Epidémiologie

Plan

1. Introduction
2. Enquête de cohorte
3. Enquête cas-témoins
4. Mesures de risques
5. Mesures d'association
6. Biais de sélection
7. Biais de classement
8. Biais de confusion
9. Stratégie d'analyse
10. Puissance
11. Modèles multivariés
12. Régression logistique

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

Epidémiologie

6. Biais de sélection

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

6. Biais de sélection

I. Définition

- Biais et paramètre étudié

II. Situations classiques de biais de sélection

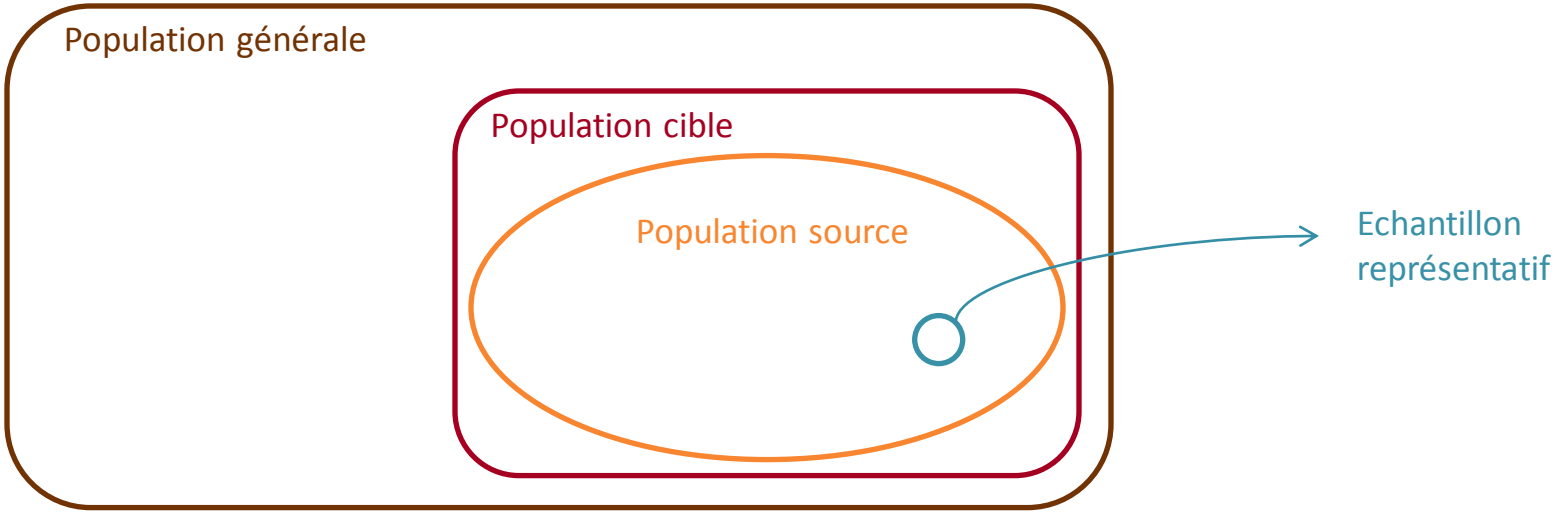
- Enquêtes cas-témoins
- Enquête de cohorte
- Healthy worker effect
- Non réponses et refus de participation

III. Limiter les biais de sélection

- Au moment de la planification
- Au moment de l'analyse

I. Définition

- Le biais de sélection résulte de la façon de choisir et de suivre les sujets de l'échantillon
- Biais de sélection si l'échantillon sélectionné n'est pas représentatif de la population cible
- Biais de sélection possible même si l'échantillon est représentatif de la population source ➡ mauvais choix de la population source



Biais et paramètre étudié

Ex: Etude de l'association entre le multi partenariat (E+ : >2 partenaires; E- : ≤1 partenaire) et une MST dans une ville de 500 000 habitants

- Population source = patients consultant gynécos et généralistes de la ville
  - On sélectionne plus de M+ (car les malades ont tendance à plus consulter)
  - On sélectionne la même proportion de E+ et de E- (parmi les M+ et les M-)

Population cible ($N = 500\,000$)

	E+	E-
M+	10 000	10 000
M-	140 000	340 000

$P = \frac{20000}{500000} = 0.04$

$OR = \frac{P_1(1-P_1)}{P_0(1-P_0)} = 2.43$

Population source ($N = 52\,000$)

	E+	E-
M+	2 000 (20%) ↘	2 000 (20%) ↘
M-	14 000 (10%) ↘↘	34 000 (10%) ↘↘

$P = 0.077$

$OR = 2.43$

- Sélection pas indépendante de la maladie ⟶ Biais sur l'estimation de la prévalence
- $P_1 = P(M+/E+)$ et $P_0 = P(M+/E-)$ différentes dans population cible et source

  $P_{E_1} = P(E+/M+)$ et $P_{E_0} = P(E+/M-)$ inchangées dans population cible et source

  ⟶ Pas de biais pour l'estimation de l'odds ratio $\left(car\ OR = \frac{P_{E_1}(1-P_{E_1})}{P_{E_0}(1-P_{E_0})}\right)$

  Biais pour l'estimation du RR

Biais et paramètre étudié

- Population source = patients consultant gynécos et généraliste de la ville après une campagne de prévention et d'information sur la MST
  - On sélectionne plus de M+ (car les malades ont tendance à plus consulter)
  - On sélectionne plus de E+ chez les M- et les M+ (campagne d'information sur les facteurs de risque)

Population cible ($N = 500\,000$)

	E+	E-
M+	10 000	10 000
M-	140 000	340 000

$P = \frac{20000}{500000} = 0.04$

$OR = \frac{P_1(1-P_1)}{P_0(1-P_0)} = 2.43$

Population source ($N = 52\,000$)

	E+	E-
M+	2 100 (21%) ↘	1900 (19%) ↘↘
M-	16 000 (11%) ↘↘	32 000 (9%) ↘↘↘

$P = 0.077$

$OR = 2.21$

- Sélection pas indépendante de la maladie ⟶ Biais sur l'estimation de la prévalence
- $P_1 = P(M+/E+)$ et $P_0 = P(M+/E-)$ différentes dans population cible et source

  $P_{E_1} = P(E+/M+)$ et $P_{E_0} = P(E+/M-)$ différentes dans population cible et source

  ⟶ Biais pour l'estimation de OR et du RR

Biais et paramètre étudié

- Population source = patients consultant gynécos et généraliste de la ville après une campagne de prévention et d'information sur une MST asymptomatique
  - On sélectionne autant de M+ que de M- (car MST asymptomatique)
  - On sélectionne plus de E+ chez les M- et les M+ (campagne d'information sur les facteurs de risque)

Population cible ($N = 500\,000$)

	E+	E-
M+	10 000	10 000
M-	140 000	340 000

$P = \frac{20000}{500000} = 0.04$

$OR = \frac{P_1(1-P_1)}{P_0(1-P_0)} = 2.43$

Population source ($N = 50\,000$)

	E+	E-
M+	1200 (12%) ↘	800 (8%) ↘
M-	16 000 (11%) ↘	32 000 (9%) ↘

$P = 0.04$

$OR = 3$

- Sélection indépendante de la maladie ⟶ Pas de biais sur l'estimation de la prévalence
- $P_1 = P(M+/E+)$ et $P_0 = P(M+/E-)$ différentes dans population cible et source

  $P_{E_1} = P(E+/M+)$ et $P_{E_0} = P(E+/M-)$ différentes dans population cible et source

  ⟶ Biais pour l'estimation de OR et du RR

II. Situations classiques de biais de sélection

1. Dans une enquête transversale
   - Echantillon non constitué par tirage au sort (Ex : échantillon de volontaire)
2. Dans une enquêtes cas-témoins
   - Constitution du groupe "témoin"
   - Biais de survie sélective (recrutement de cas prévalent)
3. Dans une enquête de cohorte
   - Constitution du groupe "exposé"
   - Perdus de vue
4. "Healthy worker effect" dans les enquêtes transversale et de cohorte
5. Refus de participation

Biais de sélection et enquête cas-témoins

- Le recrutement de témoins en milieu hospitalier (souvent impossible de recruter les cas dans un registre) qui ne sont pas représentatifs de la population cible (biais de Berkson)

- Etude de l'association entre le cancer bronchique et le tabac

  - Cas : malades hospitalisés pour un cancer broncho-pulmonaire

  - Témoins : malades hospitalisés pour d'autres pathologies pulmonaires ou cardiovasculaires (souvent liées au tabac) ⟶ les témoins fument plus que dans la population cible

