Statistique et Santé

Cécile Chouquet

Institut de Mathématiques de Toulouse

IRES - 09 avril 2018

Introduction

- Utilisation de la statistique pour répondre à des questions concrètes dans tous les domaines de la recherche biomédicale
- Recherche biomédicale
  - Activité médicale visant à améliorer la connaissance biologique ou médicale d'une maladie ou d'une thérapeutique
  - Etude des maladies, de leur fréquence, de leur répartition, des caractéristiques des personnes atteintes, des facteurs de risque associés (protecteurs ou aggravants) et des décès qui y sont liés $\Rightarrow$ Epidémiologie
- Etude de cohortes mises en place depuis l'après-guerre
  Aujourd'hui des méga-cohortes incluant les individus dès la naissance

- Spécificités des données biomédicales :
  - longitudinales : on observe l'évolution d'un processus au cours du temps
  - manquantes : on n'observe pas l'ensemble du processus et des informations

$\Rightarrow$ Méthodes de modélisation "sur mesure"

Quelques exemples de problématique en recherche biomédicale

- Dermatologie : Estimation d'un nombre de malades
  $\rightarrow$ Modèle de capture-recapture
- Fertilité humaine : quel délai pour avoir un bébé ?
  $\rightarrow$ Analyse de données censurées
- Transmission du VIH de la mère à l'enfant : quand ?
  $\rightarrow$ Rétro-calcul et modèle de Markov
- Exposition médicamenteuse au cours de la grossesse : quel type et quel risque associé ?
  $\rightarrow$ Classification de trajectoires et modélisation

Modèle de Capture-recapture en dermatologie

Dermatologie :
Estimation d'un nombre de malades

Modèle de capture-recapture

Modèle de Capture-recapture en dermatologie

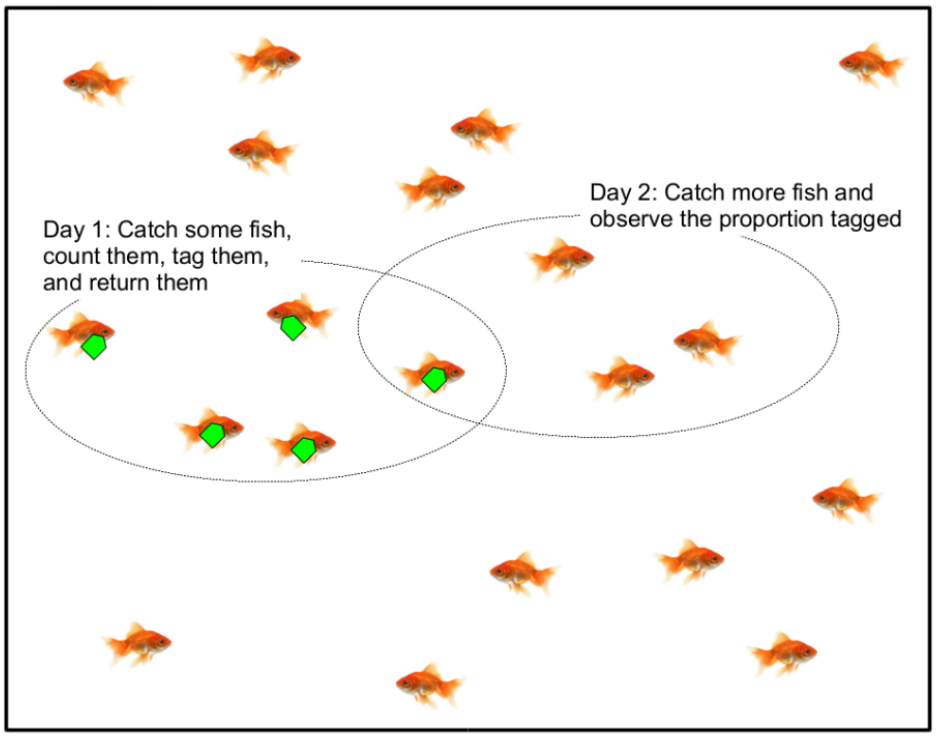
Problématique biomédicale

- Ichtyose (du grec Ichtyos "poisson) : maladie génétique rare de la peau, chronique, se caractérisant par une sécheresse et un squame de la peau
- Sévérité et suivi variable $\Rightarrow$ beaucoup de malades ne sont pas dans le système de soins.

$\Rightarrow$ Comment estimer le nombre de malades d'ichtyose en France ?
$\Rightarrow$ Mise en œuvre d'une étude épidémiologique basée sur la méthode de capture-recapture

Modèle de Capture-recapture

N = Nb total de poissons (inconnu)
S_1 = Nb de poissons capturés et marqués à la 1ère pêche
M = Nb de poissons pêchés à la 2ème pêche
S_2 = Nb de poissons marqués parmi les poissons pêchés



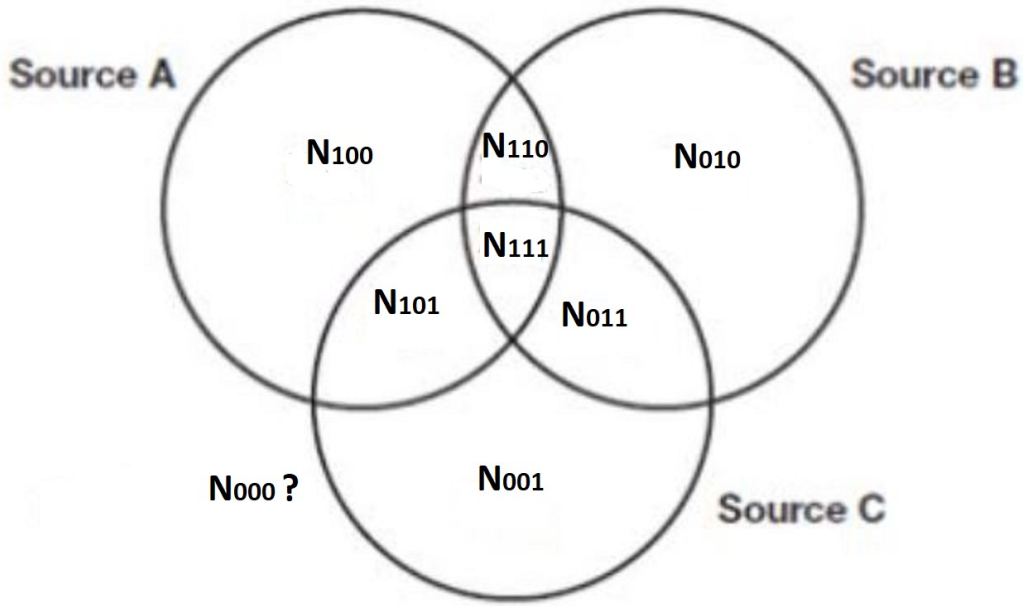
$$\frac{S_1}{N} = \frac{M}{S_2}$$

$$\Rightarrow \widehat{N} = \frac{S_1 \times S_2}{M}$$

Modèle de Capture-recapture

Trois listes incomplètes de patients :

- Hôpital
- Association
- Réseau social Facebook



- Propriétés des 3 sources :
  indépendance, homogénéité, fermeture, appariemment parfait

Modèle de Capture-recapture

- N_i représente le nombre de malades par groupe i :

$$N_i \sim Poisson(\lambda_i) \text{ avec } E(N_i) = \lambda_i$$

- On modélise N_i en fonction des sources auxquelles appartient le groupe i en supposant un modèle log-linéaire :

$$Ln(\lambda_i) = \beta_0 + \beta_1 \underbrace{\mathbb{1}_{S_{1i}=1}}_{\text{indicatrice associée à la source 1}} + \beta_2 \mathbb{1}_{S_{2i}=1} + \beta_3 \mathbb{1}_{S_{3i}=1} + e_i$$

Modèle de Capture-recapture

- On estime les paramètres β_j en fonction des effectifs observés

$$\widehat{\lambda_i} = e^{\widehat{\beta_0}+\widehat{\beta_1}\mathbb{1}_{s_{1i}=1}+\widehat{\beta_2}\mathbb{1}_{s_{2i}=1}+\widehat{\beta_3}\mathbb{1}_{s_{3i}=1}}$$

- Estimation du nombre de patients absents des 3 listes :

$$\widehat{N_{000}} = exp[\widehat{\beta_0}]$$

$$\Downarrow$$

$$\widehat{N} = \text{ Nb total de malades d'ichtyoses}$$

- Quelques extensions possibles :
  - Prendre en compte la dépendance entre les listes
  - Prendre en compte des covariables (âge, grade de sévérité, ...)

Modèle de Capture-recapture

Sur les 4 régions étudiées :
$\widehat{N} = 191$ pour 116 malades recensés

Table 3 Estimated numbers of patients and estimated capture probabilities by individual covariates

Individual covariate	Numbers of recorded patients	Estimated number of patients*	95% CI	Estimated capture probability*
Age				
Under 15 years	42	58	[49–79]	72%
Older than 15 years	74	133	[102–197]	56%
Severity grade				
Grade 1 (mild)	40	79	[55–141]	50%
Grade 2 (moderate)	40	67	[51–107]	59%
Grade 3 (severe)	26	34	[28–49]	77%
Grade 4 (very severe)	10	11	[10–17]	95%

*Estimation obtained from the best log-linear model (independent sources model with three covariates effects).

