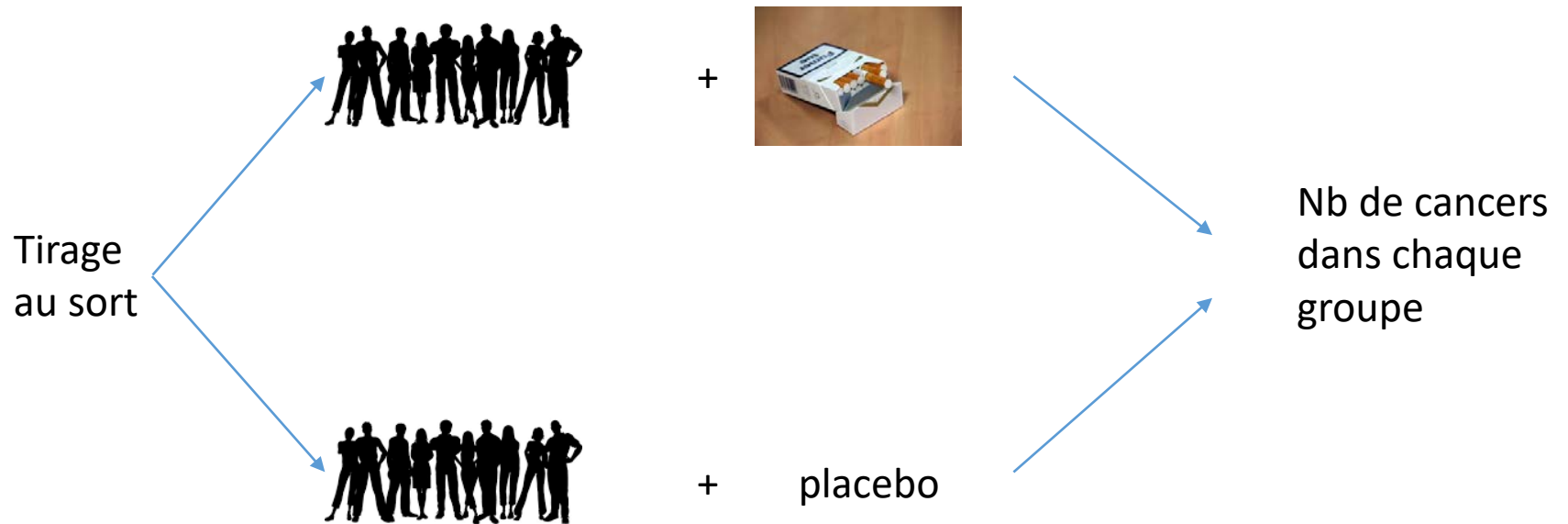
Risque (absolu)

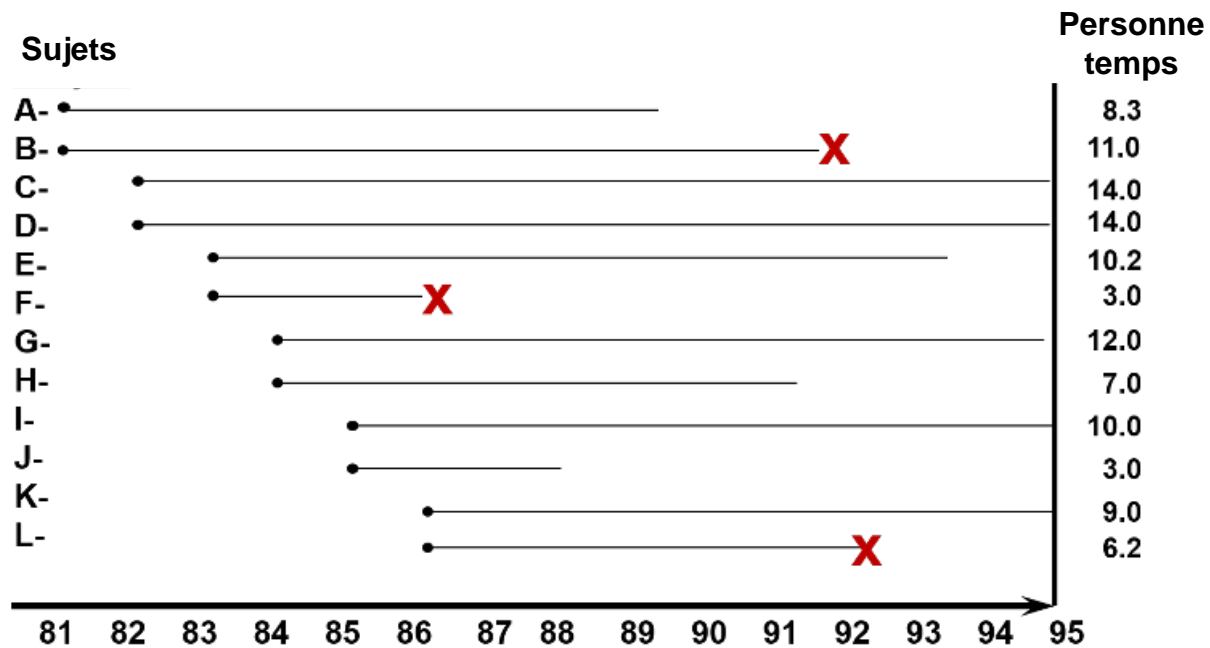
- **Définition** : probabilité de survenue d'un événement (maladie, accident..) au sein d'une population donnée au cours d'une période de temps donnée
- **Notion populationnelle**
- **Mesure du risque** : incidence

Épidémiologie analytique

- L'étude d'association entre une **exposition E** (ou facteur de risque) et la **maladie M** (ou autre événement) est une des étapes majeures de la recherche des facteurs étiologiques des maladies.