Population cible

	E+	E-
M+	a	b
M-	c	d

Population source

	E+	E-
M+	a'=a	b'=b
M-	c'>>c	d'<<d

$$OR' = \frac{ad'}{bc'} < \frac{ad}{cb} = OR$$ ⟶ sous-estimation de l'OR

Biais de sélection et enquête cas-témoins

- Autre exemple de biais de sélection lié aux choix des témoins
  - Association entre le cancer des cervicales et un faible niveau socio-économique
  - Cas : recrutés dans plusieurs hôpitaux d'une région
  - Témoins : recrutés par porte à porte autour des hôpitaux de 9h à 17h
  - Cas et témoins sélectionnés par des mécanismes différents

    ➡ voisinage des hôpitaux pour les témoins et toute la région pour les cas
  - De plus, les témoins inclus dans l'étude ont plus de chance d'être sans emploi

    ➡ les témoins ont plus de chance d'être enrôlé s'ils sont exposés

Population cible

	E+	E-
M+	a	b
M-	c	d

Population source

	E+	E-
M+	a'=a	b'=b
M-	c'>>c	d'<<d

$$OR' = \frac{ad'}{bc'} < \frac{ad}{cb} = OR$$ ➡ sous-estimation de l'OR

Biais de sélection et enquête cas-témoins

- Biais liés à une différence de surveillance
  - Association entre la thrombose et un contraceptif oral
  - Cas : femme ayant une thrombose recrutés dans un hôpital
  - Témoins : femme du même âge hospitalisées pour une autre pathologie (non associée)
  - Les résultats obtenus donne un $OR \approx 10$!!
  - Plusieurs études avaient déjà relevé ce résultat ⟶ les médecins étaient plus vigilants pour les patients exposés et les admettaient plus facilement à l'hôpital en cas de thrombose ou de signe suspect ⟶ surreprésentation des cas malades et exposés

Population cible

	E+	E-
M+	a	b
M-	c	d

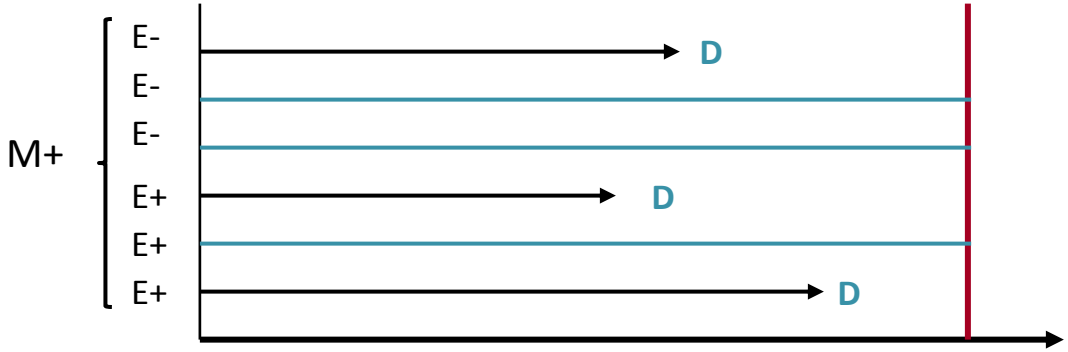
Population source

	E+	E-
M+	a'>>a	b'=b
M-	c'=c	d'=d

$$OR' = \frac{ad'}{bc'} > \frac{ad}{cb} = OR$$ ⟶ sur-estimation de l'OR

Biais de sélection et enquête cas-témoins

- Biais de survie sélective lié à la sélection de cas prévalents dans les enquêtes cas-témoins
  - Témoins : représentatifs de la population cible
  - Cas : sélection des patients qui sont toujours en vie ➡ individus les moins exposés



Début de l'étude : inclusion des cas vivants

Population cible

	E+	E-
M+	a	b
M-	c	d

Population source

	E+	E-
M+	a'<<a	b'<b
M-	c	d

$$OR' = \frac{a'd}{cb'} < \frac{ad}{cb} = OR$$

↳ sous-estimation de l'OR

Biais de sélection et enquête de cohorte

- Biais lié à sélection des exposés et non exposés dans les enquête rétrospectives
  - Ex : enquête de cohorte rétrospective sur une exposition à un polluant 15-20 ans avant dans une usine
  - Exposés et non exposés recrutés à partir des dossiers d'embauche
  - Les vieux dossiers ont tendance à être plus souvent perdus
  - Les dossiers des employés malades sont plus souvent conservés

→ On sélectionne en majorité les individus exposés qui ont développé la maladie

et moins les individus exposés qui n'ont pas développé la maladie

Population cible

	E+	E-
M+	a	b
M-	c	d

Population source

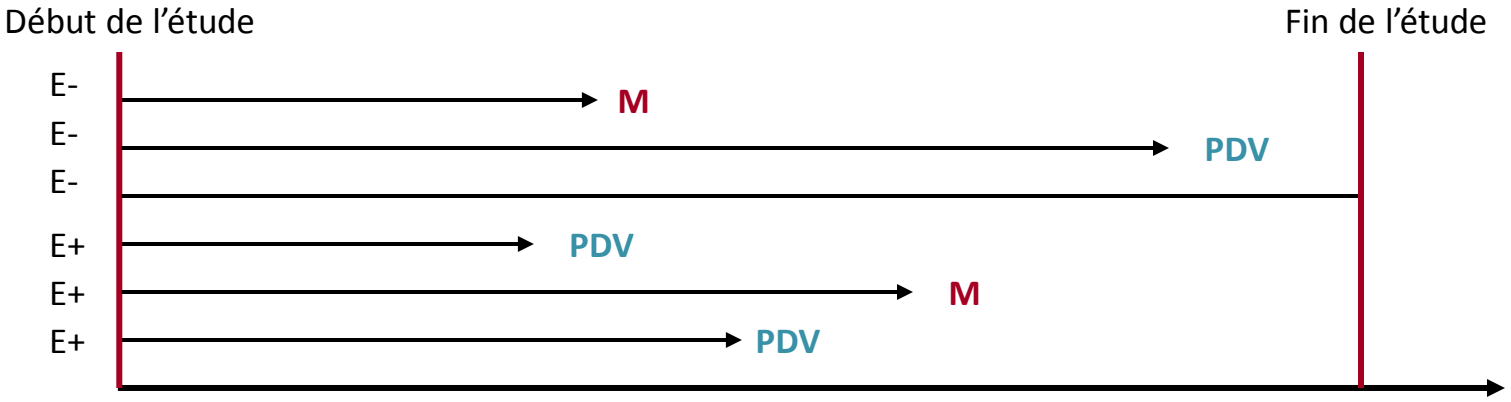
	E+	E-
M+	a'=a	b'<b
M-	c'<c	d'<d

$$OR' = \frac{a'd'}{c'b'} > \frac{ad}{cb} = OR$$

↳ sur-estimation de l'OR

Biais de sélection et enquête de cohorte

- Biais lié aux perdus de vus dans les enquêtes de cohorte (censure dépendante de M)
  - Ex : Les exposés sont plus souvent perdus de vus : les sujets exposés quittent l'entreprise pour se soustraire à l'exposition (particulièrement lors des premiers signes de la maladie)
  - Ex : Sida, asthme, effet du prozac, … → Biais d'attrition lié aux sorties d'études et interruption de traitement