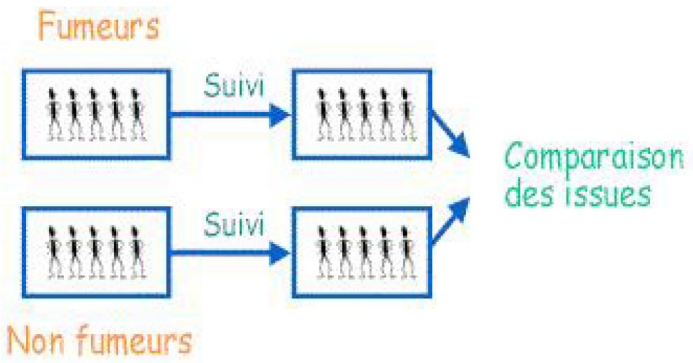
Survenue d'un évènement au cours du temps

Fertilité Humaine :
Quel délai pour avoir un bébé ?

Analyse de données censurées

Survenue d'un évènement au cours du temps

- Cohorte épidémiologique : un groupe de sujets exposés suivi pendant une période donnée et comparé à un groupe de sujets non exposés



La maladie est-elle plus fréquence chez les exposés que chez les non-exposés ?

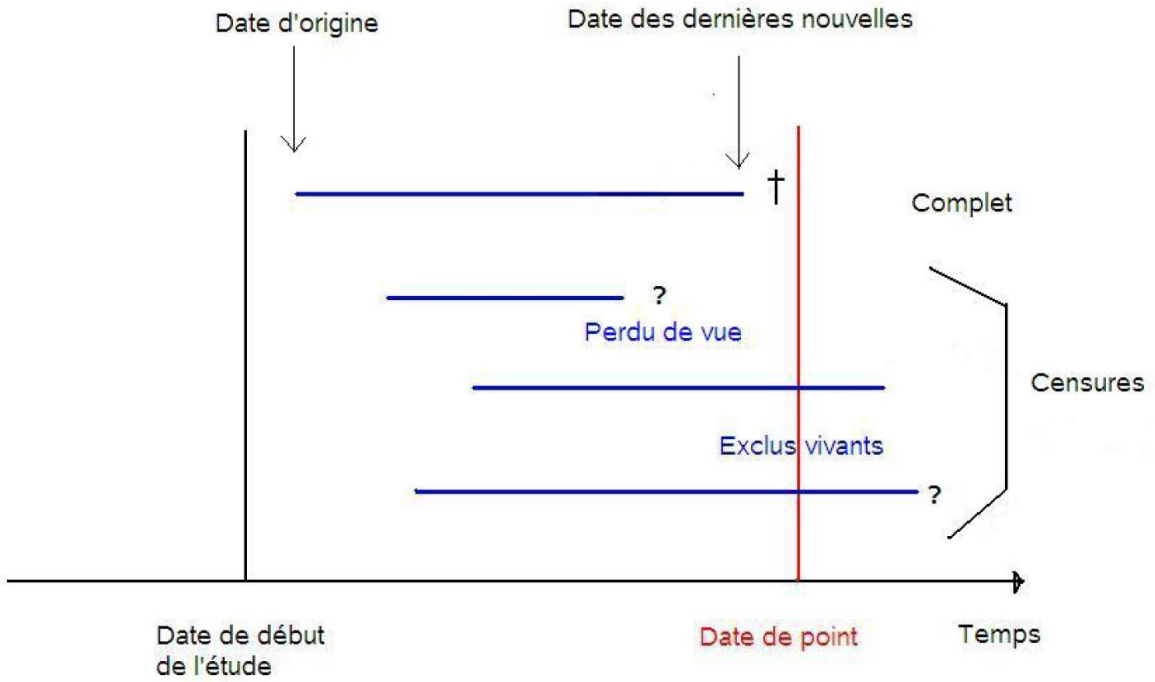
$\Downarrow$

Deux problématiques :

Données répétées, à intervalles irréguliers, incomplètes

Evènements pas toujours observables $\rightarrow$ censurés

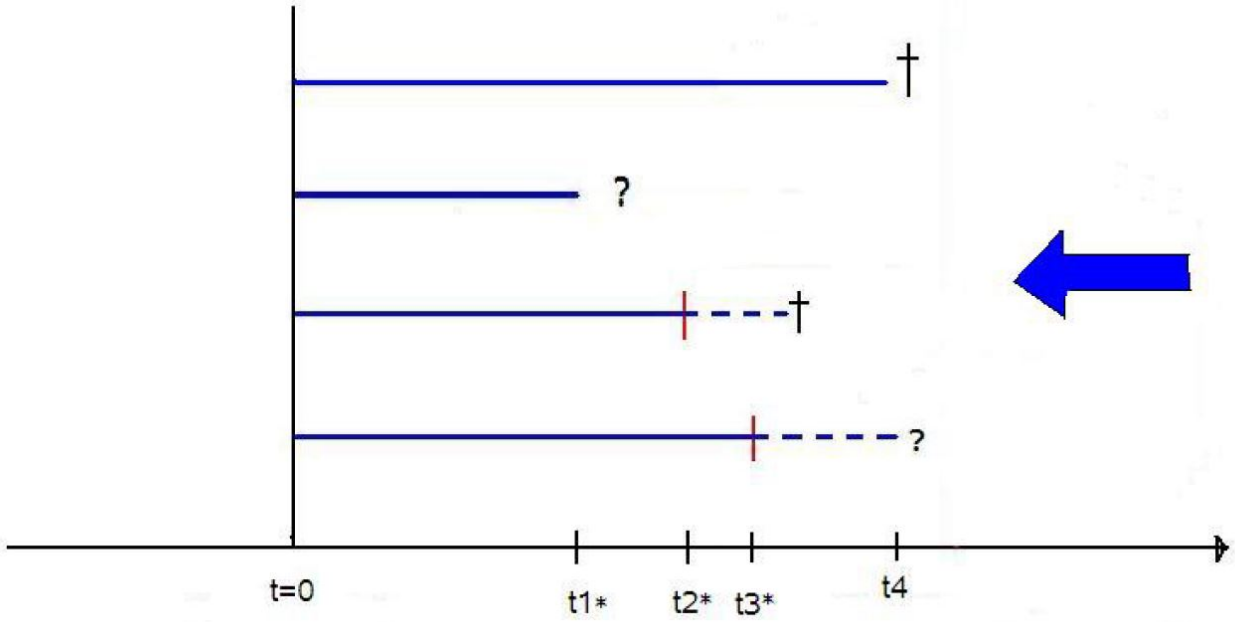
Survenue d'un évènement au cours du temps



L'absence d'évènement apporte aussi une information :
Comment prendre en compte les données censurées ?

Survenue d'un évènement au cours du temps

Etude des délais de survenue de l'évènement



⇓

Analyse des "données de survie"

Survenue d'un évènement au cours du temps

Fonction de survie

- Soit T : v.a. délai de survenue de l'évènement
- Fonction de survie $S(t)$: Probabilité qu'un patient soit encore vivant après un délai t

$$S(t) = P(T > t) = 1 - F(t)$$

- $S(t|\tau)$: probabilité de survivre après un délai t sachant que l'on est survivant après un délai τ ($\tau < t$) :

$$S(t|\tau) = Pr(T > t|T > \tau) = \frac{Pr(T > t)}{Pr(T > \tau)} = \frac{S(t)}{S(\tau)}$$

$$\Rightarrow S(t) = S(\tau)S(t|\tau)$$

Méthode de Kaplan-Meier :
On estime la fonction de survie S aux seuls temps de décès observés :

$$S(t_i) = S(t_{i-1})S(t_i|t_{i-1})$$

L'estimation proposée pour $S(t_i|t_{i-1})$ est donnée par :