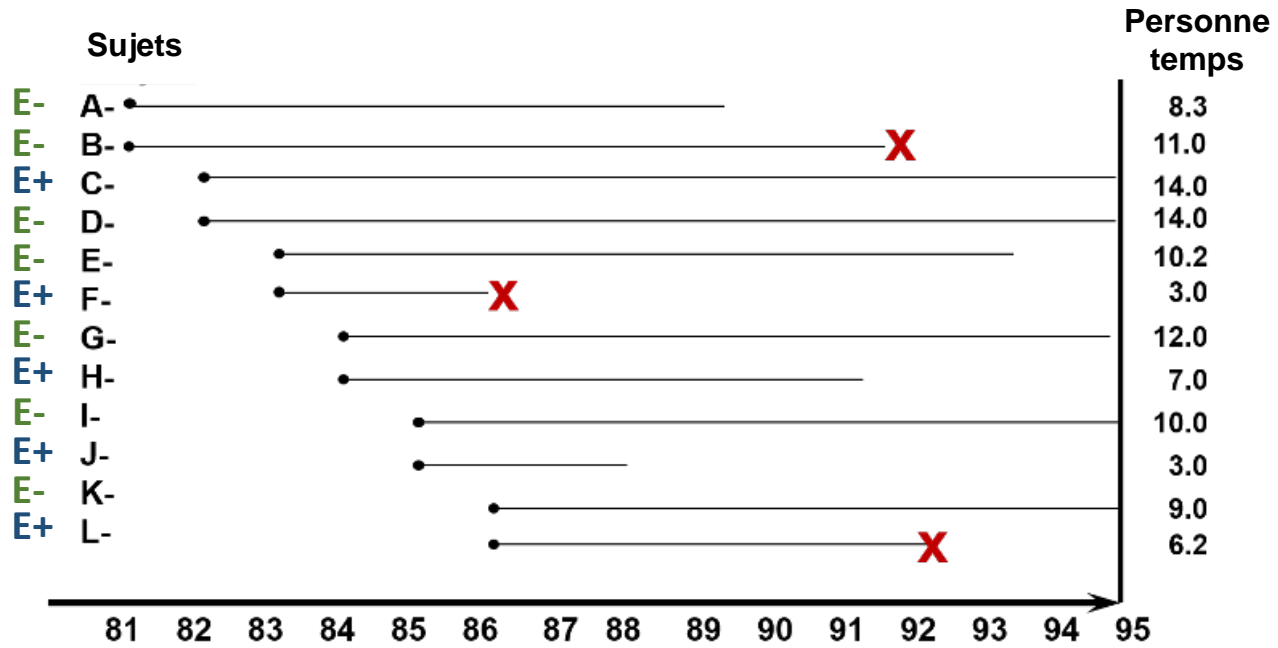
Essai randomisé

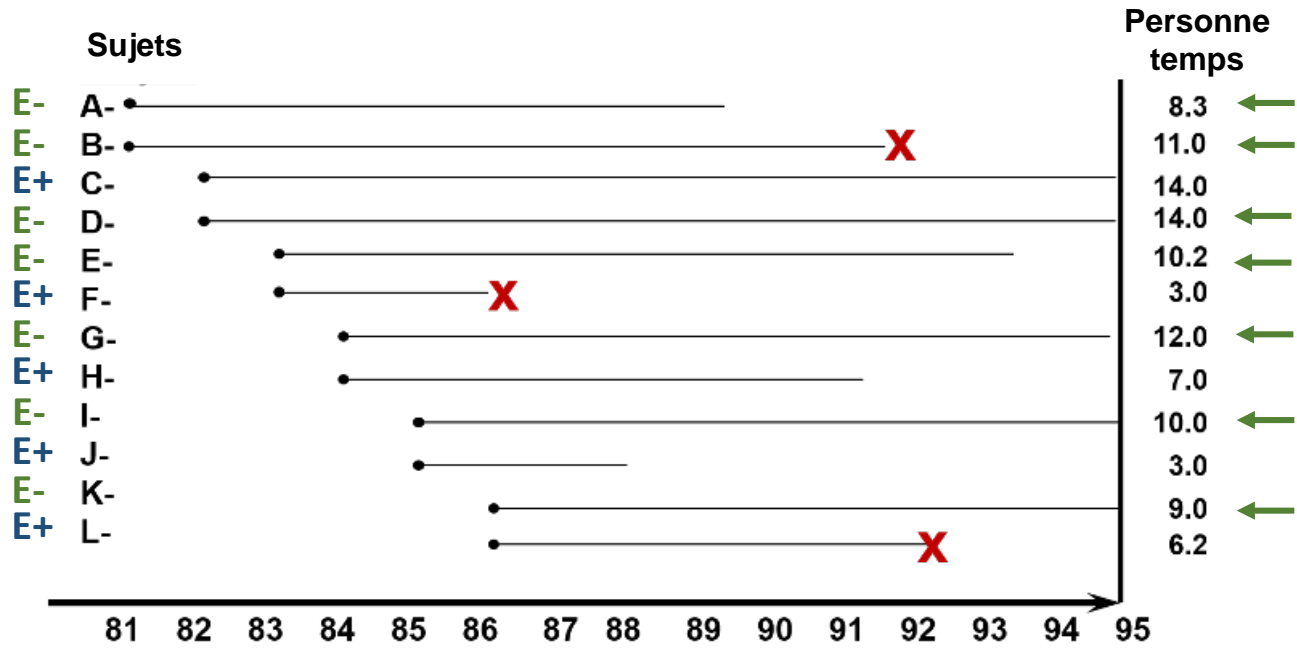




Personne-temps
total = 107.7

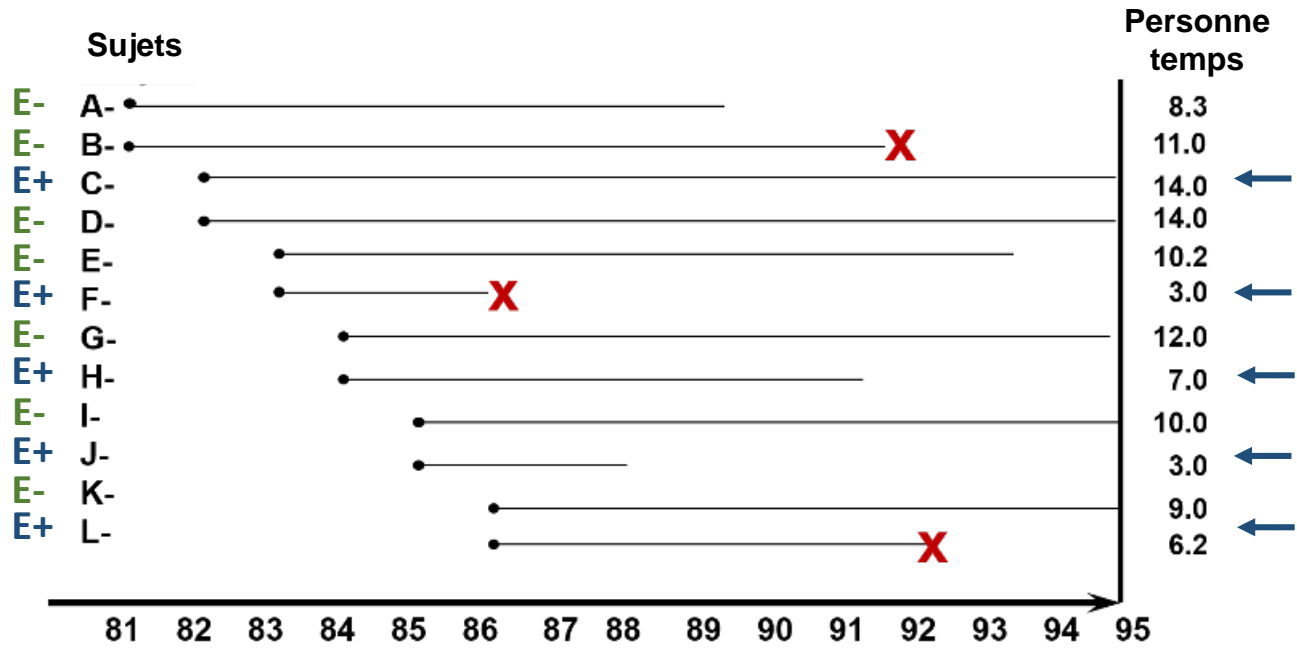
Taux d'incidence = 3 / 107.7 = 0.028 cas par pa
= 2.8 cas pour 100 pa





Personne-temps
E- = 74.5

Taux d'incidence E- = 1 / 75.4

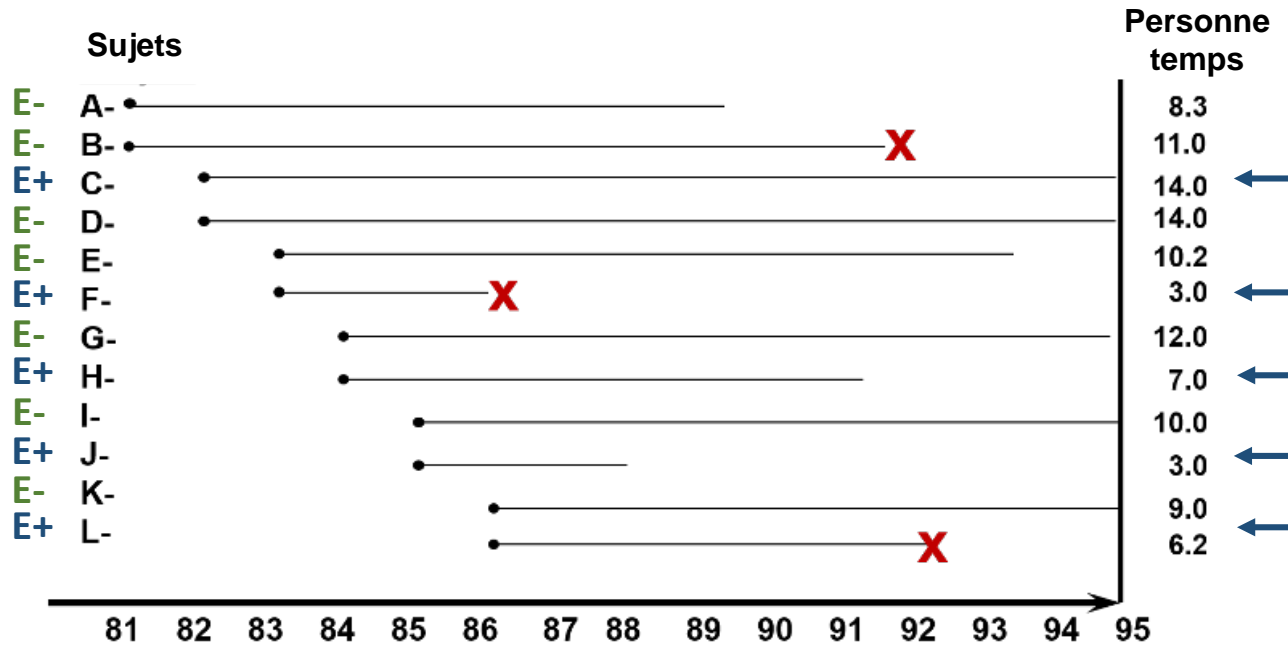


Personne-temps
E- = 74.5

Personne-temps
E+ = 33.2

Taux d'incidence E- = 1 / 75.4

Taux d'incidence E+ = 2 / 33.2



Personne-temps
E- = 74.5

Personne-temps
E+ = 33.2

Taux d'incidence E- = 1 / 75.4

Taux d'incidence E+ = 2 / 33.2

Ratio des taux d'incidence (E+ / E-) = $(2/33.2) / (1/75.4) = 4.5$

Risque relatif

- Risque relatif = rapport de risques = rapports de taux d'incidence

	Malade	Personnes (Personnes-temps)
Exposés	a	N_1
Non-exposés	b	N_0

- $RR = R1 / R0 = (a/N1) / (b/N0)$

Excès de risque

- Dans le cas d'un lien de causalité entre un facteur de risque **E** et une maladie **M**, une autre mesure d'association peut être utilisée. Il s'agit de l'« **Excès de Risque** » (ER) qui mesure **la part de responsabilité de E** quand au nombre de cas, et se calcule par une différence entre R_1 et R_0 :

$$ER = R_1 - R_0$$



ER et RR (exemple)

Dans une étude réalisée auprès d'une population masculine, les risques de décès, chez les gros fumeurs (>25 cig/j) et chez les non-fumeurs, pour 3 maladies, sont présentés dans le tableau suivant :

Cause de décès	Gros-fumeurs R_1	Non-fumeurs R_0	<i>ER</i>	<i>RR</i>
Cancer du poumon	227	7	220	32.4
Bronchite chronique	106	5	101	21.2
M cardio-vasculaires	993	732	261	1.36

Études cas - témoins

- Recrute des cas et des témoins dans une même population (bassin versant d'un hôpital)

	<i>Maladie</i>	
<i>Exposition</i>	<i>M+</i> <i>(cas)</i>	<i>M-</i> <i>(témoins)</i>
<i>E+</i>	a	b
<i>E-</i>	c	d
Total	M_1	M_0

- La mesure de risque dépend du ratio cas/témoin et n'a que peu de sens dans une étude cas-témoins



Études cas - témoins

	Cas	Témoins
E+	a	b
E-	c	d

Étude 1	Cas	Témoins
Exposés	80	50
Non-exposés	120	150
Total	200	200

$$R1 = a / (a+b) = 80 / 130$$

$$R0 = c / (c+d) = 120 / 270$$

$$RR = (80/130)/(120/270) = 1.38$$



Études cas - témoins

	Cas	Témoins
E+	a	b
E-	c	d

Étude 1	Cas	Témoins
Exposés	80	50
Non-exposés	120	150
Total	200	200

$$R1 = a / (a+b) = 80 / 130$$

$$R0 = c / (c+d) = 120 / 270$$

$$RR = (80/130)/(120/270) = 1.38$$

Étude 2	Cas	Témoins
Exposés	80	100
Non-exposés	120	300
Total	200	400

$$R1 = a / (a+b) = 80 / 180$$

$$R0 = c / (c+d) = 120 / 320$$

$$RR = (80/180)/(120/320) = 1.19$$

Études cas - témoins

- Dans une étude cas-témoin, on compare la « chance » (mesurer en *odds*, ou cote) d'être exposé parmi les cas et les témoins
- OR = estimateur non biaisé du risque relatif dans le cas