Population cible

	E+	E-
M+	a	b
M-	c	d

Population source

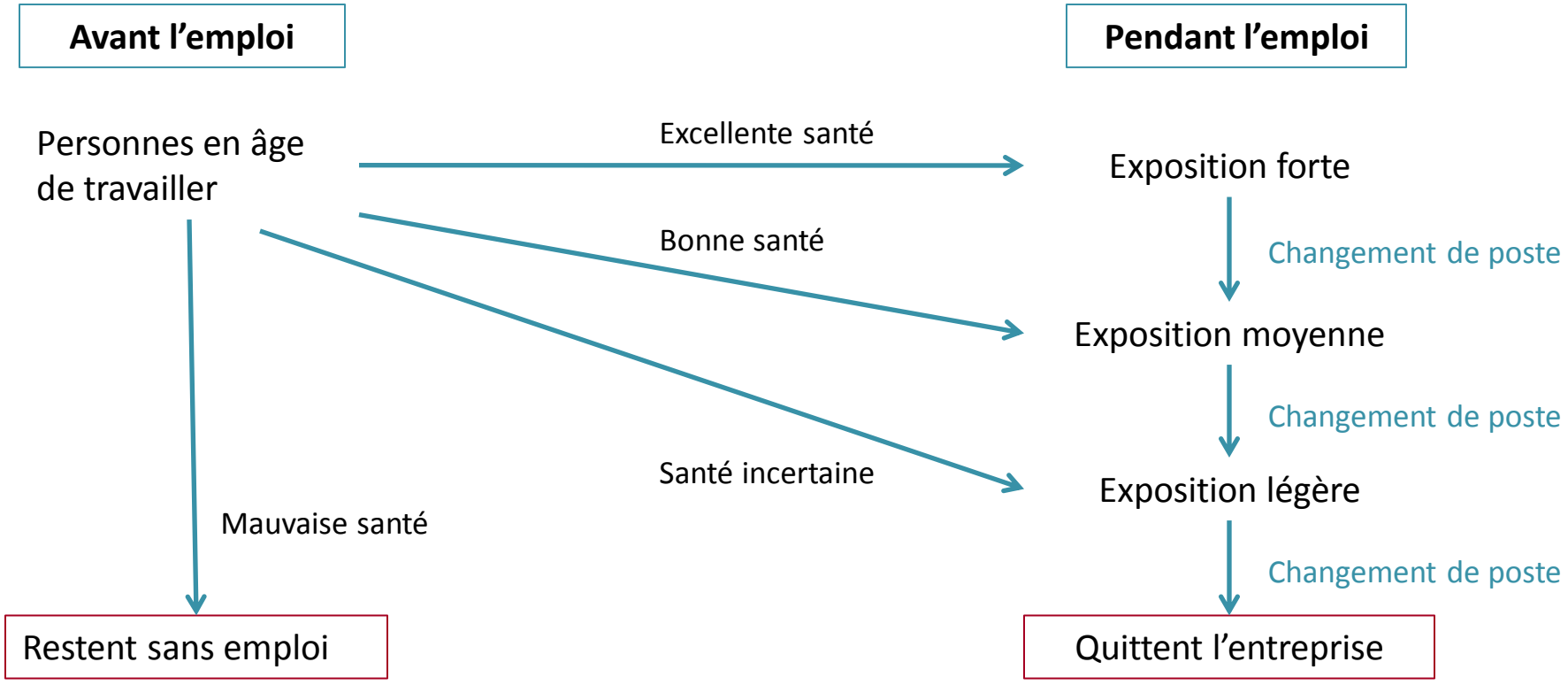
	E+ ↘↘	E- ↘
M+	a'<<<a	b'<b
M-	c'<<c	d'<d

$$OR' = \frac{a'd'}{c'b'} < \frac{ad}{cb} = OR$$

↳ sous-estimation de l'OR

Biais de sélection : "Healthy worker effect"

- Biais du travailleur en bonne santé dans les enquêtes de cohorte et transversales
- Les sujets les plus exposés sont en meilleure santé
- Biais moins important pour les maladies silencieuses



Il reste les personnes en bonne santé

Biais de sélection : "Healthy worker effect"

- Ex: mortalité de travailleurs fabriquant du PVC (E+)
  - Comparaison à la population générale avec le SMR (standardisation indirecte)
  - SMR<1 : mortalité plus faible que dans la population générale (E-)
  - SMR>1 : mortalité plus forte que dans la population générale

Cause de décès	SMR	
	Travailleurs actuels	Ex-travailleurs
Tous cancers	0.89	1.3
Cancer du poumon	0.5	1.56
Maladie cardiaques	0.63	1.11

Biais de sélection : "Healthy worker effect"

- Ex: enquête transversale pour évaluer l'effet de l'exposition à la peinture automobile en milieu professionnel et les pathologie pulmonaire

	E-	E+	Soustraits à l'exposition
Pathologie pulmonaire	5	5	9
Aucune pathologie	46	52	5

Individus ne faisant pas partie de l'échantillon

Population cible

	E+	E-
M+	a=14	b=5
M-	c=57	d=46

$OR = 2.26$

Population source

	E+ ↘	E-
M+	a'=5 <<a	b'=5
M-	c'=52 <c	d'=46

$OR = 0.88 < OR$

Il faut faire un suivi des sujets exposés

Enquête de cohorte

Refus de participation

- Les non réponses et le refus de participation à une enquête engendre un biais si les sujets qui refusent de participer sont différents des sujets qui participent.

- Ex: dans une étude sur l'alcoolisme, il y aura un biais de sélection si les sujets alcooliques de veulent pas participer à l'enquête

- La population de l'étude devient différente de la population cible

- Si la participation est trop faible, on cherche à recruter des participants (volontaires)
  - Si l'exposition ou la maladie sont liées à la participation ou non à l'enquête

→ Certaines catégories peuvent être surreprésentées

→ Biais d'auto sélection

III. Limiter les biais de sélection

- Au moment de la planification de l'enquête
  - Eviter au maximum la non participation des sujets éligibles
  - Assurer un bon suivi des sujets inclus dans une cohorte, recontacter les perdus de vue
  - Privilégier la sélection de cas incidents dans les enquêtes de cohorte
  - Rechercher des populations sources comportant des phénomènes de sélection comparables pour l'inclusion des cas et des témoins (ou des E+ et des E-)
    - Limite importante des enquêtes cas-témoins
    - Essayer de choisir plusieurs groupes témoins
    - Healthy worker effect : prendre des témoins d'une autre entreprise (sélection comparable)
- Au moment de l'analyse
  - Chercher à déterminer la direction et l'importance du biais
  - Comparer à des enquêtes indépendantes pour évaluer la reproduction des conclusions

Epidémiologie

7. Biais de classement

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

7. Biais de classement

I. Biais de classement

II. Biais de classement différentiel

- Exemples
- Situations générant des biais différentiels

III. Biais de classement non différentiel

- Exemples
- Situations générant des biais non différentiels

IV. Sensibilité, spécificité

V. Limiter les biais de classement

I. Biais de classement

- Le biais de classement ou biais d'information est une erreur systématique résultant d'une observation incorrecte d'un phénomène et conduisant à un mauvais classement
  - des sujets malades / non malades
  - des sujets exposés / non-exposés

Population de taille N
sans erreur de classement

	E+	E-
M+	a	b
M-	c	d

Même population
avec erreur de classement

	E+	E-
M+	a'	b'
M-	c'	d'

La population est identique dans les deux cas : $a + b + c + d = a' + b' + c' + d'$

Biais de classement

- Biais de classement non différentiel
  - Les erreurs de classement affectent avec la même probabilité les groupes comparés (E+/E- ou M+/M-) → les erreurs affectent identiquement les groupes
  - Erreurs de nature aléatoire et souvent dues à des imprécisions ou à une mauvaise qualité des instruments de mesure

→ Perte de puissance

- Biais de classement différentiel
  - Les erreurs de classement affectent différemment les groupes comparés
    - Erreurs de classement de la maladie sont différentes dans le groupe E+ et E-
    - Erreurs de classement de l'exposition sont différentes dans le groupe M+ et M-

→ Peut créer, renforcer ou diminuer une association

II. Biais de classement différentiel

- Biais de classement différentiel de l'exposition
  - Uniquement dans les enquêtes cas-témoins ou transversale

    ↳ dans les cohorte, le statut M+/M- est inconnu au moment de l'évaluation de l'exposition

  - Situation classique: sous évaluation de l'exposition chez les M-

    sur évaluation de l'exposition chez les M+

Population sans erreur de classement

	E+	E-
M+	9 000	7 000
M-	123 000	346 000

$OR = 3.62$

Même population avec erreur de classement

	E+	E-
M+	9 700 ←	→ 6 300
M-	98 400 →	← 370 600

$OR = 5.80$

20% des E+ sont classés E- chez les M-

10% des E- sont classés E+ chez les M+

Biais de classement différentiel

- Biais de classement différentiel de la maladie
  - Uniquement dans les enquêtes de cohorte ou transversale

    ↳ dans les cas-témoins, le statut E+/E- est inconnu au moment de l'évaluation de la maladie

  - Situation classique: sous diagnostic de la maladie chez les E-

    sur diagnostic de maladie chez les E+

Population sans erreur de classement

	E+	E-
M+	1 000	500
M-	9 000	12 000

$OR = 2.67$

Même population avec erreur de classement

	E+	E-
M+	1 900	400
M-	8 100	12 100

$OR = 7.10$

20% des M+ sont classés M- chez les E-

10% des M- sont classés M+ chez les E+

Biais de classement différentiel : situations

- Situations susceptibles de générer des biais de classement différentiel
  - Biais de subjectivité (ou biais de suspicion) apparaît quand
    - La connaissance de l'exposition peut amener à approfondir la recherche de la maladie auprès des sujets E+ et moins auprès des sujets E-
    - La connaissance du diagnostic de la maladie peut amener à approfondir la recherche de l'exposition auprès des sujets M+ et moins auprès des sujets M-
  - Biais d'enquêteur si plusieurs enquêteurs (Ex: un enquêteur pour les M+ et un pour les M-)
  - Biais de mémoire (ou biais de rappel) dû au fait qu'un sujet malade
    - se souvient davantage de ces expositions passées.
    - surestime ou sous-estime (volontairement ou déni) son exposition (Ex: alcool et cirrhose)
  - Biais de non-réponses lié au droit des participants à ne pas répondre

    Ex: Patients atteints d'une cirrhose (M+) ne veulent pas répondre sur leur consommation d'alcool (E+)
  - Biais de suivi lié à des différences de prise en charge entre les E+ et les E-