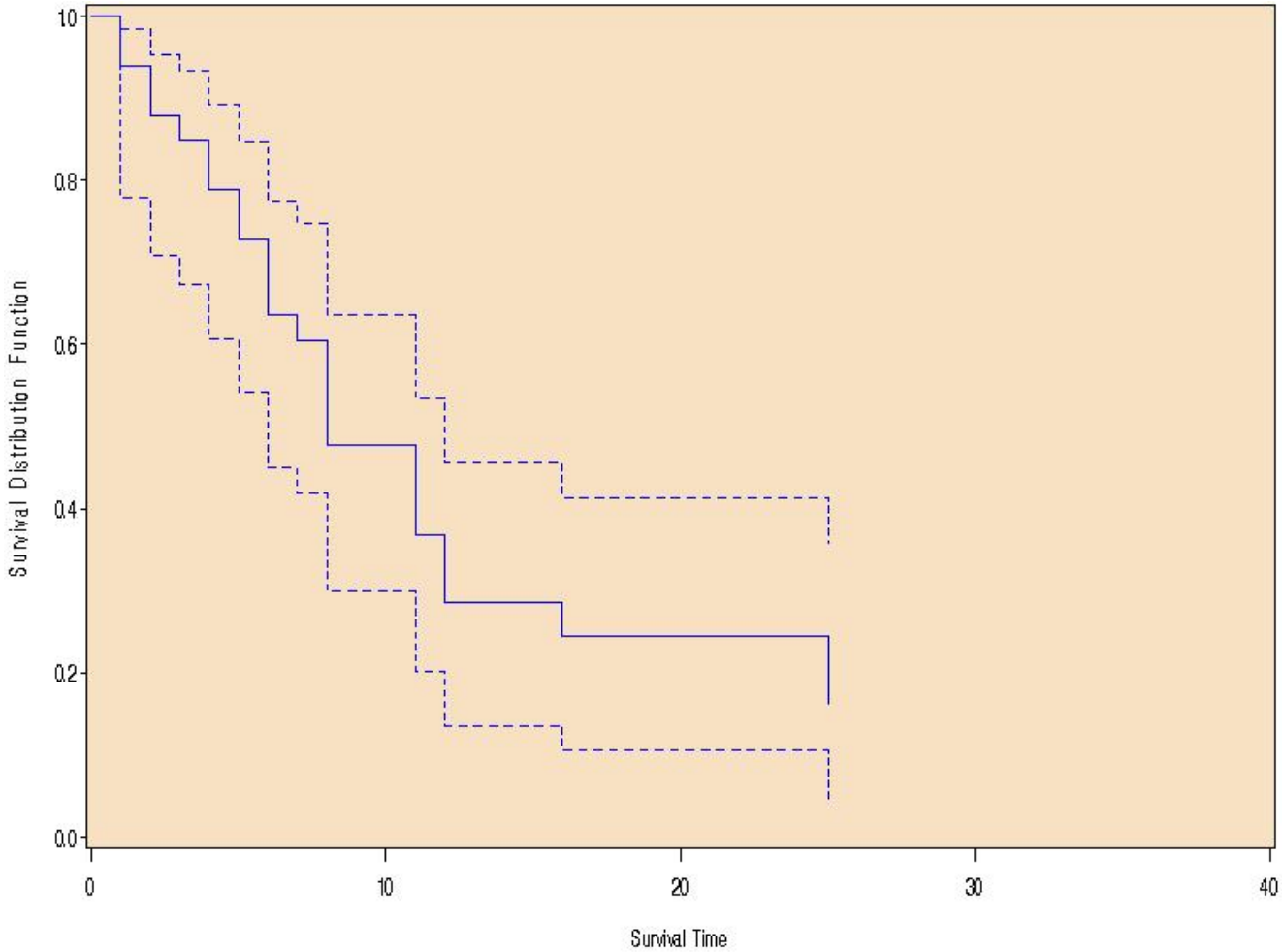
$$\widehat{S}(t_i|t_{i-1}) = 1 - \frac{D_i}{N_i - C_i}$$

où

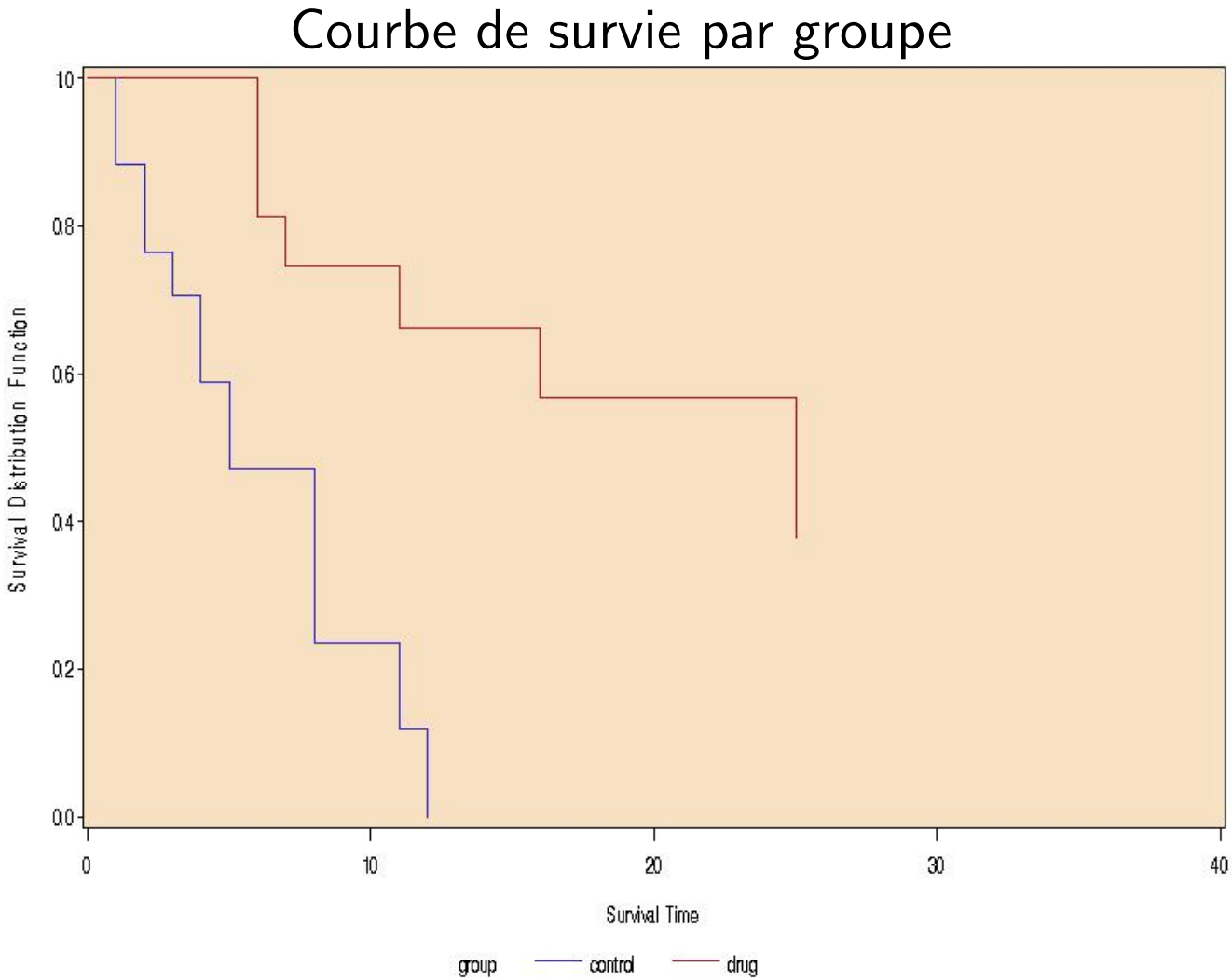
- D_i est le nombre de décès observé au temps t_i,
- C_i est le nombre de patients censurés entre t_{i-1} et t_i,
- N_i est le nombre de patients connus vivants juste après t_{i-1}.

Survenue d'un évènement au cours du temps

Courbe de survie (avec IC)



Survenue d'un évènement au cours du temps



Survenue d'un évènement au cours du temps

- Comment évaluer l'effet d'un facteur sur la "survie" ?
  $\Rightarrow$ Modèle de Cox
- Risque instantané de décès au délai t : probabilité de décéder juste après t, au cours d'un intervalle de durée $dt \rightarrow 0$) :

$$\lambda(t) = -\frac{dS}{dt}(t)/S(t)$$

- Le modèle de Cox exprime $\lambda(t, X^1, X^2, ...X^p)$ sous la forme :

$$\lambda(t, X^1, X^2, ...X^p) = \lambda_0(t)\ exp\{\sum_{j=1}^{p} \beta_j X^j\}$$