	Cas	Témoins
E+	a	b
E-	c	d

OR = $\frac{a/c}{b/d}$

Côte de l'exposition
chez les cas

Côte de l'exposition
chez les témoins

→ OR = ad/bc

Études cas - témoins

	Cas	Témoins
E+	a	b
E-	c	d

Étude 1	Cas	Témoins
Exposés	80	50
Non-exposés	120	150
Total	200	200

$$OR = (80 * 150) / (120 * 50) = 2$$

Étude 2	Cas	Témoins
Exposés	80	100
Non-exposés	120	300
Total	200	400

$$OR = (80 * 300) / (120 * 100) = 2$$

Odds - ratio

- Peut être estimé directement du coefficient d'une régression logistique
- Très fréquemment utilisé à partir de données d'enquête pour calculer des OR ajustés
- MAIS: surestime l'association dans le cas d'une fréquence élevée de la variable d'intérêt dans la pop (maladie fréquente)

Odds - ratio

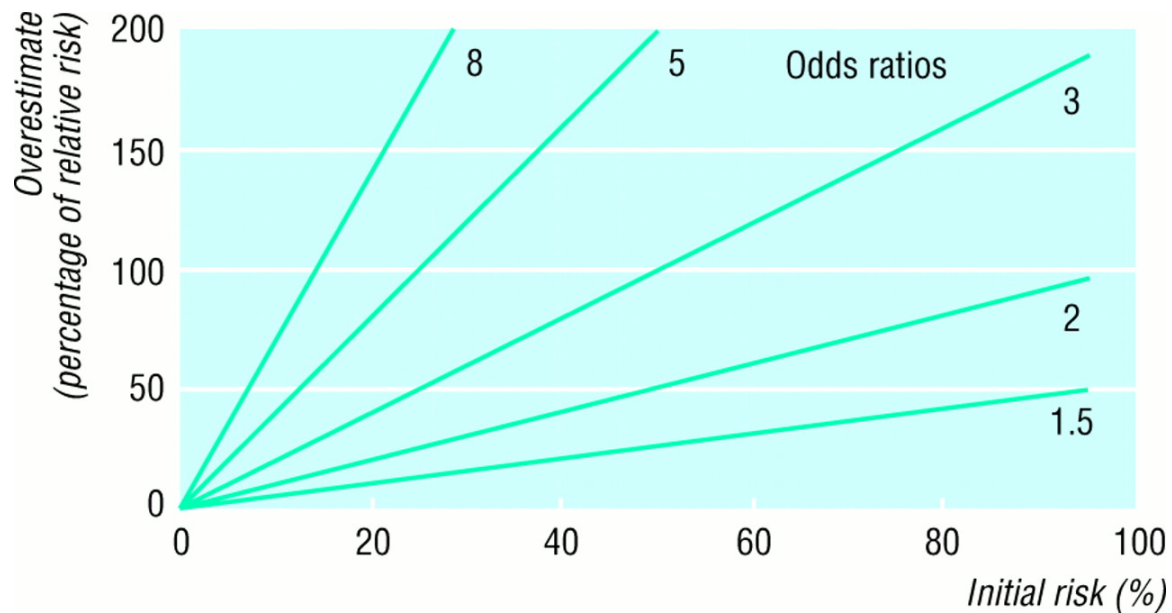


Fig 2

Amount by which odds ratios of >1 overestimate relative risk, for different odds ratios and different levels of initial risk

Davies et al, BMJ 1998

Une solution: régression de Poisson modifiée

[HTML] [Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio](#)

[AJD Barros](#), [VN Hiraakata](#) - BMC medical ..., 2003 - [bmcmedresmethodol.biomedcentral ...](#)

Cross-sectional studies with binary outcomes analyzed by logistic regression are frequent in the epidemiological literature. However, the odds ratio can importantly overestimate the prevalence ratio, the measure of choice in these studies. Also, controlling for confounding is not equivalent for the two measures. In this paper we explore alternatives for modeling data of such studies with techniques that directly estimate the prevalence ratio. We compared Cox regression with constant time at risk, Poisson regression and log-binomial regression ...

☆ 77 Cité 2743 fois [Autres articles](#) [Les 15 versions](#) ✦

Conclusions

Cox or Poisson regression with robust variance and log-binomial regression provide correct estimates and are a better alternative for the analysis of cross-sectional studies with binary outcomes than logistic regression, since the prevalence ratio is more interpretable and easier to communicate to non-specialists than the odds ratio. However, precautions are needed to avoid estimation problems in specific situations.

Illustration sous R

Comment interpréter
une association?

Les critères de causalité (B. Hill)

1. Force de l'association

- Risque relatif élevé?

2. Cohérence des résultats

- Résultats similaires dans d'autres études?

3. Spécificité

- Une cause produit un effet \Rightarrow éliminer la cause supprime l'effet

4. Relation temporelle

- La cause précède-t-elle l'effet?

5. Relation dose-réponse

- Augmentation de l'exposition \Rightarrow augmentation de l'effet ?

Les critères de causalité (B. Hill)

6. Plausibilité biologique

- Explication biologique au phénomène

7. Preuve expérimentale

- Chez l'homme ou l'animal

8. Analogie

- Phénomène similaire pour une autre maladie?

Mesures d'impact

- N'ont de sens que dans le cas de relations causales !
 - Objectifs
 - Du scientifique : Évaluer l'impact de cette exposition sur les personnes exposées (FER)
 - Du décideur : Évaluer l'impact de cette exposition sur la population générale (RA)
- ⇒ Mesurer le bénéfice potentiel de mesures préventives vis-à-vis de cette exposition

Mesures d'impact : calcul

Fraction Etiologique du Risque (FER) **chez les exposés:**

- chez un individu exposé, part du risque sanitaire que l'on peut attribuer à son exposition
- Part qui serait supprimée s'il n'était plus exposé

- $$\text{FER} = \frac{\text{RR} - 1}{\text{RR}} \quad (\text{ou} \quad \frac{\text{OR} - 1}{\text{OR}})$$

- Exemple : tabac et cancer du poumon

- RR = 11 de cancer du poumon associé au tabagisme
- FER = $10 / 11 = 91\%$

⇒ 91% du risque de cancer du poumon d'un fumeur est dû au fait qu'il fume

Mesures d'impact : calcul

Fraction de Risque Attribuable (RA) dans la population:

- Dans une population donnée, part du risque sanitaire que l'on peut attribuer à une exposition
- Part qui serait supprimée si l'exposition disparaissait
- Dépend de la fréquence f de l'exposition :

$$RA = \frac{f \times (RR - 1)}{f \times (RR - 1) + 1} \quad (\text{ou} \quad \frac{f \times (OR - 1)}{f \times (OR - 1) + 1})$$

- Exemple : tabac et cancer du poumon
 - $RR = 11$, en France en 2008 : $f = 34\%$ de fumeurs
 - $RA = 0.34 \times 10 / (0.34 \times 10 + 1) = 77\%$
- ⇒ 77% du risque de cancer du poumon en France serait supprimé s'il n'y avait plus de tabagisme*

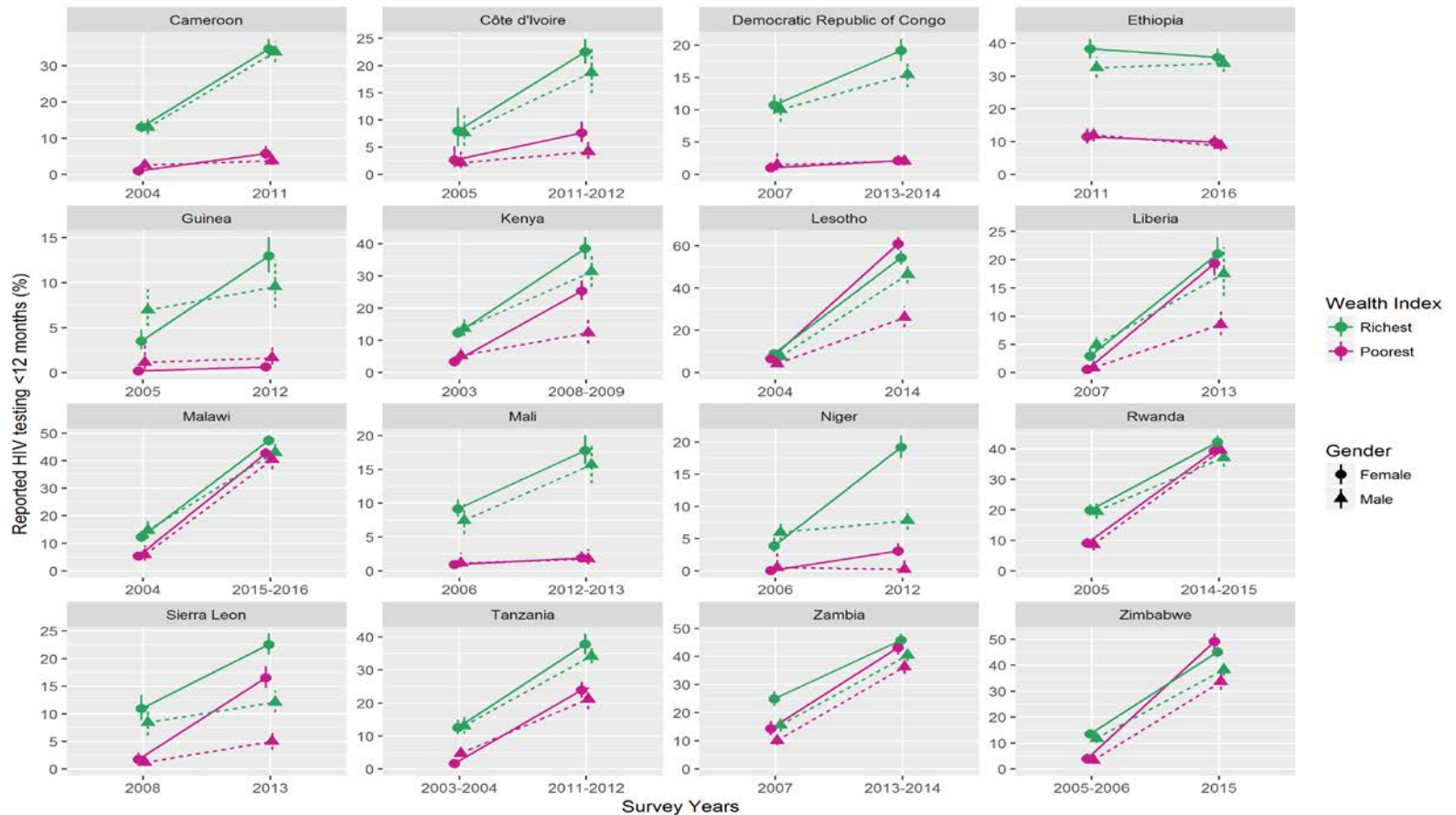
Quelques études en épidémiologie sociale