III. Biais de classement non différentiel

- Biais de classement non différentiel quand les erreurs de classement affectent identiquement les groupes comparés : E+/E- ou M+/M-

- Situations susceptibles de générer des biais de classement non différentiel
  - Imprécisions ou erreurs matérielles
  - Mauvaise qualité des instruments de mesures (dossier, questionnaire, évaluation d'experts)
    - Erreur de diagnostic de la maladie à l'aveugle (de l'exposition)
    - Erreur d'évaluation de l'exposition à l'aveugle (de la maladie)

- Un biais de classement différentiel entraine une perte de puissance

  → L'estimation de l'OR ou du RR est plus proche de la valeur 1

Biais de classement non différentiel

- Ex : biais de classement non différentiel sur l'exposition

  ↳ Expositions professionnelles évalués par des experts à l'aveugle (de la maladie)

	E+	E-
M+	70	30
M-	50	50

$OR = 2.33$

20% des E+ sont classés E-

	E+	E-
M+	56	44
M-	40	60

$\widetilde{OR} = 1.91$

- Ex : biais de classement non différentiel sur la maladie

  ↳ Evaluation de la maladie à partir des certificats de décès

	E+	E-
M+	70	30
M-	50	50

$OR = 2.33$

20% des M+ sont classés M-

	E+	E-
M+	56	24
M-	64	56

$\widetilde{OR} = 2.04$

- En pratique, les erreurs sont souvent commises dans les deux sens

IV. Sensibilité, spécificité

- La qualité d'une méthode de classement peut être mesurée par sa sensibilité et la spécificité
- Evaluation du classement de l'exposition (idem maladie) en connaissant la méthode de référence
- Sensibilité (Se) = proportion de sujet classés $\widetilde{\mathrm{E}}+$ parmi les sujets $\mathrm{E}+$
- Spécificité (Sp) = proportion de sujet $\widetilde{\mathrm{E}}-$ correctement classés $\mathrm{E}-$

		Nouvelle méthode		
		$\widetilde{\mathrm{E}}+$	$\widetilde{\mathrm{E}}-$	
Méthode de référence	$\mathrm{E}+$	a	b	a+b
	$\mathrm{E}-$	c	d	c+d

$$Se = \frac{\#(\widetilde{\mathrm{E}}+ \cap \mathrm{E}+)}{\#\mathrm{E}+} = \frac{a}{a+b}$$

$$Sp = \frac{\#(\widetilde{\mathrm{E}}- \cap \mathrm{E}-)}{\#\mathrm{E}-} = \frac{d}{c+d}$$

b est le nombre de faux négatif

c est le nombre de faux positif

- Absence d'erreur de classement $\Leftrightarrow \quad Se = Sp = 1$

Sensibilité, spécificité

- Si la maladie est rare (alors $RR \approx OR$) on montre que

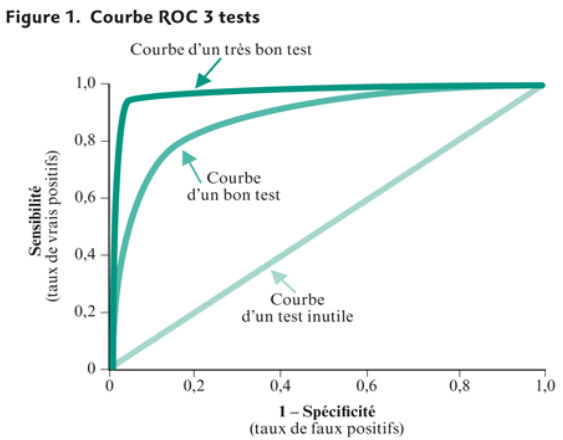
$$\widetilde{OR} = \frac{[Se \times OR \times P_E + (1-Sp)(1-P_E)] \times [(1-Se)P_E + Sp(1-P_E)]}{[Se \times P_E + (1-Sp)(1-P_E)] \times [(1-Se) \times OR \times P_E + Sp(1-P_E)]}$$

  - $\widetilde{OR}$ l'odds ratio obtenu avec la nouvelle méthode de classement
  - OR l'odds ratio obtenu avec la méthode de référence
  - P_E la probabilité d'être exposé

- Si le marqueur de l'exposition $\widetilde{E} +$ est quantitatif (Ex: volume expiratoire maximum par seconde) et si le diagnostic de l'exposition est fondé sur le fait que $\widetilde{E} +$ dépasse un certain seuil L

→ Courbe ROC (Receiver Operating Characteristics)

Pour comparer les méthodes de classement on trace (pour chaque méthode) la courbe Se en fonction de $1 - Sp$ pour plusieurs valeurs de L



V. Limiter les biais de classement

- Au moment de la planification de l'enquête
  - Limiter les erreurs de classement
    - Vérification du matériel de mesure
    - Evaluation de l'exposition et de l'état de santé par des examens objectifs et reproductibles (définition précises des critères de jugement)
    - Questionnaires standardisés, validés et testés
  - Eviter que les erreurs soient différentielles
    - Choisir des groupes (M+/M- ou E+/E-) dont la coopérativité, la mémorisation, ou la surveillance épidémiologique sont a priori comparables
    - Evaluer l'exposition (ou l'état de santé) à l'aveugle su statut de la maladie (de l'exposition)
    - Standardisation des mesures et condition d'interview identiques dans les groupes (M+/M- ou E+/E-)

Limiter les biais de classement

- Pendant l'enquête
  - Insister, informer et expliquer pour éviter les non-réponses
  - Vérifier le matériel de mesure
- Au moment de l'analyse de l'enquête
  - Décrire les caractéristiques des non répondants et les comparer aux autres
  - Discuter l'ampleur et la direction des biais de classements éventuels

Epidémiologie

8. Biais et facteurs de confusion

Philippe SAINT PIERRE

Université Paul Sabatier – Toulouse III

Institut de Mathématiques de Toulouse

philippe.saint-pierre@math.univ-toulouse.fr

8. Biais et facteurs de confusion

I. Biais de confusion

II. Phénomène d'interaction : Test du Chi-2 d'interaction

III. Méthode d'ajustement de Mantel-Haenszel

- Mesure de risque ajusté sur F
- Test de la mesure de risque ajusté

IV. Définition d'un facteur de confusion

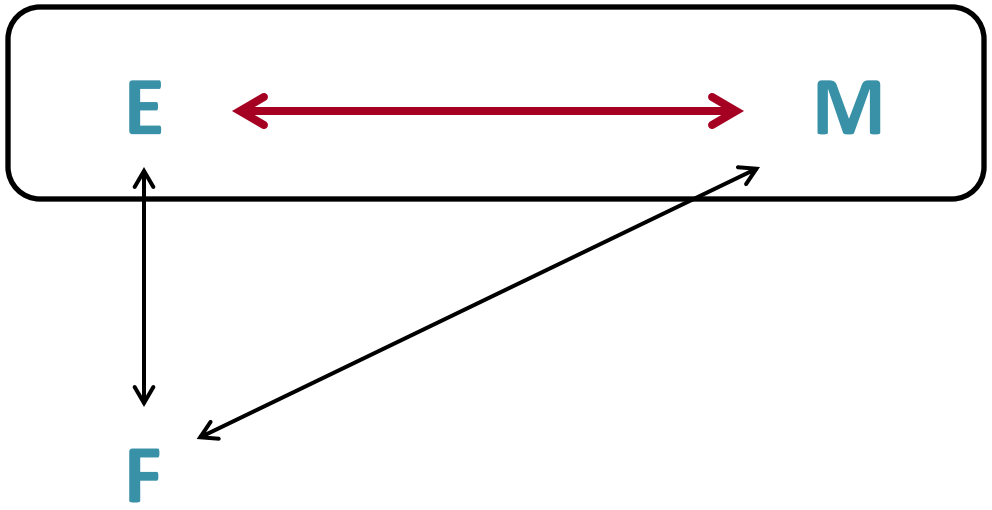
- Facteur de confusion potentiel
- Facteur de confusion : définition

V. Prise en compte d'un facteur de confusion

- Au moment de la planification de l'enquête
- Au moment de l'analyse statistique
- Difficultés pratiques

I. Biais de confusion

- On appelle biais de confusion le fait que l'effet du facteur étudié (E) sur la maladie (M) est en partie mélangé avec d'autres facteurs (F)

