POLYCOPIE

BIO-MEDECINE QUANTITATIVE
LECTURE CRITIQUE D'ARTICLES

DFGSM3 & DFASM

SUPPORT DE COURS POUR L'i-ECN

DFGSMa3

2017-2018

UFR Médecine Montpellier-Nîmes

Préambule

L'objectif de l'enseignement de Biomédecine quantitative (DFGSM3 et DFGSMa3) est de **montrer l'utilité des concepts statistiques et épidémiologiques pour la lecture critique d'articles** et, par voie de conséquence, **pour l'utilisation des connaissances en pratique clinique quotidienne.**

Cette UE **est la 1ère étape dans la préparation à l'épreuve de Lecture Critique d'Articles (LCA) de l'i-ECN** qui se déroule sur quatre ans. Cette préparation progressive a comme objectif une appropriation de la démarche critique nécessaire à la LCA.

En médecine, les UE de LCA niveau 1 (DFGSM3), niveau 2 (DFASM1) et niveau 3 (DFASM2) **prennent progressivement le relais de cet enseignement pour répondre aux objectifs de l'épreuve de LCA à l'i-ECN, et préparer les futurs praticiens à l'exercice de leur métier.**

En maïeutique, la sensibilisation à la LCA est moins formalisée mais commence à être intégrée dans le cursus au travers du module d'initiation à la recherche.

Ainsi, l'UE de Biomédecine Quantitative et les UE de LCA sont intimement liées, la 1ère constituant la base méthodologique indispensable pour bien aborder les secondes.

Ce polycopié, UNIQUE, pour couvrir l'ensemble de ces UE, a vu le jour dans ce contexte.

Il est constitué d'une 1ère partie intitulée « Bases Méthodologiques de Biostatistique et d'Epidémiologie » et d'une 2ème partie intitulée « Bases Méthodologiques de Lecture Critique d'Articles ».

ABREVIATIONS

AUC : aire sous la courbe

CJP : critère de jugement principal

CJS : critère de jugement secondaire

HR : hazard ratio

IC95% : intervalle de confiance à 95%

NSN : nombre de sujets nécessaire

OR : odds ratio

RAR : réduction absolue de risque

ROC : receiver operating curve

RR : risque relatif

RRR : réduction relative de risque

SOMMAIRE

1ERE PARTIE : BASES METHODOLOGIQUES DE BIOSTATISTIQUE ET D'EPIDEMIOLOGIE

Liste des auteurs

Sophie Bastide, PHU, CHU de Nîmes

Claire Duflos, AHU, CHU de Montpellier

Pierre Dujols, PU-PH, CHU de Montpellier

Pascale Fabbro-Peray, MCU-PH, CHU de Nîmes

Jean-Luc Faillie, PHU, CHU de Montpellier

Paul Landais, PU-PH, CHU de Nîmes

Thibault Mura, MCU-PH, CHU de Montpellier

Nicolas Nagot, PU-PH, CHU de Montpellier

Dorine Neveu, MCF, CHU de Montpellier

Marie-Christine Picot, PH, CHU de Montpellier

Fabienne Séguret, PH, CHU de Montpellier

NOTIONS STATISTIQUES DE BASE

1 PRINCIPE DES TESTS STATISTIQUES

Il s'agit de statistique inférentielle : à partir de calculs réalisés sur des données observées sur un nombre limité de sujets (**l'échantillon**), nous souhaitons émettre des conclusions sur l'ensemble des sujets (**la population**) en prenant en compte **les fluctuations d'échantillonnages** (part de hasard intervenant dans la constitution de l'échantillon), en rattachant à nos conclusions des risques de se tromper. Un test d'hypothèse est donc une démarche consistant à rejeter ou à ne pas rejeter une hypothèse au niveau d'une population (hypothèse statistique appelée hypothèse nulle), en fonction d'un jeu de données observé sur un échantillon.

Les tests statistiques peuvent avoir différentes finalités, ceux que vous rencontrerez et que nous vous décrirons ici sont les tests statistiques d'association. Ces tests évaluent l'existence d'une liaison entre deux variables, c'est-à-dire que la connaissance de la valeur de l'une donne une information sur la valeur de l'autre. Les techniques utilisées diffèrent selon que les variables analysées sont qualitatives, quantitatives ou censurées. Ce type de test va suivre une succession d'étapes définies :

1. **Formulation des hypothèses :** *hypothèse nulle* (H_0) et *hypothèse alternative* (H_1)

- **l'hypothèse nulle H_0** : généralement celle du statu quo, c'est-à-dire de l'absence de différence entre les groupes. Sous cette hypothèse les populations des deux groupes sont les mêmes et les différences observées entre les échantillons sont liées au hasard (fluctuation d'échantillonnage).

- **l'hypothèse alternative H_1 :** (ou de remplacement), généralement celle de la réalité d'une différence entre les groupes. Sous cette hypothèse les populations des deux groupes diffèrent sur le paramètre étudié, et les différences observées entre les échantillons sont la conséquence des différences réelles existant entre les populations comparées.

Remarque : H_0 est donc une hypothèse précise et unique, alors que H_1 correspond à la multitude d'hypothèses possibles différentes de H_0.

2. **Calcul d'un *degré* de signification (= p = p-value**) qui correspond (généralement) à la probabilité que l'on aurait eu d'observer une différence au moins aussi importante sous l'hypothèse où H_0 aurait été vraie.

3. Conclusion du test, en fonction d'un risque seuil, le **seuil de signification (= risque α, = risque de 1er espèce**), en dessous duquel on considèrera que H_0 est peu probable et que l'on peut la rejeter. Souvent, un risque α de 5 % est considéré comme acceptable (c'est-à-dire que dans 5 % des cas quand H_0 est vraie, l'expérimentateur se trompera et la rejettera à tort).

- si $p < \alpha$, il est peu probable que H_0 soit vraie et donc on **rejette H_0** et on accepte H_1;
- si $p > \alpha$, les observations faites sur l'échantillon sont compatibles avec H_0, il est difficile d'exclure que H_0 soit vraie et donc on ne **rejette pas H_0.**

Attention : ne pas rejeter H_0 ne signifie pas que cette dernière est vraie, car les observations faites sont vraisemblablement également compatibles avec d'autres hypothèses contenues dans H_1 (multitude de possibilité d'hypothèses différentes de H_0). La seule chose qu'on puisse démontrer par des observations est qu'une hypothèse (l'hypothèse nulle, H0) est peu probable (et donc considérée comme fausse).

Face à $p > \alpha$ (absence d'association statistiquement significative) et en l'absence de biais, il existe deux possibilités :

- soit il n'existe réellement pas de différence cliniquement intéressante entre les groupes
- soit on ne dispose pas d'une puissance suffisante (cf infra) pour mettre en évidence une différence cliniquement intéressante (NSN non calculé, ou calculé avec une différence attendue trop importante) et il est donc impossible de conclure.

2 RISQUE B, PUISSANCE ET NOMBRE DE SUJETS NECESSAIRE (NSN)

Le risque β est le risque associé à la décision de ne pas rejeter H_0 alors que H_0 est fausse. Le risque β n'est pas calculable sans données complémentaires : en effet, on ne connaît pas la valeur ou différence réelle Δ. Si on veut pouvoir calculer β, il faut spécifier une hypothèse alternative particulière H_1 ; il existe une valeur de β pour chaque valeur Δ.

La puissance d'un test est par définition égale à **1-β**. C'est la probabilité de rejeter H_0 alors que H_1 est vraie. C'est donc la capacité d'un test à reconnaître que H_1 est vraie. Comme précédemment, cette situation ne correspond à rien en pratique car H_1 représente une multitude de situations. Il faut alors fixer la valeur de la différence Δ réelle entre les groupes pour pouvoir calculer cette puissance statistique.

La puissance d'une étude dépend de plusieurs facteurs : du *nombre de sujets inclus* (NSN↑ P↑), du risque de première espèce (α↑ P↓), du caractère uni ou bilatéral de l'hypothèse alternative (hypothèse bilatérale, P↓), et de l'importance de l'effet (différence Δ entre les deux groupes pour un essai clinique, Δ↑ P↑) relativement aux autres grandeurs d'intérêt (comme la variance (V↑ P↓), ou le % dans le groupe contrôle (%↑ P↑).

Tableau récapitulatif : risques d'erreur associés à un test selon que H_0 ou H_1 est vraie :

Réalité	**Conclusion du test**	
	Rejet de H_0	Non - rejet de H_0
H_0 vraie	α	1 - α
H_1 vraie	1 - β = Puissance	β

Le nombre de sujets nécessaire (NSN) d'une étude est le nombre de sujets *à inclure dans l'étude* pour garantir une puissance suffisante (généralement 80% ou 90%) pour mettre en évidence une différence attendue fixée Δ relativement aux autres grandeurs d'intérêt (comme la variance, ou le % dans le groupe contrôle) avec un risque α fixé (généralement 5%) uni ou bilatéral. D'une façon générale le NSN dépend de l'ensemble de ces éléments et sera d'autant plus élevé que:

- Le risque α consenti sera faible et que l'hypothèse nulle sera bilatérale
- la puissance exigée (1 – β) sera élevée

- la différence attendue (Δ) entre les groupes sera faible (pour une variable qualitative, il sera nécessaire de connaître la fréquence attendue dans le groupe contrôle pour réaliser le calcul : plus elle sera faible, plus le NSN sera élevé)
- la variance sera élevée

Ce NSN peut également être majoré par le nombre attendu de données manquantes (ou de perdus de vue).

3 TEST UNILATERAL OU BILATERAL

Dans tout ce qui précède, nous avons considéré que l'hypothèse alternative H_1 était une inégalité du type H_1: $P_A \neq P_B$. On dit que nous sommes en situation bilatérale. Il est des circonstances où le problème se posera en d'autre terme :

- H_0 : $P_A = P_B$ (les deux traitements ont la même efficacité)
- H_1 : $P_A > P_B$ (le traitement A est plus efficace que le traitement B).

En pratique cela a pour conséquence de réduire la probabilité "p". Puisque la nouvelle hypothèse alternative $P_A > P_B$ est strictement incluse dans l'hypothèse alternative bilatérale : $P_A \neq P_B$, il y a moins de risque d'accepter H_1 alors que H_0 est vraie. On dit que le problème est formulé de façon unilatérale. Il est donc nécessaire de prendre en compte cette réduction de probabilité en fixant un seuil de signification plus bas (un risque α unilatéral à 5% correspond à un risque α bilatéral à 10%), **on choisira donc plutôt un risque α unilatéral à 2.5%**.

4 ANALYSES UNIVARIEES / MULTIVARIEES ET CHOIX DES TESTS STATISTIQUES

Les analyses statistiques présentées dans les articles peuvent être dites :

- **« univariées » = « non ajustées » = « brutes »*** lorsqu'elles ne prennent pas en compte l'effet de facteurs de confusion potentiels (une seule variable « explicative » dans le modèle ou l'analyse). **(univariate or crude en anglais)*
- **« multivariées » = « ajustées » = « indépendantes »*** qui permettent de prendre en compte d'autres variables et leurs effets de confusion potentiels (Voir Fiche « analyses Multivariées »). **(multivariate or adjusted or independent)*

Le choix des tests statistiques est essentiellement lié à la nature des variables à analyser et des conditions de validité des différentes méthodes statistiques. Les différents tests sont résumés dans les 3 tableaux suivants :

A. VARIABLE A EXPLIQUER QUALITATIVE

Problématique	**Outil statistique**	**Hypothèses**
Estimation d'une fréquence théorique	$IC_{1-\alpha}$	NA
Comparaison d'une fréquence observée à une fréquence théorique	Test Chi2 de Pearson, test exact de Fisher, $IC_{1-\alpha}$	H0 : $F = F_{th}$ H1 : $F \neq F_{th}$

Comparaison de fréquences observées dans groupes indépendants	Test Chi2 de Pearson, test exact de Fisher	H0 : $F_A = F_B = F_C$ H1 : au moins une F est ≠
Comparaison de fréquences observées dans séries appariées	Test Chi2 de Mc Nemar	H0 : Egalité $F_{T1} = F_{T2}$ H0 : Différence $F_{T1} \neq F_{T2}$
Liaison de deux variables qualitatives	Test Chi2 de Pearson, test exact de Fisher, $IC_{1-\alpha}$ de OR	H0 : Indépendance, OR = 1 H1 : Liaison, OR ≠ 1
Concordance (accord)	Indice kappa	H0 : pas d'accord au delà du hasard, kappa = 0 H1 : accord au delà du hasard, kappa ≠ 0
Liaison var à expliquer qualitative sur séries indépendantes * var explicatives	Modèle de Régression logistique multivariée (=> OR ajusté)	H0 : indépendance; OR = 1 H1 : liaison; OR ≠ 1
Liaison var à expliquer qualitative sur séries appariées * var explicatives *Ex: cas-témoins appariées*	Modèle de Régression logistique conditionnelle (=> OR ajusté)	H0 : indépendance; OR = 1 H1 : liaison; OR ≠ 1

$IC_{1-\alpha}$= Intervalle de confiance au niveau de confiance 1-α ; NA : non applicable

B. VARIABLE A EXPLIQUER QUANTITATIVE

Problématique	**Outil statistique**	**Hypothèses**
Estimation d'une moyenne théorique	$IC_{1-\alpha}$	NA
Comparaison d'une moyenne observée à une moyenne théorique	Test Student, $IC_{1-\alpha}$	H0 : $\mu = \mu_{th}$ H1 : $\mu \neq \mu_{th}$
Comparaison de 2 moyennes observées dans 2 groupes indépendants	Test Student, *Mann-Whitney-Wilcoxon*	H0 : $\mu_A = \mu_B$ H1 : $\mu_A \neq \mu_B$
Comparaison de 2 moyennes observées dans 2 séries appariées	Test Student apparié, *Wilcoxon apparié*	H0 : Egalité $\mu_{T1} = \mu_{T2}$ H0 : Différence $\mu_{T1} \neq \mu_{T2}$
Liaison de deux variables quantitatives	Coefficient de corrélation linéaire ρ de Pearson, *ρ de Spearman*	H0 : Indépendance, ρ = 0 H1 : Liaison, ρ ≠ 0
Concordance (accord)	Coefficient de corrélation intraclasse,	H0 : pas d'accord, ICC= 0 H1 : accord, ICC ≠ 0
Concordance (accord)	Graphe de Bland et Altman	
Liaison var à expliquer (Y) quantitative * 1 var explicative X qualitative à 2 classes ou plus	ANOVA, *test Kruskall-Wallis*	H0 : indépendance, $\mu_A = \mu_B = \mu_C$ H1 : liaison, au moins une μ est ≠
Liaison Y quant * X_i quantitatives	Modèle* de Régression linéaire multiple	H0 : indépendance, β = 0 H1 : liaison, β ≠ 0
Liaison Y quant * X_i qualitatives et X_i quantitatives	Analyse de covariance	H0 : indépendance, β = 0 H1 : liaison, β ≠ 0

Liaison Y quant à mesures répétées* X_i	Modèle mixte	H0 : indépendance, $\beta = 0$ H1 : liaison, $\beta \neq 0$
Capacité d'un test diagnostique à discriminer malades des non malades	$IC_{1-\alpha}$ de l'aire sous la courbe ROC (AUC)	H0 : AUC = 0,5 H1 : AUC ≠ 0,5

Les outils statistiques en italiques sont des tests non paramétriques.

C. VARIABLE A EXPLIQUER CENSUREE

Problématique	**Outil statistique**	**Hypothèses**
Estimation d'un taux de survie	Courbe de Kaplan Meier, courbe actuarielle	NA
Comparaison de k courbes de survie	Test du Logrank (k-1 ddl)	H0 : égalité des distributions de survie H1 : au moins une courbe de survie est ≠
Liaison Y censurée * var explicatives qualitatives ou quantitatives	Modèle* multivarié de Cox (semi-paramétrique, à risques proportionnels) (=> RRI ajusté)	H0 : indépendance; RRI = 1 H1 : liaison; RRI ≠ 1

RRI = rapport de risques instantanés (en anglais HR: hazard ratio) ; *tous les modèles peuvent produire des résultats univariés (Beta, OR ou HR bruts, « crude » en anglais) ou multivariés (Beta, OR ou HR ajusté, «adjusted » en anglais).

Rq : d'une façon générale, l'analyse principale d'une étude randomisée sera **brute** (car on suppose l'absence de facteurs de confusion) et celle d'une étude observationnelle sera **ajustée** (car on suppose l'analyse brute biaisée par des facteurs de confusion).

5 INTERVALLE DE CONFIANCE A 95% (IC95%)

Un intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique sur un échantillon. Plus précisément, il permet de définir une marge d'erreur liée aux fluctuations d'échantillonnage à partir des résultats observés sur un échantillon.

Il s'agit donc d'un intervalle dans lequel il est raisonnable (probable à 95% pour l'IC95%) de penser que la valeur réelle du paramètre estimé dans l'échantillon se situe. Cet intervalle est centré autour de l'estimation dite « ponctuelle » du paramètre. Sa largeur dépend de la variabilité du paramètre et du nombre de sujets analysés. Plus le nombre de sujets étudiés sera grand, plus l'estimation sera précise et plus l'intervalle sera étroit.

Dans certain cas, l'IC95% peut se substituer à un test statistique. Par exemple si l'IC95% d'un risque relatif ne contient pas la valeur 1 (exemple RR=2, IC95% [1.8-2.2]), alors il est peu probable (<5%) que la valeur réelle (mesurée dans la population) de ce RR soit égale à 1. On peut donc rejeter l'hypothèse nulle d'un RR=1 qui correspond à l'absence d'association entre les deux variables. (Attention, ce raisonnement basé sur l'absence de la valeur 1 dans l'IC95% ne s'applique qu'aux Odds Ratio, Risque

Relatif et Hasard Ratio pour lesquels l'hypothèse nulle correspond à une valeur de paramètre=1). L'intervalle de confiance d'une différence de pourcentage entre deux groupes s'interprétera quant à lui par rapport à la valeur 0 car l'hypothèse d'absence de différence correspond à une valeur de paramètre=0.

6 CORRELATION STATISTIQUE ET REPRODUCTIBILITE D'UNE MESURE

Deux autres indicateurs statistiques peuvent être rencontrés au cours d'une lecture de LCA et nécessitent d'être connus :

- **Le coefficient de corrélation** (de Pearson ou de Spearman) : permet de mesurer la force d'une association entre deux variables quantitatives (ex : « apport de sel/j » et « tension artérielle »). Ce coefficient corrélation varie **de 1** (corrélation positive parfaite : les variations de la première variable expliquent entièrement les variations de la seconde. V1 et V2 varie dans le même sens), **à -1** (corrélation négative parfaite (idem que précédemment mais V1 et V2 varient dans des sens opposés), **en passant par 0** (aucune corrélation, les variations de V1 ne sont pas associées aux variations de V2). L'IC95% de ce coefficient permet de rejeter l'hypothèse nulle H_0 d'absence d'association (coef=0) si celui-ci ne contient pas 0.
- **Le coefficient Kappa** : permet de mesurer la concordance entre deux jugements qualitatifs. Par exemple, si deux radiologues doivent classer une même série d'images en « tumeur maligne », « tumeur bénigne », donnent-ils généralement la même réponse ? Le coefficient Kappa varie de 1 (concordance parfaite entre les évaluateurs) à < 0 (« 0 » correspond à la concordance dite « aléatoire » c'est-à-dire celle observée sur des jugements portés au hasard), et s'interprète à l'aide de l'échelle de Landis et Koch :

7 COMPARAISONS MULTIPLES ET INFLATION DU RISQUE ALPHA

Les comparaisons multiples correspondent à la réalisation simultanée de plusieurs tests statistiques pour répondre à une même question. Plus le nombre de tests considérés est grand, plus le risque de conclure à une différence à tort (rejeter H0 à tort) est important, ce qui conduit donc à une augmentation (inflation) du risque α.

Accord	*Kappa*
Excellent	≥ 0,81
Bon	0,80 - 0,61
Modéré	0,60 - 0,41
Médiocre	0,40 - 0,21
Mauvais	0,20 - 0,0
Très mauvais	< 0,0

Par exemple, supposons que nous considérions l'efficacité d'un médicament en fonction de la réduction de l'un ou l'autre des symptômes d'une maladie. Plus le nombre de symptômes considérés est élevé, plus la probabilité que le médicament apparaisse comme efficace à tort sur au moins un symptôme est importante.

En cas de comparaison multiple, il existe des procédures qui permettent de prendre en compte l'inflation du risque alpha (procédure de Bonferroni, de Tukey…), afin de conserver un risque global égal à 5%. Le principe est soit de déterminer un nouveau seuil de signification

(plus bas, par exemple par la méthode de Bonferonni pour 3 tests : $\alpha_{corrigé} = \alpha/3 = 0.05/3 = 0.016$), soit de corriger le calcul du degré de signification (plus haut). Dans tous les cas il sera toujours plus difficile de conclure à une différence significative après correction.

Dans les essais contrôlés randomisés, les corrections doivent être **utilisées systématiquement,** pour pouvoir conclure sur une relation causale, dans **3 cas particuliers** : (1) lors de la réalisation d'analyse intermédiaire, (2) si le critère de jugement principal est multiple, (3) si le critère de jugement principal est analysé dans des sous-groupes (par exemple une analyse séparée selon le sexe après avoir réalisé l'analyse dans l'ensemble de l'échantillon).

ANALYSES MULTIVARIEES

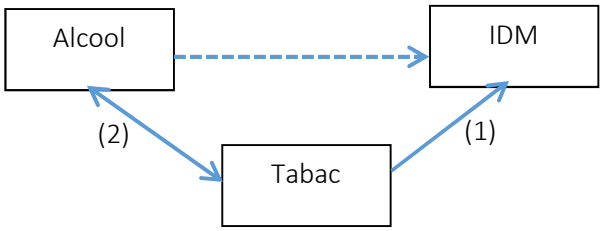
1. INTRODUCTION

Lors de l'analyse statistique d'une étude, faire une analyse multivariée (c'est-à-dire utiliser un modèle statistique multivarié) est un moyen pour neutraliser les facteurs de confusion.