ISS en lien avec le VIH

- Depuis ~2008, de large programmes de dépistage et traitement du VIH ont été déployés en Afrique subsaharienne
- Des exemples au Nord ont montré que des interventions de santé peuvent accroître ou générer des ISS (hypothèse de l'équité inverse de C. Victora)
- Quelles évolutions des ISS en lien avec le VIH?

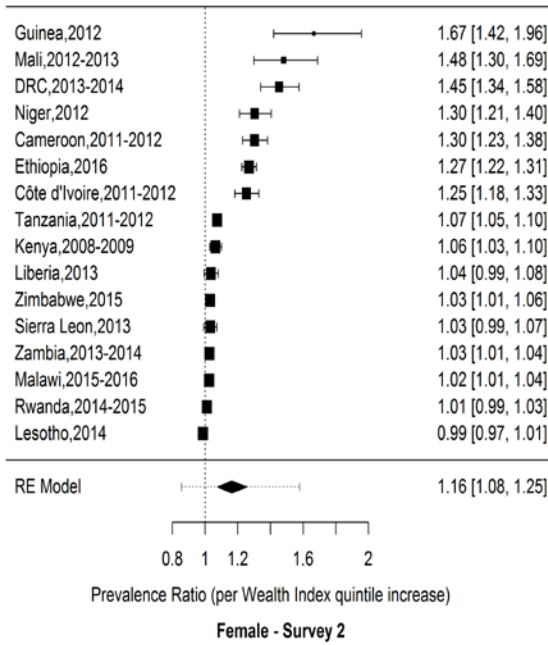
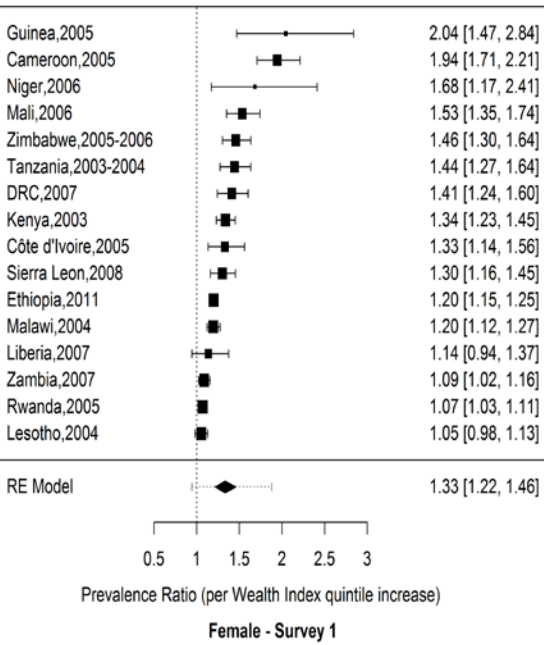
Recours récent au dépistage VIH

TESLA results: % of reported HIV testing in < 12 months among different educational levels by gender



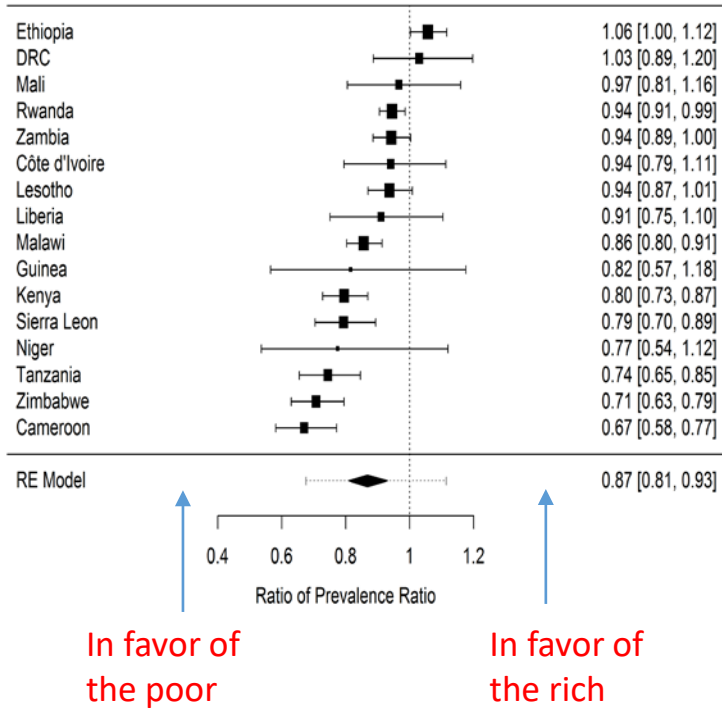
TESLA results: Meta-analysis (females)

PR of recent HIV testing (< 12months) per 1 quintile increase in wealth in earlier & later surveys among 16 countries



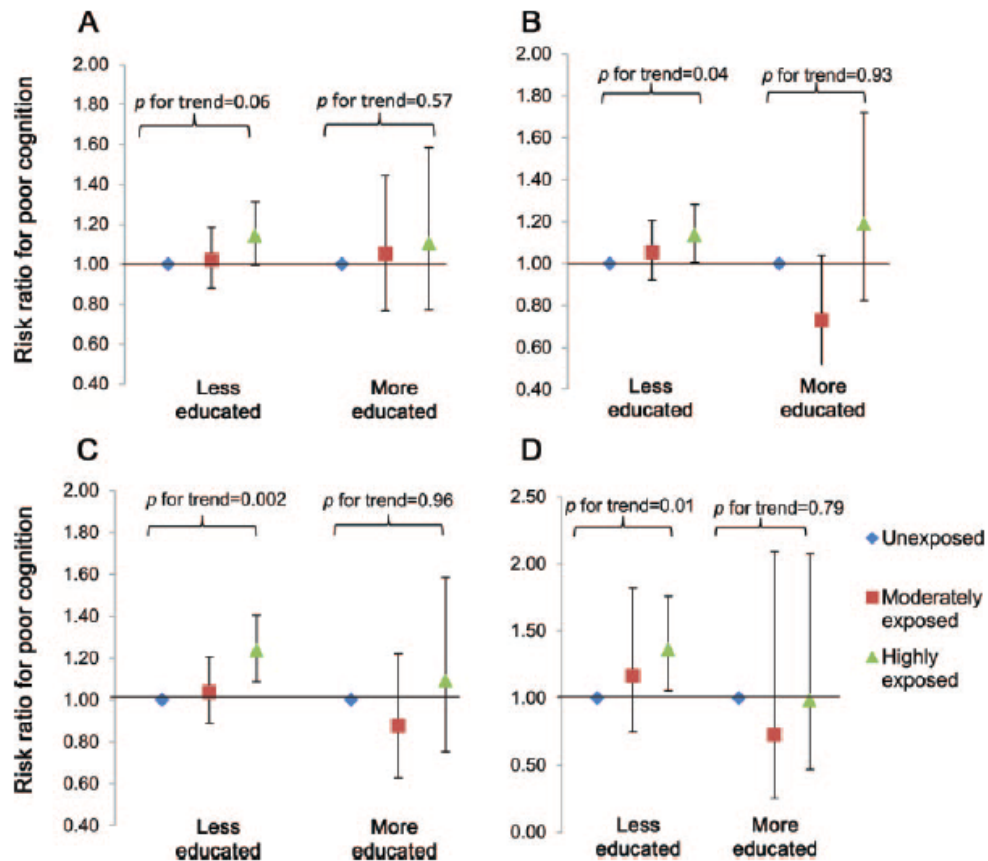
$$RPR = PR_{post-2008} / PR_{pre-2008}$$

RPR of recent HIV testing (< 12months) per 1 quintile increase in wealth among 16 countries