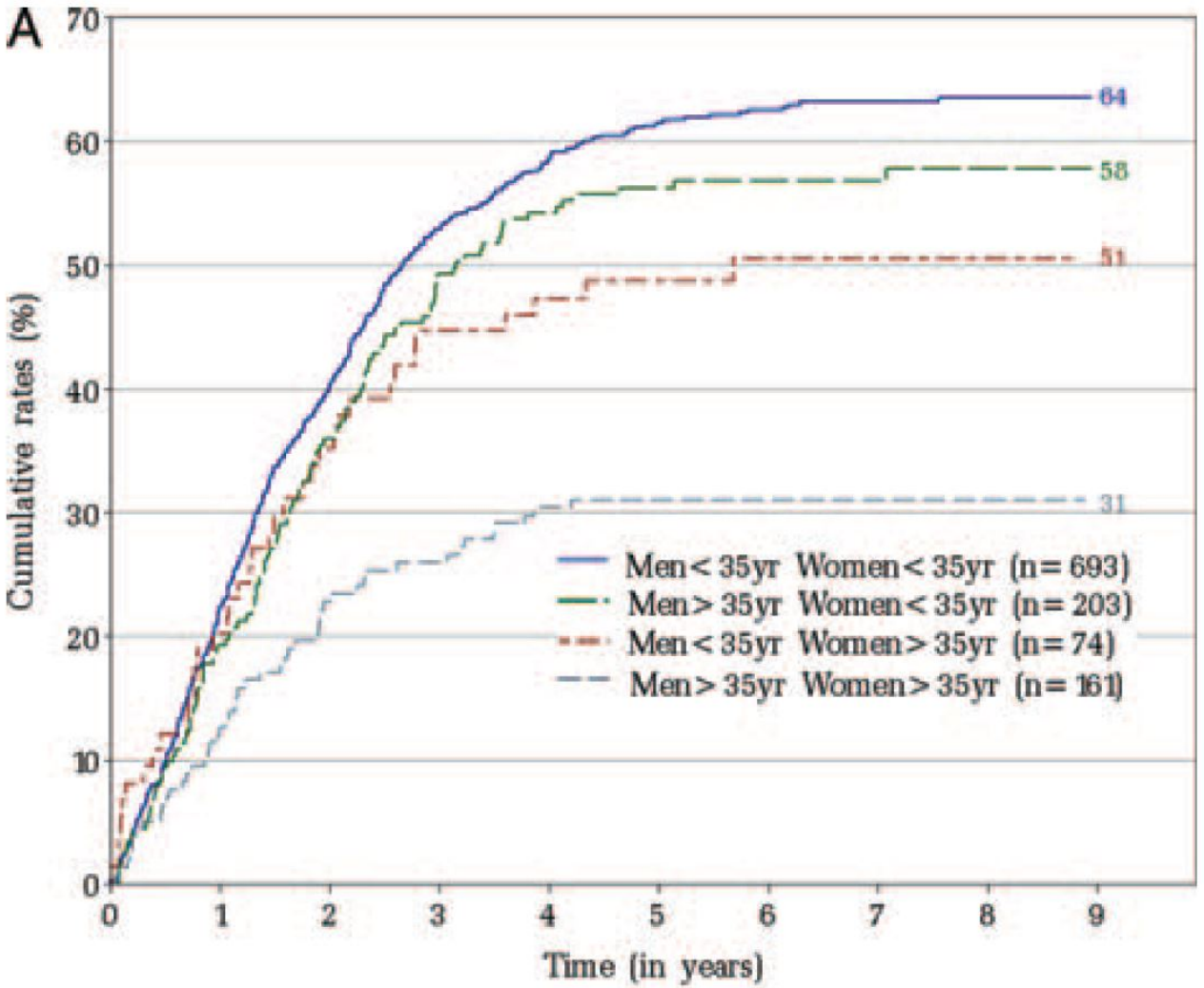
- Le cœfficient β_j représente l'effet de la caractéristique X^j sur la survenue de l'évènement. $exp(\beta_j)$ représente le facteur par lequel le risque de décès est multiplié en cas de présence de la caractéristique X_j.

Survenue d'un évènement au cours du temps

- Etude menée par le groupe de recherche sur la fertilité humaine à l'hôpital Purpan
- 1131 couples suivis au moins 4 années jusqu'à l'accouchement, l'arrêt du traitement ou la fin de l'étude (2008)
  - 56% des couples ont eu un enfant $\rightarrow$ délai jusqu'à la grossesse
  - pour les autres couples $\rightarrow$ délai de suivi sans évènement

- Deux questions de recherche :
  - Dans quel délai les couples ont-ils un enfant ?
  - Quels sont les facteurs qui peuvent influencer ce délai ?

Survenue d'un évènement au cours du temps



Transmission du VIH de la mère à l'enfant :
estimation du moment

Modèle de markov

Modélisation par processus de Markov

- Cohorte où l'on ne peut pas dater la survenue de l'évènement d'intérêt
- Données longitudinales incomplètes :
  - Suivi partiel de la maladie
  - Suivi intermittent avec des visites plus ou moins régulières
  - Datation des évènements inconnue

  Au lieu d'observer une date d'évènement, on observe un intervalle de temps au cours duquel l'évènement s'est réalisé

$$\Downarrow$$

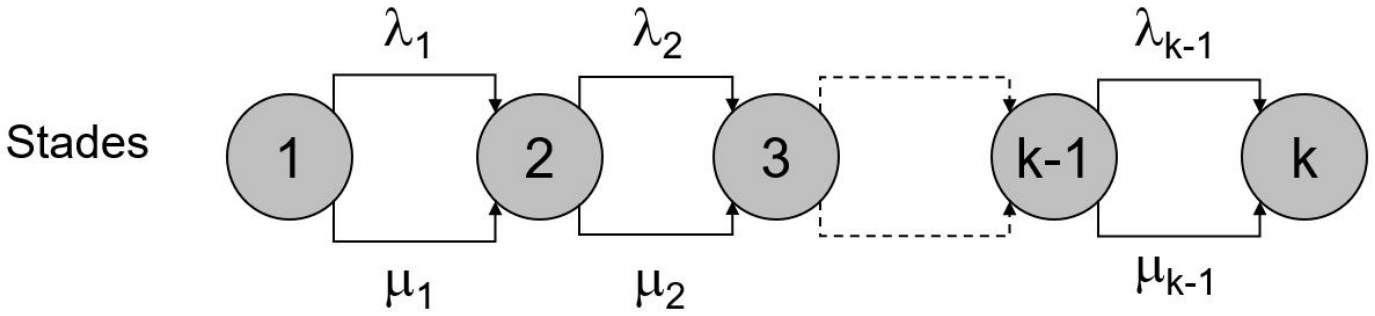
Données censurées par intervalle

Modélisation par processus de Markov

Modèles de Markov

- permettent d'étudier l'évolution d'un processus clinique à travers différents stades de la maladie
- peuvent être homogènes dans le temps ou dépendants du temps
- peuvent être à stades réversibles ou irréversibles

Modèle de Markov homogène dans le temps à k stadesréversibles



- Vraisemblance du modèle

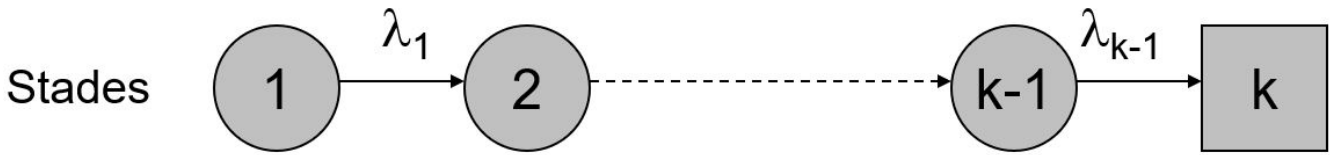
Les données (X_{ij}, D_{ij}) pour N sujets indépendants ayant n_i prélèvements

- D_{ij} : durée entre les prélèvements j-1 et j
- X_{ij} : stade observé pour le sujet i au $j^{ème}$ prélèvement

$$L(\lambda,\mu) = \Pi_i \Pi_j \underbrace{Pr\ x_{ij-1},x_{ij}(D_{ij})}$$

probabilité de transition du stade X_{ij-1} au stade X_{ij} sur un intervalle de durée D_{ij}

Modèle de Markov homogène dans le temps à k stades irréversibles



- On peut écrire les probabilités de transition à partir des distributions des temps d'attente dans chaque stade du processus :

$$\Pr \{ X(t_0+D) = k \mid X(t_0) = k \} = \Pr (T_k > D)$$

- Les distributions de T_k doivent vérifier la propriété suivante qui décrit un processus homogène dans le temps :