2. RAPPELS

2.1. Définition d'un facteur de confusion

Un facteur de confusion (X) est un facteur qui modifie l'association statistique entre une exposition (E) et une maladie (M). Il peut être de nature variée (ex. âge, sexe, antécédent, exposition environnementale, comédications, etc.)
Il doit répondre à certaines conditions :
X est un facteur de risque (ou protecteur) de M
X est associé à E mais n'est pas une conséquence de E
L'estimation du risque lié à E (risque relatif [RR], odds ratio [OR] ou hazard ratio [HR]) sous sa forme brute (« *crude* », c'est-à-dire non ajustée sur X) est modifiée lorsqu'on ajuste sur X. On considère généralement que la différence entre risque brut et ajusté doit être supérieure à 10-15%.
Un facteur de confusion peut conduire à la surestimation ou à la sous-estimation d'une association entre un facteur d'exposition et une maladie et peut même changer la direction de l'effet observé. Si le facteur de confusion est à la fois lié positivement et de façon indépendante au facteur de risque et à la maladie, la liaison entre le facteur de risque et la maladie est surestimée. En revanche, si les liaisons entre facteur de confusion et maladie d'une part, et facteur de confusion et facteur de risque d'autre part, sont en sens inverse, alors la liaison entre le facteur de risque et la maladie est sous-estimée par rapport à la réalité éventuelle de cette liaison
Exemple :
On étudie la relation entre consommation d'alcool (E) et la survenue d'infarctus du myocarde (M). Si on ne tient pas compte du tabac, l'HR mesuré est significativement supérieur à 1 : on observe donc un risque d'IDM associé à la consommation d'alcool. Si on tient compte du tabac (c'est-à-dire si on ajuste ce résultat sur la consommation de tabac), l'HR mesuré devient non significatif. Le tabac est donc un facteur de confusion de la relation alcool-IDM car le tabac est un facteur de risque d'IDM (1) et tabac et alcool sont souvent associés (2).



2.2. Prise en compte des facteurs de confusion :

Négliger les facteurs de confusion dans l'analyse des liens entre exposition et maladie conduit à des interprétations erronées dues à des **biais de confusion.**

La prise en compte des facteurs de confusion se fait :

- soit lors de la planification des enquêtes (sélection des sujets) par :
  - **randomisation** (uniquement dans les essais cliniques)
  - **restriction de la population :** restreindre l'étude aux sujets qui ne présenteraient pas le facteur de confusion donné. *Exemple : une étude réalisée uniquement chez les femmes élimine l'effet du sexe*
  - **appariement** sur le facteur. *Exemple : sujets exposés et non exposés (ou cas et témoins) appariés sur l'âge et le sexe*
- soit lors de l'analyse statistique des enquêtes par :
  - **stratification** de l'analyse sur le facteur. *Exemple : L'analyse réalisée uniquement chez les femmes élimine l'effet du sexe*
  - **ajustement** sur le ou les facteur(s) de confusion

2.3. Principe de l'ajustement

Ajuster sur un facteur de confusion consiste en l'estimation d'un risque (RR, OR, HR) « pondéré » sur les risques estimés au sein de chaque niveau/strates du facteur de confusion.

Exemple :

Lors d'une étude de cohorte sur le lien entre prothèse totale de hanche (E) et cancer du colon (M), on observe une incidence du cancer de colon plus importante chez les personnes porteuse de PTH avec un RR brut significatif à 1,75.

	K colon (+)	K colon (-)	Inc.
PTH (+)	121	2979	3,9%
PTH (-)	270	11930	2,2%

RR brut=1,75

L'âge étant un facteur de confusion potentiel, on mesure le risque ajusté sur l'âge et calculant l'âge moyen pondéré selon les différentes classes d'âge.

On observe que l'incidence augmente avec l'âge mais que le port d'une PTH n'est pas associé au cancer du colon (RR=1) dans toutes les classes d'âge. Le RR ajusté vaut donc 1 est l'âge doit être considéré comme un facteur de confusion à prendre en compte dans cette étude.

Age<50

	K colon (+)	K colon (-)	Inc.
PTH (+)	1	99	1%
PTH (-)	10	990	1%

RR=1

50<âge<65

	K colon (+)	K colon (-)	Inc.
PTH (+)	20	980	2%
PTH (-)	200	9800	2%

RR=1

Age>65

	K colon (+)	K colon (-)	Inc.
PTH (+)	50	950	5%
PTH (-)	30	570	5%

RR=1

RR ajusté=1

On peut expliquer le résultat brut supérieur à 1 en considérant que les porteuses de PTH étaient plus âgées donc plus à risque de cancer du colon que les non-porteuses de PTH.

En pratique, on utilise un modèle statistique effectuant les calculs pour ajuster sur un ou plusieurs facteurs de confusion simultanément : ce modèle est appelé **modèle multivarié**.

3. PRINCIPE DU MODELE MULTIVARIE

Le modèle multivarié s'exprime sous la forme d'une fonction reliant une variable à expliquer (Y) (souvent la survenue d'une maladie) à plusieurs variables dites explicatives ou covariables pouvant être des facteurs de risque (E) ou des facteurs de confusion (X) :

Y = f (E1 , E2 , …. , En , X1 , X2 , … , Xn)

Exemple : Valvulopathie = f (benfluorex[E1], âge[X1], sexe[X2], antécédent [X3])

Le modèle est dit univarié s'il n'y a qu'une seule covariable.

Le modèle multivarié permet d'estimer **l'effet propre de chaque facteur E_i sur Y, indépendamment de l'effet des autres facteurs inclus dans le modèle**. Autrement dit, à chaque facteur E_i est associé un paramètre permettant l'estimation d'un risque quantifiant l'association entre E_i et Y ajustée sur les autres variables du modèle. Afin d'être interprétées, les **estimations de risque doivent toujours données avec leur intervalle de confiance à 95%.**

4. TYPES DE MODELE MULTIVARIE

Il existe plusieurs types de modèle multivarié selon la nature de la variable à expliquer (Y).

- Si Y est une **variable binaire** (ex. présence ou absence d'une maladie), le modèle sera une **régression logistique** et les estimations de risque seront données sous forme d'**Odds Ratio (OR) ajustés** que l'on peut interpréter comme des Risque Relatifs (RR) si la maladie est rare.
  Exemple :
  Etude cas témoin sur l'association entre tabagisme et présence d'un cancer du poumon en tenant compte de l'âge et du sexe des individus. L'OR pour le tabac ajusté sur l'âge et le sexe est égal à 3,5 [IC95% : 2,3 – 5,2] : le tabac multiplie par 3,5 le risque de cancer du poumon indépendamment de l'effet de l'âge et du sexe.

- Si Y est une **variable censurée** (ex. survenue d'un événement, survie), le modèle sera un **modèle de Cox** (appelé aussi semi-paramétrique, « à risques proportionnels») et les estimations de risque seront données sous forme d'**Hazard Ratio (HR) ajustés** que l'on pourra interpréter comme des RR.
  Exemple :
  Cohorte de patients atteint de cancer et étude de la survenue de métastase en fonction de la consommation d'aspirine en prenant en compte l'âge, le sexe, le tabagisme. Le HR pour l'aspirine ajusté sur l'âge, le sexe et le tabagisme est égal à 0,64 [IC95% : 0,48 – 0,84] : l'aspirine diminue le risque de métastase de 36% indépendamment de l'effet de l'âge, du sexe et du tabagisme.

- Si Y et les covariables sont des **variables quantitatives** (ex. valeurs biologiques, âge en années, etc), le modèle sera une **régression linéaire multiple**.
  Exemple :

  Etude de l'association entre l'âge maternel et le poids des enfants à la naissance (Y) tout en tenant compte de l'âge gestationnel. L'âge maternel présente un coefficient négatif et statistiquement

différent de zéro (p-value < 0,05) : l'âge maternel est négativement associé au poids de naissance indépendamment de l'effet de l''âge maternel.

5. VARIABLES A INCLURE DANS LE MODELE MULTIVARIE

Pour éviter les biais de confusion, tous les facteurs de confusion doivent être inclus dans le modèle multivarié. Ceci nécessite d'avoir prévu dans le protocole de recueillir les facteurs de confusion.

En pratique, il y a des facteurs de confusion connus dans la littérature mais les chercheurs sont souvent confrontés à des facteurs susceptibles d'être des facteurs de confusion sans qu'il soit possible de l'affirmer : ce sont de facteurs de confusion potentiels. Parmi ces facteurs, il y a les facteurs de risque ou pronostiques connus ou supposés de la maladie étudiée. Les facteurs de risque ou pronostiques connus seront inclus dans le modèle multivarié. Pour les facteurs de risque supposés, une première analyse univariée évaluera leur lien avec la maladie pour décider de leur inclusion dans le modèle multivarié.

6. DANS QUELLES ETUDES A-T-ON RECOURS A UN MODELE MULTIVARIE ?

Dans les études observationnelles analytiques de type cohorte ou cas-témoins, l'analyse multivariée permettant l'ajustement sur les facteurs de confusion connus ou potentiels est cruciale et doit être systématique. Son absence expose à des biais de confusion. En utilisant un modèle multivarié, les résultats peuvent au final être encore biaisés par une confusion résiduelle. C'est la confusion qui demeure après des tentatives infructueuses de la contrôler.

Dans les essais cliniques, les biais de confusion sont moins présents car, en dehors de l'intervention étudiée, tous les autres facteurs sont censés être distribués de façon équilibrée entre les groupes par la randomisation. C'est pourquoi l'analyse principale sur laquelle se base les conclusions d'un essai randomisé est généralement l'analyse brute (=univariée). Il est possible toutefois d'utiliser un modèle multivarié lorsque des facteurs pronostiques ou des traitements concomitants pouvant influencer le critère de jugement semblent malencontreusement déséquilibrés entre les groupes lors de la comparaison des caractéristiques initiales des groupes.

Analyses de survie

(ou analyses de données censurées)

Les 6 mots-clés de la survie

Variable censurée
Analyse de survie
Courbe de Kaplan-Meier
Test du Log-Rank
Modèle de Cox
Hazard Ratio

1. INTERET

En médecine, la variable d'intérêt servant de critère de jugement est souvent un délai d'apparition: temps jusqu'à la récidive, jusqu'au décès, jusqu'à la guérison. Ainsi, le « temps jusqu'à un évènement » n'est ni une variable quantitative (on ne mesure pas que le temps), ni une variable qualitative (on ne mesure pas que l'évènement), c'est une variable censurée.

Prenons pour exemple un essai qui évalue l'efficacité d'un nouveau traitement de la phase aiguë de l'infarctus du myocarde (IDM). Le critère de jugement est le délai jusqu'au décès. Notre étude dure 3 ans. Cependant tous les patients ne seront pas suivis sur la même durée. Prenons deux cas : les patients qui ont présenté l'évènement et ceux qui ne l'ont pas présenté.

- Les patients qui ont présenté l'évènement seront suivis jusqu'à l'évènement, qui peut survenir à différents délais au cours de l'étude.
- Les patients qui n'ont pas présenté l'évènement sont appelés « patients censurés » ou encore « censures » car, à partir d'un certain moment, leur suivi s'arrête même s'ils n'ont pas présenté l'évènement. Cela peut se produire pour deux raisons :
    - L'étude est arrivée à sa « date de point », c'est-à-dire à la date où le protocole prévoyait d'arrêter le suivi pour analyser les résultats. Les patients encore vivants à cette date ont des durées de suivi différentes. En effet, tous les patients n'ont pas nécessairement été inclus le premier jour de l'étude. Les sujets encore suivis et vivants à la date de point sont appelés « exclus vivants », et ils sont dits « censurés à la date de point ».
    - Au cours du suivi le sujet peut être « perdu de vue ». La dernière visite réalisée est alors appelée la « date des dernières nouvelles », et le patient est dit « censuré à la date des dernières nouvelles ».

Pour les sujets censurés, nous ne connaissons donc pas le temps exact jusqu'à leur décès : nous savons seulement que ce temps est supérieur au temps de participation à l'étude.

Comment analyser les résultats de notre étude sur l'IDM ? Si l'on ne faisait pas d'analyse de survie, il faudrait simplement analyser le taux d'évènement. Dans notre exemple sur l'IDM, on pourrait avoir 14% de décès à un an dans le groupe « nouveau traitement », contre 25% dans le groupe « traitement de référence ». Dans ce cas, tous les perdus de vue avant un an auraient été inutilisables : comme on ne sait pas s'ils ont présenté l'évènement durant la première année ou pas, on ne peut pas les analyser.

Ce phénomène constitué par les patients qui quittent l'étude (les perdus de vue) s'appelle l'attrition. Si cette attrition ne se fait pas au hasard (par exemple, si ce sont les patients les plus graves qui sont perdus de vue), cela va provoquer un biais d'attrition: l'échantillon que l'on va analyser ne va pas être comparable à la population initiale de l'étude, et donc à la population cible. Le biais d'attrition fait donc partie des biais de sélection.

Les analyses de survie sont une réponse partielle à ce problème. En effet, elles permettent de prendre en compte toute la durée d'observation de chaque patient, quelle que soit cette durée, et même pour

les sujets censurés : par exemple, on prend en compte l'information qu'un sujet perdu de vue au bout de 6 mois a eu une durée de survie supérieure à 6 mois. Les analyses de survie permettent donc d'analyser plus de patients, y compris les perdus de vue, ce qui a deux conséquences : elles sont plus puissantes, et elles diminuent les conséquences du biais d'attrition.

2. TYPES D'ETUDES CONCERNES

Les analyses de survie sont utilisées dans les études longitudinales, c'est-à-dire principalement : essais thérapeutiques, cohortes à but étiologique ou pronostique, évaluation de programmes de dépistage.

3. PRINCIPAUX ELEMENTS D'ANALYSE

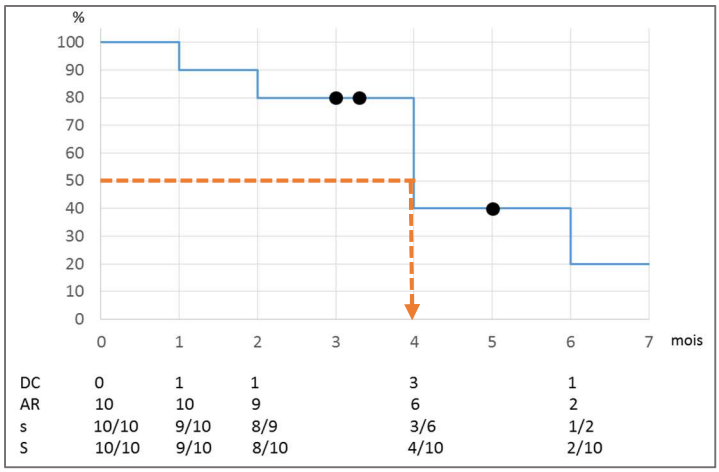
3.1. ANALYSE DESCRIPTIVE : COURBE DE KAPLAN-MEIER

La courbe de survie[1] représente la probabilité de survivre (ou de ne pas présenter l'évènement) jusqu'au temps t. Il existe plusieurs méthodes pour les estimer, la plus fréquente étant la méthode de Kaplan-Meier. Cette méthode consiste à calculer, à chaque instant, une probabilité de survie instantanée, c'est-à-dire une probabilité de survivre sachant que l'on était vivant jusque-là. Il faut donc connaître, à chaque instant, le nombre d'évènements et le nombre de personnes à risque de présenter l'évènement à ce moment-là. On calcule ensuite, à chaque instant, la probabilité de survivre jusqu'à cet instant-là, depuis le début de l'étude : c'est cette probabilité qui est représentée par la courbe de Kaplan-Meier.

La courbe de Kaplan-Meier suivante présente la probabilité de survie d'un groupe de 10 patients, suivis pendant 7 mois maximum. Les censures sont représentées par les points noirs. Les lignes de chiffres présentent les éléments nécessaires au calcul de la probabilité de survie à chaque temps :

- DC est le nombre de décès à cet instant
- AR est le nombre de patients à risque de décéder à cet instant
- s est la probabilité de survie instantanée (probabilité de survivre au temps t, sachant que l'on est vivant juste avant t)
- S est la probabilité de survivre jusqu'au temps t.

[1] Le terme « courbe de survie » est utilisé pour n'importe quel évènement (récidive, guérison …), et pas seulement pour le décès.

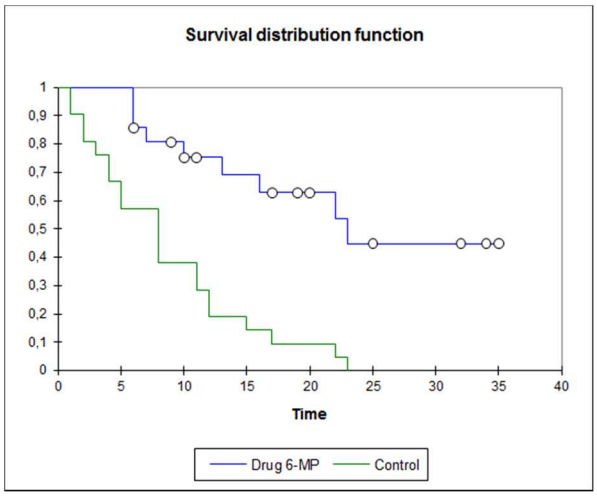


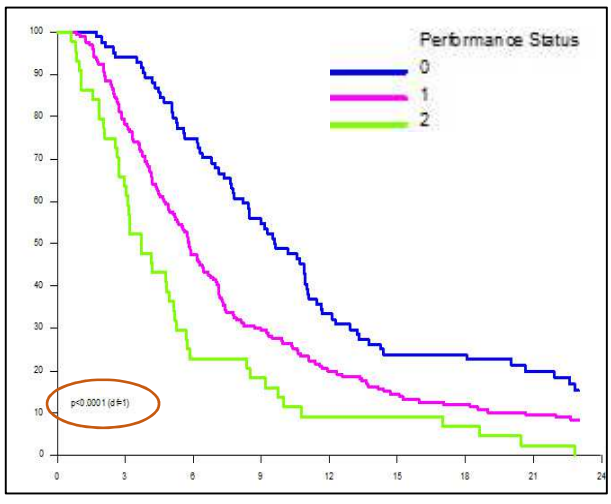
Juste avant 1 mois, il y a 10 patients à risque. L'un décède, donc la probabilité de survie instantanée s est de (AR-DC)/AR = (10-1)/10. Juste avant 2 mois, il y a donc 9 patients à risque ; l'un décède, s = 8/9. Juste avant 4 mois, 2 patients sont décédés et deux patients sont censurés (entre 2 et 4 mois). Il reste donc 6 patients dans notre échantillon, à risque de décéder. On observe 3 décès, donc la probabilité de survie instantanée est de (AR-DC)/AR = (6-3)/6. Pour obtenir la probabilité de survie S, qui est la probabilité de survivre jusqu'au temps t, il faut multiplier toutes les probabilités de survie instantanées : survivre jusqu'à t est en effet survivre à 1 mois et à 2 mois et à 4 mois … A 4 mois, S = 9/10*8/9*3/6 = 4/10.

Il faut noter qu'au cours de l'étude, AR diminue en raison des patients qui présentent l'évènement et des patients perdus de vue. La précision des estimations de s diminue donc au cours de l'étude. Or, c'est sur les valeurs de s que sont réalisés les tests statistiques. Il faut donc présenter, en dessous des courbes de survie, le nombre de personnes à risque à chaque instant : cela donne une indication sur la confiance que l'on peut accorder aux résultats portant sur les suivis tardifs.

Un paramètre descriptif classique est la médiane de survie : temps au bout duquel 50% des patients sont décédés. Ici, la médiane de survie est de 4 mois. L'intervalle de confiance de cette médiane doit être calculé et présenté avec la médiane.

3.2. ANALYSE UNIVARIEE : TEST DU LOG-RANK, MODELE DE COX





Les patients prenant le traitement à l'essai survivent-ils significativement plus longtemps que les patients ayant le traitement contrôle ? La durée de survie est-elle significativement liée à l'état général mesuré par le Performance Status ? On peut répondre à ces questions en comparant ces courbes par un test du Log-Rank. Le résultat du test (p-value) est souvent présenté sur le graphique.

Le traitement à l'essai diminue-t-il le risque par 2, 3, 4 … ? On cherche ici à connaître la taille d'effet, et non la significativité statistique de cet effet, comme ci-dessus. En survie, la mesure de risque est le Hazard Ratio (HR, ou rapport de risque instantané RRI), que l'on calcule grâce à un modèle de Cox (ou modèle semi paramétrique à risques proportionnels, ou « proportional hazard model » en anglais). Le HR s'interprète comme un risque relatif : s'il est supérieur à 1, la variable est un facteur de risque ; s'il est inférieur à 1, la variable est un facteur protecteur ; s'il est égal à 1, il n'y a pas d'association entre

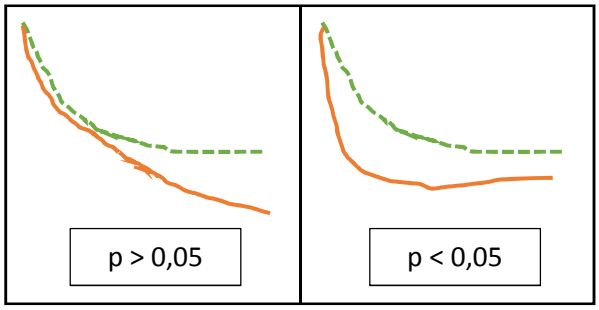
la variable et le critère de jugement. Il doit être accompagné de son intervalle de confiance et du résultat du test statistique testant si sa valeur est significativement différente de 1. Pour une analyse univariée, on fait un modèle de Cox univarié : par exemple, S = f(groupe) (probabilité de survie en fonction du groupe de traitement), ou S = f(Performance Status).

3.3. ANALYSE MULTIVARIEE : MODELE DE COX

Pour une analyse multivariée, on fait un modèle multivarié : par exemple, S = f(âge, Performance Status, stade de la maladie). (Voir fiche sur l'analyse multivariée).

4. VALIDITE DES ANALYSES DE SURVIE

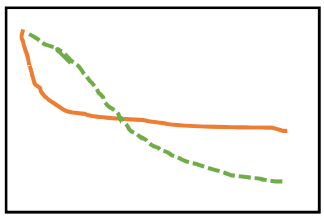
4.1. PUISSANCE



Comme dans toute analyse, la puissance dépend de l'effectif. Contrairement aux analyses classiques, l'effectif diminue tout au long du suivi. A la fin du suivi, les estimations sont moins précises et les tests moins puissants. Il est donc important de ne pas avoir trop de perdus de vue au début du suivi, et de regarder, sur les courbes de survie, quel est l'effectif à risque. Sur le premier graphique ci-contre, les courbes sont parallèles au début et se séparent à la fin, là où le test manque de puissance : il n'est donc pas significatif. Sur le second graphique, les courbes se séparent au début, puis restent à peu près parallèles. Le test est donc significatif.

4.2. RISQUES PROPORTIONNELS

Pour que l'analyse de survie soit valide, l'effet d'un facteur de risque doit rester le même tout au long du suivi : cela s'appelle la proportionnalité des risques. Cette hypothèse peut être vérifiée mathématiquement, ce qui est généralement rapporté par les auteurs : « un modèle de Cox a été construit après vérification de l'hypothèse de proportionnalité des risques ».