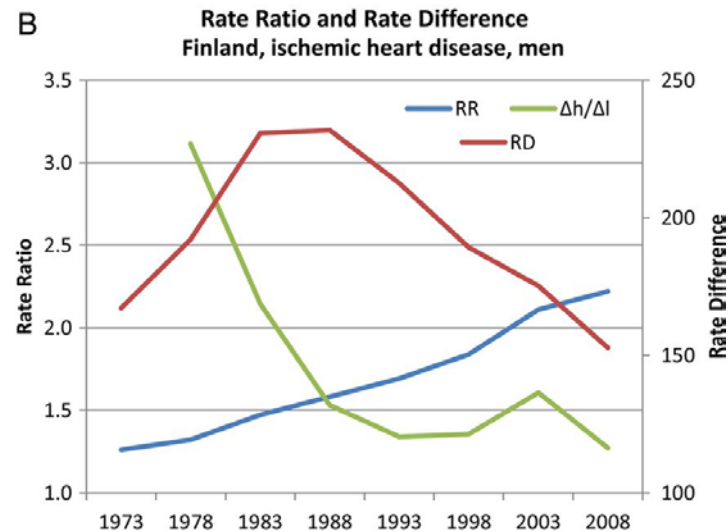
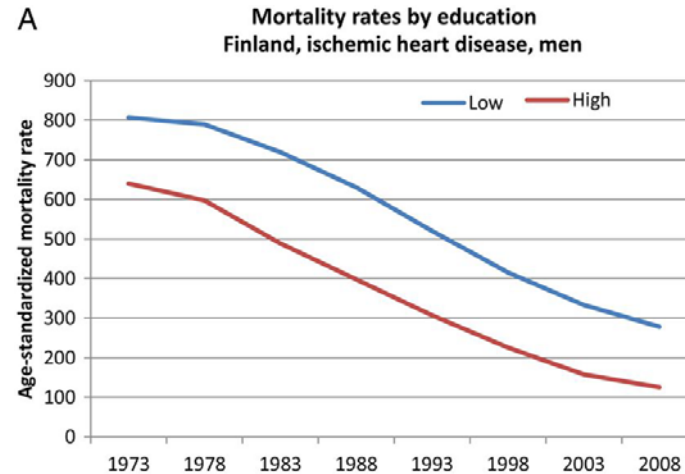


Mettre en évidence des vulnérabilités

Figure Risk ratio for poor cognition by educational attainment and solvent exposure status



Arithmétique des mesures d'inégalités



Enjeux socio-comportementaux du traitement VIH

- Effets sociaux de l'initiation des ARV sous critères CD4:

Dévoilement du statut



- Mise en place de solidarités
- Soutien psycho-social

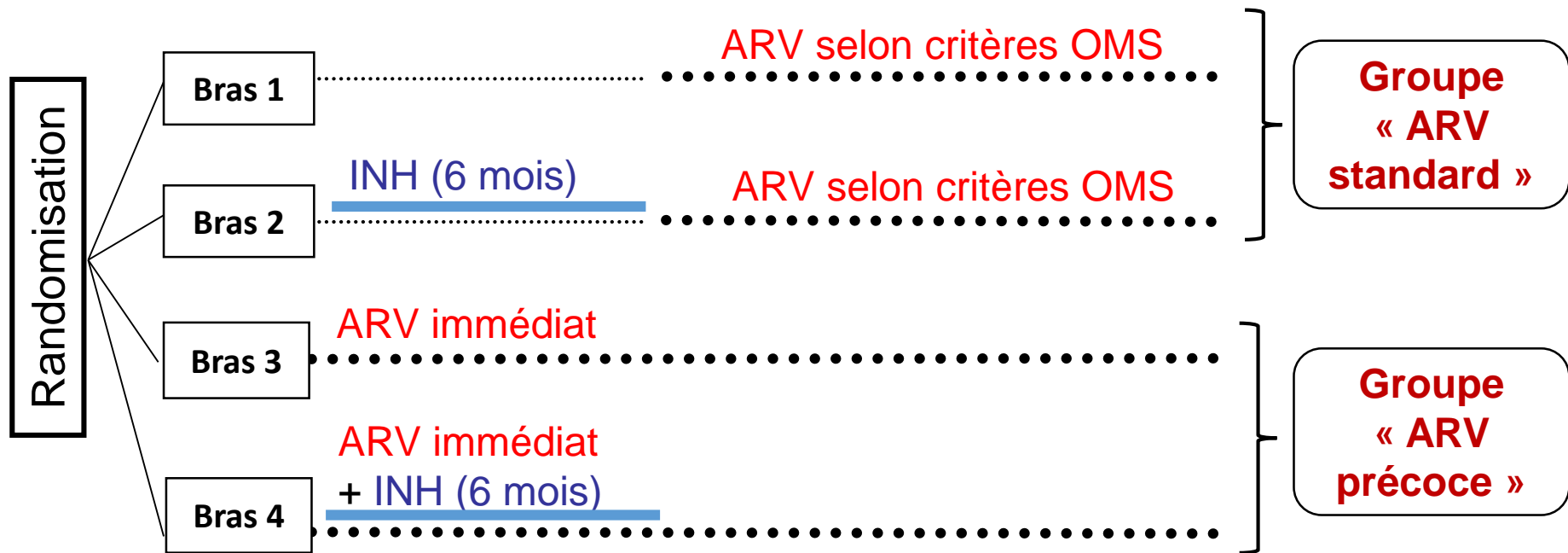
- Rupture d'union (femmes)
- Discrimination

Documenté chez patients initiant ART à un stade **symptomatique**

- Effet comportemental: diminution des comportements à risque
- **Enjeux:** - acceptabilité et adhésion au traitement
- effet préventif

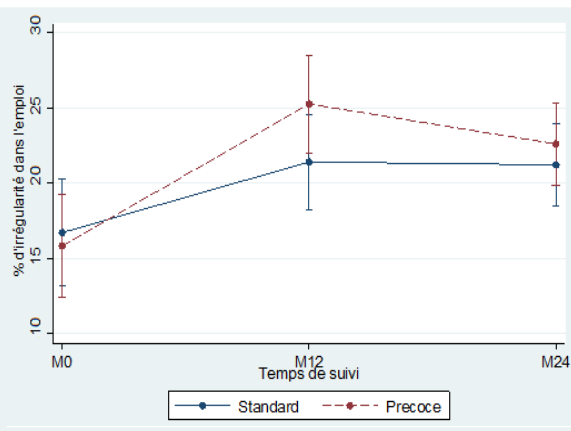
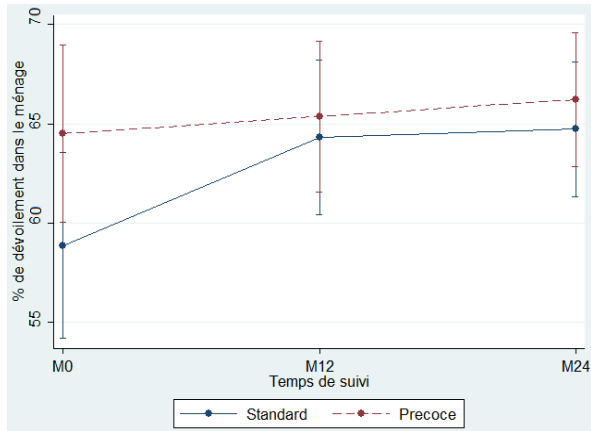
Objectif: Mesurer les impacts sociaux et comportementaux du traitement ARV précoce

Temprano Social



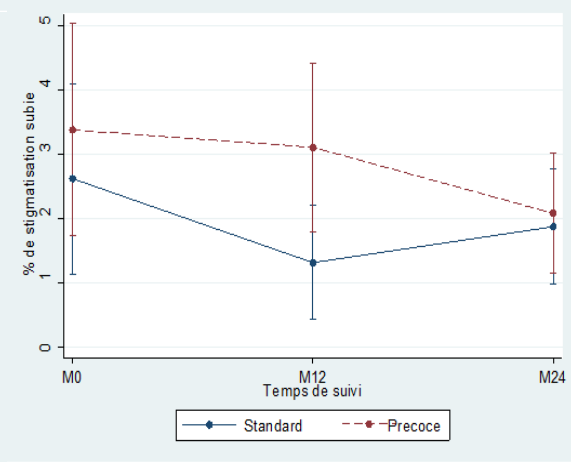
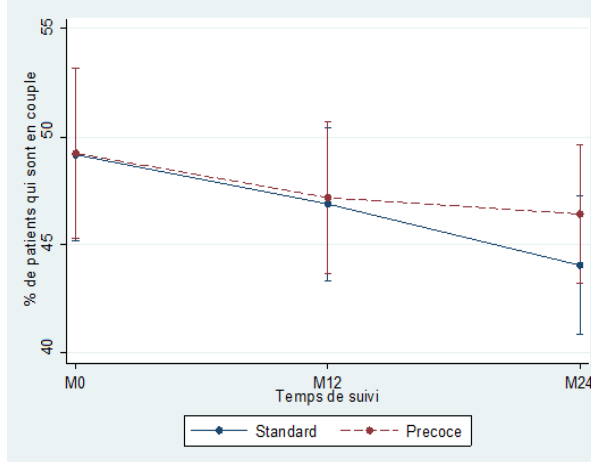
Evolution des indicateurs sociaux

Dévoilement du statut au sein du ménage



Irrégularité dans l'activité professionnelle

En couple



Discriminations subies en raison du statut VIH

Quand l'épidémiologie s'inspire d'autres champs

- Régression par discontinuité
- Analyse de trajectoire

Estimation par discontinuité

- Application:
 - L'exposition dépend (au moins en partie) d'une variable continue mesurée (**variable d'affectation**), et d'une valeur seuil pour cette variable
 - Variable < seuil fixé → pas d'exposition
 - Variable > seuil fixé → exposition
- Hypothèse:
 - Si les individus ne peuvent pas manipuler la valeur de cette variable alors, **autour de la valeur seuil, l'exposition est allouée de manière aléatoire** (aussi bien que si randomisé)

Justification

- L'exposition ou le traitement est basé sur une règle de seuil par rapport à une variable d'affectation continue
 - Commun en médecine: cholestérol, glucose, poids à la naissance, BMI
- Variabilité aléatoire (« bruit ») dans la mesure de cette variable continue
 - Erreur de mesure, variabilité naturelle...
- Autour de la valeur seuil, du fait de ce bruit, une personne se situera au dessus/en dessous du seuil de manière aléatoire
 - La seule \neq entre patients juste en dessous vs. juste au dessus du seuil \rightarrow exposition

Contexte randomisé!