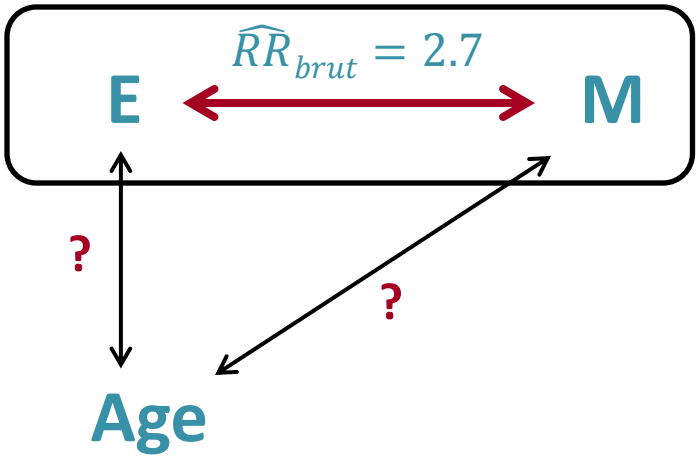


Exemple 1

Exposition

Maladie	E+	E-	
M+	184	680	764
M-	816	9320	10136
	1000	10000	

$$\widehat{RR} = \frac{184/1000}{680/10000} = 2.7$$



Exemple 1

Relation Age - Exposition

Age	Exposition E+	Exposition E-
<25 ans	200	5000
25-40 ans	300	3000
>40 ans	500	2000
	1000	10000

$P_{E+} = 0.038$

$P_{E+} = 0.091$

$P_{E+} = 0.2$

→ L'exposition est d'autant plus fréquente que les sujets sont âgés

Relation Age – Maladie

Age	Maladie M+	Maladie M-
<25 ans	216	4984
25-40 ans	248	3012
>40 ans	360	2140
	1000	10000

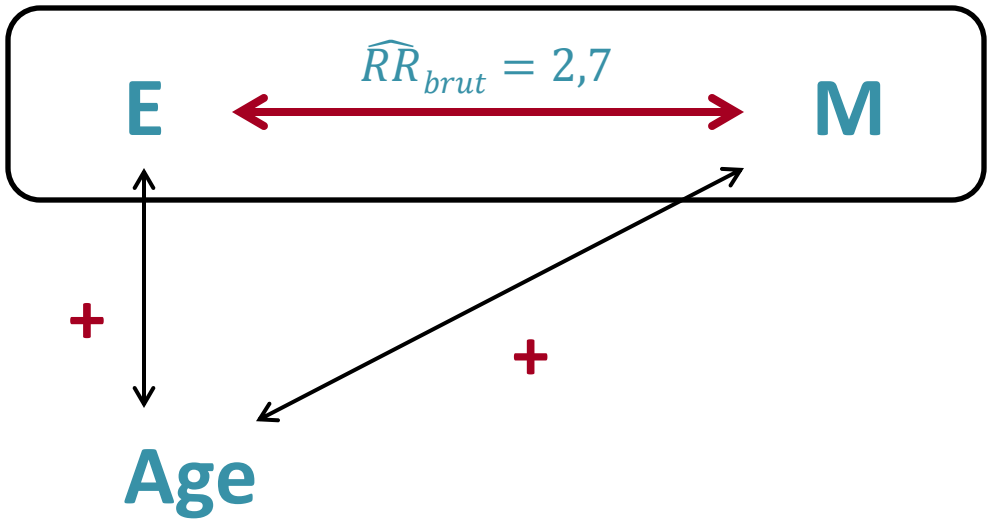
$P_{M+} = 0.042$

$P_{M+} = 0.076$

$P_{M+} = 0.144$

→ La maladie est d'autant plus fréquente que les sujets sont âgés

Exemple 1



- Les sujets exposés sont plus âgés et les sujets âgés sont plus malades

↳ effet de l'exposition sur la maladie perturbé par l'effet de l'âge

→ L'âge (**F**) est un **facteur de confusion potentiel** pour la relation entre **E** et **M**

Exemple 1

Exposition

Maladie	E+	E-	
M+	184	680	764
M-	816	9320	10136
	1000	10000	

$$\widehat{RR}_{brut} = \frac{184/1000}{680/10000} = 2.7$$

< 25 ans

	E+	E-
M+	16	200
M-	184	4800
	200	5000

$\widehat{RR}_1 = 2$

25- 45 ans

	E+	E-
M+	48	240
M-	252	2760
	300	3000

$\widehat{RR}_2 = 2$

> 40 ans

	E+	E-
M+	120	240
M-	380	2760
	200	5000

$\widehat{RR}_3 = 2$

$\widehat{RR}_1 = \widehat{RR}_2 = \widehat{RR}_3 \neq \widehat{RR}_{brut}$ ⟹

L'âge (F) est un **facteur de confusion** pour la relation entre E et M

Exemple 2

Exposition : Tabac

Maladie	Brun : E+	Blond : E-
M+	353	54
M-	253	73

$$\widehat{OR}_{brut} = \frac{353 \times 72}{253 \times 54} = 1.89$$

Sujets inhalant la fumée

	Brun	Blond
M+	267	32
M-	134	39

$$\widehat{OR}_1 = 2.43$$

Sujets n'inhalant pas la fumée

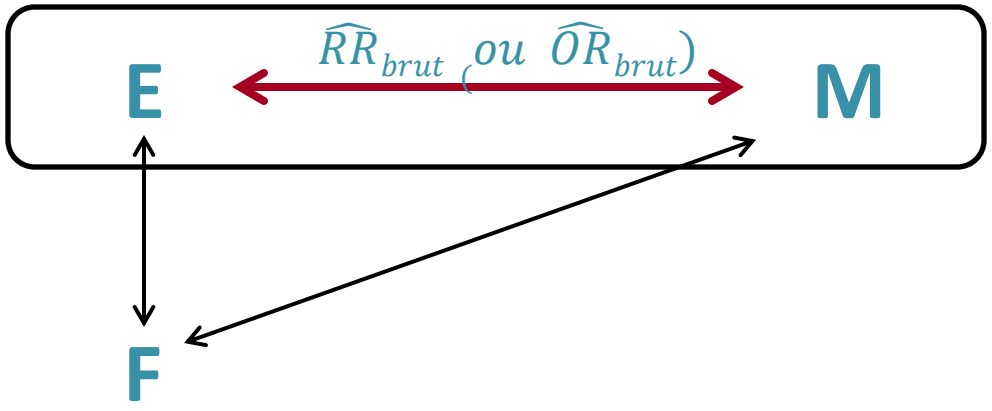
	Brun	Blond
M+	86	22
M-	119	34

$$\widehat{OR}_2 = 1.07$$

$$\widehat{OR}_1 \neq \widehat{OR}_2$$

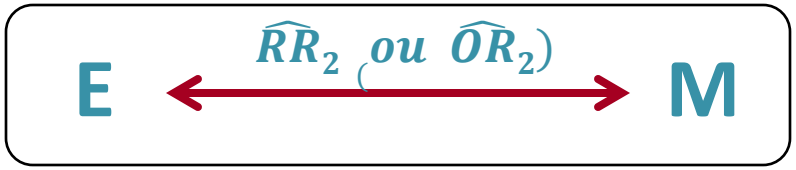
Il y a **interaction** entre **E** et **F** pour la relation avec **M**

II. Phénomène d'interaction



Strate 1 du facteur **F**

Strate 2 du facteur **F**



F est un **facteur d'interaction** avec **E** vis-à-vis de **M** si

$$\widehat{RR}_1 \neq \widehat{RR}_2 \ (ou\ \widehat{OR}_1 \neq \widehat{OR}_2)$$

Test du Chi-2 d'interaction

- Soit un facteur **F** à k classes, pour tout i $= 1, \dots, k$

Strate i du facteur F

	E+	E-	
M+	a_i	b_i	m_{1i}
M-	c_i	d_i	m_{0i}
	n_{1i}	n_{0i}	n_i

$$\widehat{RR}_i = \frac{a_i/n_{1i}}{b_i/n_{0i}}, \qquad \widehat{Var}(Ln(\widehat{RR}_i)) = \frac{c_i}{a_i \times n_{1i}} + \frac{d_i}{b_i \times n_{0i}}$$

$$\widehat{OR}_i = \frac{a_i \times d_i}{c_i \times b_i}, \qquad \widehat{Var}(Ln(\widehat{OR}_i)) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

Test d'égalité des RR_i

H_0: $RR_1 = \cdots = RR_k$

H_1: $\exists$ *au moins une différence*

Test d'égalité des OR_i

H_0: $OR_1 = \cdots = OR_k$

H_1: $\exists$ *au moins une différence*

Test du Chi-2 d'interaction

Test d'égalité des RR_i

$$\begin{cases} H_0 : \ RR_1 = \cdots = RR_k \\ H_1 : \exists \ au \ moins \ une \ différence \end{cases}$$