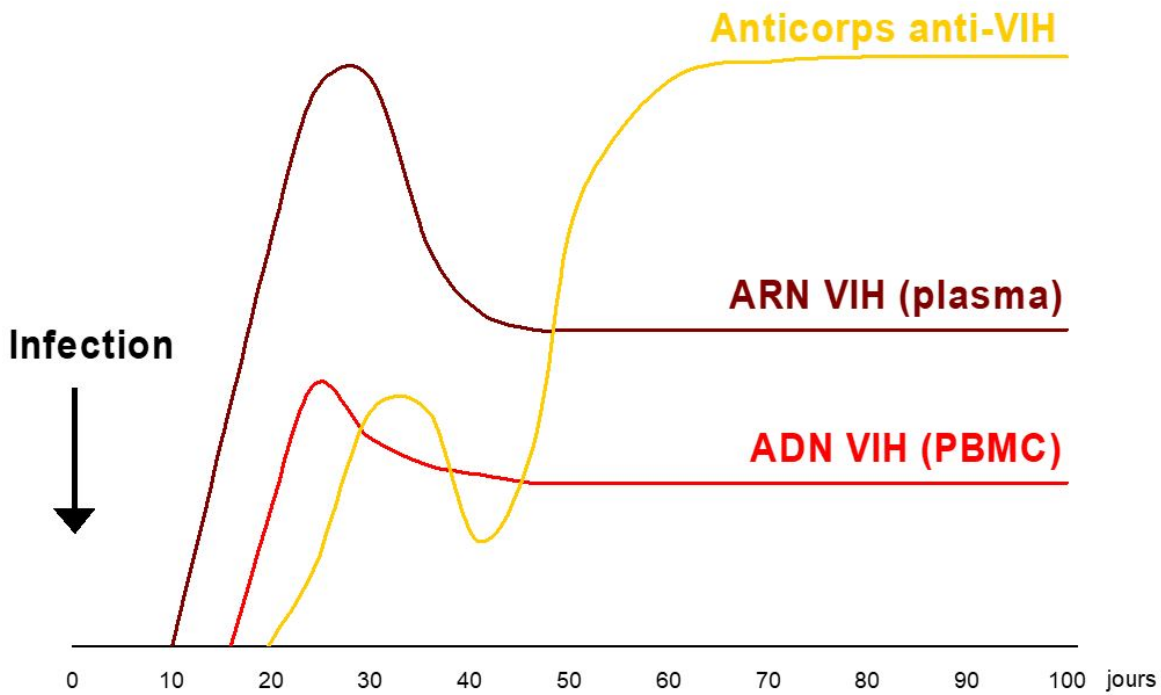
$$\Pr (T_k > t_0+D \mid T_k > t_0) = \Pr (T_k > D)$$

=> T_k est distribuée selon une loi exponentielle de paramètre λ_k

Transmission du VIH de la mère à l'enfant

- La transmission peut se produire in utero, intrapartum ou par allaitement
- Moment de la transmission
  - ✓ Non observable, risque de contamination du foetus
  - ✓ Eléments en faveur d'une transmission tardive
  - ✓ Axe de recherche important pour le développement de nouvelles stratégies de prévention
- Objectif de ce travail :
  estimer le moment de la transmission materno-fœtale du VIH

Schéma d'apparition des marqueurs viraux et immunologiques de la contamination au début de la séroconversion

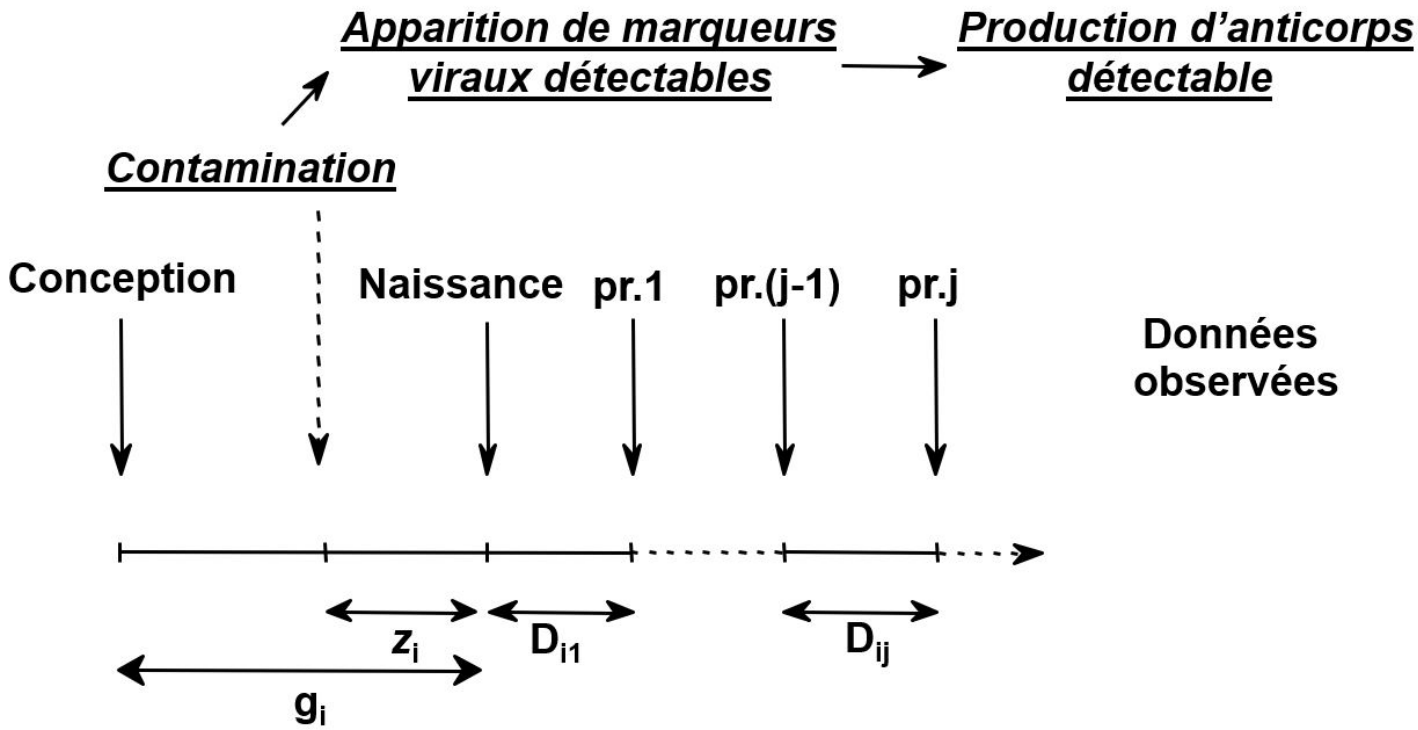


Busch et al (1997). Applications of molecular biology to blood transfusion medecine. American Association of blood banks.

Population

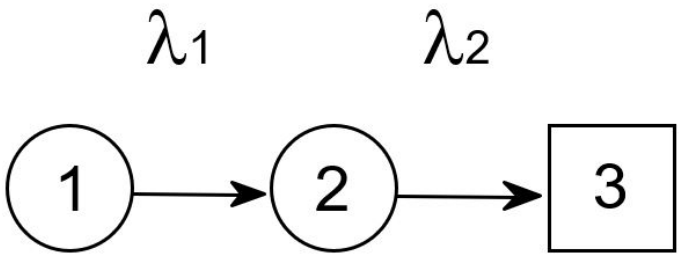
- 700 enfants nés de mère séropositive inclus entre mai 1988 et juin 1993 dans la Cohorte Pédiatrique Française
- 135 enfants diagnostiqués infectés par le VIH-1
  - ✓ production d'anticorps à l'âge de 18 mois
  - ✓ mort par SIDA avant cet âge
- Prélèvements de sang dans les trois premiers mois de vie pour les enfants infectés
  - ✓ détection des marqueurs viraux
  - ✓ production d'anticorps anti-VIH

Calendrier des évènements pré et postnataux pour un enfant infecté



Un modèle de Markov à 3 stades

caractérise l'évolution conjointe des marqueurs viraux et des anticorps chez l'enfant infecté



Infection	+	+	+
Marqueurs viraux	-	+	+
Production d'anticorps	-	-	+

Par exemple : { (1,3), (2,32), (2,15), (3,16) }

Processus observés des 135 enfants infectés

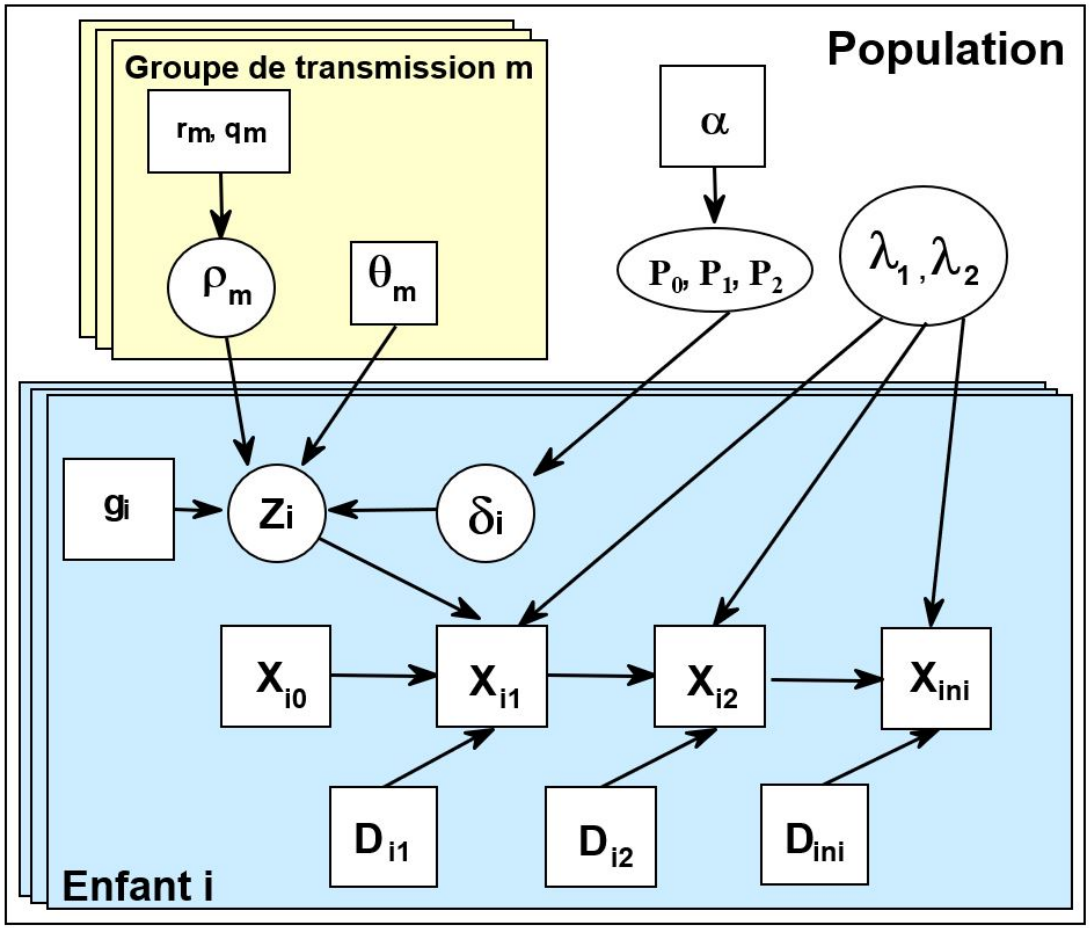
Stade observé au 1er prélèvement	Processus observé sur les prélèvements suivants	Nombre de cas	
1		22	87
1	→ 1	6	
1	→ 1/2	1	
1	→ 2	10	
1	→ 2 → 2	15	
1	→ 2 → 2 → 3	1	
1	→ 2 → 3	6	
1	→ 2/3	16	
1	→ 3	10	
1/2		4	
2		3	30
2	→ 2	19	
2	→ 2 → 3	2	
2	→ 3	6	
2/3		12	
3		2	