Régression sur discontinuité

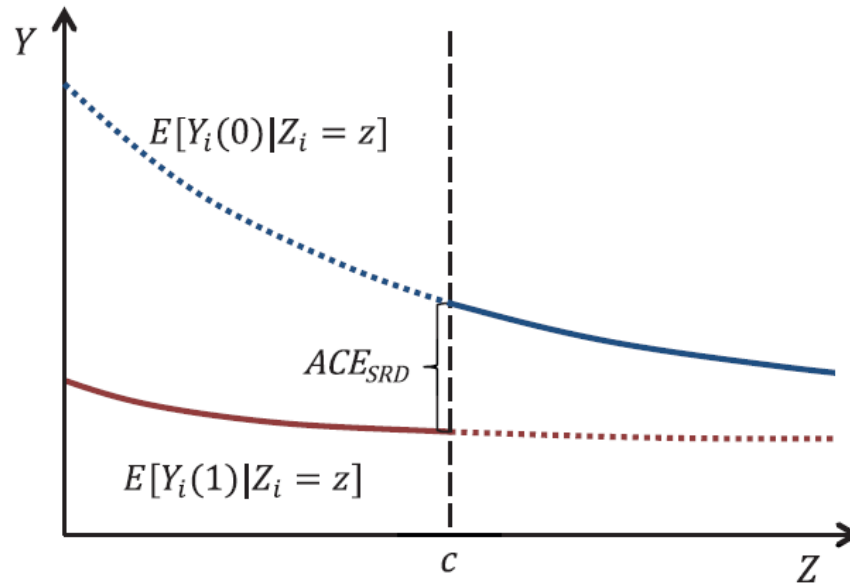


FIGURE 1. Sharp regression discontinuity design. This figure shows the conditional expectation functions for each of the potential outcomes $E[Y_i(1) | Z_i = z]$ and $E[Y_i(0) | Z_i = z]$. The solid lines show the conditional expectation function of the observed data, $E[Y_i | Z_i = z]$.

ACE = Average Causal Effect

SRD = Sharp Regression discontinuity (règle seuil déterministe)

≠

FRD = Fuzzy Regression Discontinuity (règle seuil probabiliste)

$$ACE_{SRD} = \lim_{\{z \uparrow c\}} E[Y_i | Z_i = z] - \lim_{\{z \downarrow c\}} E[Y_i | Z_i = z]$$

Exemple VIH – CD4

Regression Discontinuity Designs in Epidemiology *Causal Inference Without Randomized Trials*

Jacob Bor,^{a,b,c} Ellen Moscoe,^c Portia Mutevedzi,^b Marie-Louise Newell,^{b,d} and Till Bärnighausen^{b,c}

Abstract: When patients receive an intervention based on whether they score below or above some threshold value on a continuously measured random variable, the intervention will be randomly assigned for patients close to the threshold. The regression discontinuity design exploits this fact to estimate causal treatment effects. In spite of its recent proliferation in economics, the regression discontinuity design has not been widely adopted in epidemiology. We describe regression discontinuity, its implementation, and the assumptions required for causal inference. We show that regression discontinuity is generalizable to the survival and nonlinear models that are mainstays of epidemiologic analysis. We then present an application of regression discontinuity to the much-debated epidemiologic question of when to start HIV patients on antiretroviral therapy. Using data from a large South African cohort (2007–2011), we estimate the causal effect of early versus deferred treatment eligibility on mortality. Patients whose first CD4 count was just below the 200 cells/ μ L CD4 count threshold had a 35% lower hazard of death (hazard ratio = 0.65 [95% confidence interval = 0.45–0.94]) than patients presenting with CD4 counts just above the threshold. We close by discussing the strengths and limitations of regression discontinuity designs for epidemiology.

(Epidemiology 2014;25: 00–00)

Causal inference in nonexperimental studies typically requires a strong, untestable assumption: that no unobserved factors confound the relationship between the exposure and the outcome.¹ Violations of this assumption will lead to biased estimation of causal effects. The regression discontinuity design is one important quasi-experimental study design in which this assumption is not required for causal inference. Regression discontinuity designs can be implemented when the exposure of interest is assigned—at least in part—by the value of a continuously measured random variable and whether that variable lies above (or below) some threshold value. Provided that subjects cannot precisely manipulate the value of this variable, assignment of the exposure is as good as random for observations close to the threshold, and valid causal effects can be identified.²

The regression discontinuity design first appeared in the educational psychology literature in 1960,^{3–5} was further developed in the 1970s and 1980s,^{6,7} and has become well established in economics over the last 2 decades.^{2,8,9} In recent years, a number of clinical and population health studies have been published in economics journals using regression discontinuity designs.^{10–17} These studies have used regression

Exemple VIH – CD4

- Quand débiter un traitement antirétroviral (ARV)?
 - A partir de 2006: traiter si CD4 <200/mL
 - 2010: RCT en Haïti: 75% de réduction de mortalité si ARV débuté entre 350 et 200 CD4/mL versus <200 CD4/mL (Severe et al, NEJM 2010)
- En Afrique du Sud, jusqu'en 2011:
 - Initiation du traitement ARV au seuil CD4<200 /mL
 - Quel bénéfice sur la mortalité du traitement ARV en situation réelle?
- Données: zone de surveillance sanitaire et démographique du district de Hlabisa, Afrique du Sud
 - Données des cliniques VIH
 - Croisement avec des données démographiques pour tous les habitants du district
 - Echantillons: 4391 patients en centre de santé de 2007 à 2011 (13,139 personnes-années)

Conditions de validité de la régression sur discontinuité

1) La variable d'affectation doit être continue autour du seuil de décision (pas de manipulation de la décision de traitement)

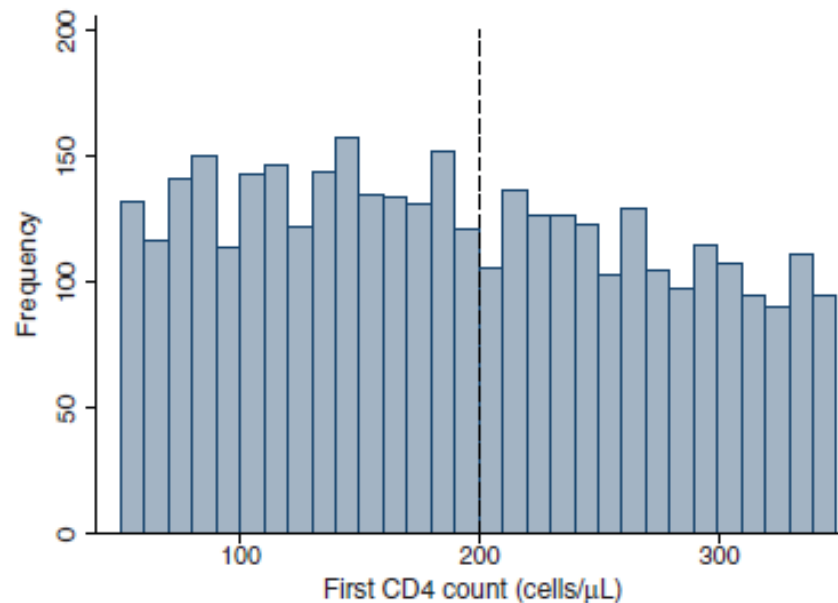


FIGURE 2. Distribution of first CD4 counts in the HIV treatment and care program.

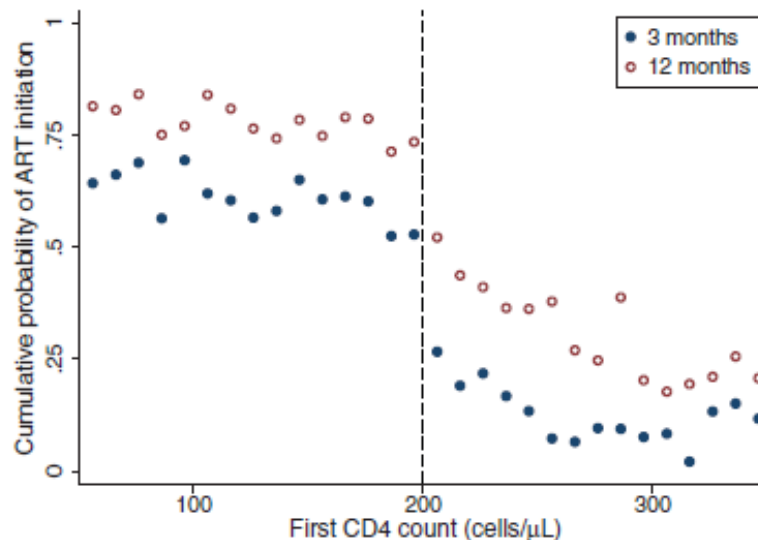
Conditions de validité de la régression sur discontinuité

2) Vérifier la condition de « randomisation locale » = les individus de part et d'autre du seuil sont comparables pour la plupart des covariables (âge, sexe...)

« Further analysis revealed balance in variables observed at baseline (age and sex) at the cutoff (not shown). »

Conditions de validité de la régression sur discontinuité

3) Spécifier la règle de décision pour le traitement: déterministe ou probabiliste?



→ *probabiliste*
(« *fuzzy*
« *discontinuity*)

FIGURE 3. First CD4 count and ART initiation. Kaplan–Meier estimates of the probability that a patient initiated ART within 3 and 12 months of first CD4 count in the HIV treatment and care program.