Test d'égalité des OR_i

$$\begin{cases} H_0 : \ OR_1 = \cdots = OR_k \\ H_1 : \exists \ au \ moins \ une \ différence \end{cases}$$

$$X_I = \sum_{i=1}^{k} \omega_i \ (Y_i - \bar{Y})^2 = \sum_{i=1}^{k} \omega_i {Y_i}^2 - \frac{\left(\sum_{i=1}^{k} \omega_i Y_i\right)^2}{\sum_{i=1}^{k} \omega_i}$$

avec $\boldsymbol{X_I \equiv \chi^2(k-1)}$ et $\bar{Y} = \frac{\sum_{i=1}^{k} \omega_i Y_i}{\sum_{i=1}^{k} \omega_i}$

$$\boldsymbol{Y_i = Ln(\widehat{RR}_i)}$$

$$\boldsymbol{\omega_i = \frac{1}{\widehat{Var}(Ln(\widehat{RR}_i))} = \frac{1}{\frac{c_i}{a_i \times n_{1i}} + \frac{d_i}{b_i \times n_{0i}}}}$$

$$\boldsymbol{Y_i = Ln(\widehat{OR}_i)}$$

$$\boldsymbol{\omega_i = \frac{1}{\widehat{Var}(Ln(\widehat{OR}_i))} = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}}$$

Exemple 2

Exposition : Tabac

Maladie	Brun : E+	Blond : E-
M+	353	54
M-	253	73

$$\widehat{OR}_{brut} = \frac{353 \times 72}{253 \times 54} = \mathbf{1.89}$$

Sujets inhalant la fumée

	Brun	Blond
M+	267	32
M-	134	39

$\widehat{OR}_1 = 2.43$ $\quad Ln(\widehat{OR}_1) = 0.887$

$$\omega_i = \frac{1}{\widehat{Var}(Ln(\widehat{OR}_1))} = 14.684$$

Sujets n'inhalant pas la fumée

	Brun	Blond
M+	86	22
M-	119	34

$\widehat{OR}_2 = 1.07$ $\quad Ln(\widehat{OR}_2) = 0.066$

$$\omega_i = \frac{1}{\widehat{Var}(Ln(\widehat{OR}_2))} = 10.762$$

$X_I = 4.19 > \chi^2{}_\alpha(1) = 3.84$

On rejette l'hypothèse H_0: $\widehat{OR}_1 = \widehat{OR}_2$

Il y a **interaction** entre **E** et **F** pour la relation avec **M**

Test du Chi-2 d'interaction

- Le test d'interaction est peu puissant, il rejette difficilement H_0
- Interprétation du test du Chi-2 d'interaction
  - On rejette H_0: $OR_1 = \cdots = OR_k$

    ➡ Le facteur **F** est un facteur d'interaction avec l'exposition **E** pour la relation avec le risque de maladie **M**

    ➡ Analyse séparée des associations entre **E** et **M** dans chaque strate du facteur de confusion **F**

  - On ne rejette pas H_0: $OR_1 = \cdots = OR_k$

    ➡ Il n'y a pas interaction

    ➡ Méthode d'ajustement de Mantel-Haenszel

III. Méthode d'ajustement de Mantel-Haenszel

- Uniquement après avoir éliminé un phénomène d'interaction
- Démarche générale

1. Estimer une valeur du RR (ou OR) entre **E** et **M** ajusté sur **F**

   ⟹ RR ajusté de Mantel-Haenszel

2. Tester si l'association entre **E** et **M** reste significative après ajustement sur **F**

   ⟹ Test du Chi-2 de Mantel-Haenszel

3. **F** est-il un facteur de confusion ?

   ⟹ Evaluation qualitative de la différence entre RR et RR ajusté
   (pas de test statistique)

RR ajusté de Mantel -Haenszel

- Soit un facteur **F** à k classes, pour tout i $= 1, \dots, k$

Strate i du facteur F

	E+	E-	
M+	a_i	b_i	m_{1i}
M-	c_i	d_i	m_{0i}
	n_{1i}	n_{0i}	n_i

$$\widehat{RR}_i = \frac{a_i/n_{1i}}{b_i/n_{0i}}, \qquad \widehat{Var}(Ln(\widehat{RR}_i)) = \frac{c_i}{a_i \times n_{1i}} + \frac{d_i}{b_i \times n_{0i}}$$

$$\widehat{OR}_i = \frac{a_i \times d_i}{c_i \times b_i}, \qquad \widehat{Var}(Ln(\widehat{OR}_i)) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

$$\omega_i = \frac{1}{\widehat{Var}(Ln(\widehat{RR}_i))} = \frac{1}{\frac{c_i}{a_i \times n_{1i}} + \frac{d_i}{b_i \times n_{0i}}}$$

$$Ln(\widehat{RR}_{MH}) = \frac{\sum_{i=1}^{k} \omega_i \, Ln(\widehat{RR}_i)}{\sum_{i=1}^{k} \omega_i}$$

$$IC(\alpha): \; Ln(\widehat{RR}_{MH}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{\sum_{i=1}^{k} \omega_i}}$$

$$\omega_i = \frac{1}{\widehat{Var}(Ln(\widehat{OR}_i))} = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}$$

$$Ln(\widehat{OR}_{MH}) = \frac{\sum_{i=1}^{k} \omega_i \, Ln(\widehat{OR}_i)}{\sum_{i=1}^{k} \omega_i}$$

$$IC(\alpha): \; Ln(\widehat{OR}_{MH}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{\sum_{i=1}^{k} \omega_i}}$$

Test du Chi-2 de Mantel -Haenszel

- Test de l'absence d'association : $\begin{cases} H_0 : \ RR_{MH} = 1 \text{ ou } OR_{MH} = 1 \\ H_1 : \ RR_{MH} \neq 1 \text{ ou } OR_{MH} \neq 1 \end{cases}$

Strate i du facteur F

	E+	E-	
M+	a_i	b_i	m_{1i}
M-	c_i	d_i	m_{0i}
	n_{1i}	n_{0i}	n_i

$$X_{MH} = \frac{\left(\sum_{i=1}^{k}[a_i - E(A_i)]\right)^2}{\sum_{i=1}^{k} V(A_i)} \quad \text{avec} \quad X_{MH} \equiv \chi^2(1)$$

$$E(A_i) = \frac{n_{1i} \times m_{1i}}{n_i} \quad et \quad V(A_i) = \frac{n_{0i} \times n_{1i} \times m_{0i} \times m_{1i}}{{n_i}^2 (n_i - 1)}$$

Exemple 3

Prématurité	Activité professionnelle: E+ (jamais)	Activité professionnelle: E- (oui)
M+	21	37
M-	256	868

$\widehat{RR}_{brut} = 1.9$

$IC(\alpha)$: $[1.1 \ -3.1]$

$X(1) = 5.55 \quad p = 0.019$

< 25 ans

	E+	E-
M+	13	13
M-	108	174

$\widehat{RR}_1 = 1.6 \ [0.7 \ -3.2]$

≥ 25 ans

	E+	E-
M+	8	24
M-	148	694

$\widehat{RR}_2 = 1.5 \ [0.7 \ -3.4]$

1. Test d'interaction : l'âge n'est pas un facteur d'interaction : $X_I(1) < 0.0001$
2. Estimation de $\widehat{RR}_{MH}$ ajusté sur l'âge $\widehat{RR}_{MH} = 1.5 \ [0.9 \ -2.6]$
3. Test de Mantel-Haenszel pas significatif : $X_{MH}(1) = 2.52 \quad p = 0.15$

- $\widehat{RR}_{brut} \neq \widehat{RR}_{MH}$ ⇒ **l'âge est un facteur de confusion**
- Après ajustement sur l'âge, l'activité professionnelle n'est plus associé significativement au risque de prématurité

Démarche générale (synthèse)

F facteur à k classes, pour tout i $= 1, \ldots, k$

Test d'interaction ($\boldsymbol{X_I} \equiv \boldsymbol{\chi^2}$ à $k-1$ ddl) : H_0: $RR_1 = \cdots = RR_k$

Rejet de H_0 : RR_i différents

- **F** est en interaction avec **E** vis-à-vis de **M**
- Analyse séparée des RR_i

Non rejet de H_0 : " RR_i égaux "

- Mesure ajustée : $\widehat{RR}_{MH}$ $(ou\ \widehat{OR}_{MH})$
- Test du Chi-2 de Mantel-Haenszel