Il existe un cas extrême de non proportionnalité des risques : lorsque les courbes de Kaplan-Meier sont clairement séparées au début et à la fin, et se croisent au milieu. Dans ce cas, le test du Log-Rank et le Hazard Ratio ne sont pas interprétables. En effet, on ne sait pas s'ils concernent l'effet délétère avant le croisement ou l'effet protecteur après le croisement …

4.3. CENSURE ALEATOIRE

L'hypothèse de censure aléatoire énonce qu'un sujet perdu de vue a le même risque qu'un sujet suivi jusqu'à la date de point.

Si la censure n'est pas aléatoire, c'est-à-dire si la probabilité d'être censuré (perdu de vue) dépend de l'évènement que l'on mesure, on peut avoir un biais lié aux perdus de vue (attrition) dans une analyse de survie. Par exemple, dans notre étude sur le temps de survie jusqu'au décès après infarctus du myocarde, si les décès ne sont recueillis que par enquête téléphonique, les patients qui ne répondent

pas au téléphone seront censurés. Or, c'est précisément les patients décédés qui ne répondent pas. Ils avaient donc un risque plus élevé que les sujets suivis jusqu'à la date de point.

L'hypothèse de censure aléatoire est importante, mais elle ne peut pas être vérifiée mathématiquement. Elle doit être discutée cliniquement et méthodologiquement.

ESSAIS THERAPEUTIQUES

1 INTRODUCTION

La mise sur le marché d'un nouveau médicament pour une indication donnée est l'aboutissement d'un long processus de recherche qui a démarré *in vitro,* s'est poursuivi *in vivo* chez l'animal, puis chez l'homme. La chronologie des étapes est à respecter strictement puisque le préalable indispensable à l'introduction du médicament chez l'homme est l'évaluation et le contrôle de sa toxicité, ainsi qu'une première évaluation du bénéfice thérapeutique attendu. Ce processus a pour objectif final d'évaluer l'efficacité et la tolérance du médicament dans une indication donnée. Après une brève présentation de ces différentes étapes, nous insisterons plus particulièrement sur la méthodologie des essais comparatifs. Volontairement, l'accent est mis sur les essais de médicaments puisque ce sont les plus contraignants d'un point de vue législatif mais les essais comparatifs de thérapeutiques non médicamenteuses obéissent dans leurs grandes lignes aux mêmes règles méthodologiques.

Pour des raisons didactiques, nous prendrons comme support la comparaison de deux traitements dans des groupes parallèles mais tous les concepts évoqués s'appliquent à la comparaison de plus de deux traitements.

2 LES ETAPES D'INTRODUCTION D'UN NOUVEAU TRAITEMENT EN PRATIQUE MEDICALE (DOMAINE DU MEDICAMENT)

1. La phase préclinique

Il s'agit d'une phase réalisée chez l'animal qui vise à déterminer les conditions d'efficacité et de sécurité d'un produit médicamenteux.
L'efficacité potentielle est approchée par des études pharmacodynamiques (évaluation de la dose efficace, du type et de la durée de l'effet, du mécanisme et des effets secondaires, etc.) et des études pharmacocinétiques (évaluation des conditions d'absorption, de diffusion, d'élimination et de métabolisme du produit).
Les conditions de sécurité sont approchées par des études toxicologiques, en conditions aiguë, subaiguë et chroniques, des études des effets sur la reproduction, des études de mutagenèse (recherche de modification du matériel génétique), des études de cancérogenèse.
Ces différents types d'études sont réglementaires et doivent obéir aux bonnes pratiques cliniques de laboratoire qui décrivent avec précision leurs conditions de réalisation.

2. Les phases cliniques

Lorsqu'un bénéfice thérapeutique est espéré au terme de cette phase, moyennant des conditions de sécurité d'emploi suffisantes, le passage à l'expérimentation chez l'homme est possible. Les conditions d'expérimentation des médicaments chez l'homme ont été définies initialement en France par la loi du 20 décembre 1988, dite loi Huret-Sérusclat, dont le contenu a été modifié en août 2004 en vue de s'adapter à la directive 2001/20/CE du Parlement européen concernant le rapprochement des

dispositions législatives, réglementaires et administratives des Etats membres relatives à l'application de bonnes pratiques cliniques dans la conduite d'essais cliniques de médicaments à usages humains (*cf. infra*).
Il existe 4 phases qui sont mises en œuvre successivement.

a. Phase I

Cette phase correspond aux **premières administrations du produit chez l'homme**. Les essais de phase 1 ont pour but de connaître le **comportement du médicament chez l'homme** (pharmacocinétique) et de déterminer ma posologie entraînant les premiers effets indésirables, et la posologie entraînant les premiers effets pharmacodynamiques souhaités. Ils sont conduits chez des **volontaires sains** sauf lorsque la toxicité du produit est trop importante (ex : antimitotiques) et sont donc dénués d'intérêt thérapeutique pour ces sujets ; ils sont ainsi dénommés « sans bénéfice individuel direct ».

b. Phase II

Succédant aux précédents, ces essais visent à déterminer les **modalités optimales d'administration du produit pour obtenir l'effet thérapeutique escompté** : voie d'administration (nombre de prises journalières, posologie, rythme, durée, etc.). Ces essais sont le plus souvent conduits chez des **malades présentant la maladie** que le médicament est censé traiter. Ils sont généralement menés sur des groupes parallèles, se distinguant entre eux soit par la dose administrée, soit par les modalités de l'administration. A l'issue de cette phase, les conditions optimales de prescription sont connues. Ce sont elles qui seront utilisées en phase III.

c. Phase III

C'est la phase de **l'étude de l'efficacité thérapeutique réelle**. Les essais de phase III vont préciser l'efficacité et la tolérance du médicament dans les **indications** pour lesquelles on le suppose actif. Ces essais sont réalisés sur un **grand nombre de patients**, contrairement aux essais de phase I et II ; et sont toujours **comparatifs**, par rapport à un **groupe témoin**. Ils sont menés chez des **malades pour l'indication revendiquée**.
C'est à l'issue de ces essais que les autorités sanitaires délivreront **l'Autorisation de mise sur le marché (AMM)**, indispensable pour que le médicament puisse être prescrit par les médecins et éventuellement remboursé par l'assurance maladie.

d. Phase IV

Ces études succèdent à la mise sur le marché d'un médicament pour une indication donnée. Elles peuvent permettre de préciser l'activité d'un médicament dans un groupe particulier de sujets (personnes âgées, enfants, etc.), l'efficacité (effectiveness), certains effets indésirables rares, les modalités à grande échelle de prescription et d'utilisation.
En revanche, tout essai d'un médicament réalisé après l'obtention de l'AMM dont l'objectif est d'obtenir **une extension** de celle-ci (nouvelle indication, nouvelle posologie, etc.) est **un essai de phase II ou III**. Il faut alors considérer le produit comme un nouveau médicament.

3 LES ESSAIS THERAPEUTIQUES COMPARATIFS : DE LA NECESSITE ET DE LA DIFFICULTE DE COMPARER DES THERAPEUTIQUES

Pour un médecin et son patient, l'objectif d'un traitement est d'obtenir une amélioration de la qualité ou de la quantité de vie. Les objectifs précis sont variables selon que la pathologie à traiter est bénigne (raccourcissement du délai de guérison, diminution de la douleur, etc.) ou grave (prolongation de la survie, prévention des récidives ou des complications, etc.). L'atteinte de cet objectif passe par une phase préalable d'**évaluation expérimentale du traitement** aussi bien en termes d'efficacité que de tolérance (rapport bénéfice-risque). Or, l'observation de « guérisons » ou d'améliorations sous traitement ne suffit pas pour attribuer la causalité de la guérison au traitement : en effet, une grippe guérit presque toujours spontanément, c'est son évolution naturelle. **Il est donc nécessaire pour établir la causalité entre un traitement et un bénéfice clinique de comparer un groupe de sujets soumis à ce traitement à un groupe de sujets comparables non soumis à ce traitement**. Cette comparaison comporte des difficultés pour différentes raisons qui sont énoncées ci-après.

1. Concept de variabilité

Des patients

Même atteints d'une même maladie, les patients sont le plus souvent très différents les uns des autres. Cette **variabilité** concerne aussi bien la sémiologie, le pronostic, la sensibilité au traitement, etc.

Des observateurs

Dans l'évaluation des critères pour juger de l'efficacité d'un traitement appelés critères de jugement, intervient la variabilité des observateurs (intra-observateur ou inter-observateur), le plus souvent quand le critère est l'évaluation d'un signe subjectif (douleur, etc.) ou d'une mesure (prise de tension artérielle, **etc.)**

2. Rôle de la suggestion

La connaissance par le patient ou par le médecin de la nature du traitement pris par le patient risque d'instaurer des différences de comportement du médecin et/ou du patient selon le groupe auquel il appartient. Par exemple, un patient sachant qu'il prend un placebo va peut-être sous-évaluer l'effet antalgique du produit, ou un médecin sachant que son patient prend un médicament ayant des effets adverses potentiellement important va avoir tendance à instaurer une surveillance plus soutenue. C'est le rôle de la suggestion qui risque donc **d'introduire entre les deux groupes une différence de suivi et de compromettre leur comparabilité** même si elle était acquise initialement.

3. Choix des critères d'efficacité

Ils n'ont **pas tous la même pertinence** et il peut être tentant de choisir un critère facile à mesurer, reproductible mais qui ne reflète pas le meilleur bénéfice clinique possible. Par exemple, en cancérologie, le critère le plus pertinent pour juger de l'efficacité d'un traitement est la survie à long terme mais la mesure de ce critère impose des contraintes de temps, de long suivi des patients et de difficulté à recueillir ce critère. Il serait tentant de choisir comme critère la présence d'une réponse complète mesurée sur l'imagerie. Or on sait bien que la réponse complète n'est pas un critère superposable à la survie car certains patients rechutent.

Dans l'absolu, **ils peuvent être multiples**. Rares sont les pathologies dont l'évolution se mesure par un critère unique. Il peut être tentant de les multiplier en arguant le fait que *« qui peut le plus peut le moins »*. Mais ce raisonnement entraîne une multiplication de tests statistiques qui majorent les risques de conclure à tort à l'efficacité d'un traitement. On s'attachera donc à choisir **un critère de jugement principal** qui sera utilisé pour répondre à la question principale et des critères de jugement secondaires.

4. Nécessités éthiques

Elles sont au premier plan dans la mesure où l'essai thérapeutique est une expérimentation chez l'homme. La nécessité éthique est donc de *« tout faire pour arriver au mieux à la conclusion correcte »*.

4 PRINCIPES METHODOLOGIES DE L'ESSAI THERAPEUTIQUE COMPARATIF

Ces difficultés ont conduit à établir des règles méthodologiques strictes à respecter dans les essais thérapeutiques comparatifs.
La première difficulté évoquée concerne la variabilité des patients. Cette remarque met en exergue la nécessité de définir précisément la population à laquelle seront extrapolés les résultats de l'essai, dite **population cible**. En effet, pour une même pathologie, la prise en charge thérapeutique sera variable en fonction de certaines caractéristiques du patient (âge, sexe, etc.), de certains facteurs pronostiques (stade, forme clinique, etc.). Il convient donc de prévoir le recrutement de patients de façon à ce qu'ils soient **représentatifs** de cette population cible, à travers la définition du **mode de recrutement** (ambulatoire, hospitalier, etc.) et de **critères d'inclusion** concernant la pathologie et les patients. Des **critères de non-inclusion** sont également définis et sont généralement liés à des impossibilités de suivre le traitement de façon optimale (contre-indication aux traitements testés, mauvaise observance supposée, etc.).
Une fois la population définie et le recrutement des patients planifié, l'objectif de la comparaison de l'effet du traitement entre le groupe « traité » (qui reçoit le traitement objet de la recherche) et le groupe contrôle ou groupe témoin (qui reçoit le traitement de référence ou le placebo) va édicter un certain nombre de règles méthodologiques à respecter pour permettre d'émettre un jugement à trois niveaux :

- Jugement de **signification statistique** ;
- Jugement de **signification clinique** ;
- Jugement de **causalité**

Ce sont les jugements de signification clinique et de causalité qui permettent de conclure à l'efficacité d'un traitement mais l'étape statistique est un préalable indispensable.

1. Jugement de signification statistique

On répond à la question : « la différence est-elle réelle ? ».
C'est le test statistique qui permet d'y répondre.

Principe du test statistique ou test d'hypothèse

Il existe de nombreux tests statistiques choisis en fonction de la nature des paramètres recueillis moyenne, pourcentage, courbe de survie, etc.) et de leur distribution. Ils reposent tous sur le même principe (cf Chapitre Notions statistiques de base).
Le test statistique va permettre de tenir compte du rôle du hasard dans les différences qui seront observées entre les deux traitements. Le rôle du hasard (fluctuations d'échantillonnage) sera d'autant plus difficile à éliminer que la différence entre les deux traitements sera petite, que les effectifs seront faibles, que la variabilité des caractéristiques mesurées sera grande.
Il repose sur la formulation d'une hypothèse, dite hypothèse nulle que l'on va chercher à rejeter au bénéfice d'une hypothèse alternative. L'hypothèse nulle est l'hypothèse d'absence de différence entre les deux groupes à comparer. Rejeter l'hypothèse nulle, c'est admettre l'existence d'une différence significative entre les deux groupes.
Pour chaque test, on calcule un « paramètre » tenant compte des différences observées, de la variabilité des mesures, des effectifs des échantillons de l'étude. Ce paramètre est comparé à une table statistique standardisée et permet de déterminer si la différence observée est suffisamment « grande » pour être expliqué par autre chose que le hasard.

Conclusion du test statistique : valeur du « p »
Cette quantification du hasard dans les différences observées est fournie par la valeur du « p » : plus p est petit, moins le hasard peut expliquer les différences observées, donc plus ces différences peuvent s'expliquer par autre chose que le hasard. Un consensus scientifique existe pour considérer qu'au-delà de 5 %, on considère que le hasard est potentiellement trop important pour expliquer les différences observées et on préfère conclure que la différence n'est pas significative. Pour les valeurs de « d » inférieures à 5 %, la différence est dite significative

Le risque d'erreur
La décision prise à l'issue d'un test statistique est entachée de deux risques d'erreur possibles.
Lorsqu'on conclut à **une différence significative** (rejet de l'hypothèse nulle), le risque est de conclure à tort à une différence. C'est le risque de première espèce, α, que l'on fixe toujours *a priori* (le consensus scientifique fait qu'on le fixe généralement à 5 %).
Lorsqu'on ne conclut pas à une différence significative (non-rejet de l'hypothèse nulle), le risque est de passer à côté d'une différence réelle. C'est le risque de deuxième espèce, β, qu'on appelle aussi manque de puissance du test. 1-β est la puissance du test, c'est-à-dire la capacité du test à mettre en évidence une différence qui existe.
On peut conclure à un rejet de l'hypothèse nulle (différence significative) ou à un non-rejet de l'hypothèse nulle (absence de différence significative) en l'absence de connaissance du risque β. Dans ce cas-là, la seule conclusion informative est la conclusion de différence significative. Ce n'est pas parce qu'on conclut à une différence non significative, qu'il n'y a pas de différence. L'absence de preuve n'est pas la preuve de l'absence.

Le calcul du nombre de sujets nécessaires
Si l'on ne contrôle pas le risque β, et si la différence est non significative, l'essai ne permettra pas de trancher entre « absence de différence » ou « manque de puissance pour mettre en évidence une différence ». Si l'on veut minimiser le risque β (par exemple à 5 ou 10 %) ou assurer au test une puissance suffisante (90 ou 95 %), il est nécessaire de mettre des contraintes en termes d'effectif de sujets.
C'est pourquoi on calculera dans tout essai thérapeutique comparatif un nombre de sujets nécessaires pour mettre en évidence une différence jugée « cliniquement intéressante » qu'il faudra déterminer, tout en minimisant les deux risques d'erreur α et β (cf chapitre notions statistiques de base).

Plus une différence est « importante réellement », plus le nombre de sujets nécessaires pour le mettre en évidence sera faible.

2. Jugement de signification clinique

On répond à la question : « la différence est-elle cliniquement intéressante ? ».
On trouve donc ici le concept de différence « cliniquement intéressante », c'est-à-dire qui apporte un bénéfice au patient par rapport à un traitement de référence. Une différence très modeste entre deux paramètres pourra être statistiquement significative au prix de l'inclusion d'un nombre très élevé de patient ; en revanche, une différence importante cliniquement pourra ne pas être statistiquement significative en raison d'un effectif trop faible (*cf. Infra*).
C'est la connaissance de la pathologie qui permet de déterminer ce qu'est une différence « cliniquement intéressante ». Par exemple, pour une pathologie très grave, un bénéfice même modeste peut être intéressant pour le patient.

3. Jugement de causalité

On répond à la question : « la différence est-elle imputable au seul traitement ? »
Si une différence est statistiquement significative, cela n'implique pas obligatoirement que la différence soit liée au seul traitement. Encore faut-il s'assurer que l'essai a été réalisé dans des conditions permettant d'affirmer que les deux groupes à comparer n'étaient différents que par le traitement à tester : c'est le principe de la méthode expérimentale.

Principe de la comparabilité initiale des groupes
On doit tout faire pour éviter que s'introduise entre les deux groupes une différence systématique concernant les caractéristiques des malades, qu'elles soient connues ou non. Pour effectuer la répartition des malades entre deux groupes, il faut éviter les procédures telles que :

- Groupe traité parmi les patients d'un hôpital et groupe témoin parmi ceux d'un autre hôpital ;
- Groupe traité parmi les patients dont le nom commence par A à L et groupe témoin parmi ceux dont le nom commence par M à Z ;
- Groupe traité parmi les patients vus du lundi au mercredi et groupe témoin parmi ceux vu de jeudi au samedi etc.

Des procédures de ce type introduisent la possibilité d'une différence systématique (appelée biais) entre les deux groupes comme la différence de recrutement de deux hôpitaux ou la différence d'origine ethnique selon le nom ou la différence de catégorie socioprofessionnelle liée au jour de la consultation. De la même manière, le jugement du médecin peut entraîner un biais même s'il est inconscient.
La seule procédure acceptable pour obtenir la comparabilité initiale des groupes sur l'ensemble des caractéristiques du patient, connues ou inconnues, est l'attribution aléatoire des traitements aux patients ou **randomisation**.
Ainsi la règle d'or de tout essai thérapeutique comparatif est d'être randomisé.

Principe de maintien de la comparabilité des groupes au cours de l'essai
Même si un essai est parfaitement randomisé et si les deux groupes ont toutes les chances d'être comparables en début d'essai, la comparaison finale peut être faussée parce que la comparabilité n'a pas été maintenue tout au long de l'essai. *Plusieurs raisons peuvent mettre en péril le maintien de la comparabilité.*

La connaissance du groupe dans lequel est randomisé le patient par le patient et/ou par le médecin peut, par l'effet de la *suggestion*, entraîner des différences de suivi de la part du médecin ou d'évaluation de la part du médecin ou du patient notamment lorsque le critère de jugement est subjectif. *Pour éviter cet écueil* :

- *Il est fondamental* que le groupe témoin reçoive un « traitement » même si ce traitement n'est pas pharmacologiquement actif : l'abstention médicamenteuse n'est pas l'abstention thérapeutique, d'où la **notion de placebo**. Un placebo est une substance dénuée d'effet pharmacologique mais qui peut cependant être à l'origine, par exemple par suggestion psychique, d'une amélioration de l'état de santé ou du bien-être du malade (effet placebo), ou encore d'une manifestation indésirable (effet nocebo). L'utilisation d'un placebo dans le groupe témoin permet de mesurer l'importance des effets induits par la prescription, son environnement ou le fait de participer à une recherche. Ainsi les deux groupes de patients seront-ils à égalité quant à l'effet « psychique » du traitement :
- *Il est fondamental*, dans la mesure du possible et pour éviter que la suggestion n'entache l'évaluation de l'effet du traitement, de travailler **en « aveugle » ou en « insu ».** Dans l'essai en **simple aveugle**, les patients ignorent lequel des deux traitements ils reçoivent, mais le médecin évaluateur et le personnel soignant le savent. Dans l'essai en **double aveugle**, ni les patients, ni le médecin évaluateur, ni le personnel soignant ne connaissent lequel des traitements les patients reçoivent. La pratique d'un essai en aveugle implique donc que les traitements soient parfaitement identiques en termes d'apparence, de goût, etc. C'est ainsi que les laboratoires pharmaceutiques fabriquent spécifiquement des placebos identiques aux médicaments testés. Lorsque les traitements à comparer n'ont pas la même forme galénique (par exemple comprimé *versus* gélule) ou la même voie d'administration (voie orale *versus* voie injectable), la technique du double placebo est utilisée : le groupe « traitement de référence » reçoit le traitement de référence et un placebo du traitement à tester et le groupe « traitement à tester » reçoit le traitement à tester un placebo du traitement de référence. Il va de soi que dès lors qu'on utilise un placebo pour le groupe témoin, il faut travailler en aveugle et de préférence en double aveugle ;
- *Il est fondamental* que les médecins évaluateurs et tout le personnel impliqué dans le suivi des patients travaillent de façon homogène. A cette fin, une **standardisation des procédures** d'évaluation et de suivi, une **formation préalable** des personnes impliquées et une **surveillance régulière** de déroulement de l'essai sont nécessaires. Ceci est particulièrement vrai dans les essais **multicentriques**.

Les **écarts au protocole** peuvent être de plusieurs natures :

- Les **inclusions à tort**, c'est-à-dire le non-respect des critères d'inclusion ou d'exclusion ;
- Le **non-respect strict de la prise du traitement** (suspension temporaire ou interruption du traitement par le médecin, non-compliance du patient) à cause d'effets indésirables, de contraintes jugées trop importantes ;
- Le **non-respect du suivi et du recueil des critères d'évaluation** aboutissant à une absence de données.

Tous ces écarts peuvent faire remettre en cause le bien-fondé de la conservation des patients concernés dans l'analyse de l'essai et il est tentant de les en exclure pour ne conserver dans l'analyse que les patients ayant respecté scrupuleusement le protocole : c'est **l'analyse en « per protocole ».** Cette attitude est à proscrire pour la raison essentielle qu'elle abolit l'intérêt de la randomisation initiale en introduisant potentiellement une différence entre les deux groupes qui seront analysés. *le biais sera d'autant plus important que ces écarts au protocole seront nombreux et liées au type de traitement reçu* ; par exemple, la non-observance peut être plus importante avec un traitement présentant des effets indésirables marqués*. A contrario*, dans une maladie chronique l'absence

d'efficacité ressentie par le patient prenant un placebo pourra être un facteur de mauvaise observance.
La seule attitude à adopter est le principe de **l'analyse en intention de traiter** : ***tous les patients randomisés seront analysés dans leur groupe de randomisation*** même s'ils n'ont pas suivi le traitement correctement, voire même s'ils ont suivi le traitement de l'autre groupe. Cela suppose que l'on recueille *le critère de jugement pour tous les patients, donc que l'on évite au maximum d'avoir des perdus de vue*.
L'analyse en intention de traiter sert à maintenir la comparabilité initiale des groupes mais ne résout pas tous les problèmes. S'il y a beaucoup d'écarts au protocole, malgré l'analyse en intention de traiter, les résultats et la puissance de l'essai seront affectés. S'il y a beaucoup de perdus de vue, l'analyse en intention de traiter perd son bénéfice et n'est même plus réalisable *stricto sensu*.
C'est pourquoi les écarts au protocole doivent être évités autant que faire se peut (limitation des inclus à tort et des perdus de vue, etc.) tout en gardant en mémoire la notion de bénéfice-risque pour le patient et que tout évènement indésirable grave justifie l'arrêt du traitement.