Représentation et modélisation de la discontinuité

HR = 0.59 [0.42-0.83]

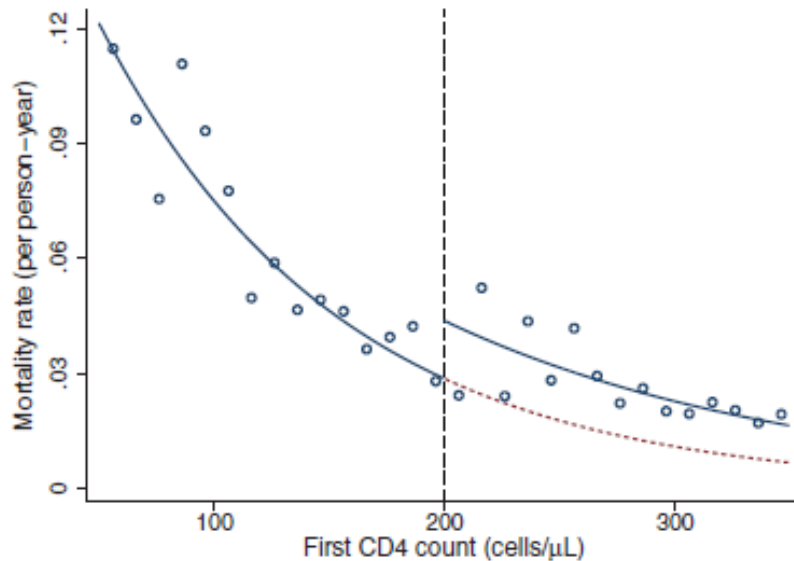


FIGURE 4. First CD4 count and mortality hazard rate. Predicted hazards from the Table, model 2a are displayed as solid lines. Dashed line shows extrapolated prediction if all patients were treatment eligible at first CD4 count. Dots are hazards predicted for CD4 count bins of width 10 cells.

Modèle local:

- Linéaire
- Polynomial
- Exponentiel
- Cox

Explorer les résultats sur \neq intervalles (\neq distances au seuil)

Effet de l'éducation sur la santé

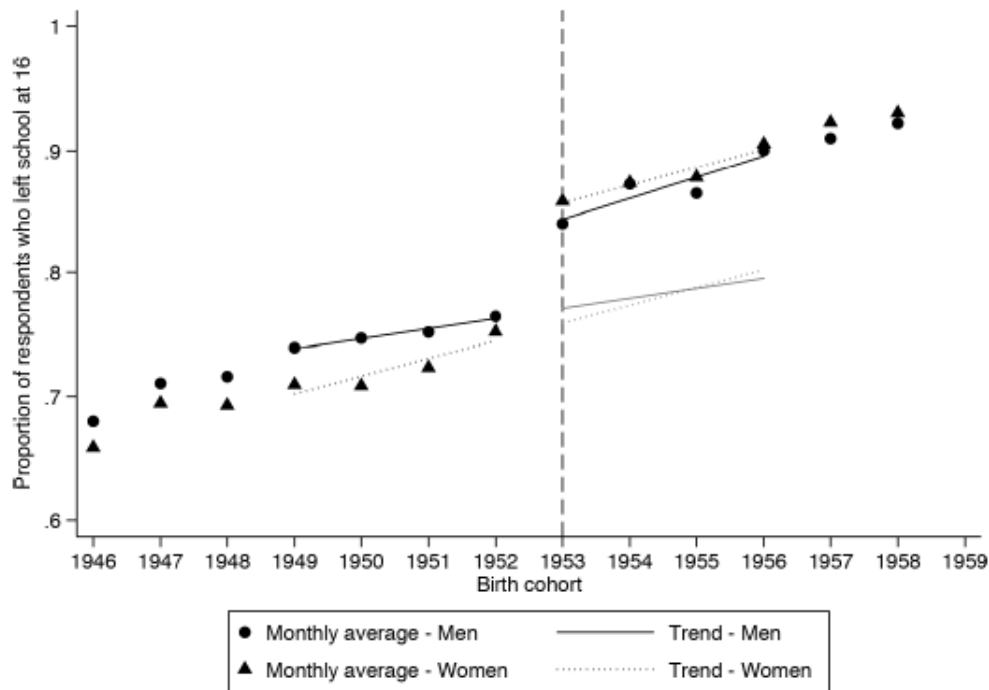
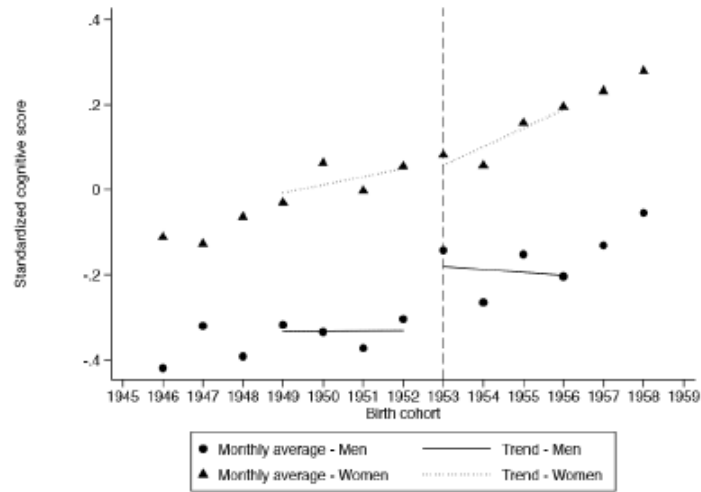


Figure 1 Effect of eligibility to the 1959 Berthoin reform on the proportion of respondents who left school after 16 by gender in the 2012 French Labor Force Survey (n=72 133), Birth cohorts 1946-1959. The dots and triangles show the average school leaving age, respectively, for men and women and for each birth cohort. The dashed line represents the cut-off for eligibility to the reform (1 January 1953). The fitted lines represent the linear trends for our analytical sample: respondents born up to 48 months before or after the reform, separately for men (black lines) and women (dotted lines). The grey fitted lines show the predicted school leaving age in the absence of the reform.

A Standardized cognitive score

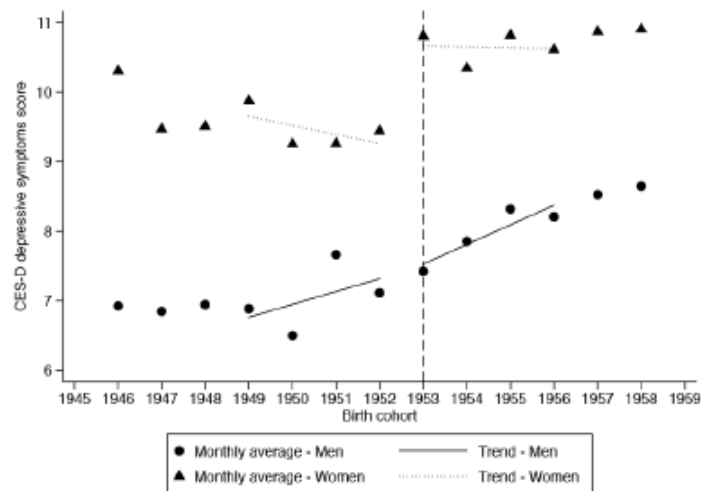


C Physical functioning measures

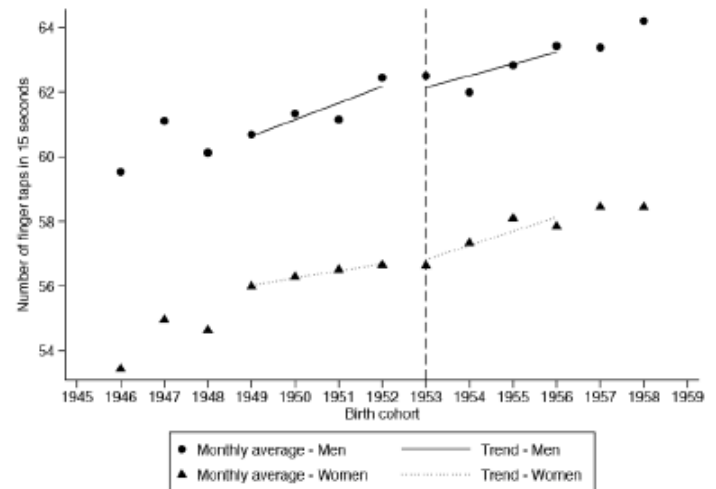
Hand grip strength



B CES-D depressive symptoms score



Finger tapping



How to use sequence analysis for life course epidemiology? An example on HIV-positive Sub-Saharan migrants in France

Anne Gosselin,^{1,2} Annabel Desgrées du Loû,^{1,2} Eva Lelièvre,³ The PARCOURS Study Group

¹IRD, Institut de Recherche pour le Développement, Marseille, France

²SAGESUD Team, CEPED (Université Paris Descartes, IRD, Inserm), Paris, France

³Mobilité, Logement et Entourage Unit, National Institute for Demographic Studies (INED), Paris, France

Correspondence to

Dr Anne Gosselin, CEPED – Paris Descartes University, Paris, France; anne.gosselin@ceped.org

Received 18 July 2017

Revised 10 January 2018

Accepted 11 January 2018

ABSTRACT

Background Life course epidemiology is now an established field in social epidemiology; in sociodemography, the quantitative analysis of biographies recently experienced significant trend from event history analysis to sequence analysis. The purpose of this article is to introduce and adapt this methodology to a social epidemiology question, taking the example of the impact of HIV diagnosis on Sub-Saharan migrants' residential trajectories in the Paris region.

Methods The sample consists of 640 migrants born in Sub-Saharan Africa receiving HIV care. They were interviewed in healthcare facilities in the Paris region within the PARCOURS project, conducted from 2012 to 2013, using life event history calendars, which recorded year by year their health, family and residential histories. We introduce a two-step methodological approach consisting of (1) sequence analysis by optimal matching to build a typology of migrants' residential pathways before and after diagnosis, and (2) a Cox model of the probability to experience changes in the residential situation

demographers had developed life history calendars to collect life histories.^{6,7}

Sequence analysis—among which optimal matching—was first designed by biologists to analyse DNA strings; it was imported in social sciences in the 1980s to take entire individuals' trajectories into account in the analyses of social phenomena.^{8,9} Optimal matching consists of measuring similarities between pairs of sequences by computing the 'cost' of transforming a sequence into another.⁹ These measures of dissimilarity form a distance matrix used as the input for clustering that produces typologies of life sequences. Since then, optimal matching has been successfully applied to individual life courses and family histories in demography.^{8,10,11} Although these developments are most valuable for life course epidemiology, to our knowledge there are yet no examples of applying these methods to epidemiology research questions. The purpose of this article is to use this new method and combine it with life event history analysis to address a social epidemiology question.