($\boldsymbol{X_{MH}} \equiv \boldsymbol{\chi^2}$ à 1 ddl), $H_0: RR_{MH} = 1$

$\widehat{RR}_{MH} \approx \widehat{RR}_{brut}$

F pas facteur de confusion

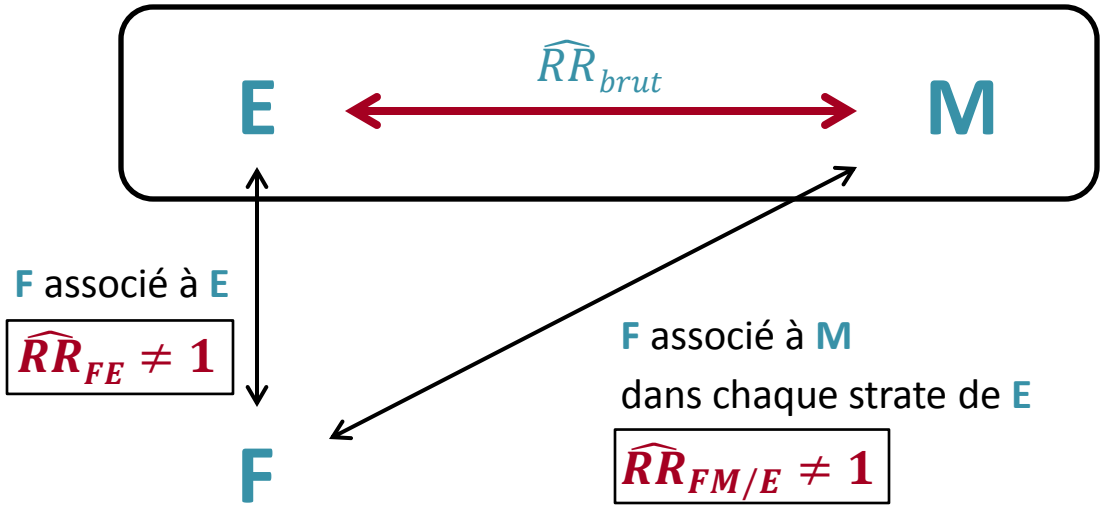
$\widehat{RR}_{MH} \neq \widehat{RR}_{brut}$

F facteur de confusion

IV. Facteur de confusion : définition

- Non rejet de H_0: $RR_1 = \cdots = RR_k$ ⟹ Il n'y a pas d'interaction
  - $RR_1 = \cdots = RR_k = RR_{MH} \approx RR_{brut}$ ⟹ F n'est pas un facteur de confusion
  - $RR_1 = \cdots = RR_k = \boldsymbol{RR_{MH} \neq RR_{brut}}$ ⟹ F est un **facteur de confusion** pour la relation entre E et M

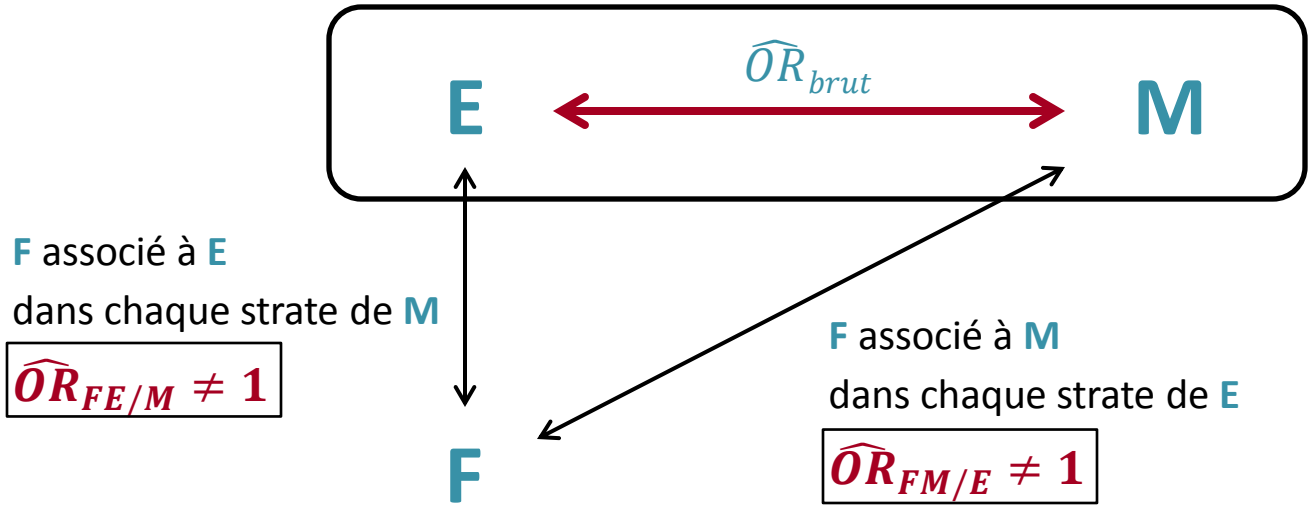
- F est un facteur de confusion pour la relation entre E et M



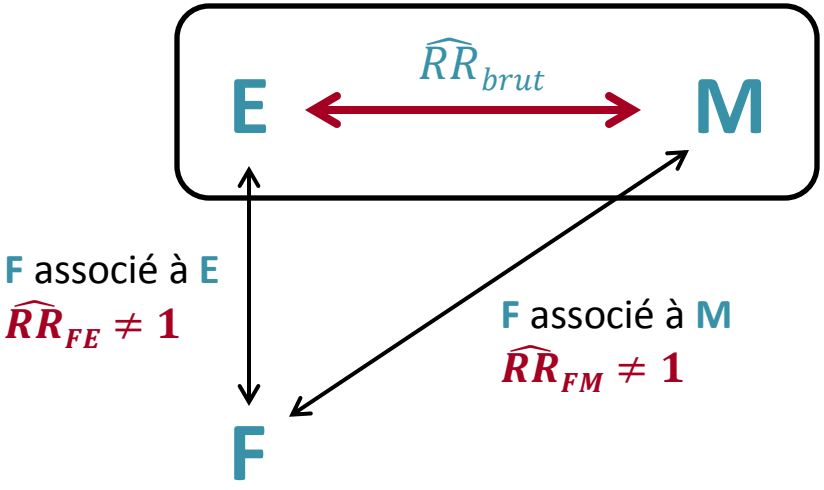
Facteur de confusion : définition

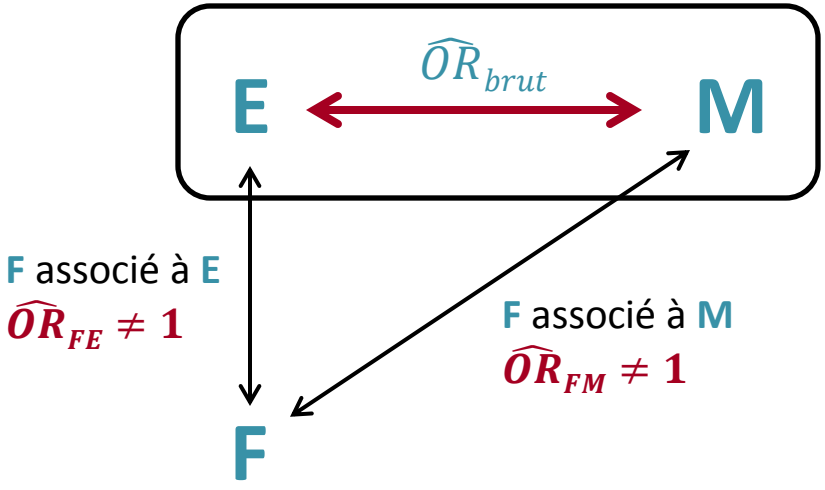
- Non rejet de H_0: $OR_1 = \cdots = OR_k$ ⟹ Il n'y a pas d'interaction
  - $OR_1 = \cdots = OR_k = OR_{MH} \approx OR_{brut}$ ⟹ F n'est pas un facteur de confusion
  - $OR_1 = \cdots = OR_k = \boldsymbol{OR_{MH} \neq OR_{brut}}$ ⟹ F est un **facteur de confusion** pour la relation entre E et M

- F est un facteur de confusion pour la relation entre E et M



Facteur de confusion potentiel





F est un **facteur de confusion potentiel** pour la relation entre E et M

- En pratique, on ne veut pas rater de facteur de confusion. Les facteurs de confusion potentiels sont donc répertoriés et supprimés par la suite si le rôle de confusion est mineur.