5 COMMENT REPRESENTER L'EFFET D'UN TRAITEMENT ? LES INDICES D'EFFICACITE

1. Critère de jugement binaire

Lorsque le résultat du traitement se mesure en termes de survenue d'un évènement (guérison/non-guérison, rémission/non-rémission, décès/non-décès, etc.), les indices d'efficacité quantifient l'efficacité à partir des modifications qu'il induit dans la fréquence de survenue de cet évènement. Ils expriment la « distance » entre les **risques absolus** observés dans le groupe expérimental (R_e) et le groupe témoin (R_t).

	Critère de jugement	
	Evènement (effet -) Ex : décès	Pas d'évènement (effet +) Ex : non-décès
E (expérimental	a	b
T (témoin)	c	d

R_E = a/a + b = incidence du critère de jugement dans le groupe expérimental
R_T = c/c + d = incidence du critère de jugement dans le groupe témoin

Indices d'effet multiplicatif

- Risque relatif (cf. chapitre épidémiologique)
  $RR = R_E / R_T$

- Réduction relative du risque
  $RRR = | R_E - R_T| / R_T = |1 – RR| \times 100\ \%$

Indices d'effet additif

- Différence des risques ou réduction absolue du risque (cf. excès de risque dans le chapitre épidémiologie)

RAR = | R_E - R_T|

- Nombre de sujets nécessaires à traiter
  NST = 1 / RAR

Remarque : tous ces indices sont des estimations donc doivent être assortis de leur intervalle de confiance.

Exemple d'expression des résultats de l'essai : Ohkubo Y, Kishikawa H, Araki E **et al**. Intensive insulin therapy prevents the progression of diabetic microvascular complications in Japanese patients with non+-insulin-dependent diabetes mellitus : a randomized prospective 6-years study. **Diabetes Res Clin Pract** 1995. May ; 28 : 103-17

Critère de jugement	RE	RT	RRR	RAR	NST
Aggravation de la rétinopathie diabétique	13 %	38 %	66 %	25 %	4

Source : EBM édition française
– Plusieurs années de traitement intensif réduisent de 38 à 13 % le pourcentage de patients dont la rétinopathie s'aggrave.
– **RRR** (réduction relative du risque) : c'est la réduction proportionnelle du taux d'événements défavorables entre le groupe expérimental et le groupe témoin de l'essai, calculée par la formule $|(R_E - R_T)|/ R_T$, et accompagnée de son IC à 95 %. Dans l'exemple de l'aggravation d'une rétinopathie diabétique, $|(R_E - R_T)|/ R_T = |13\ \% - 38\ \%|/38\ \% = 66\ \%$.
– **RAR** (réduction absolue du risque) : c'est la différence arithmétique absolue entre les taux d'événements défavorables dans le groupe expérimental et dans le groupe témoin de l'essai, calculée par la formule $|R_E - R_T|$. Dans le présent exemple, $|R_E - R_T| = |13\ \% - 38\ \%| = 25\ \%$.
– **NST** (nombre de sujets à traiter) : c'est le nombre de patients qu'il faut traiter pour obtenir 1 résultat favorable supplémentaire, calculé par la formule 1/RAR, et accompagné de son IC à 95 %. Dans le présent exemple, $1/RAR = 1/25\ \% = 4$.

2. Critère de jugement quantitatif

Lorsque le résultat du traitement se mesure par un critère continu, on peut exprimer l'efficacité en termes de différence des valeurs moyennes obtenues sous traitement expérimental et sous traitement contrôle. Elle peut être exprimée soit en différence absolue soit en différence relative. On peut également ramener le critère de jugement à un critère binaire par rapport à un seuil mais il faut argumenter le choix de ce seuil qui doit être validé. On se retrouve ainsi dans le cas de figure précédent (tableau 3.I).

Différence absolue en fin d'essai
On calcule la différence des mesures en fin d'essai entre les deux groupes : dans notre exemple, on peut ainsi dire que le traitement entraîne une diminution moyenne absolue de la PAS de 20 mmHg.

Tableau 3.I. Exemple d'expression des résultats d'un essai thérapeutique ayant pour critère de jugement la pression artérielle systolique (PAS)

Critère de jugement PAS (mmHg)	Valeur de base (avant traitement)	Valeur moyenne au moment de l'évaluation (après traitement)	Différence avant-après	Différence absolue	Différence des différences avant-après	Différence relative
Groupe contrôle	160	150	– 10	– 20	– 22	(130 – 150)/ 150 = – 13 %
Groupe expérimental	162	130	– 32			

Toutes les valeurs du tableau sont des valeurs moyennes qu'il faut assortir de leur intervalle de confiance.

Cette expression de l'effet du traitement suppose une bonne comparabilité des groupes sur la mesure de base.

Différence absolue des différences avant-après

On calcule la moyenne des différences avant-après dans chaque groupe et on compare ces moyennes entre les deux groupes. Ceci permet de tenir compte de la valeur de base qui n'est pas obligatoirement identique entre les deux groupes même si la randomisation a été correcte. Dans notre exemple, la diminution de PAS a été plus importante en moyenne de 22 mmHg dans le groupe expérimental.

Différence relative en fin d'essai

On calcule la différence des moyennes des critères mesurés en fin d'essai par rapport à la moyenne du groupe contrôle : $(m_e - m_c) / m_c$: dans notre exemple, le traitement expérimental entraîne une diminution relative de PAS de 13 % en moyenne par rapport au groupe contrôle.

D'une manière générale, les critères de jugement continus apportent plus d'information donc permettent plus facilement de mettre en évidence des différences significatives. Le piège est que ces différences « statistiques » ne soient pas pertinentes cliniquement. Ils sont plus souvent sujets à des difficultés d'interprétation.

ETUDES DE COHORTES

Une cohorte est une enquête ***longitudinale*** (c'est-à-dire qui comporte un suivi au cours du temps). Les participants, groupe d'individus de caractères définis (âge, sexe,…) échantillonnés à partir d'une *population source* sont suivis au cours du temps pour mesurer l'évolution de caractéristiques prédéfinies (cliniques, paracliniques,..). L'objectif d'une étude de cohorte peut être simplement *descriptif* (incidence d'une maladie, description de l'histoire naturelle d'une affection), *étiologique* (recherche d'un lien entre un facteur d'exposition et une maladie), ou *évaluatif* (exploration de l'effet d'une intervention). La ***durée du suivi*** d'une étude de cohorte doit être cohérente avec son objectif (ex. délai naturel entre exposition et maladie).

Une cohorte ***descriptive*** peut être relative à des données de statistiques de santé d'une population : description des principaux indices de mortalité ou de morbidité d'une population au fil du temps ou statistiques d'exposition à certains facteurs de risque, en particulier en milieu professionnel. Il peut s'agir d'enquêtes de cohortes spécifiques : de natalité ou de mortalité par cause et par catégorie socio-professionnelle. Il peut s'agir d'études des variations temporelles ou spatiales d'indices de morbidité ou de mortalité telles des variations saisonnières, des tendances décennales ou séculaires ; ou encore des études de variations géographiques et spatiales, d'études d'agrégats spatio-temporels (ou clusters). Une étude de cohorte peut décrire l'incidence d'une maladie, son histoire naturelle, ses complications, son évolution. L'incidence peut être exprimée comme la proportion de personnes qui développent la maladie (***incidence cumulée***) ou sous forme de taux par personne-temps de suivi (***taux d'incidence***). Des termes spécifiques sont utilisés pour décrire différents types d'incidences : entre autres, le taux de mortalité, le taux de natalité, le taux d'attaque ou le taux de létalité. La population étant issue d'un échantillonnage, les estimateurs d'incidence doivent être présentés avec leur intervalle de confiance à 95%.

Les enquêtes de cohortes ***étiologiques*** (appelées aussi ***analytiques***) ont pour objet de rechercher une association entre un facteur d'exposition donné et une maladie. Il peut s'agir d'identifier des facteurs de risque d'une maladie en population générale (ex. relation entre tabac et cancer bronchique ou entre traitement hormonal substitutif (THS) et maladies cardiovasculaires) ou d'explorer des éléments pronostiques ou prédictifs d'une affection dans une population de malades (ex. : pronostic d'une maladie de Hodgkin selon la présence de ganglions au-dessus et en dessous du diaphragme). Les cohortes étiologiques sont de ***type exposés - non exposés***.

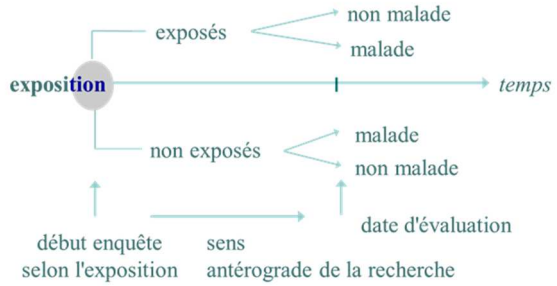


Figure. Principe d'une étude exposé / non exposé

Le début de l'enquête de cohorte étiologique est défini par la présence ou l'absence d'une exposition. Les participants sont suivis au fil du temps jusqu'à la date d'évaluation ou date de point. A cette date

on évalue la proportion de participants qui présentent ou non la maladie étudiée en fonction de l'exposition ou de son absence (ou entre des groupes de personnes présentant différents niveaux d'exposition). Les investigateurs peuvent évaluer plusieurs critères de jugement différents, et examiner l'exposition et les critères de jugement à plusieurs moments au cours du suivi. Les études étiologiques explorent des relations de causalité. Leur objectif est la comparaison du risque de survenue de l'évènement d'intérêt, entre les exposés et non exposés au facteur étudié. Ces études de cohorte estiment souvent un « ***risque relatif*** » (RR), qui peut représenter des ratios de risques (ratios d'incidence cumulée), ainsi que des ratios de taux (ratios de taux d'incidence). Le RR, correspondant ainsi au rapport de l'incidence chez les exposés sur l'incidence chez les non exposés, représentera le facteur par lequel est multiplié le risque basal de survenue d'une maladie lorsqu'un facteur d'exposition est présent.

RR = (Incidence chez les exposés) / (Incidence chez les non exposés) = I_e / I_{ne}

Des enquêtes de cohortes peuvent être appelées ***évaluatives*** lorsqu'elles portent sur l'analyse observationnelle de l'efficacité d'une intervention thérapeutique (le groupe exposé à l'intervention est comparé au groupe non exposé). Ce type d'étude peut être envisagé lorsqu'un essai thérapeutique randomisé ne peut pas être mis en œuvre.

Si les patients sont inclus puis suivis au cours de l'étude, il s'agit d'une cohorte ***prospective*** (ex. un groupe de femmes traitées par radiothérapie pour traitement de cancers gynécologiques et un groupe sans radiothérapie sont suivis chaque année pendant 10 ans pour étudier la survenue de leucémies). Lorsque l'étude consiste à analyser des données de suivi dans le temps qui ont déjà été collectées, il s'agit d'une étude de cohorte ***historique*** (ex. Etude dans la base de données de la sécurité sociale du lien entre exposition au Benfluorex (Mediator®) et la survenue d'une valvulopathie cardiaque chez des patients diabétiques entre 2006 et 2010)

On distingue des cohortes fermées et des cohortes ouvertes. Les cohortes ***fermées*** (les cohortes de naissance, par exemple) incluent un nombre défini de participants au début de l'étude suivis à partir de ce moment, souvent à intervalles donnés, jusqu'à une date de fin d'étude déterminée. Dans les cohortes ***ouvertes***, la population de l'étude est dynamique : les sujets peuvent entrer et sortir de la cohorte à différents moments au cours du temps (exemple : le suivi de la cohorte composée des habitants du comté de Framingham aux USA). Les cohortes ouvertes changent en fonction des décès, des naissances ou des migrations, mais la composition de la population, au regard de variables comme l'âge et le sexe, peut rester à peu près constante, en particulier sur une courte période.
Dans une cohorte fermée, les incidences cumulées et les taux d'incidence peuvent être estimés. Lorsque des groupes de sujets exposés et non exposés sont comparés, cela conduit à des estimations de ratios d'incidences cumulées ou à des ratios de taux d'incidence. Pour les cohortes ouvertes, on estime seulement des taux d'incidence et des ratios de taux d'incidence.

Dans l'étude de cohorte, les ***procédures de suivi*** pour tous les participants doivent être décrites, y compris les procédures minimisant les ***perdus de vue***. Une perte de vue se produit lorsque des participants se retirent d'une étude avant cette date. Le taux de perdus de vue est un critère de qualité d'une étude de cohorte. Le nombre de participants qui ont été perdus de vue doit être indiqué. En plus d'une perte de puissance, les perdus de vue entrainent un ***biais de sélection*** et gêner la validité d'une étude si la perte de vue survient de façon sélective chez les individus exposés, ou chez des personnes à risque élevé de développer la maladie. On parle alors de « *censure informative* ».
Lors du suivi, la surveillance doit être identique dans les deux groupes. Dans des études où les procédures de suivi différent entre les groupes exposés et non-exposés, il faut anticiper un biais qui

pourrait être dû à l'inégale détermination des évènements ou à des différences concernant les perdus de vue par exemple. Un biais de réponse est un autre type de biais de sélection qui se produit si les différences de caractéristiques entre ceux qui répondent et ceux qui refusent la participation à une étude affectent les estimations de l'incidence d'une affection ou son évolution. En général, un biais de sélection affecte la *validité interne* de l'étude. Ce point est différent des problèmes qui peuvent survenir lors de la sélection des participants qui affecte sa *validité externe* (*i.e.* sa *généralisabilité*), plutôt que sa validité interne.

La façon dont les expositions, les facteurs de confusion et les critères de jugement ont été mesurés affecte la fiabilité et la validité d'une étude. Une *erreur de mesure* ou une *classification erronée* des expositions ou des critères de jugement peuvent rendre plus difficile la détection des relations de cause à effet, ou peuvent engendrer des relations trompeuses. Ce sont les ***biais de mesure***, appelés aussi biais de classement ou d'information. Il est donc utile de rendre compte de la validité ou de la fiabilité des évaluations ou des mesures, y compris en fournissant des détails sur le gold standard qui a été utilisé. En outre, il est important de savoir si les groupes comparés diffèrent selon la façon dont les données ont été recueillies. Ainsi, les groupes comparés doivent bénéficier d'un ***recueil standardisé et identique des informations, avec des outils validés, par des enquêteurs formés et si possible en insu (aveugle) de l'état d'exposition du sujet***. Par exemple, si les transplantés ont des examens cutanés plus réguliers et plus approfondis que les dialysés, alors la mise en évidence d'un cancer de la peau sera plus complète chez les transplantés. Des patients recevant un médicament qui provoque des maux d'estomac non spécifiques peuvent être soumis plus souvent à une fibroscopie gastrique et ainsi, présenter un nombre plus important d'ulcères détectés que des patients ne prenant pas ce médicament, même si ce médicament ne provoque pas plus d'ulcère. Ce type de biais de mesure est également appelé « biais de détection » ou « biais de surveillance ». Une façon d'évaluer son influence est de mesurer l'intensité de la surveillance médicale dans les différents groupes d'étude, et d'ajuster les résultats dans les analyses statistiques.

Comme dans toute étude étiologique observationnelle, la prise en compte des facteurs de confusion est cruciale. La prévention d'un ***biais de confusion*** peut être envisagée soit à la conception du protocole de l'étude ou soit à lors de l'analyse. A la conception du protocole de l'étude, les données identifiant les facteurs de confusion pour lesquels doivent être recherchées. A ce stade, il est possible de réaliser une *restriction* ou un *appariement*. Ainsi, on pourrait vouloir restreindre une étude aux femmes qui ne présenteraient pas un facteur de confusion donné, par exemple une pression artérielle élevée. On pourrait aussi le cas échéant apparier les patients exposés et non exposés sur leur niveau de pression artérielle. Lors de l'analyse, une *stratification* ou une *analyse multivariée* peuvent être utilisées pour réduire l'effet des facteurs de confusion. Même après toutes ces précautions, les résultats peuvent au final être encore biaisés par une ***confusion résiduelle***. C'est la confusion qui demeure après des tentatives infructueuses de la contrôler. Il faut envisager les facteurs de confusion résiduels dus à des variables non mesurées ou à des mesures imprécises des facteurs de confusion. Par exemple, le statut socioéconomique est associé à de nombreux critères de jugement de santé et diffère souvent entre les groupes comparés. Les variables utilisées pour mesurer les statuts socioéconomiques (revenus, éducation, ou profession) sont des substituts à d'autres expositions non définies et non mesurées, et le facteur de confusion vrai sera, par définition, mesuré avec une erreur.

L'***interprétation des résultats*** d'une étude doit tenir compte des facteurs de confusion, des résultats des analyses de sensibilité pertinentes, de la multiplicité des tests et des analyses en sous-groupes. Pour guider la réflexion et les conclusions sur la *causalité*, les critères de Bradford Hill sont utiles.

Le ***Risque Relatif*** mesure la force de l'association entre une exposition et une maladie. Il représente le facteur par lequel est multiplié le risque basal de survenue d'une maladie lorsqu'un facteur d'exposition est présent. Si le RR est égal à 1, il n'y a pas de lien entre l'exposition et la maladie ; si le RR est supérieur à 1, l'exposition est un facteur de risque ; si le RR est inférieur à 1, l'exposition est un facteur protecteur. Si le risque relatif est éloigné de 1, il est moins probable que l'association soit due à un biais de confusion. Les effets relatifs ou les associations ont tendance à être plus cohérents entre études et entres populations que les mesures absolues. Le Hazards Ratio des études de survie s'interprète de la même manière que le RR.

Le calcul du RR est une estimation, il doit toujours être présenté avec son ***intervalle de confiance à 95%*** (IC 95%). Si l'IC 95% ne contient pas la valeur 1, alors le lien entre le facteur et la maladie est statistiquement significatif au risque de 5% et on peut interpréter le RR. En revanche, si l'IC 95% contient la valeur 1, alors le facteur n'est pas statistiquement lié à la maladie au risque de 5%. Dans ce cas, il y a deux cas de figure : i) soit la puissance statistique est suffisante (respect du NSN calculé *a priori*) et on pourra conclure que l'exposition n'est pas associée à la maladie (en l'absence de biais) ; ii) soit la puissance statistique est insuffisante ou inconnue et on ne pourra uniquement conclure qu'il n'est pas mis en évidence d'association statistiquement significative entre l'exposition et la maladie (on ne peut pas conclure à une indépendance).

Dans de nombreux cas, le risque absolu associé à une exposition est plus intéressant que le risque relatif. Basées sur le risque absolu, ***les mesures d'impact*** telles que le *risque attribuable* ou la *fraction attribuable* dans la population peuvent être utiles pour évaluer la façon dont beaucoup de cas pourraient être évités si l'exposition était supprimée. Des hypothèses fortes sont faites dans ce cadre, y compris une relation causale entre le facteur de risque et la pathologie étudiée.

- Le ***Risque Attribuable*** (RA) (appelé aussi *« excès de risque »*) exprime la variation du risque absolu due à la présence du facteur d'exposition.

  RA = (Incidence chez les exposés) – (Incidence chez les non exposés) = I_e - I_{ne}

  Le RA fournit une estimation du nombre de cas qui pourraient être évités si l'exposition était supprimée. Ex. Association entre contraception oestroprogestative (COP) et maladie thrombo-embolique veineuse (MTEV) : Incidence en présence de COP =15/100 000 femmes-années, Incidence en l'absence de COP =5/100 000 femmes-années, RR=3, Risque attribuable=10/100.000 femmes-années. Ainsi, dans une population de 100.000 femmes sous COP on aura 10 MTEV attribuables à la COP en plus, par rapport à 100.000 femmes non exposées. Si on supprime la COP chez 100.000 femmes pendant un an, on évite 10 MTE.
- Le ***Pourcentage de Risque Attribuable dans la population (PRA)*** (appelé aussi *« fraction attribuable »*) correspond à la proportion de cas dans la population imputables au facteur de risque, c'est-à-dire la proportion de cas que l'on éviterait dans la population générale si on éradiquait le facteur de risque.

  Le PRA dépend à la fois du **RR** et de la fréquence du facteur de risque dans la population, notée **FR** ici. Ex. Association entre tabac et cancer de vessie : RR = 2,6 et il y a 50% de fumeurs dans la population (FR=0,5). Le PRA est égal à 0,44. Ainsi, dans la population, 44% des cancers de vessie sont dus au tabac.

  $$PRA = \frac{I_{pop} - I_{ne}}{I_{pop}} = \frac{FR \times (RR-1)}{1 + FR \times (RR-1)}$$
- La ***Fraction Etiologique du Risque*** correspond au PRA chez les exposés (*i.e.* si FR=100%)

  FER = (RR-1) / RR

  Il correspond à la proportion de cas exposés imputable au facteur de risque. Dans l'exemple précédent, FER=1,6/2,6=61,5%. Ainsi, chez les fumeurs, 61,5% des cancers de vessie sont dus au tabac.

La ***généralisabilité des résultats*** d'une étude, appelée aussi *validité externe*, indique dans quelle mesure les résultats d'une étude peuvent être appliqués à d'autres circonstances. Les résultats

peuvent-ils être appliqués à des populations qui diffèrent de ceux inclus dans l'étude en ce qui concerne l'âge, le sexe, l'origine ethnique, la gravité de la maladie, la nature et le niveau d'exposition ou les définitions des critères de jugement ? Est-ce-que les données qui ont été recueillies dans des études longitudinales de nombreuses années auparavant sont toujours d'actualité ? Les résultats provenant des services de santé d'un pays sont-ils applicables aux systèmes de santé d'autres pays ? Ainsi, il est essentiel d'obtenir des informations adéquates les sources ou les sites de recrutement (par exemple, les listes électorales, les cliniques ambulatoires, les registres de cancer ou les centres de soins tertiaires), les lieux de l'étude, les critères de sélection, les expositions et la manière dont elles ont été mesurées, la définition des critères de jugement, la période de recrutement et le suivi ou le taux de non-participation et la proportion de participants non exposés qui présente l'évènement d'intérêt. La connaissance du risque absolu et de la prévalence de l'exposition, qui varient souvent entre les populations, sont utiles lors de l'application des résultats à d'autres contextes et à d'autres populations.