Gosselin et al, JECH 2018

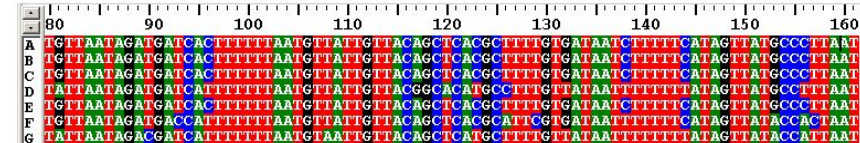
Analyse de séquence

- Séquence = suite ordonnée d'événements/d'items tirées parmi un nombre fini d'événements/d'items possibles

		Characters																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20				
Types																									
1	G	A	T	G	T	T	G	G	C	A	G	T	C	C	G	A	T	G	T	A	A	G	C	-	-
2	B	A	T	G	T	T	G	G	C	A	G	T	C	C	G	A	T	G	T	A	A	G	C	-	-
3	C	A	T	G	T	T	G	G	C	A	G	T	C	C	G	A	T	G	T	A	A	C	C	-	-
4	D	A	C	G	G	T	A	G	C	A	G	T	C	T	G	A	T	G	T	A	T	C	C	-	-
5	A	A	C	G	G	T	A	G	C	A	G	T	C	T	G	A	T	G	T	A	T	C	C	-	-
6	F	C	T	G	C	T	G	G	T	A	G	T	C	G	T	T	T	G	T	A	A	C	C	-	-
7	I	A	T	G	G	T	G	-	C	A	G	T	C	G	G	G	T	G	T	C	A	C	C	-	-
8	H	A	T	G	C	T	G	G	C	A	G	T	C	G	G	G	T	G	T	A	A	C	C	-	-
9	J	A	C	G	G	T	A	G	C	A	G	A	C	T	G	A	T	G	T	A	T	C	C	-	-
10	E	C	T	G	C	T	G	G	T	A	G	T	C	G	T	T	T	G	T	A	A	C	C	-	-
11	K	C	T	G	C	T	G	G	C	A	G	T	C	G	G	T	T	G	T	A	A	C	C	-	-

Analyse de séquence

- Optimal matching = mesurer des similarités entre séquences en calculant un « coût » de passage d'une séquence à l'autre
- « Coûts » totaux calculés comme somme des coûts de modifications individuelles (substitution / insertion / délétion)
- Exprimés sous forme d'une matrice de distance
- Clustering / typologies / phylogénies à partir des distances

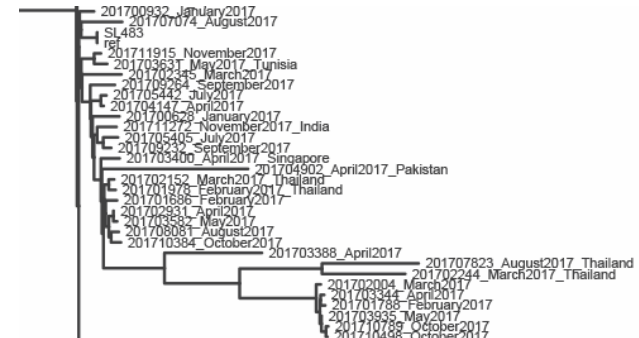


Distance:

Example: observed percent sequence difference

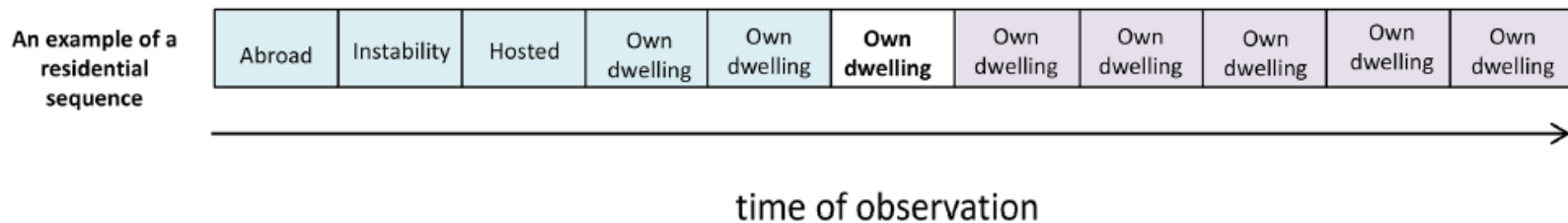
Distance matrix:

	A	B	C	D	E	F	G
A	-						
B	63	-					
C	94	79	-				
D	111	96	47	-			
E	67	20	83	100	-		
F	23	58	89	106	62	-	
G	107	92	43	16	96	102	-



Adaptation aux données biographiques

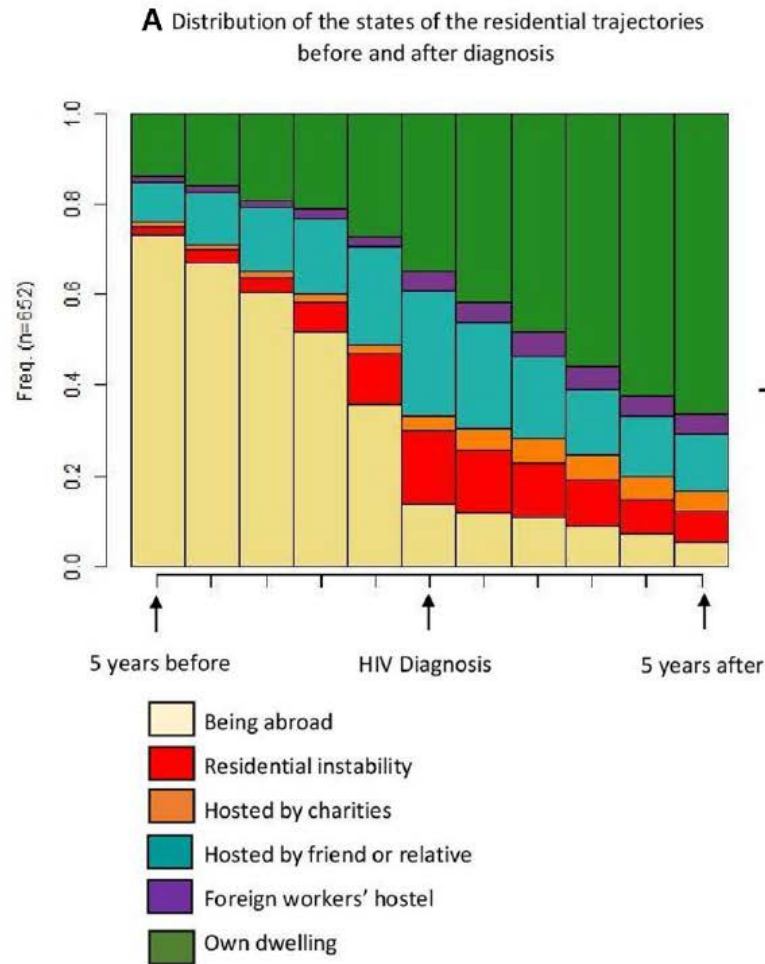
- Ex: séquence résidentielle chez des migrants d'Afrique subsaharienne



6 possible states: being abroad; residential instability (frequently changes of accommodation during the year), hosted by a friend or relative, charities, foreign workers' hostel, own dwelling.

- 6 états possibles
- N= 640 participants / séquences

Résultats



B Typology of pathways before and after diagnosis, sequence plots

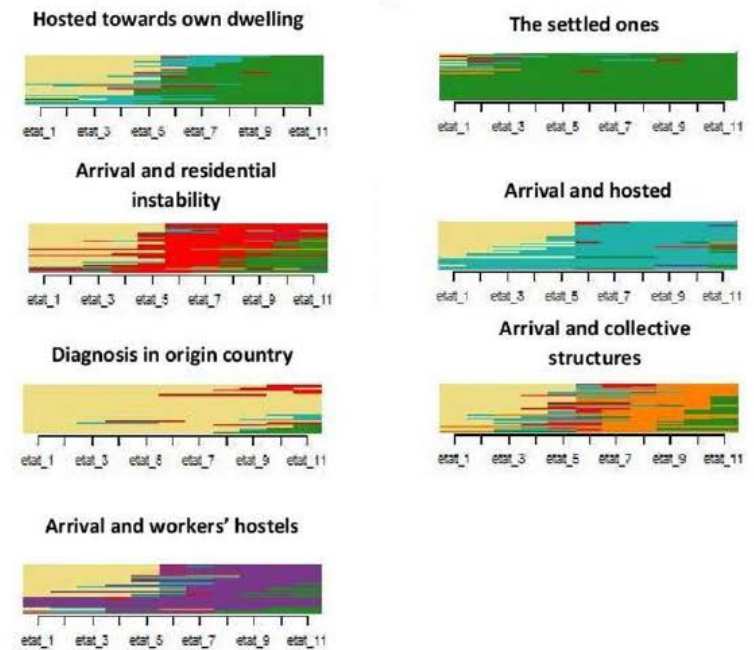


Figure 2 Residential trajectories before and after HIV diagnosis among Sub-Saharan migrants.

Résultats

- Hypothèse: les personnes hébergées sont rejetés au moment du diagnostique
 - Modèle de Cox sur la probabilité de transition « être hébergé » -> « foyer propre » chez les participants hébergés l'année précédent le diagnostique (n=141)
 - Var. explicatives: âge, sexe, niveau d'instruction, en emploi, mise en couple, naissance
 - (Non testé: type de trajectoire!)
- Résultats: l'acquisition d'un foyer propre n'est pas associée à l'attitude du réseau social, mais avec la mise en couple (HR = 2.99, $p < 10^{-3}$)