V. Prise en compte d'un facteur de confusion

1. **Au moment de la planification de l'enquête**
   - Relever les facteurs de risque connus du phénomène étudié **M** (littérature)
   - Relever les facteurs de risque potentiels (plausibilité, intuitions)
   - Relever les facteurs d'interaction potentiels
2. **Au moment de la conception de l'enquête**
   - Recueillir les informations sur les facteurs de confusion et d'interaction
   - Définir les modalités d'échantillonnage
     a) Randomiser l'exposition
     b) Restreindre la population d'étude à une catégorie du facteur de confusion
     c) Stratifier ou apparier sur un ou plusieurs facteurs de confusion
3. Au moment de l'analyse statistique
4. Les difficultés pratiques

Au moment de la conception de l'enquête

a) Randomiser l'exposition

- Consiste à répartir au hasard les sujets qui recevront l'exposition
- Les facteurs de confusion potentiels ont en moyenne la même distribution dans les groupes exposés et non exposés
- Les facteurs de confusion potentiels ne sont pas associés à l'exposition

➡ Interprétation causale

- Possible en situation expérimentale : traitements ou intervention
- Impossible en situation d'observation : tabagisme, expositions professionnelles, précarité sociale

Au moment de la conception de l'enquête

b) Restreindre la population d'étude

- Effectuer l'étude dans **une catégorie** du facteur de confusion **F**

  ↳ Femme, enfants, sportifs, …

- **Exclure** certains sujet appartenant à une **catégorie rare** du facteur de confusion **F**

  ↳ Dans l'étude des facteurs de risque du cancer de la vessie, exclusion des patients atteints de bilharziose (rare en Europe et facteur de risque connu du cancer de la vessie)

→ Limite la portée de l'étude à un sous groupe de **F**

Au moment de la conception de l'enquête

c) **Stratifier ou apparier** sur certains facteurs de confusion

- **Appariement** (Ex : âge, sexe, CSP, zone de résidence, …)
  - Pour chaque cas, on sélectionne un (ou plusieurs témoins) du même âge, sexe, CSP
  - Pour chaque $E+$, on sélectionne un $E-$ du même âge, sexe, CSP

- **Stratification, appariement par classe** (Ex : pays, régions, hôpital de recrutement, …)
  - Cas et témoins sélectionnés au sein de chaque pays, région, hôpital
  - $E+$ et $E-$ sélectionnés au sein de chaque pays, région, hôpital

- **Objectif** : Equilibrer la répartition du facteur de confusion **F** dans les groupe comparés

  ↳ **F** n'est pas lié à **M** (ou **E**) dans une enquête cas-témoins (ou cohorte)

- **Difficultés** : – Difficile de trouver un témoin quand appariement sur plusieurs facteurs

  – Le lien entre **F** et **M** (ou **E**) ne peut pas être étudié

Au moment de l'analyse statistique

1. Au moment de la planification de l'enquête
2. Au moment de la conception de l'enquête
3. **Au moment de l'analyse statistique**
   - Analyse univariée, étudier l'association entre **M**, **E** et différents facteurs **F**
   - Analyse stratifiée (méthode de Mantel-Haenszel) ⇒ analyse préliminaire
     - **M** et **E** doivent être binaire (2 classes)
     - Le facteur d'ajustement **F** doit être qualitatif (k classes)
     - A cause de la stratification, analyse de peu de facteurs de confusion **F** en même temps
   - Analyse multivariée
     - Possibilité de prendre plusieurs facteurs de confusion simultanément
     - Possibilité d'utiliser des variables qualitatives et quantitatives
     - Régression linéaire, régression logistique, modèle de survie (Cox, …)
4. Les difficultés pratiques

Les difficultés pratiques

1. Au moment de la planification de l'enquête
2. Au moment de la conception de l'enquête
3. Au moment de l'analyse statistique
4. **Les difficultés pratiques**

   a) Sur-appariement

   b) Facteur de confusion déséquilibré dans les groupes de **F**

   c) Sur-ajustement : Variable prise en compte à tort dans le modèle

   d) Valeurs manquantes

   e) Erreurs de classement sur le facteur de confusion **F**

   f) Facteur intermédiaire

Les difficultés pratiques

a) Sur-appariement

- L'appariement sur des facteurs de confusion entraîne sans le savoir un appariement sur l'exposition

↳ Peut faire disparaître artificiellement une association

- Exemple: choisir les témoins parmi les amis ou la famille des cas (cancer du seins et pilule)

↳ Cas et témoins peuvent se ressembler pour les facteurs de confusion (niveau social, mode de vie) mais aussi pour l'exposition (contraception)

Les difficultés pratiques

b) Facteur de confusion déséquilibré dans les groupes de **F**

< 30 ans

	E+	E-
M+	50	2
M-	150	15
	150	17

≥ 30 ans

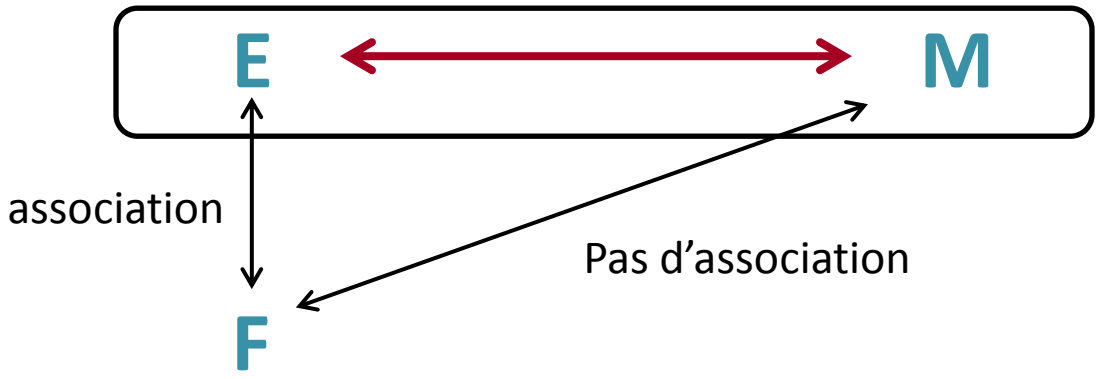
	E+	E-
M+	0	28
M-	4	201
	4	**229**

L'exposition est confondue avec l'âge

↳ Ajustement sur l'âge est impossible

Les difficultés pratiques

c) Sur-ajustement



Ajustement sur **F** ➡ Perte de puissance

Les difficultés pratiques

- Exemple sur-ajustement : ajustement sur l'âge

< 30 ans

	E+	E-
M+	9	1
M-	36	9
	45	10

$\widehat{OR}_1 = 2.25$

≥ 30 ans

	E+	E-
M+	1	4
M-	4	36
	4	**40**

$\widehat{OR}_2 = 2.25$

- Age est très lié à l'exposition ($E+$ sont jeunes et $E-$ sont vieux)
- $\widehat{OR}_{MH} = \widehat{OR}_{brut} = 2.25$ ➡ l'âge n'est pas un facteur de confusion
- $Var\left(Ln(\widehat{OR}_{brut})\right) = 0.347 < 0.886 = Var\left(Ln(\widehat{OR}_{MH})\right)$

➡ Ajustement sur l'âge entraine une **perte de puissance**

$\widehat{OR}_{brut} = 2.25 \quad [0.7 - 7.1] \qquad \widehat{OR}_{MH} = 2.25 \quad [0.4 - 11.4]$

Les difficultés pratiques

d) Données manquantes pour le facteur de confusion

Activité professionnelle

Prématurité	E+ (jamais)	E- (oui)
M+	21	37
M-	256	868
n = 1182	277	905

$\widehat{RR}_{brut} = 1.9\ [1.1\ -3.1]$

na

Activité professionnelle

Prématurité	E+ (jamais)	E- (oui)
M+	15	35
M-	238	777
n = 1065	2534	812

$\widehat{RR}_{brut} = 1.4\ [0.8\ -2.5]$

La salubrité est un facteur de confusion?

Logement salubre

	E+	E-
M+	11	21
M-	189	559
n_1 = 780	200	580

$\widehat{RR}_1 = 1.5\ [0.7\ -3.1]$

Logement insalubre

	E+	E-
M+	4	14
M-	39	218
n_2 = 285	53	232

$\widehat{RR}_2 = 1.3\ [0.4\ -3.6]$

Chi-2 d'interaction ($X_I = 0.08$), pas d'interaction

$\widehat{RR}_{MH} = 1.4\ [0.8\ -2.6]$

→ Biais possible

Manque de puissance

Les difficultés pratiques

e) Erreur de classement sur le facteur de confusion

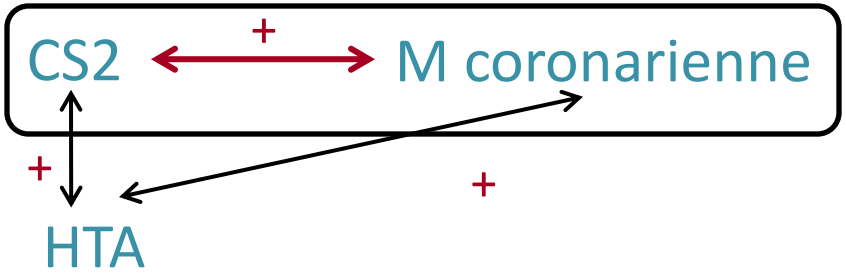
Biais de classement non différentiel ➡ manque de puissance

Biais de classement non différentiel ➡ peut créer une association

Les difficultés pratiques

f) Facteur intermédiaire

CS2 ⟷ M coronarienne



$\widehat{OR}_{brut} \neq 1$

$\widehat{OR}_{MH} = 1$ ➡ HTA est un facteur de confusion !!

CS2 ⟶ HTA ⟶ M coronarienne

L'effet de CS2 s'explique par la survenue d'une HTA chez les CS2+

➡ **HTA est un facteur intermédiaire**

Mettre en évidence le phénomène

conclure avec des données externes (biologiques, mécaniques, …)