Annexes

Avantages et inconvénients des études de cohorte versus cas-témoin

Enquête exposé / non exposé	Enquête cas / témoin
Avantages • Étude des expositions rares • Étude du FR sur plusieurs M simultanément • Calcul d'incidence, du RR • Moins de biais	Avantages • Étude des maladies rares • Étude des maladies à long délai d'apparition • Étude de plusieurs E simultanément • Durée courte (pas de suivi), coût moindre
Inconvénients • Étude longue et coûteuse (prospective) • Non adaptée aux maladies rares ou à long délai d'apparition • Problème de suivi (perdus de vue, changement d'exposition au cours du temps, modification des critères diagnostiques) • Biais de confusion	Inconvénients • Pas de calcul d'incidence (l'OR est une estimation du RR) • Biais de mesure (de mémorisation++, biais lié à l'enquêteur si pas d'aveugle) • Biais de sélection (représentativité des groupes, sélection des témoins) • Biais de confusion

Interaction

Une interaction existe lorsque l'association d'une exposition au risque de la maladie diffère en présence d'une autre exposition. L'effet d'une exposition peut être mesurée de deux façons: comme un risque relatif (ou rapport des taux) ou comme une différence de risque (ou la différence de taux). L'action conjointe de deux facteurs peut être caractérisée de deux façons : sur une échelle *additive*, en termes de différences de risques, ou sur une échelle *multiplicative*, en termes de risque relatif. L'utilisation du risque relatif conduit à un modèle multiplicatif, tandis que l'utilisation de la différence de risque correspond à un modèle additif. Il y a un consensus pour considérer que l'échelle additive, qui utilise des risques absolus, est plus appropriée pour la santé publique et pour la prise de décision clinique. Une distinction est parfois faite entre une «interaction statistique » qui peut être un écart soit à un modèle multiplicatif soit à un modèle additif, et une « interaction biologique » qui est mesurée par l'écart à un modèle additif. Toutefois, ni les modèles additifs ni les modèles multiplicatifs n'indiquent un mécanisme biologique particulier.

Il est aussi utile de comprendre comment l'effet conjoint de deux (ou plusieurs) expositions diffère de leurs effets distincts (en l'absence de l'autre exposition). Les données de l'étude sur les contraceptifs oraux et la mutation du facteur V Leiden illustrent cette problématique. Les contraceptifs oraux et la mutation du facteur V Leiden augmentent le risque de thrombose veineuse; leurs effets séparés et conjoints peuvent être représentés par des odds ratios avec comme valeur de référence (un OR de 1) représente la référence des femmes sans facteur V Leiden qui n'utilisent pas de contraceptifs oraux.

Données manquantes

Une approche fréquente de traitement des *données manquantes* est de restreindre les analyses à des personnes ayant des données complètes sur toutes les variables nécessaires requises pour une analyse particulière. Bien que ces analyses '*cas-complets*' ne soient pas biaisées dans de nombreuses circonstances, elles peuvent l'être et être ininterprétables. Un biais survient si des individus avec des données manquantes ne sont pas représentatifs de l'ensemble de l'échantillon. L'inefficacité résulte d'une taille réduite de l'échantillon à l'analyse. Une technique d'imputation des données manquantes

consiste à remplacer chaque valeur manquante par une valeur supposée ou estimée. Le choix des valeurs imputées doit reposer sur une méthodologie claire pour éviter d'atténuer ou d'exagérer l'association d'intérêt. Il existe une typologie des situations liées aux *données manquantes*, fondée sur un modèle de probabilité qu'une observation soit manquante. Il existe plusieurs méthodes pour traiter les données manquantes au hasard qui ne seront pas abordées ici compte tenu de leur complexité.

Analyses de sensibilité

Les *analyses de sensibilité* sont utiles pour déterminer si oui ou non les principaux résultats sont cohérents avec ceux obtenus avec des stratégies d'analyse ou des hypothèses différentes. Les questions qui peuvent se poser concernent notamment les critères d'inclusion dans les analyses, les définitions des expositions ou des critères de jugement, quelles variables de confusion nécessitent un ajustement, quelle modalité de traitement des données manquantes, l'existence de possibles biais de sélection ou biais de mesure dû à une mesure inexacte ou inappropriée d'une exposition, de la maladie ou d'autres variables, et les choix d'analyse spécifiques, tels que le traitement des variables quantitatives. Des méthodes sophistiquées sont de plus en plus utilisées pour modéliser simultanément l'influence de plusieurs biais ou de plusieurs hypothèses.

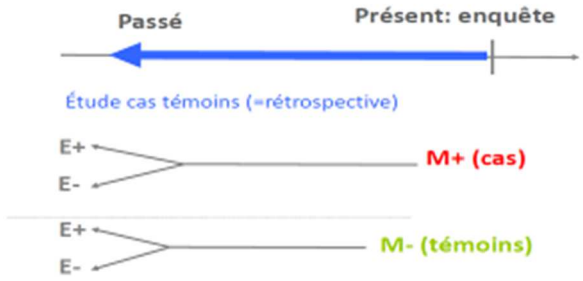
ETUDES CAS-TEMOINS

1 PRINCIPE DES ETUDES CAS - TEMOINS

Les études cas - témoins sont utilisées dans le cadre des recherches en épidémiologie étiologique (ou analytique). Elles ont pour principal objet d'explorer l'existence d'une relation entre une exposition à un facteur de risque et la survenue d'une maladie.

	Descriptif	Analytique/ étiologique	Évaluatif
Épidémiologie de population (population générale)	Répartition et fréquence d'une maladie	Recherche de facteurs de risque	Évaluation d'action s de santé publique (action de prévention, de dépistage …)
Epidémiologie clinique (patients)	Nosographie de la maladie	Recherche de facteurs pronostiques	Évaluation de: •Méthodes diagnostiques •Méthodes thérapeutiques

Il s'agit d'études basées sur la comparaison de la fréquence d'exposition à un facteur de risque (mesuré de façon rétrospective) entre des sujets atteints de la maladie (***les cas***) et des sujets indemnes de la maladie (***les témoins***). L'enquête commence donc ***après*** l'apparition de la maladie et bien entendu après l'exposition (enquêtes rétrospectives).



Remarque : Les études cas-témoins peuvent également être réalisées en ***épidémiologie clinique*** dans le cadre d'étude ***pronostique*** en comparant la fréquence d'exposition à un facteur pronostique entre des sujets malades ayant développé un évènement pronostique péjoratif (***les cas*** : par exemple des patients VIH + au stade SIDA) et des sujets malades n'ayant pas développé cet évènement pronostique péjoratif (***les témoins*** par exemple des patients VIH + n'étant pas au stade SIDA).

2 SELECTION DES CAS

Le groupe des cas est constitué de sujets malades devant être représentatif de l'ensemble des personnes atteintes de la maladie considérée.

- Source de recrutement des cas :
    - Tirage au sort sur une liste exhaustive de tous les cas : les registres nationaux de maladie (par exemple les registres de cancers). De tels registres n'existent cependant pas pour toutes les pathologies.
    - Recrutement hospitalier

- Définition précise de la maladie :
    - Définir les critères diagnostiques qui doivent être objectifs, standardisés pour tous les cas et validés (ex. diagnostic histologique)
    - Préciser le lieu et la période du recrutement des cas
    - Préciser l'ancienneté de la maladie : **prendre des cas incidents (nouvellement diagnostiqués**) **plutôt que des cas prévalents (file active de patients**). En effet, les cas prévalents sont les cas vivants au moment de l'inclusion, que la pathologie ait été diagnostiquée hier ou il y a 10 ans. Pour certaines pathologies, seules certaines formes de la maladie ou certains types de patients seront « survivants » 5 ou 10 ans après le diagnostic. Les patients inclus ne seront pas représentatifs de l'ensemble des patients mais de ceux ayant survécus. Il s'agit donc d'un biais de sélection potentiel (le groupe n'est pas représentatif de l'ensemble des personnes atteintes de la maladie).

3 SELECTION DES TEMOINS

Le groupe des sujets témoins est constitué de sujets indemnes de la maladie considérée et doit être représentatif de la population d'où sont issus les cas.

- Sources de recrutement des témoins (deux principalement) :
    - En population générale : les témoins sont tirés au sort sur des listes : par exemple les listes électorales ou des listes de numéro de téléphone (« random digit dialing »). Ces listes ne sont cependant généralement pas exhaustives (Problème de représentativité).
    - Milieu hospitalier si les cas sont des patients hospitalisés : plus faciles d'accès, plus motivés. Mais risque de biais de sélection si le motif d'hospitalisation est lié au facteur étudié ou si la gravité de la maladie est différente.
- Equilibre des groupes cas et témoins :
    - Les groupes « cas » et « témoins » peuvent présenter des effectifs ***équilibrés*** (1 cas pour 1 témoin) ou ***déséquilibrés*** (1 cas pour 2 à 4 témoins). Cette pratique permet, lorsque l'on dispose d'un nombre limité de sujets « cas », d'augmenter la puissance de l'étude en augmentant le nombre de sujets « témoins ». Le gain de puissance statistique est cependant limité au-delà de 4 témoins pour 1 cas, rendant l'augmentation du nombre de témoins au-delà de ce seuil peu pertinent.
- **Appariement**

Il s'agit d'une technique permettant de sélectionner pour chaque cas, un ou plusieurs témoins présentant des caractéristiques communes (âge, sexe ou autres facteurs). Par exemple, si un homme de 46 ans est inclus dans le groupe des cas, on inclura également un homme de 46 ans dans le groupe témoins (appariement sur l'âge et le sexe). L'appariement permet de limiter les biais de confusion liés à ces caractéristiques. On choisira donc généralement comme facteur d'appariement des facteurs particulièrement susceptibles de jouer un rôle de confusion (par exemple un facteur lié de façon forte à la maladie).

Le nombre de facteurs sur lesquels il est possible de réaliser un appariement est cependant ***limité à 3 (ou 4)*** en raison de la difficulté technique à identifier des témoins présentant à l'identique l'ensemble des caractéristiques de cas. De plus, un appariement sur trop de facteur, ou le recrutement de témoin trop proche des cas (par exemple le voisin, concubin, frère, sœur…), risque également de conduire à un ***sur-appariement***, c'est-à-dire à l'identification de témoins ressemblant « trop » au cas, y compris sur le facteur de risque étudié, et susceptible de masquer une association potentielle.

4 RECUEIL DES DONNEES

Il existe différents mode de recueil de données : sur dossiers médicaux (souvent incomplets), par interview en face à face (++),par téléphone, par courrier… à l'aide d'hétéro-questionnaire ou d'auto-questionnaire. Ce recueil est toujours rétrospectif et donc une source potentiel d'erreur.

Pour être de qualité, ce recueil doit être :

- Standardisé (systématique et identique pour tous les sujets cas et témoins)
- Avec des enquêteurs formés au questionnaire de recueil
- Réalisé en insu (aveugle) de l'état de santé du sujet
- Si possible *en évitant de faire appel à la mémoire des sujets (sinon risque de biais de **mémorisation (biais de classement différentiel)** ➔ les cas se souviennent souvent mieux de leur exposition que les témoins)*

Sinon risque de ***biais d'information = biais de classement***

5 ANALYSE STATISTIQUE DANS LES ETUDES CAS - TEMOINS

Les études cas - témoins ne permettent pas de calculer ni les incidences ni les prévalences de maladie puisque les sujets sont inclus dans l'étude sur la base de la présence ou de l'absence de maladie, dans des groupes dont l'effectif est défini *a priori*. Il n'est donc pas possible de calculer directement un risque relatif (puisqu'il s'agit d'un rapport de l'incidence de la maladie chez les exposés et chez les non exposés). Il est cependant possible de calculer un ***rapport de côte*** (ou ***Odds-Ratio***) :

Calcul des fréquences d'exposition :

- chez cas : $FE_K = a/(a+c)$
- chez témoins : $FE_T = b/(b+d)$

	Cas	Témoins
Exposition présente	a	b
Exposition absente	c	d
Total	a+c	b+d

$$OR = \frac{FE_K / (1 - FE_K)}{FE_T / (1 - FE_T)} = \frac{ad}{bc}$$

Le « rapport des côtes » : l'Odds Ratio (OR) :

L'OR est une estimation qui doit toujours être présentée et interprétée avec son intervalle de confiance à 95%.

Cet OR est une bonne approximation du risque relatif si la maladie est rare (< 1% chez les non exposés au facteur de risque).

Démonstration : Si la fréquence de la maladie est faible dans la population :

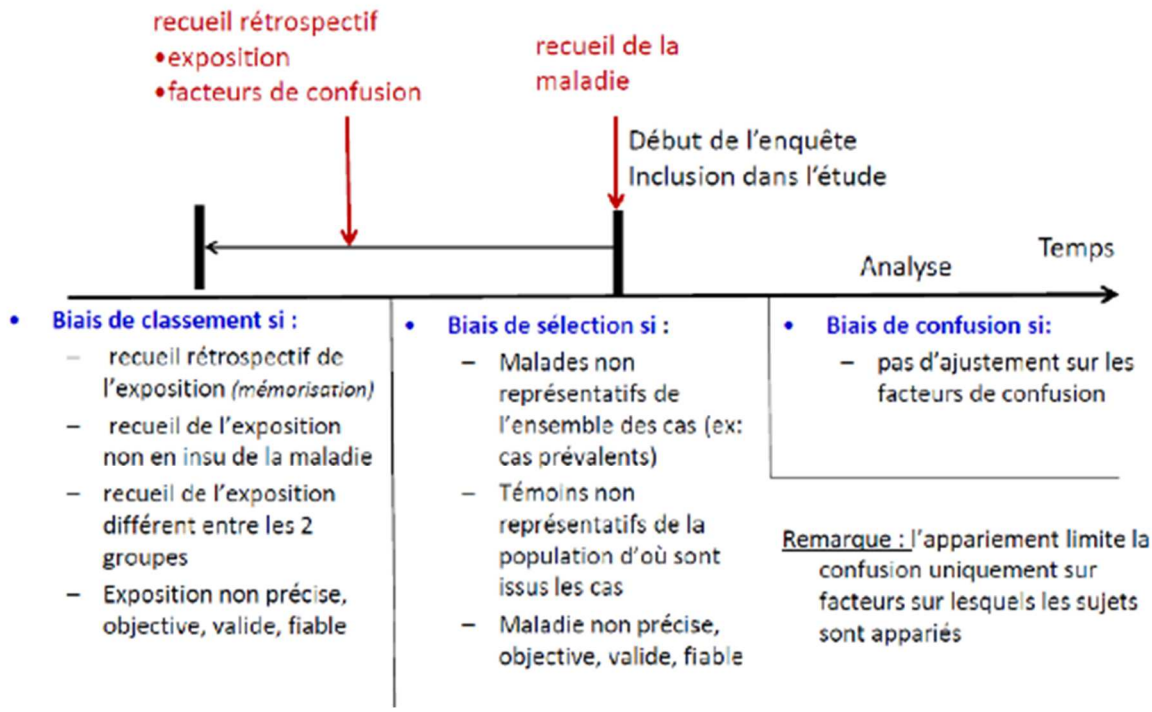
$$RR = \frac{A/A+B}{C/C+D}$$

- si A << B et C < <D, alors A+B ≈ B et C+D ≈ D
- Donc RR ≈ (A/B)/(C/D) = AD/BC = OR

Les études cas - témoins sont des études observationnelles. Ces études sont donc particulièrement sujettes aux ***biais de confusion***. L'analyse principale de l'étude permettant de répondre à l'objectif de l'étude est donc l'***analyse multivariée*** prenant en compte les facteurs de confusion potentiels (= analyse ***ajustée*** sur les facteurs de confusion, analyse de l'***association « indépendante »*** entre le facteur de risque et la maladie).

Le modèle statistique permettant le calcul des Odds-Ratio (bruts ou ajustés) est le modèle de régression logistique.

6 LES BIAIS DANS LES ETUDES CAS - TEMOINS



1. INTERET ET NIVEAU DE PREUVE DES ETUDES CAS - TEMOINS

Les études cas - témoins ont plusieurs avantages :

- Etudes adaptées aux maladies rares
- Étude possible de maladies à long délai d'apparition
- Étude d'une ou plusieurs E simultanément
- Durée courte (pas de suivi)
- Coût moindre

Mais plusieurs inconvénients :

- Pas de calcul d'incidence possible
- Pas de calcul de prévalence possible
- Biais de classement fréquent et important (***mémorisation*** ++)
- Biais de sélection fréquent et important (représentativité des groupes, sélection des témoins difficile)
- Séquence temporelle non assurée

Le niveau de preuve de ces études est donc généralement faible (***Niveau de preuve 3, grade de recommandation C*** selon la grille de la HAS).

7 UN DESIGN PARTICULIER : LES ETUDES CAS-TEMOINS NICHEES DANS UNE COHORTE

(cas-témoins nichées dans une cohorte

Il s'agit au départ d'une étude de cohorte traditionnelle (les cas incidents sont identifiés). Ces cas incidents seront comparés à des témoins sélectionnés dans cette même cohorte. Les cas et les témoins sont donc recrutés à partir d'une unique cohorte. L'intérêt de l'approche est que les informations sur les caractéristiques des sujets et leur niveau d'exposition ont généralement été recueillies lors du recrutement dans la cohorte, c'est-à-dire avant la survenue de la maladie (prospectivement). En particulier, si des prélèvements biologiques ont été réalisés à l'inclusion dans la cohorte et congelés, ceux-ci peuvent être utilisés après la survenue des cas pour doser un biomarqueur d'exposition. Si la maladie est rare, le coût total est bien plus faible que si les dosages avaient été réalisés sur l'ensemble de la cohorte.

Ce design présente l'avantage d'éviter le biais de sélection lors du choix des témoins, évite les biais de classement potentiellement différentiels liés au recueil rétrospectif de l'exposition. Permet d'utiliser les données d'incidence de la cohorte et d'estimer ainsi des risques relatifs.

EVALUATION DIAGNOSTIQUE

1 INTRODUCTION

Le diagnostic d'une maladie repose sur l'observation de signes cliniques, et/ou la réalisation d'examens complémentaires (biologiques, radiologiques ou autres). C'est de la valeur de ces signes ou de ces examens que dépend la validité du diagnostic porté.

La prescription d'examens complémentaires doit être basée sur leur utilité. En effet, toute prescription d'examen complémentaire a un coût. Ce coût comprend le coût financier, le coût lié au désagrément ou au risque qu'entraîne l'examen et le coût du préjudice subi par le malade lorsque le résultat de l'investigation conduit à des décisions inadéquates. Imposer au malade de supporter ce coût doit se justifier par l'espérance d'un bénéfice. Ce bénéfice peut résulter d'une appréciation plus exacte du diagnostic ou du pronostic pour un choix plus pertinent des investigations ultérieures et une meilleure décision thérapeutique. Le choix de retenir ou non un nouvel examen de diagnostic ne peut donc se faire qu'après une évaluation rigoureuse de ses avantages et de ses inconvénients par rapport à ceux des examens qui existaient jusque là.

L'examen doit avoir fait la preuve notamment de :

• **Sa qualité technique** : par exemple un électrocardiogramme ne pourra pas être interprété si les contractions des muscles de la paroi thoracique apparaissent sur le tracé.

• **Sa valeur spécifique** : l'examen mesure t-il ce qu'il est censé mesurer ? Si l'on veut doser la glycémie, la méthode permet-elle de mesurer spécifiquement le glucose ou dose t-elle l'ensemble des sucres sanguins?

• **Sa fiabilité** : y a t-il une erreur systématique de la mesure (exactitude)? La précision de celle ci est-elle satisfaisante (variance faible) ?

• **Sa reproductibilité :** s'assurer que la variabilité de l'interprétation des résultats de l'examen, inter et intra observateur, est acceptable.

• **Sa valeur diagnostique** : Si le test a une valeur diagnostique, le pourcentage de positifs parmi les malades sera statistiquement supérieur au pourcentage de positifs parmi les non malades. La valeur diagnostique (d'un test ou d'un signe clinique) est sa capacité à discriminer les sujets malades des non malades. Les deux paramètres qui caractérisent la valeur diagnostique d'un test T pour une maladie M sont sa sensibilité et sa spécificité pour la maladie. Ce sont les propriétés intrinsèques du test dans le diagnostic différentiel Malade/Non Malade.

• **Sa valeur prédictive** Un individu positif au test est-il affecté par la maladie? Un individu négatif au test est-il indemne de cette même maladie ? La valeur prédictive d'un test est la probabilité qu'un sujet soit malade ou non selon le résultat du test. On appelle valeur prédictive positive, la probabilité d'être malade si le test est positif et valeur prédictive négative, la probabilité de ne pas être malade si le test est négatif.

2 VALEUR DIAGNOSTIQUE D'UN SIGNE OU D'UN EXAMEN COMPLEMENTAIRE (NOUS L'APPELLERONS « TEST ») : LES INDICATEURS

La sensibilité d'un test (pour une maladie donnée) est sa capacité à donner un résultat positif quand la maladie est présente ou sa capacité à éviter les faux négatifs (FN).

La spécificité d'un test est sa capacité à donner un résultat négatif quand la maladie est absente, ou sa capacité à éviter les faux positifs (FP).

L'estimation de ces paramètres ne peut se faire que si l'on connaît l'état réel des sujets de l'échantillon testé: Malades (M) ou non malades ($\overline{M}$). Les diagnostics doivent avoir été affirmés préalablement, sans utilisation du test étudié, par une procédure dite **« Gold standard »** servant de référence (meilleure procédure diagnostique connue jusqu'alors). Le Gold Standard ou étalon-or n'est pas obligatoirement un examen unique ; il peut être représenté par une combinaison d'examens (algorithme ou stratégie).

2.1. CAS D'UN TEST QUALITATIF BINAIRE

2.1.1 ESTIMATION DE LA SENSIBILITE ET DE LA SPECIFICITE

Soit un échantillon de N sujets : N1 malades et N2 = (N – N1) non malades. A ces N sujets on applique le test dont on veut éprouver la valeur diagnostique pour la maladie M, et dont on suppose le résultat normal (test négatif T-) ou anormal (test positif T+). La distribution des sujets selon leur état réel et selon le résultat du test est présentée dans le tableau suivant :

	M	$\overline{M}$
T+	VP	FP
T-	FN	VN

VP: Vrai Positif: Sujet malade présentant un test positif

VN: Vrai Négatif: Sujet non malade présentant un test négatif

FP: Faux Positif: Sujet non malade présentant un test positif

FN: Faux Négatif : Sujet malade présentant un test négatif

Si le test était parfait, c'est à dire s'il détectait tous les cas de la maladie, et s'il ne détectait qu'eux, on aurait FN=FP=0

• Sensibilité : C'est la probabilité que le test soit positif parmi les malades : $\boldsymbol{Se = Pb(T^{+}/M)}$

(probabilité conditionnelle de la présence du test positif si la maladie M est présente)
[NB : La probabilité conditionnelle est traduite par la barre verticale entre T et M, probabilité de la réponse positive si la maladie est présente.]
Si l'échantillon est représentatif de la population des malades, on peut l'estimer par la fréquence de la réponse positive parmi les sujets atteints de la maladie soit : $\widehat{Se} = \frac{VP}{VP+FN}$
La sensibilité varie entre 0 et 1, on l'exprime en % (0 – 100 %).
Elle exprime l'aptitude du test à détecter tous les cas de la maladie, elle évalue sa capacité à bien classer les malades. Ainsi, plus le nombre de FN est faible, plus la sensibilité est élevée. Dans le cas où un test a une sensibilité égale à 1 (100 %), l'absence du signe ou la négativité du test exclut la maladie : il n'y a aucun faux négatif .