Modélisation de la durée entre la contamination et la naissance

- Mélange de distributions à trois composantes prenant en compte les “mécanismes” de transmission :
  - ✓ in utero, précoce ou tardive
  - ✓ intrapartum, pendant l’accouchement
- Chaque composante m est caractérisée par :
  - ✓ une proportion P_m d’enfants infectés dans ce groupe
  - ✓ une distribution gamma de la durée z_i
- Le modèle pour z_i s’écrit :

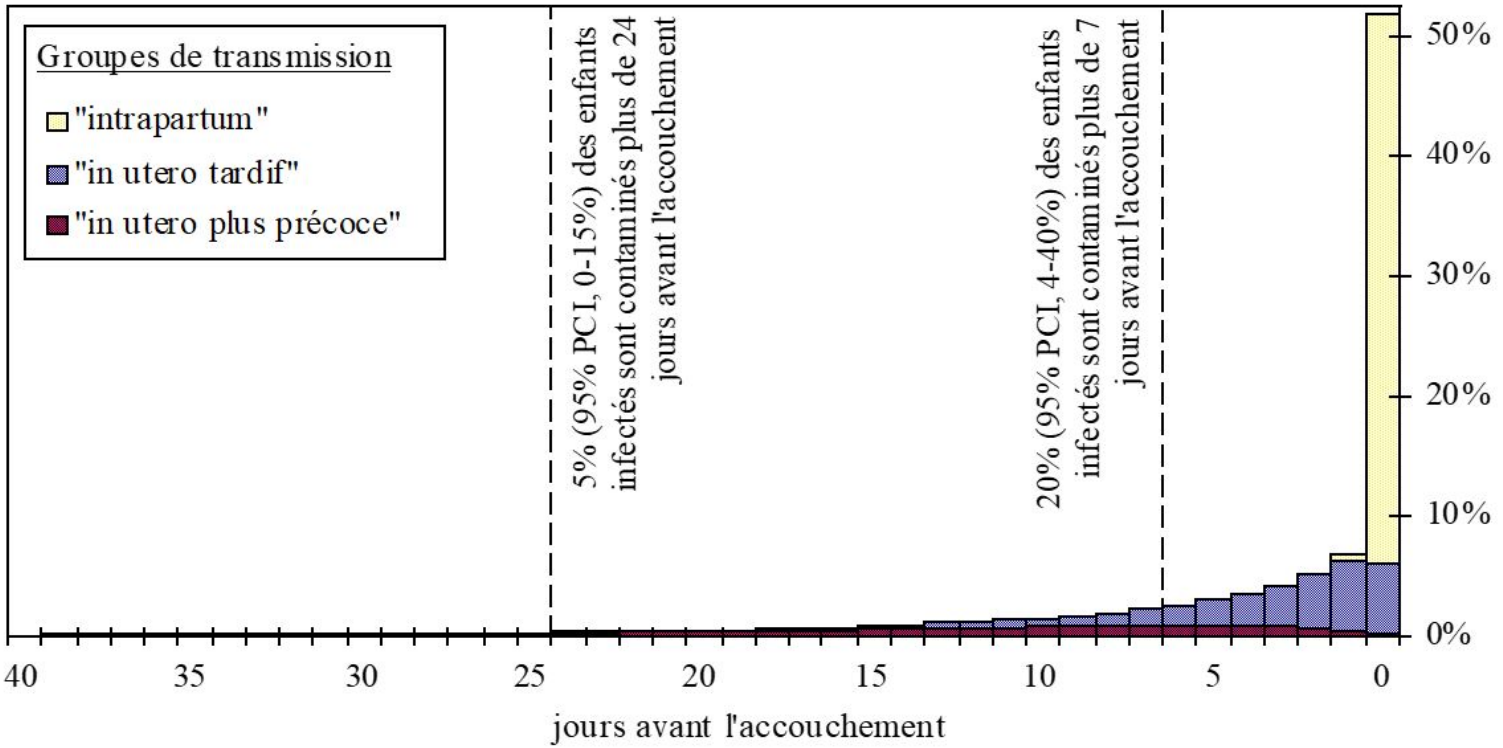
$$z_i \sim P_0.\Gamma_0(\bullet|\theta_0, \rho_0) + P_1.\Gamma_1(\bullet|\theta_1, \rho_1) + P_2.\Gamma_2(\bullet|\theta_2, \rho_2)$$

Graphe du modèle

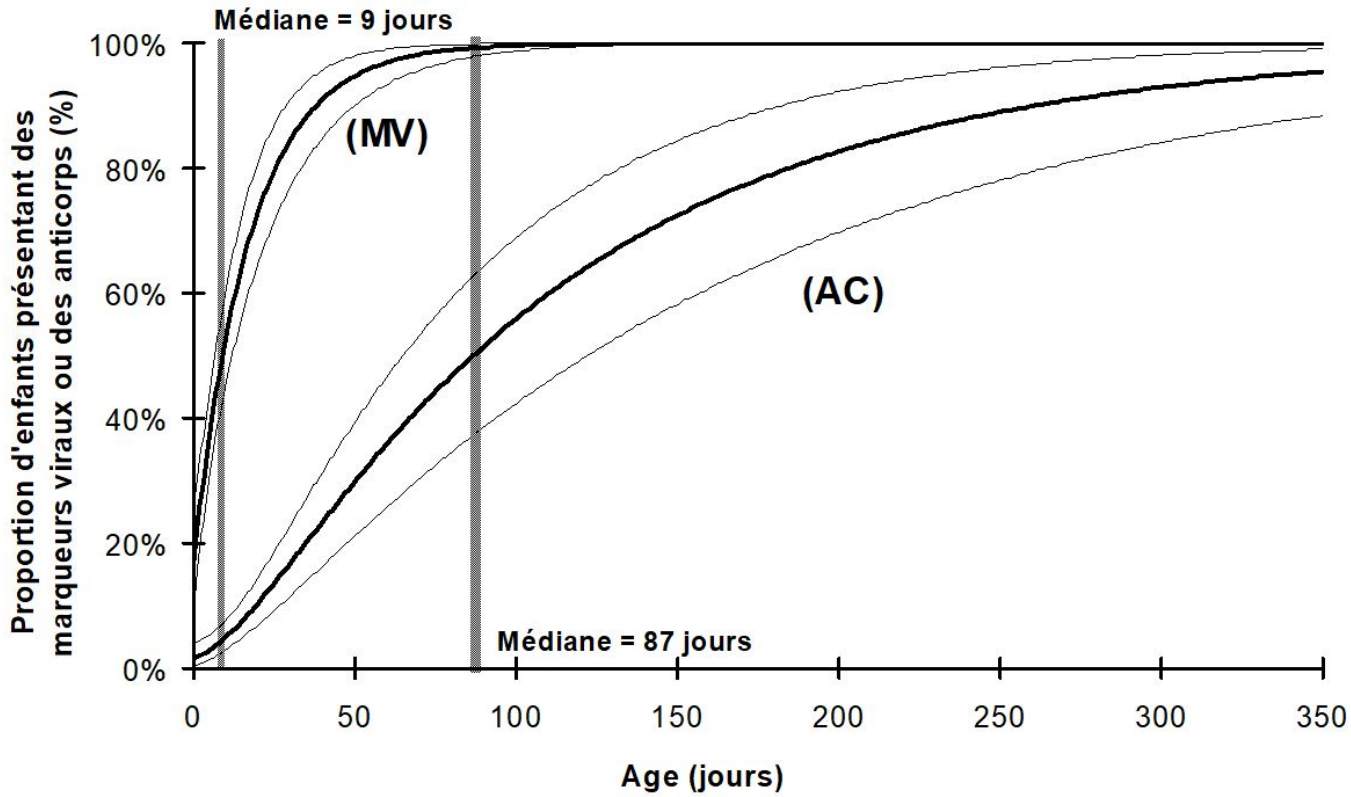


$$P(z, \delta, \lambda_1, \lambda_2, \Theta, P, X, D, g)$$
$$= P(\lambda_1) \cdot P(\lambda_2) \cdot P(\Theta) \cdot P(P) \cdot P(\delta|P) \cdot P(Z|\Theta,\delta,g) \cdot P(X|D,z,\lambda_1,\lambda_2)$$