• Spécificité : C'est la probabilité que le test soit négatif parmi les non malades : $\boldsymbol{Sp = Pb(T^{-}/\overline{M})}$

(probabilité conditionnelle du résultat négatif du test si la maladie M est absente)

Si l'échantillon est représentatif de la population des non malades, on peut l'estimer par la fréquence de la réponse négative parmi les sujets non atteints de la maladie : $\widehat{Sp} = \frac{VN}{VN+FP}$

La spécificité exprime l'aptitude du test à ne détecter que les cas de la maladie, elle évalue sa capacité à bien classer les non malades. Ainsi, plus le nombre de FP est faible, plus la spécificité est élevée. Quand la spécificité d'un test pour une maladie vaut 1 (100%), ce test n'est positif qu'en présence de cette maladie : On dit que le signe est pathognomonique de la maladie (ex : le signe de Köplick dans la rougeole) ; cette éventualité est rare.

L'estimation de la sensibilité et de la spécificité doivent toujours s'accompagner d'un intervalle de confiance, considéré en général à 95% (α=5%).

2.1.2. RELATION ENTRE SENSIBILITE ET SPECIFICITE : LES RAPPORTS DE VRAISEMBLANCE

Un bon test doit avoir une bonne sensibilité et une bonne spécificité.

On appelle Rapport de Vraisemblance (likelihood ratio) positif (RV+) le rapport : $\frac{Pb(T+/M)}{Pb(T+/\bar{M})} = \frac{Se}{(1-Sp)}$.

Si RV+=k, un sujet a k fois plus de risque d'avoir un test positif s'il est malade que s'il est non malade. Le RV+ Indique dans quelle mesure un test positif augmente la probabilité que le patient est malade. L'intérêt diagnostique d'un test positif est d'autant plus fort que RV+ est grand.

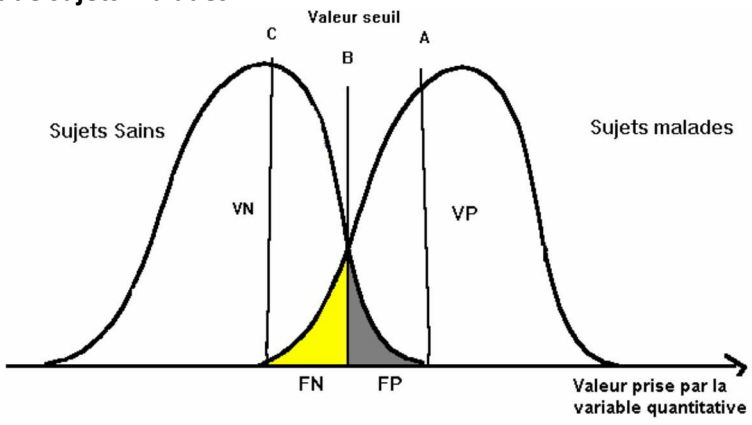
On appelle Rapport de Vraisemblance négatif (RV-) le rapport : $\frac{Pb(T-/M)}{Pb(T-/\bar{M})} = \frac{(1-Se)}{Sp}$.

Si RV-=k, un sujet a k fois plus de risque d'avoir un test négatif s'il est malade que s'il est non malade. Le RV- Indique dans quelle mesure un test négatif diminue la probabilité que le patient soit indemne. L'intérêt diagnostique d'un test négatif est d'autant plus fort que RV- est proche de 0.

2.2. CAS D'UN SIGNE QUANTITATIF ORDINAL OU CONTINU

Le plus souvent le résultat d'un test n'est pas binaire (positif ou négatif) mais s'exprime par une variable ordinale (ex : 0 , +, ++, +++, ++++) ou continue (résultats quantitatifs). Il faut déterminer une valeur « seuil » qui sépare les valeurs dites « normales » des valeurs dites « pathologiques ». Ce choix est arbitraire car les distributions des résultats chez les non malades et chez les malades se recouvrent. Si le paramètre mesuré est augmenté en cas de maladie (exemple : mesure de la glycémie chez des sujets sains et des diabétiques) on peut représenter la distribution de la variable de la façon suivante :

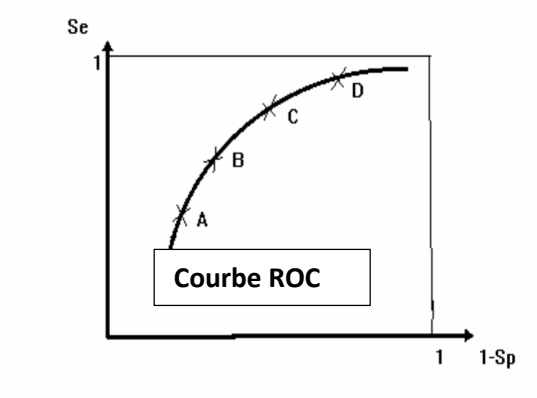
Distribution des résultats d'un test dans une population de sujets sains et de sujets malades



Si on déplace le seuil vers la gauche : on augmente le nombre de FP et on diminue le nombre de FN, le test est plus sensible et moins spécifique. Au seuil C, tous les malades sont positifs mais nombre de non malades sont eux aussi positifs et donc mal classés. Inversement si on déplace le seuil vers la droite (seuil A) : on diminue le nombre de FP et on augmente le nombre de FN, le test est moins sensible et plus spécifique. La meilleure valeur seuil, celle qui entraîne le minimum de

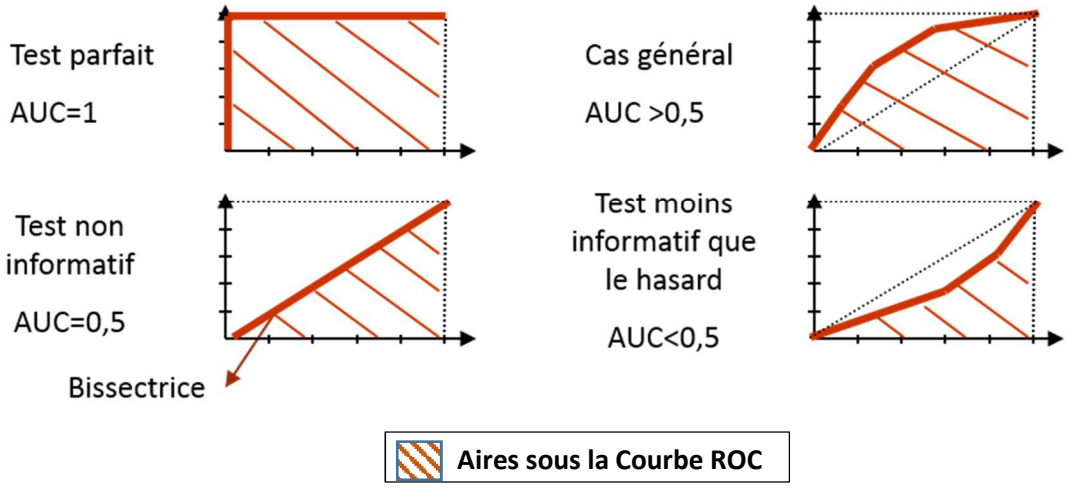
mauvaise classification dans un sens ou dans un autre, est B. Quand on fait varier le seuil,la sensibilité et la spécificité varient en sens inverse.

• **Courbe ROC** : On représente en général les résultats d'un test quantitatif au moyen d'une courbe dite courbe ROC (Receiver Operating Characteristic Curve), sur laquelle chaque seuil possible est représenté par un point ayant pour abscisse le taux de faux positifs (1-Sp) et pour ordonnée le taux de vrais positifs (Se). La forme de la courbe est en général concave en bas et à droite.



Chaque point de la courbe de A à D correspond à une valeur seuil. La capacité discriminante du test pour la maladie est évaluée par l'aire sous la courbe ROC (AUC).

.



L'AUC quantifie l'aptitude du test à discriminer les malades des non malades. Plus elle est proche de 1, plus le test est discriminant. Elle peut être estimée (entre 0 et 1) avec un $IC_{1-\alpha}$ et des tests de comparaison existent :

- Comparaison à la valeur 0,5 (intérêt discriminant du test)
- Comparaison de deux AUC (pour données appariées (2 tests réalisés chez le même sujet) ou indépendantes)).

Choix de la valeur seuil : Pour un test ordinal ou quantitatif, la décision de considérer un test positif ou négatif nécessite de choisir une valeur seuil pour laquelle la sensibilité et la spécificité seront suffisamment élevées; ce choix est guidé par le coût imputable aux erreurs (FP et FN) ; Il ne s'agit pas seulement du coût financier, mais aussi du coût en mortalité, en perte de qualité de vie, en souffrance... Le clinicien, en choisissant le seuil d'anormalité détermine, dans une certaine mesure, les gravités relatives des faux négatifs (ne pas diagnostiquer la maladie chez un malade) et des faux positifs (diagnostiquer à tort la maladie). Aussi, ce seuil doit-il être fixé en tenant compte de la gravité relative de chaque type d'erreur et du contexte (dépistage, démarche diagnostique classique,...).

*Exemple : La **phénylcétonurie** est une maladie génétique rare et grave en relation avec un trouble du métabolisme, entraînant une accumulation de cette dernière dans l'organisme. Elle*

est responsable d'une déficience intellectuelle progressive en l'absence de traitement approprié. En France on procède à la naissance à un dépistage systématique de la phénylcétonurie en dosant la phénylalaninémie. Lorsque le test est négatif, la famille de l'enfant est rassurée ; lorsqu'il est positif, le diagnostic est confirmé (ou infirmé) par un examen plus complexe (analyse chromatographique). En cas de diagnostic confirmé, un régime strict pauvre en phénylalanine est entrepris immédiatement et prolongé pendant la petite enfance ; il permet le développement intellectuel normal de l'enfant. Dans ce cas précis, le cout d'un FN est infiniment plus grand (déficience intellectuelle) que le cout d'un FP (cout de l'examen de confirmation + cout psychologique pour les parents) : on a donc privilégié un seuil de très haute sensibilité ; cependant la spécificité doit être également élevée.

3 PERFORMANCES D'UN TEST EN SITUATION REELLE – LES VALEURS PREDICTIVES

3.1. ESTIMATIONS

Dans la pratique médicale, on s'intéresse avant tout à la façon dont le résultat du test modifie la probabilité d'avoir ou non la maladie. Il ne suffit pas qu'un test ait une bonne sensibilité et une bonne spécificité pour qu'il soit un bon instrument de dépistage ou de diagnostic. Il importe de connaître la probabilité qu'un sujet soit ou non malade en fonction des résultats de l'examen complémentaire: Les valeurs prédictives (positive ou négative) sont des probabilités dites *a posteriori* ou *probabilités post test* (c'est-à-dire conditionnelles au résultat du test).

La valeur prédictive positive (VPP) d'un test est égale à la probabilité d'être malade dans le groupe des sujets ayant un résultat positif au test. $\boldsymbol{VPP = Pb(M/T+)}$

La valeur prédictive négative (VPN) d'un test est égale à la probabilité d'être non malade dans le groupe des sujets ayant un résultat négatif au test. $\boldsymbol{VPN = Pb(\bar{M}/T-)}$

Le théorème de BAYES* permet de calculer ces probabilités en partant de la sensibilité (Se), de la spécificité (Sp) et de la probabilité *a priori* ou *probabilité pré-test* de la maladie (c'est à dire avant que le test ne soit exécuté) Pb(M) ; cette probabilité est la prévalence de la maladie dans la population à laquelle on veut appliquer le test.

$$VPP = Pb(M/T+) = \frac{Se\ x\ Pb(M)}{\big(Se\ x\ Pb(M)\big) + \big((1-Sp)\ xPb(\bar{M})\big)}$$

$$VPN = Pb\left(\bar{M}/T-\right) = \frac{Sp\ x\ Pb(\bar{M})}{\big(Sp\ x\ Pb(\bar{M})\big) + \big((1-Se)\ xPb(M)\big)}$$

Comment varient les valeurs prédictives ?

• **Influence de la spécificité** : Quand la spécificité Sp varie de 0 à 1, la VPP augmente.

• **Influence de la sensibilité** : Quand la sensibilité Se varie de 0 à 1, la VPN augmente.

• **Influence de la prévalence** : Mais, pour un même examen et une même maladie, c'est à dire à sensibilité et spécificité constantes, ces valeurs prédictives varient en fonction de la prévalence de la maladie que l'on cherche à dépister ou à diagnostiquer dans la population qui va subir le test.

Si l'on est dans un contexte de dépistage, la prévalence de la maladie n'est autre que la fréquence de la maladie dans la population qui sera soumise au dépistage.

Si au contraire il s'agit d'un examen à visée diagnostique, Pb(M) est la fréquence de la maladie dans la population des patients qui seront soumis au test diagnostique

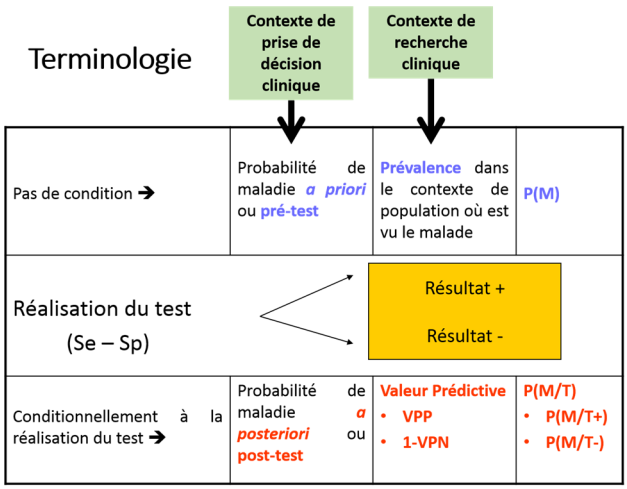
$$* \ Théorème\ de\ BAYES: Pb(^{A}/_{B}) = \frac{Pb(^{B}/_{A})xPb(A)}{\left(Pb(^{B}/_{A})xPb(A)\right) + \left(Pb\left(^{B}/_{\bar{A}}\right)xPb(\bar{A})\right)}$$

Exemple: dépistage hépatite C par le Test ELISA (Se=99% et Sp=90%)

Groupe cible	*Prévalence hépatite C*	*VPP*	*VPN*
Toxicomanes	*80%*	*97.5%*	*95.7%*
Transfusés	*7%*	*42.7%*	*99.9%*
Population générale	*0.9%*	*8.2%*	*≈100%*
Donneurs de sang nouveaux	*0.3%*	*2.9%*	*≈100%*
Donneurs de sang connus	*0.01%*	*0.1%*	*≈100%*

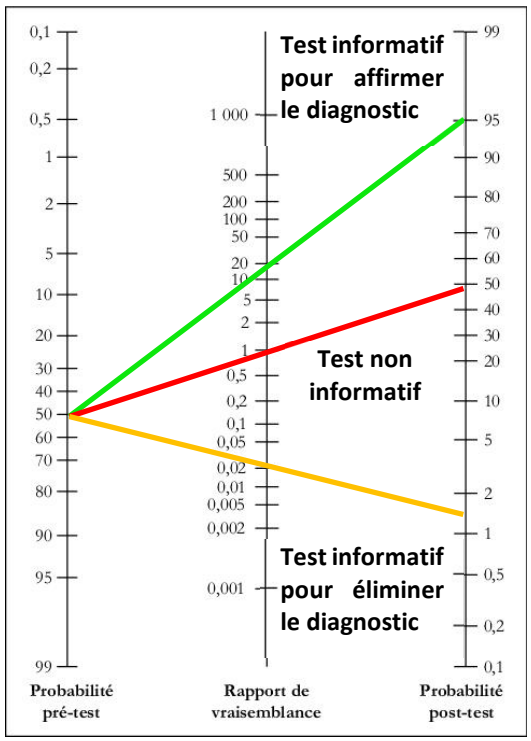
3.2. UTILISATION EN PRATIQUE CLINIQUE – LE NOMOGRAMME DE FAGAN

Ces indicateurs de validité diagnostique d'un test pour une maladie sont destinés à être utilisés en pratique clinique. La terminologie utilisée dans ce contexte figure dans le schéma ci-dessous.



Le Nomogramme de Fagan ci contre permet de visualiser la probabilité post-test à partir d'une ligne reliant la probabilité

Nomogramme de Fagan



pré-test et le rapport de vraisemblance d'un test. Par exemple, si la probabilité pré-test de maladie est de 50 % et le rapport de vraisemblance de 20, on trouve une probabilité post-test de maladie de 95% en reliant les deux premières valeurs entre elles. Le gain diagnostique est représenté par la différence entre la probabilité post-test et la probabilité pré-test. En pratique, si un patient est positif au test, on utilisera le rapport de vraisemblance positif pour évaluer sa probabilité post test d'être malade et s'il est négatif au test, on utilisera le rapport de vraisemblance négatif.

4 . CHOIX D'UN TEST DIAGNOSTIQUE

Le choix d'un test diagnostique dépend du contexte et des conséquences des erreurs à éviter.

Dans le cas d'une maladie dont le traitement est facile et efficace : on souhaite diagnostiquer TOUS les malades : favoriser la Se

- *Dans un essai thérapeutique: on ne veut diagnostiquer QUE des malades : favoriser la Sp*
- *Dans le cas d'un dépistage de masse (en population générale) : on choisit un test qui a une meilleure sensibilité que spécificité pour éviter les faux négatifs. Puis chez les sujets positifs, on réalisera un test qui aura alors une meilleure spécificité pour affirmer le diagnostic. IL faut cependant garder un équilibre acceptable des « coûts » des FP et FN : importance de minimiser aussi les FP (sinon effets délétères, problème de faisabilité).*

Si la maladie à dépister est rare dans la population (prévalence faible) la VPP sera médiocre même avec des Se et Sp élevées.

5 . LES ETUDES D'EVALUATION DIAGNOSTIQUE

5.1 Etude de la reproductibilité du test= étude de la fiabilité du test

Avant l'évaluation de la valeur diagnostique d'un test, il convient de vérifier la reproductibilité des mesures et observations du test (étude de la fiabilité). En effet, en raison de la variation des procédures de laboratoire, des appareils, des observateurs ou de l'évolution de l'état de santé des sujets, le test peut ne pas toujours donner le même résultat lorsqu'il est répété. On distingue :

- Variabilité inter-observateur : pour un même patient le résultat diffère selon l'observateur (l'appareil) qui interprète le test
- Variabilité intra-observateur : pour un même patient le résultat diffère entre les 2 interprétations faites par le même observateur (l'appareil) à deux moments différents (ex : 1 mois)
- Variabilité intra-sujet : le résultat du test est modifié par une modification de l'état de santé du sujet sur une courte période (ex : TA le matin ou le soir)

La mesure de la reproductibilité est réalisée avec un indicateur différent selon la nature de la variable :

- **Variable qualitative ou ordinale**: **coefficient de concordance Kappa de Cohen** (compris entre <0 et 1)
- Exemple : deux radiologues A et B codent (normal, anormal) un même ensemble de mammographies).

Radiologue		A		
B		Anormal	Normal	
	Anormal	53	0	53
	Normal	0	47	47
		53	47	100

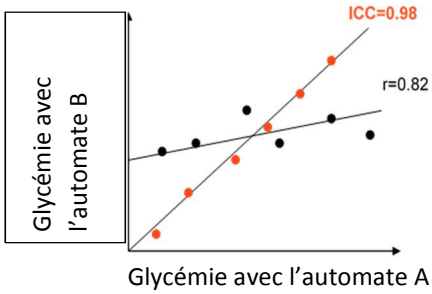
Dans cet exemple, la concordance est parfaite : il n'y a pas de discordance entre les deux radiologues.

Accord	*Kappa*
Excellent	≥ 0,81
Bon	0,80 - 0,61
Modéré	0,60 - 0,41
Médiocre	0,40 - 0,21
Mauvais	0,20 - 0,0
Très mauvais	< 0,0

- **Variable quantitative** : **le coefficient de corrélation intra-classe (ICC)** évalue la proximité des deux mesures (plus il est proche de 1, meilleure est la concordance). Une concordance parfaite (ICC=1) est constitué par la bissectrice entre les deux axes.

- Exemple : Cas d'une variable **quantitative** (on compare la concordance du dosage de glycémie entre deux automates différents).

Attention : ici le coefficient de corrélation linéaire r est inadapté : il n'évalue que l'existence d'un lien entre les mesures et pas leur adéquation.



5.2 ETUDE DE VALIDATION DIAGNOSTIQUE D'UN TEST POUR UNE MALADIE :

a. Points-clés du protocole

- **Schéma** : Etude transversale++++ (on cherche à faire un diagnostic à un moment donné), cohortes (seulement si un suivi est nécessaire pour affirmer ou infirmer le diagnostic).
  L'étude cas-témoins est inadaptée, en raison des erreurs de classement++, des valeurs manquantes et du fait qu'en général l'évaluation diagnostique d'un test relève de l'expérimentation et est une étude interventionnelle au sens de la loi.
- L'état du sujet ne doit pas avoir évolué entre le moment où on réalise le test à l'étude et le moment où on recueille le diagnostic par le gold standard.
- L'échantillon de l'étude doit être **représentatif** de la population à laquelle on veut appliquer le test
  - Malades : formes cliniques, …impact sur la prévalence
  - Non malades
- Le calcul du **NSN** : garantit la **précision** des estimations des paramètres
- La **fiabilité** du test doit avoir été évaluée: précision, reproductibilité (référence à étude antérieure)
- Les **critères de positivité** du test doivent être définis (selon le contexte)
  - Test qualitatif à l'aide d'une grille de lecture standardisée
  - Test quantitatif : choix du seuil

- **La procédure de référence : gold standard : c'est elle qui permet de différencier les malades des non malades**
    - Doit être éthique, acceptable, objective, consensuelle, Indépendante du test (ne doit pas utiliser les résultats du test)
    - En pratique elle est difficile à mettre en œuvre car souvent lourde/non éthique
- Tous les sujets doivent subir **le test à évaluer et la procédure de référence**
- L'Interprétation du test et de la procédure de référence doivent être réalisées en **insu** l'une de l'autre pour éviter que l'interprétation de l'un ne soit influencée par la connaissance de l'autre (biais de classement).
    - indépendance lecture du test à l'étude/gold standard et inversement
- **Opérateurs/lecteurs/outils**
    - expérience / caractéristiques comparables à ceux de la pratique réelle
- **Analyse en intention de tester** (tous les sujets inclus doivent être inclus dans l'analyse)
- Les **réponses douteuses** doivent être classées en positif ou négatif de façon à pénaliser le paramètre (Se, Sp..) que l'on souhaite évaluer de manière optimale.
- Les indices doivent être estimés (Se, Sp, VPP, VPN) avec leur **intervalle de confiance**
- **Cas particulier**
    - Comparaison de 2 (ou plusieurs) tests :
        - Privilégier un design où les sujets puissent avoir les deux examens
        - Si nécessaire, prévoir une randomisation pour le mode d'attribution et/ou l'ordre des examens :
        - Si nécessaire, déterminer le délai entre les tests
        - Prévoir l'Insu entre les tests comparés (+++)
        - Estimer le NSN pour mettre en évidence la différence attendue

b. <u>Les biais dans les études de validation diagnostique</u>

- Biais de sélection
    - Echantillon d'étude non représentatif de la population cible
    - inadéquation des formes cliniques présentes
    - prévalence de la maladie différente comparée avec la situation courante
    - Perdus de vue (si un suivi long est nécessaire pour confirmer ou infirmer le diagnostic)
- Biais de classement
    - Absence d'insu entre test et examen de référence
    - Critères de positivité imprécis/seuil inadéquat (grille de lecture standardisée +++)
    - Réponses douteuses
    - Expérience des lecteurs (qualité des appareils) trop différente de la pratique courante
    - Erreurs de l'examen de référence (GS imparfait)
    - GS non indépendant du nouveau test
    - Modification des techniques au cours du temps (en cas de suivi long)
    - Examen de référence non réalisé chez tous les sujets
    - État du sujet qui a évolué entre les deux tests

c. <u>Conclusion</u>

Les études de validation diagnostique sont une illustration du concept d'« Evidence-Based Medicine » et de la nécessité de fonder les décisions cliniques sur les résultats les plus probants.

Au terme de l'évaluation diagnostique, il reste à évaluer :

- **la validité du test en pratique courante** c'est-à dire si les résultats observés au cours de la validation sont similaires dans les études pragmatiques en population
- **l'utilité réelle du test en pratique courante pour le médecin, le malade, en termes de santé publique** (Influence sur les stratégies diagnostiques, thérapeutiques, impact de l'utilisation du test sur la morbidité, mortalité, la qualité de vie, impact sur l'état de santé de la population, les coûts).

2EME PARTIE : BASES METHODOLOGIQUES DE LECTURE CRITIQUE D'ARTICLES

Liste des auteurs

Dorine Neveu, MCF

Pierre Dujols, PU-PH

Pascale Fabbro-Peray, MCU-PH

Et les membres de l'équipe pédagogique de lecture critique d'articles pour leur contribution à la relecture et à la mise à jour du document :

Camille Agostini, AHU

Sophie Bastide, PHU

Claire Duflos, AHU

Jean-Luc Faillie, PHU

Grégoire Mercier, PH

Thibault Mura, MCU-PH

Nicolas Nagot, PU-PH

Marie-Christine Picot, PH

Fabienne Séguret, PH

INTRODUCTION

Ce document est un support de cours de lecture critique d'article (LCA). Il vous donne une **méthode de travail** que vous mettrez en pratique dans les enseignements dirigés de LCA. Il fait appel à un raisonnement et doit être complété par vos connaissances cliniques et par les exemples vus en enseignements dirigés de LCA. Il fait appel à des notions définies dans le glossaire du CNCI que vous trouvez sur le site du CNCI : http://www.cnci.univ-paris5.fr/medecine/

Il traite les objectifs pédagogiques de l'épreuve de LCA formulés par le CNCI , listés ci-dessous.

SAVOIR IDENTIFIER

1) Savoir identifier l'objet d'un article médical scientifique parmi les suivants: évaluation d'une procédure diagnostique, d'un programme de dépistage, d'un traitement, estimation d'un pronostic, enquête épidémiologique

2) Savoir identifier la question posée par les auteurs (hypothèse)

SAVOIR CRITIQUER LA METHODOLOGIE

3) Identifier les caractéristiques (données démographiques) de la population étudiée, à laquelle les conclusions pourront être appliquées.

4) Analyser les modalités de sélection des sujets, critères d'inclusion et de non-inclusion et d'exclusion.

5) Identifier la technique de randomisation et vérifier sa cohérence, le cas échéant.

6) Discuter la comparabilité des groupes soumis à comparaison

7) Discuter l'évolution des effectifs étudiés et leur cohérence dans la totalité de l'article; savoir si le calcul du nombre de sujets nécessaires a été effectué à priori.

8) S'assurer que la méthode employée est cohérente avec le projet du travail, et qu'elle est effectivement susceptible d'apporter "une" réponse à la question posée dans l'introduction.

9) Vérifier que les analyses statistiques (en fonction des notions élémentaires) sont cohérentes avec le projet; connaître les limites de l'analyse par sous-groupes; connaître la notion de perdus de vue.

10) Vérifier le respect des règles d'éthique.

ANALYSER LA PRESENTATION DES RESULTATS

11°) Analyser la présentation, la précision et la lisibilité des tableaux et des figures, leur cohérence avec le texte et leur utilité.

12°) Vérifier la présence des indices de dispersion permettant d'évaluer la variabilité des mesures et de leurs estimateurs.

CRITIQUER L'ANALYSE DES RESULTATS ET DE LA DISCUSSION

13°) Discuter la nature et la précision des critères de jugement des résultats.

14°) Relever les biais qui ont été discutés. Rechercher d'autres biais d'information et de sélection éventuels non pris en compte dans la discussion et relever leurs conséquences dans l'analyse des résultats.

15°) Vérifier la logique de la discussion et sa structure. Reconnaître ce qui relève des données de la littérature et ce qui est opinion personnelle de l'auteur.

16°) Discuter la signification statistique des résultats.

17°) Discuter la pertinence clinique des résultats.

18°) Vérifier que les résultats offrent une réponse à la question annoncée.

19°) Vérifier que les conclusions sont justifiées par les résultats.

20°) Indiquer le niveau de preuve de l'étude (grille de l'HAS).

EVALUER LES APPLICATIONS CLINIQUES

21°) Discuter la ou les applications potentielles proposées par l'étude

ANALYSER LA FORME DE L'ARTICLE

22°) Identifier la structure IMRAD (Introduction, Matériel et méthode, Résultats, Discussion) et s'assurer que les divers chapitres répondent à leurs objectifs respectifs.

23°) Faire une analyse critique de la présentation des références

24°) Faire une analyse critique du titre.

Chaque objectif est traité de la façon suivante :

- Les questions essentielles à se poser,
- Où chercher les éléments de réponse dans l'article
- Rappels théoriques relatifs à chaque objectif avec exemples.

OBJECTIF 1 : SAVOIR IDENTIFIER L'OBJET D'UN ARTICLE MEDICAL SCIENTIFIQUE

1. LES QUESTIONS A SE POSER

L'objet de la publication est-il:

- L'évaluation d'une procédure diagnostique
- L'évaluation d'un programme de dépistage
- L'évaluation d'un traitement
- L'estimation d'un pronostic
- Une enquête épidémiologique

Ces 5 objets étant les cadres des études sur lesquelles peut porter l'épreuve de l'ECN

2. OU CHERCHER

Dans le titre, en fin de section introduction, dans la section méthode.

L'objet sera déduit de la question étudiée et du schéma de l'étude (encore appelé schéma expérimental). Ces trois points, objet - question - schéma, devront être cohérents.

3. RAPPELS THEORIQUES :

L'**objet d'un article** est, pour l'épreuve de l'ECN, **le grand cadre de l'étude** qui a donné lieu à la publication analysée.

4. ATTENTION :

Faire attention: l'objet n'est pas le plan expérimental, même si la connaissance du plan expérimental est un des éléments qui permettra, avec la question étudiée de déterminer l'objet.

Exemple : Estimation d'un pronostic : cet objet s'applique à toute étude portant sur des sujets malades dont on veut déterminer le pronostic en termes de survie, rechute, rémission, guérison ….

OBJECTIF 2 : SAVOIR IDENTIFIER LA QUESTION ETUDIEE

1. LES QUESTIONS A SE POSER

a. La question, énoncée dans l'objectif (les objectifs), est-elle une question de recherche?
b. La question est-elle énoncée dans toutes ses composantes ?

2. OU CHERCHER

Pour a: dans l'introduction

Pour b: en fin de section Introduction et, parfois, début de section Méthodes et/ou paragraphe considérations statistiques

3. RAPPELS THEORIQUES :

Il faut tout d'abord comprendre et se souvenir qu'une question étudiée dans une publication:

A. Doit être une vraie question de recherche (cf ci-dessous §A) donnant une avancée des connaissances / confirmation scientifique
B. Doit être bien énoncée, c'est à dire avec toutes ses composantes (cf ci-dessous §B)
C. Doit être hiérarchisée
D. Est la pierre angulaire de l'étude. La méthode doit permettre de répondre à la question. Les conclusions, et les éventuelles recommandations se rapportent à la question principale.
E. Dont l'Observé (le phénomène que l'on étudie) est mesurable par un critère de jugement adéquat selon un Schéma d'Etude adapté.

A. Ce qu'est une question de recherche

Toute question bio-médicale n'est pas une question de recherche.

Une question de recherche doit être pertinente et argumentée, c'est à dire:

- Bien énoncée (cf §B)
- Face à une problématique d'intérêt, argumentée sur l'état actuel des connaissances, s'appuyant sur des références publiées, portant sur les 3 points suivants (**SCL**)
  - ✓ **S**anté Publique (fréquence de la maladie / du problème en santé, augmentation de fréquence, coût pour la société, mortalité importante …)
  - ✓ **C**linique ou physiopathologique (pronostic grave ou invalidant, séquelles fréquentes / importantes, difficulté du diagnostic par les moyens habituels …)
  - ✓ **L**acunes dans les connaissances (mécanismes physiopathologiques, arguments génétiques) ou limites dans les études précédentes de même objet.

B. Composantes d'une question de recherche

Les composantes d'une question de recherche sont (**PICO**):

- **P**opulation : ensemble d'unités, le plus souvent des personnes définies sur des critères précis
- **I**ntervention : facteur étudié (action, programme, facteur ou thérapie que le groupe de comparaison reçoit ou auquel il est exposé)
- **C**omparateur : action, facteur, programme ou thérapie à qui on compare l'intervention et auquel le groupe de comparaison, quand il existe, est exposé.
  - E*xemples*

- *un traitement « classique »« ou l'absence de traitement, ou un placebo dans un essai médicamenteux*
- *la « non exposition » au facteur de risque dans une enquête épidémiologique analytique*

- **O**bservé *(Outcome en anglais)* : phénomène d'intérêt (clinique, physiopathologique, …), sur lequel l'intervention est supposée avoir un impact.
  Ce phénomène sera mesuré par 'un critère de jugement ou d'évaluation *(ex: le diabète est mesuré au travers de la glycémie) (cf §C et objectif 13)*

Exemple : Comparer l'efficacité (O) et la tolérance (O) de l'enoxaparine (I) et de l'héparine non fractionnée (C) chez des patients avec un syndrome coronarien aigu (P)

Selon la recherche, certains des éléments constitutifs de la question sont sans objet.

Exemple : dans les enquêtes descriptives, les composantes C et I sont absentes. Estimer la prévalence des infections nosocomiales (O) parmi les patients des établissements hospitaliers français publics et privés (P)

C. Mesure de l'observé

La mesure de l'Observé est effectuée:

- Par un **Critère de Jugement** qui est le paramètre que l'on va mesurer pour représenter l'Observé. Ce critère, aussi appelé Critère d'Evaluation, sert au calcul du nombre de sujets nécessaires à l'étude *[cf Objectif-LCA n° 13]*
- Selon un **Schéma d'Etude**, ou type d'étude (study design) *[cf Objectif-LCA n° 8]*
- *Remarque : parfois l'expression "plan expérimental" est utilisée pour le schéma de l'étude. Cette expression ne signifie pas que l'étude est expérimentale, ni que l'intervention/facteur étudié est contrôlée.*

4. ATTENTION :

1. Toutes les composantes de la question ne sont pas obligatoires selon les objets d'étude [cf Objectif-LCA n°1]
2. Si l'observé est directement mesurable, il peut être aussi le critère de jugement (exemple: survie).
3. Une étude ne devrait correspondre qu'à une question et donc un critère de jugement, un résultat et une conclusion. Cependant, il arrive souvent que plusieurs questions soient posées dans une étude. On parlera alors d'objectif principal et d'objectifs secondaires **L'objectif principal** (= la question principale) est celle qui a été à la base de l'étude et **fonde le résultat principal de l'étude**. Elle est énoncée en premier. Elle a permis (quand licite) de calculer le nombre de sujets nécessaires. Les **objectifs secondaires**, questions secondaires auxquelles l'étude essaie de répondre, **ont une valeur surtout exploratoire** et doivent être peu nombreux. Comme ils n'ont pas participé au calcul du nombre de sujets nécessaires, les analyses faites sur leurs mesures risquent de ne pas avoir une puissance suffisante pour une conclusion étayée.
4. **Ne pas confondre objectif et hypothèse**. Dans l'hypothèse, on formule la réponse escomptée. Dans l'objectif, on reste neutre.
   - **Exemple:**

Objet : Evaluation de l'efficacité thérapeutique d'une intervention médicale

Objectif : Evaluer l'efficacité d'un médicament antivitamine K pour prévenir la survenue d'accident thrombo-embolique chez des sujets de plus de 65 ans, en arythmie cardiaque par fibrillation atriale.

Hypothèse : les antivitamines K diminuent le risque de survenue d'accident thrombo-embolique chez des sujets de plus de 65 ans, en arythmie cardiaque par fibrillation atriale.

OBJECTIF 3 : IDENTIFIER LES CARACTERISTIQUES (DONNEES DEMOGRAPHIQUES) DE LA POPULATION ETUDIEE, A LAQUELLE LES CONCLUSIONS POURRONT ETRE APPLIQUEES ET OBJECTIF 4 : ANALYSER LES MODALITES DE SELECTION DES SUJETS, CRITERES D'INCLUSION ET CRITERES DE NON-INCLUSION

1. LES QUESTIONS A SE POSER

a. A quelle population souhaite-on a priori proposer l'intervention si elle s'avère efficace et bien tolérée ou extrapoler les résultats? (population cible)
b. A partir de quelle population a-t-on extrait le (les) échantillon(s) de l'étude ? (population source)
c. Le mode de recrutement des sujets est-il précisé ?
d. Le lieu de l'étude est-il précisé ?
e. Les critères d'inclusion et de non inclusion des sujets sont-ils précisés ?
f. L'échantillon est-il représentatif de la population source ?
g. La base de recrutement (sondage) permet-elle d'atteindre toute la population définie par les critères d'inclusion et de non inclusion ?
h. A quelle population les conclusions peuvent-elles être appliquées ?

2. OU CHERCHER

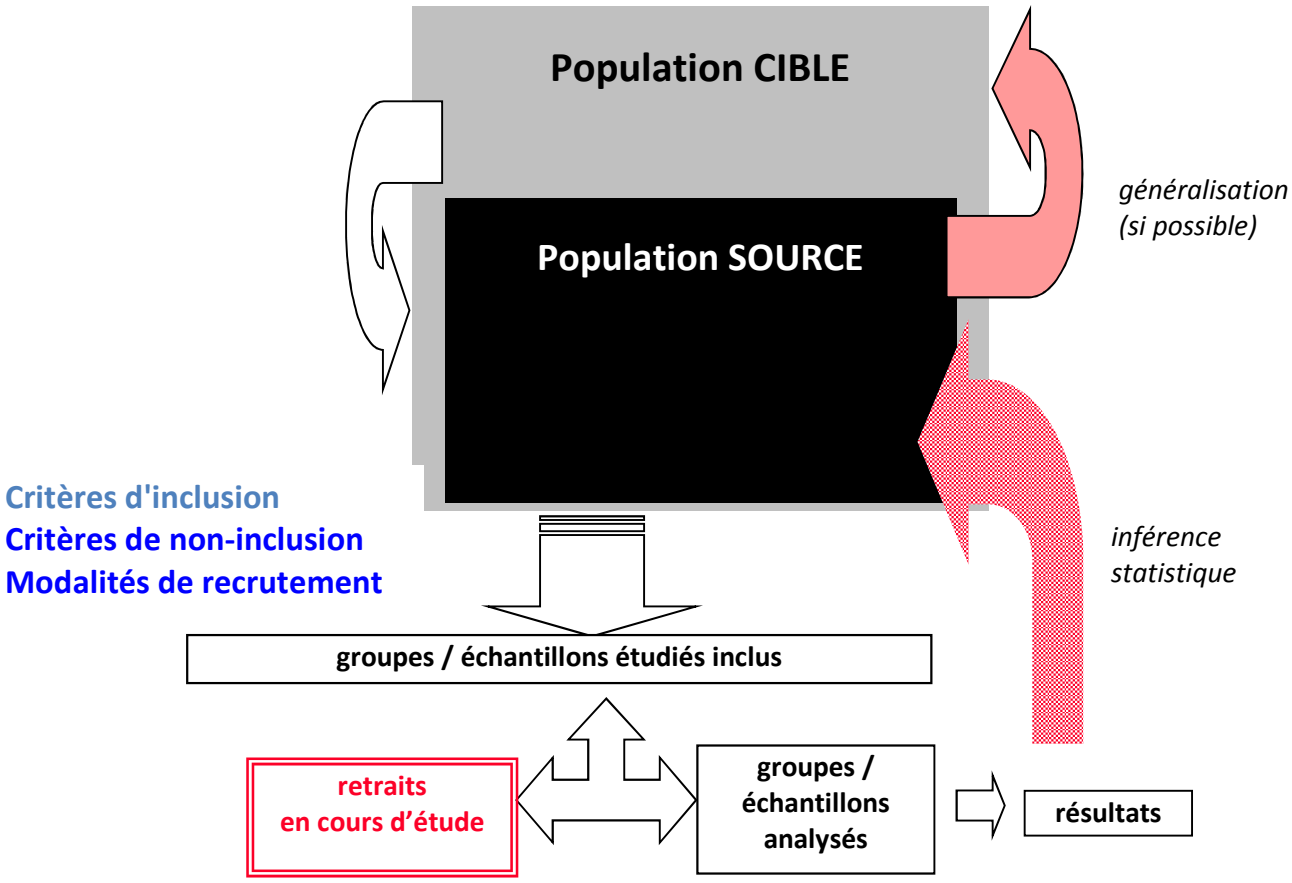
Population cible: dans le titre, à la fin de l'introduction (la question), au début de la partie méthodes, au début de la partie discussion

Population source: partie méthodes

Population à laquelle les conclusions peuvent être appliquées : méthode, résultats, diagramme de flux, fin de la partie discussion

3. RAPPELS THEORIQUES

Toute population incluse dans une étude a des caractéristiques spécifiques, qu'il est important de décrire, en particulier pour interpréter les résultats de l'étude et leur caractère généralisable (validité externe).



Population-Cible: population à laquelle les résultats d'une étude pourront a priori être étendus.

Population-Source: population au sein de laquelle l'échantillon a été tiré . Avec le lieu d'étude elle permet la reproductibilité de l'étude (possible en prospectif mais plus difficile en rétrospectif).

2 niveaux : critères d'inclusion et de non inclusion, et modalités de recrutement (*ex: listes électorales …*).

Modalités de recrutement des sujets (critères d'inclusion et de non inclusion) puis modalités de sélection des sujets analysés permettent de déterminer les individus éligibles pour l'étude et d'argumenter à quelle population on peut généraliser des résultats (population cible ou une autre à définir). Utiliser pour cela le diagramme de flux.

A. Modalités de recrutement

Qui : malades, sujets sains, volontaires sains ou malades, cotisants d'assurance sociale …

Comment : par publicité, sur liste électorale, sur liste d'individus présents ….

Où : un/des service(s) spécialisé(s), un hôpital, des hôpitaux, médecine de ville, entreprise, école, ville, région …, niveau national ou international …

Quand : ***date***

Il convient de regarder si les modalités de recrutement ne ciblent pas des sujets à profil particulier donc non représentatifs de la population cible. Si les modalités de recrutement ne permettent pas d'atteindre toute la population source, la généralisation est problématique.

B. Lieu(x) ou site(s) d'étude

Le(s) lieu(x) où l'étude s'est déroulée est un élément important de l'évaluation de la représentativité de l'échantillon et de l'applicabilité des résultats de l'étude dans le contexte particulier du lecteur.

Il est en particulier important de savoir si l'étude:

- s'est déroulée dans un centre de référence (plateau technique sophistiqué) ou dans les hôpitaux généraux …
- portait sur des soins de santé primaires (en médecine ambulatoire)
- portait sur des non malades (essais d'intervention), ciblés ou non …
- se déroulait dans un lieu géographique particulier (avec population particulière) …
- était mono ou multicentrique

C. Critères d'inclusion des sujets

Définition: Ensemble de critères qui définissent de façon précise les caractéristiques des sujets qui peuvent entrer dans une étude.

Ils reflètent la population cible de façon positive. Ils peuvent être:

- des critères sociodémographiques: origine ethnique, origine sociale, âge, sexe …
- des critères géographiques: origine, habitation, hospitalisation …
- des critères cliniques

D. Critères de non-inclusion des sujets

Définition: ensemble des critères faisant que les sujets ne peuvent pas être inclus dans une étude ou un essai.

Ce sont des critères de limitation de la sélection.