Modélisation par processus de Markov



Evolution de l'émergence des marqueurs viraux et de la production d'anticorps au cours de la première année de vie des enfants infectés



Conclusions

- Notre modèle suggère que pour une grande majorité d'enfants infectés nés de mère séropositive, la transmission de la mère à l'enfant se produit au cours du dernier mois de la grossesse :
  - ✓ la moitié des cas le jour de l'accouchement attribuables à une contamination in utero tardive ou intrapartum,
  - ✓ l'autre moitié des enfants infectés in utero en moyenne 10 jours avant l'accouchement.
- La transmission in utero dans les premiers mois de vie est extrêmement rare :
  - ✓ seulement 1% des enfants infectés ont été contaminés plus de 2 mois avant l'accouchement

Classification de trajectoires

Exposition médicamenteuse au cours de la grossesse :
Quel type d'exposition et quel risque associé pour l'enfant ?

Classification de trajectoires et modélisation

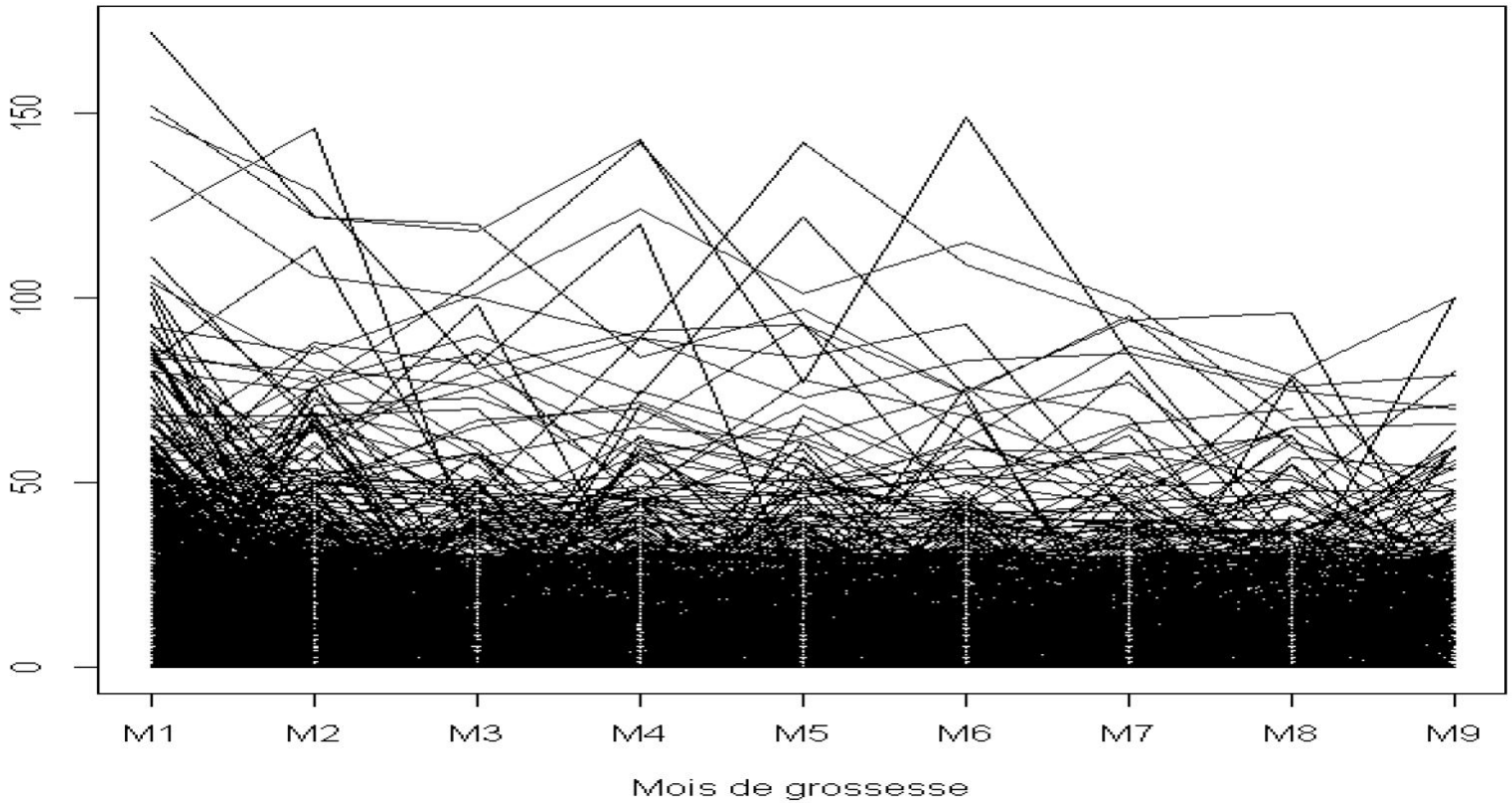
Classification de trajectoires

- Contexte :
  Effets indésirables médicamenteux liés à la dose et à la durée de l'exposition :
  $\Rightarrow$ Exemple au cours de la grossesse avec le risque sur l'enfant variable en fonction des paramètres (dose, durée, période)
- Problématique :
  Proposer une nouvelle méthode statistique en pharmaco-épidémiologie
  - prenant en compte l'aspect longitudinal et quantitatif de l'exposition médicamenteuse
  - de regroupement de patients proches en terme d'exposition médicamenteuse

Classification de trajectoires

- Application à l'exposition aux psychotropes au cours de la grossesse dans la cohorte EFEMERIS
- 79000 couples mère-issue de grossesse
  dont 3700 mères sous psychotropes à un moment de la grossesse
  - Informations sur la grossesse
  - Informations sur la prescription de médicaments : nom du médicament, dosage, date de délivrance
- Mesure de l'exposition par les DDD (defined daily dose) au cours de la grossesse

- Construction des trajectoires individuelles d'exposition
  - Par patient et par unité de temps



Classification de trajectoires

- Méthode des K-means :
  méthode de classification non-supervisée
  - Objectif : diviser la population en clusters homogènes vis-à-vis d'une ou plusieurs caractéristiques
  - Minimiser l'inertie intra-groupe
    $\rightarrow$ Concentration des points autour du centre de gravité
  - Maximser l'inertie inter-groupe
    $\rightarrow$ Eloignement des centres de groupes entre eux
- Dans le cas de données longitudinales :
  classification de trajectoires par la méthode KmL

Classification de trajectoires

- Soit y_{it} une variable mesurée pour un sujet i $(i = 1...n)$ à différents instants (t=1...T)
  $\Rightarrow$ y_i trajectoire du sujet i définie par les T mesures
- Algorithme
  - Initialisation : tirage au sort de K centres de gravité
  - Affectation des sujets au centre de gravité le plus proche
  - Calcul de la distance euclidienne entre 2 trajectoires individuelles des sujets i et j :

$$dist^E(y_i, y_j) = \sqrt{\sum_{t=1}^{T}(y_{it} - y_{jt})^2}$$

  - Calcul des nouveaux centres de gravité et rè-allocation des sujets aux nouveaux clusters
  - Convergence quand plus aucun sujet ne change de cluster

Qualité de la classification

- Critère de Calinski et Harabatz :

$$C(k) = \frac{Trace(B)/(k-1)}{Trace(W)/(n-k)}$$

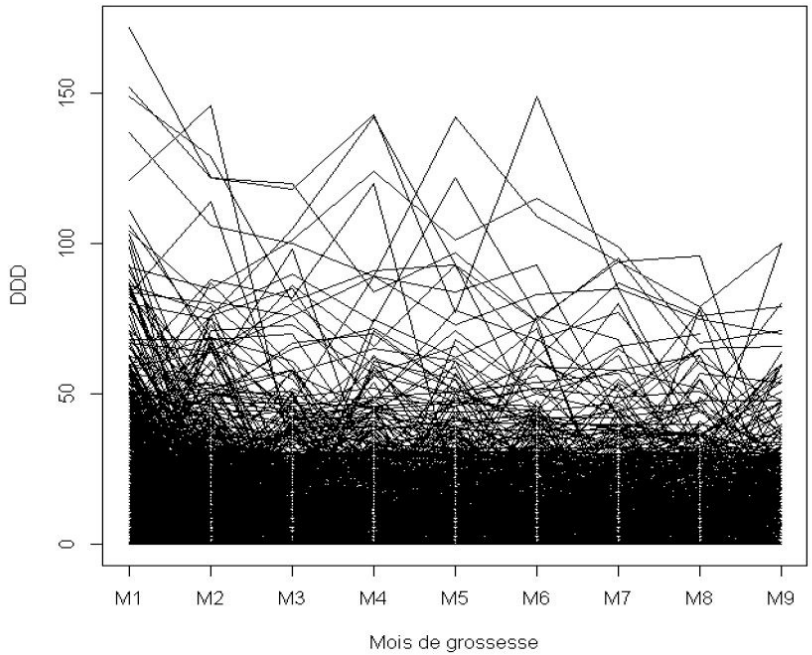
où

- $Trace(B)$: inertie inter-cluster (si possible, élevée)
- $Trace(W)$: inertie intra-cluster (si possible, faible)

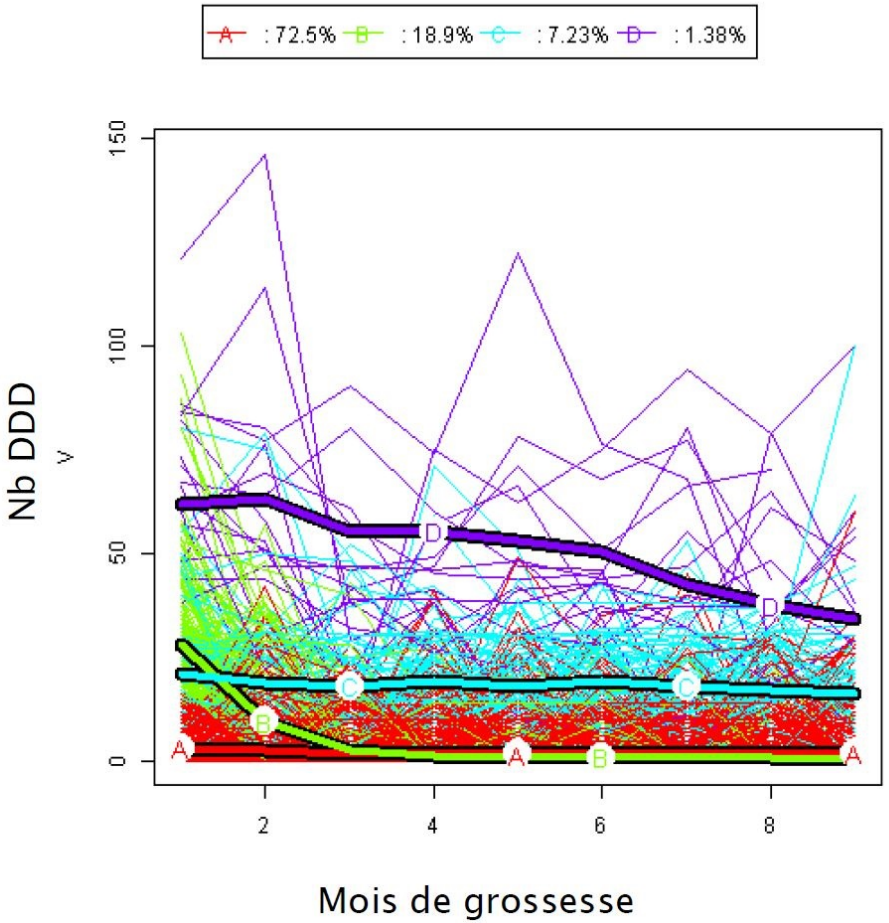
$\Rightarrow$ $C(k)$ élevé : clusters homogènes et bien séparés

Classification de trajectoires

Trajectoires d'exposition aux psychotropes
(N = 3708)



Classification



Classification de trajectoires

Le risque de pathologies néonatales est-il différent selon les 4 profils d'exposition ?

- Non-exposés : 6%
- Cluster A (faible exposition) : 5.3%
- Cluster B (exposition décroissante) : 7.1%
- Cluster C (exposition modérée) : 11.6%
- Cluster D (exposition élevée) : 29.6%

Classification de trajectoires

Modélisation du risque de pathologie néonatale par une régression logistique

- On explique une variable Y_i binaire (évènement/non-évènement), distribuée selon une loi de Bernoulli :

$$Y_i \sim Bernoulli(p_i) \text{ avec } E(Y_i) = p_i$$

où p_i est la probabilité de survenue de l'évènement : $p_i = P(Y_i = 1)$

- Pour étudier les effets de variables sur la survenue d'un évènement, on modélise la probabilité de survenue de l'évènement p_i en fonction de ces variables.

- Le modèle de régression logistique exprime une fonction de p_i en fonction des variables explicatives :

$$logit(p_i) = \ln(\frac{p_i}{1 - pi}) = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + e_i$$

β_j exprime l'effet de la variable X^j sur la probabilité de survenue de l'évènement.
Sa nullité est synonyme d'absence d'effet de la variable X^j.

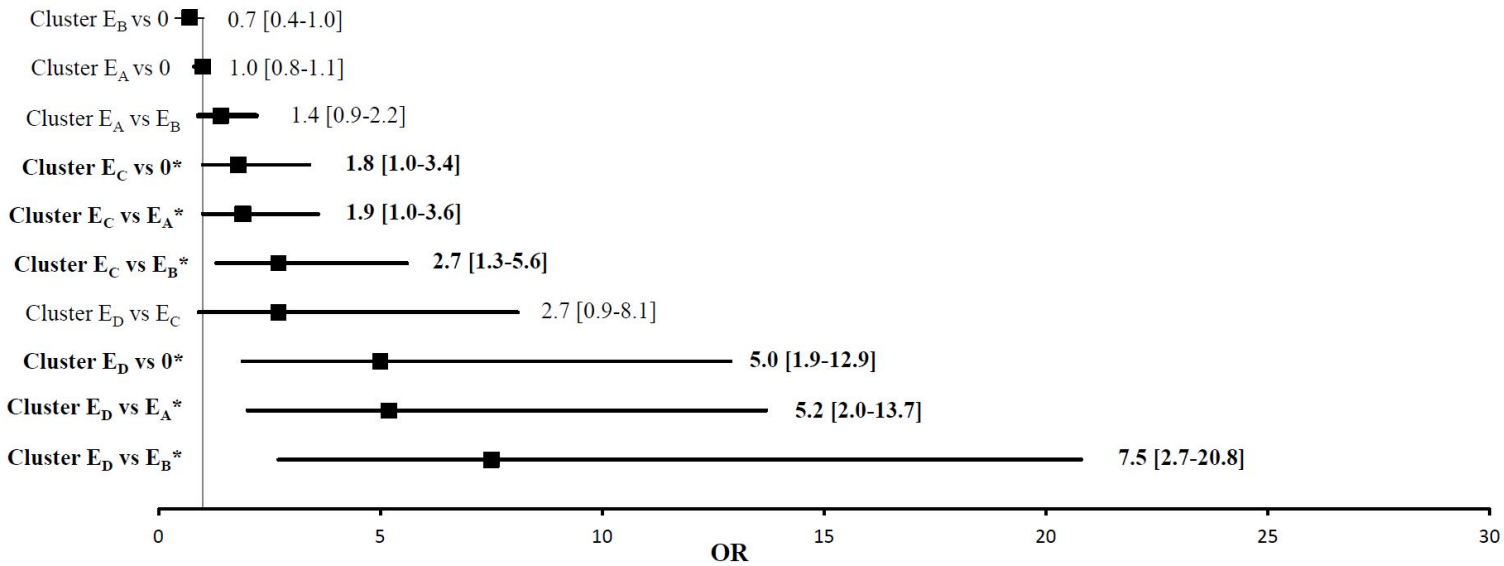
- Interprétation de l'effet par l'odds-ratio :

$$OR_j = e^{\widehat{\beta_j}}$$

Le risque d'évènement est multiplié par OR_j quand le facteur est présent.

Classification de trajectoires

Figure 3 Association between anxiolytic-hypnotic burden exposure (classified in clusters) and risk of neonatal pathologies (Analysis 1_b)



* Significant association: p <0.05 – Adjusted on congenital anomalies, multiple births, sex, hypertension during pregnancy, duration of pregnancy and exposure to antipsychotic, antidepressant and antiepileptic drugs

Et encore bien d'autres problématiques !

- Evaluation des tests diagnostiques :
  Sensibilité, spécificité $\Rightarrow$ probabilités conditionnelles
- Modèles pharmaconétiques :
  étude des actions d'un médicament et de son évolution dans l'organisme
- Analyse de données génétiques :
  identification de gènes responsables de maladies