Ils appartiennent à une des 3 classes suivantes:

- *raison de prudence* (prévention d'événements indésirables)
  - ✓ Contre-indication à l'intervention étudiée (grossesse, autre traitement, maladies associées...)
  - ✓ Contre-indication à une exploration nécessaire au critère de jugement
- *difficulté potentielle d'évaluation des critères de jugement (on veut que l'effet mesuré dans chaque groupe via le critère de jugement soit dû à la pathologie et/ou au traitement et non à une co-morbidité ou à un traitement concomitant)*
  - ✓ Traitement associé interférent (interactions médicamenteuses, …)
  - ✓ Maladie associée ou handicap pouvant fausser les évaluations
- *difficultés potentielles de suivi*
  - ✓ Pathologie associée prioritaire dans les soins
  - ✓ Motivation insuffisante
  - ✓ Possibilité de perte de vue

- ✓ Co-morbidités ou caractéristiques favorisant une mauvaise observance ou un abandon d'étude

E. Population à laquelle les conclusions peuvent être appliquées

- **Population cible** (généralement caractérisée par les critères d'inclusion).
  - o Si l'échantillon est représentatif de la population définie par les critères d'inclusion et de non inclusion (pas de biais majeur de sélection lié à des exclusions des analyses, des perdus de vue, des données manquantes, des modalités de recrutement).
  - o Et si le diagramme de flux est donné avec le processus de sélection entre la pré-sélection (pré-sélection des sujets selon les critères d'inclusion) et l'inclusion,
    - ▪ Si le taux de sujets non inclus parmi les pré-sélectionnés est peu élevé et les motifs de non-inclusion (refus de participer, critères de non inclusion restrictifs) suggèrent que ces non-inclus sont comparables aux sujets analysés. (Chercher aussi dans les sections résultats et discussion des informations complémentaires.)
    - ▪ Attention : Si ce taux de sujets non inclus est élevé (de l'ordre de 20-30%), la généralisation des conclusions est problématique. Argumenter en vous basant sur les motifs de non inclusion.
- **Population source** (définie par les modalités de recrutement, les critères d'inclusion et de non inclusion)
  - o Si échantillon analysé représentatif de la population source.
  - o Attention aux modalités de recrutement.

ATTENTION SI ETUDE AVEC ANALYSES EN SOUS-GROUPES. CF § OBJECTIF 9.

OBJECTIF 5 : IDENTIFIER LA TECHNIQUE DE RANDOMISATION ET VERIFIER SA COHERENCE, LE CAS ECHEANT.

1. LES QUESTIONS A SE POSER

a. L'étude devrait-elle être randomisée ?
b. Si oui, l'étude comporte elle une randomisation ?
c. Si oui, la méthode est-elle décrite ?
d. Si oui, la randomisation est-elle de qualité c'est à dire garantit-elle l'imprévisibilité du traitement qui sera alloué au moment de l'inclusion, (allocation du traitement non prédictible, avec des modalités adaptées à l'objectif de l'étude, à un moment adéquat dans le déroulement de l'étude) ? = respect de la clause d'ignorance
e. Si oui, le mode de randomisation est-il particulier ?
f. Si l'essai clinique est multicentrique, la randomisation est-elle stratifiée sur les centres ?
g. Si des analyses en sous-groupes sont prévues, la randomisation a t-elle été stratifiée sur les variables qui définissent les sous-groupes ?

2. OU CHERCHER

Dans la partie méthode pour la description du mode et du moment de la randomisation

(Les auteurs de l'article doivent donner suffisamment d'information afin que le lecteur puisse s'assurer que la méthode employée est bien une allocation au hasard et qu'il n'y a pas eu de biais dans l'allocation des sujets)

3. RAPPELS THEORIQUES :

A. Définitions

Randomisation : tirage au sort des patients permettant une répartition au hasard, aléatoire, des patients dans deux ou plusieurs groupes (*définition du CNCI*).

Attention :

Le terme randomisation est consacré à un tirage au sort qui aboutit à la constitution de plusieurs groupes.

Quand on a un seul groupe, on utilise le terme tirage au sort.

Clause d'ignorance : Fait de ne pas révéler à un sujet le médicament qu'il va recevoir dans un essai thérapeutique, et pour un médecin qui inclut un sujet dans un essai, de ne pas savoir *(avant l'inclusion)* quel traitement ce sujet va recevoir *(selon le groupe où il sera alloué)*. Sinon, l'inclusion des sujets dans l'essai risque d'être influencée par la conviction intime du médecin de l'efficacité de l'un ou l'autre traitement réellement efficace. Le tirage au sort respecte la clause d'ignorance. (*définition du CNCI*).

B. Principe et intérêt de la randomisation

Procédé d'allocation au hasard (*random en anglais*) des sujets dans les groupes d'étude, elle permet:

- ✓ que chaque sujet ait la même chance, connue *a priori*, d'être assigné à un groupe donné
- ✓ de constituer des groupes comparables, en moyenne, sur tout sauf l'intervention, en particulier sur les facteurs pouvant avoir un effet sur le critère de jugement principal.

- ✓ donc d'éviter les biais de confusion
- ✓ d'éviter les biais de sélection, en ne permettant pas que l'affectation à un traitement soit le fruit d'un arrangement systématique et/ou soit liée au jugement de l'investigateur qui inclut les sujets.
- ✓ d'interpréter correctement les tests d'hypothèse; H0 "pas de différence entre intervention et contrôle(s)" présuppose que seules les interventions différencient les groupes; on impute au traitement l'origine de la différence observée si H0 est rejetée.
- ✓ et donc de respecter les contraintes éthiques de la méthode expérimentale en Sciences de la Vie.

C. Faire une randomisation

- ***Méthode*** : tirage au sort de l'allocation des sujets à un groupe ou à un autre par utilisation d'une table de nombres au hasard
- ***Mode*** :
  - ✓ Randomisation simple : allocation aléatoire des sujets dans les groupes.
  - ✓ Randomisation stratifiée : le but est d'équilibrer les effectifs de chaque groupe vis à vis de certaines caractéristiques et ainsi de forcer la comparabilité des groupes pour ces caractéristiques
    - on détermine des catégories de sujets, appelées strates, en fonction d'une ou de quelques caractéristiques (maximum 2 ou 3), connues pour modifier potentiellement l'effet du traitement - (*ex: centre dans un essai clinique multicentrique, classes d'âge, service clinique …). A*u sein d'une strate, les individus sont répartis par tirage au sort entre les groupes de traitement.
    - *exemple :*
      - o *Stratifier **par centre** revient à faire une liste de randomisation indépendante et différente pour chaque centre. Par exemple, dans un essai avec deux groupes de traitement et un rapport 1:1 entre les groupes, la randomisation stratifiée par centre permet d'équi-répartir à l'inclusion les patients d'un même centre sur les deux groupes de traitement. La randomisation stratifiée par **centre et par sexe** permet d'équi-répartir à l'inclusion, sur les deux groupes de traitement, les patients du même sexe et du même centre. S'il y a n centres, il y aura 2n strates (n pour les hommes et n pour les femmes).*
    - dans un essai clinique multicentrique, la randomisation **doit** être stratifiée systématiquement sur le centre (*par exemple: si le rapport est 1:1, dans chaque centre on aura autant de patients recevant le traitement A que le traitement B*) pour forcer la comparabilité a priori des groupes de traitement pour les facteurs pronostiques au sein de chaque centre. On neutralise théoriquement l'effet centre. Généralement on ajuste aussi sur le centre dans l'analyse.
    - Quand des analyses en sous-groupes sont prévues, la randomisation est stratifiée sur les variables qui définissent les sous-groupes (cf § objectif 9)
  - ✓ Randomisation par blocs: le but est d'équilibrer les effectifs de chaque groupe au fur et à mesure de l'allocation des sujets. Un bloc est un sous-groupe de taille déterminée à l'avance (exemple 6) à l'intérieur duquel on a une allocation aléatoire des patients. *Ex pour 2 traitements A et B et des blocs de 6: on a 3 sujets dans chacun des groupes A et B, ABAABB, BAAABB, ABABAB….Ainsi au bout de 6 sujets inclus, 3 sont alloués au traitement A et 3 au traitement B. L'investigateur* ne doit pas connaître la taille du bloc afin de ne pas pouvoir prévoir le groupe du ou des derniers sujets du bloc. Pour ce type de randomisation, **la meilleure imprévisibilité est obtenue avec des blocs de taille variable définie aléatoirement avec permutation de séquence**. **La randomisation par blocs contribue à neutraliser l'effet cohorte.**
- ***Equilibre de taille des groupes:***

- ✓ Randomisation équilibrée: les groupes doivent avoir la même taille
- ✓ Randomisation déséquilibrée: les groupes sont de tailles différentes; par exemple pour 1 patient randomisé dans le groupe A, 3 sont randomisés dans le groupe B.

- ***Organisation***: la liste (séquence) de randomisation doit être élaborée dans un centre indépendant (procédure centralisée) avant toute inclusion de sujet dans l'étude et son accessibilité se fait à l'heure actuelle le plus souvent soit par un serveur vocal, soit par internet, avec traçabilité de l'inclusion des patients pour empêcher des retraits d'inclusion.

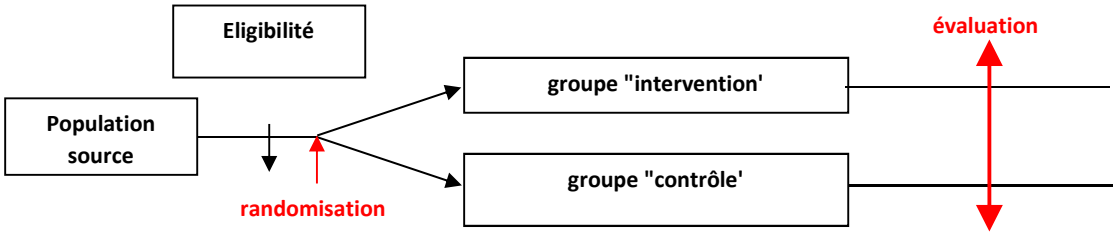
Moment de l'allocation du sujet: **après confirmation de l'éligibilité +++** et le plus tard possible c'est à dire juste avant la première administration de l'intervention (pour éviter des exclusions post randomisation)

Cas particuliers

- ✓ Les essais en grappe (essais en cluster): on randomise les grappes (groupes homogènes quant à la question étudiée *ex des classes d'école*) et tous les sujets d'une même grappe (*ex tous les élèves de la même classe*) reçoivent la même intervention
- ✓ Essai croisé (ou en cross-over) : le patient est son propre témoin. On randomise la séquence de traitements.

Les détails concernant toute particularité de cette randomisation (stratification) doivent être précisés car il devra en être tenu compte dans l'analyse.

D. Validité d'une randomisation



- L'allocation du traitement ne doit pas pouvoir être prédite (respecter la clause d'ignorance)
- les modalités doivent être adaptées à l'objet et à la question
- La mise en œuvre doit être faite au moment opportun

E. Selon l'objet de l'étude:

La randomisation n'a lieu d'être **que dans les études interventionnelles comprenant plusieurs groupes soumis à comparaison**.

Dans les essais thérapeutiques comparatifs, la randomisation est la règle.

Dans les études où on veut comparer l'efficacité ou l'utilité d'une procédure de dépistage

Remarque : Dans certaines études le tirage au sort des patients (ou sondage) effectué pour obtenir un échantillon représentatif de la population à laquelle on veut extrapoler les résultats (ex : témoins dans une étude cas-témoins) est parfois appelé randomisation. Le terme de randomisation est utilisé alors à mauvais escient.

OBJECTIF 6 : DISCUTER LA COMPARABILITE DES GROUPES SOUMIS A LA COMPARAISON

1. LES QUESTIONS A SE POSER

a. Les groupes analysés sont-ils comparables initialement en ce qui concerne les facteurs pronostiques ou de confusion connus ou suspectés? (un tableau présente-t-il les données permettant de s'assurer de cela ?)
b. Les groupes analysés sont-ils comparables, sur ces mêmes facteurs, à la fin de l'étude ?
c. Les perdus de vue, les sorties d'étude ou données manquantes sont-ils nombreux et déséquilibrés ?
d. Dans un essai clinique, les groupes analysés ont-ils reçu la même prise en charge en dehors des interventions étudiées, au cours du suivi ?

2. OU CHERCHER

- Comparabilité initiale: dans le tableau donnant les résultats (nombre de sujets, fréquences et/ou moyennes) dans chaque groupe concernant les caractéristiques initiales des sujets étudiés, facteurs pronostiques majeurs ou de confusion.
- Suivi d'étude : section résultats et dans le diagramme de flux (flow chart) - ou son équivalent textuel - au début de la partie résultats, parfois à la fin de la partie méthode (perdus de vue, sorties d'étude).

3. RAPPELS THEORIQUES :

A. Suivi du nombre de sujets:

Il est important de connaître le nombre de sujets, éligibles (remplissant les critères d'inclusion), le nombre de sujets inclus et les raisons de non inclusion (refus, oubli, exclusion volontaire…).*[cf objectif n°7]*

B. Tableau de comparabilité des groupes (comparabilité initiale)

La comparabilité doit être clinique et non statistique car le critère de jugement principal sur lequel est fondée la conclusion de l'étude n'est pas un des facteurs pronostiques retenus pour évaluer la comparabilité. L'effectif n'a pas été calculé pour détecter une différence entre les groupes pour les facteurs pronostiques.

Quand l'étude est randomisée:

La randomisation des sujets entre les groupes garantit en théorie la comparabilité de ces groupes et prévient les biais de sélection.

Cependant, cette garantie n'est pas totale et la "comparabilité des groupes avant mise en œuvre de l'intervention" doit être vérifiée, en particulier sur les variables susceptibles d'intervenir sur les résultats (facteurs pronostiques).

S'il existe des différences dans les caractéristiques initiales (données de base), les auteurs doivent indiquer s'ils en ont tenu compte dans l'analyse (ajustement, analyse multifactorielle ou multivariée).

Quand l'étude n'est pas randomisée:

Les différences observées entre les groupes comparés peuvent s'expliquer par de nombreux facteurs autres que celui que l'on étudie.

Il est normal dans de telles études que les groupes ne soient pas comparables pour certaines variables. Pas d'imputation causale directe (voir faisceau d'arguments).

Il faut donc s'assurer que ces variables, lorsqu'elles interfèrent avec les expositions qui font l'objet de l'étude et avec la maladie étudiée, ont été prises en compte dans l'analyse : ajustement, analyse multifactorielle).

C. Diagramme de flux (maintien de la comparabilité initiale)

Il renseigne sur :

- le type d'analyse effectuée dans le cas d'un essai thérapeutique (analyse en intention de traiter) (cf objectif 9) : nombre de sujets exclus de l'analyse et motifs d'exclusion de l'analyse
- les sorties d'étude (nombre et motifs)

Si les motifs d'exclusion de l'analyse ou de sortie d'étude sont liés à l'intervention et/ou au phénomène mesuré, on a un biais de sélection. Ce biais est faible si ces exclusions ou sorties d'étude sont peu nombreuses (10% est la limite communément admise) et équilibrées entre les groupes. Il est élevé si elles dépassent 20%, et l'est d'autant plus qu'elles sont déséquilibrées.

- **Les traitements concomitants et différences de prise en charge pendant le suivi, le cas échéant**

Toute prise en charge ne faisant pas partie de l'intervention et concomitante à l'intervention, susceptible d'avoir un effet sur le phénomène étudié engendre un biais (cas des traitements concomitants).

Le phénomène observé ne sera pas dû uniquement à l'intervention ou au facteur étudié. Ce biais est d'autant plus important que la prise en charge est différente entre les groupes comparés.

OBJECTIF 7 : DISCUTER L'EVOLUTION DES EFFECTIFS ETUDIES ET LEUR COHERENCE DANS LA TOTALITE DE L'ARTICLE ; SAVOIR SI LE CALCUL DU NOMBRE DE SUJETS NECESSAIRES A ETE EFFECTUE A PRIORI

1. LES QUESTIONS A SE POSER

a. Le nombre de sujets : éligibles, inclus, suivis … peut-il être suivi dans l'article ? Le diagramme de flux et les tableaux de résultats permettent-ils de suivre les effectifs ?
b. Un calcul du nombre de sujets nécessaire a-t-il été fait ?
c. Ce calcul porte t-il sur le critère de jugement principal ? Si non, le choix du critère de jugement est-il justifié?
d. Dans une étude comparative, la différence cliniquement "intéressante" (Δ) de valeur du critère de jugement principal entre les groupes est-elle fournie et argumentée ? Le risque de première espèce (α) est-il fourni ainsi que la puissance (1-β) ? Sont-ils satisfaisants ? Si le critère de jugement est qualitatif ou censuré, le taux dans le groupe contrôle est-il précisé ?
e. Dans une étude non comparative dont l'objectif est l'estimation de paramètres (voir le critère de jugement principal), il permet de garantir une précision de cette estimation (ex : estimation de prévalence d'une maladie dans une enquête transversale, estimation des Se, Sp, VPP d'un examen dans le cadre d'une étude diagnostique). Le taux attendu, la précision et le risque α sont-ils donnés ?
f. Le nombre de sujets sélectionnés tient compte du % potentiel de perdus de vue.
g. Le nombre de sujets analysés est-il supérieur ou égal au nombre de sujets nécessaires

2. OU CHERCHER

Dans la partie méthode, pour le calcul

Dans la partie résultats, pour le suivi des effectifs et le diagramme de flux.

3. RAPPELS THEORIQUES :

Importance du calcul du nombre de sujets nécessaire (NSN)

- ✓ Le NSN est calculé en fonction du critère de jugement principal (CJP).
- ✓ Il permet de garantir à l'étude une puissance pour détecter une différence cliniquement intéressante du CJP entre les groupes compte tenu de la dispersion du CJP et des éventuelles sorties d'étude.
- ✓ La puissance doit au moins être égale à 80%
- ✓ Si le résultat n'est pas significatif, cela peut être dû à un manque de puissance, un biais ou à une différence plus petite que celle qu'on voulait détecter.
- ✓ Dans une étude non comparative, il permet de garantir une précision de l'estimation du CJP.

La méthode de calcul du nombre de sujets doit être indiquée avec la différence que l'on veut mettre en évidence entre les groupes comparés, (ou la précision des indicateurs évalués) et les risques d'erreur acceptés (en général α = 5% et β entre 5 et 20%).

La "plausibilité" de la différence "intéressante" repose sur une argumentation.

« Le CONSORT statement » propose un diagramme-type (flow chart) permettant de s'assurer que le lecteur retrouve clairement le nombre de sujets concernés à chaque étape - recrutement des patients, inclusion, randomisation, suivi, analyse de l'étude - ainsi que les motifs d'exclusion/non suivi de l'étude.

OBJECTIF 8 : S'ASSURER QUE LA METHODE EMPLOYEE EST COHERENTE AVEC LE PROJET DU TRAVAIL, QUE LA METHODOLOGIE EST EFFECTIVEMENT SUSCEPTIBLE D'APPORTER UNE REPONSE A LA QUESTION POSEE DANS L'INTRODUCTION.

1. LES QUESTIONS A SE POSER

Le schéma d'étude est-il cohérent avec l'objet ?Le schéma d'étude est-il cohérent avec la question étudiée?(Pour la justification voir articles vus en enseignements dirigés).

A.

2. OU CHERCHER

En général cité dans le titre de l'article puis précisé dans le chapitre méthode. Parfois en fin d'introduction avec le (les) objectif(s) de l'étude.

3. RAPPELS THEORIQUES :

A. Définition du Schéma d'étude ou Type d'étude.

Le schéma d'étude est la méthode stratégique mise en œuvre pour répondre à la question posée. C'est sur lui, en adéquation avec la question posée, que repose toute la valeur de l'article.

Selon qu'il s'agit d'un problème diagnostique, thérapeutique, pronostique ou épidémiologique, le schéma d'étude varie.

[*voir PARTIE 1 pour le détail des schémas d'étude*]

QUESTION		**TYPE d'ENQUETE / ETUDE les plus adaptées**
Epidémiologie descriptive	Incidence	Cohorte
	Prévalence	Transversale
Epidémiologie analytique (causalité)	Phénomène non contrôlable fréquent	Cohorte exposés / non exposés
	Phénomène rare	Cas - Témoins
Traitement	Efficacité	Essai contrôlé randomisé
Traitement	Sécurité	Essai contrôlé randomisé Cohorte exposés / non exposés
action de dépistage	Efficacité	Essai contrôlé randomisé Cohorte exposés / non exposés
Procédure diagnostique	Fiabilité,(reproductibilité/variabilité)	Transversal comparatif avec répétition de mesure
	Validité (sensibilité/spécificité)	Transversal comparatif avec référence (Gold standard, étalon or) (sujet propre témoin)
Pronostic	Maladie fréquente	Cohorte exposés / non exposés
	Maladie rare	Cas - Témoins

B. Schéma d'étude par Objet d'article

a. Le schéma d'étude des enquêtes épidémiologiques

1. ***Enquêtes descriptives***:

but : décrire des indicateurs mesurant l'importance de la maladie : incidence, prévalence, mortalité.

Schémas : transversal pour la prévalence ou longitudinal (pour l'incidence ou la mortalité): cohorte. Les registres sont des enquêtes à recueil permanent , prospectif et dont le but est l'exhaustivité et l'exactitude, utilisés pour obtenir des incidences.

Sujets : soit échantillon représentatif d'une population définie (sondage), soit population entière.

2. Enquêtes analytiques

but : mettre en évidence des facteurs de risque de maladies

Schéma : cohorte, cohorte exposés-non exposés, cas-témoins, registre (recueil permanent)

Sujets : 2 groupes (exposés / non exposés: cohorte exposés / non exposés ou malades / non malades: cas témoins). Dans les enquêtes de cohorte où on recherche l'association entre plusieurs facteurs d'exposition et une maladie, on recueille les informations concernant les facteurs d'exposition chez tous les sujets (on ne distingue pas de groupes).

b. Le schéma d'étude des études diagnostiques

Au sein des études diagnostiques, deux grands types:

(1) évaluation de la fiabilité d'un examen ou procédure

(2) évaluation de la validité d'un examen ou procédure

Evaluation de la fiabilité d'un examen ou procédure

Etude de concordance dans laquelle on fait évaluer le même examen par plusieurs juges et on étudie la concordance des résultats entre les différentes évaluations. C'est une étape préalable à l'examen de la validité.

Evaluation de la validité d'un examen ou procédure

On peut soit :

- ✓ vouloir estimer la sensibilité et la spécificité d'un nouveau test,
- ✓ déterminer un seuil à partir duquel d'autres examens ou un traitement seront mis en oeuvre,
- ✓ comparer un test diagnostique à un autre (ou une séquence à une autre).

Point clé à vérifier : le nouveau test et le gold standard doivent mesurer le même état du sujet, c'est pourquoi le meilleur schéma est une étude expérimentale **transversale**.

c. Le schéma d'étude des études d'estimation d'un pronostic

Ce sont des études à visée analytique.

On distingue deux types d'étude pronostique : prospectives et rétrospectives.

Les études prospectives consistent à suivre un groupe de malades (cohorte) pendant une durée définie, d'enregistrer les facteurs que l'on veut étudier (facteurs potentiellement pronostiques) et de noter les événements (complications, décès) au fur et à mesure de leur apparition.

Les études rétrospectives :

- Ce sont le plus souvent des cohortes : elles consistent à reconstituer une cohorte *a posteriori* à partir d'une file active et à aller chercher dans les dossiers pour une période antérieure les éléments (facteurs pronostiques et évènements) nécessaires à l'étude. Elles pâtissent souvent de la qualité des dossiers …
- Plus rarement ce sont des études cas-témoins

d. Le schéma d'étude des études d'évaluation d'un traitement

C'est l'essai thérapeutique contrôlé randomisé en double aveugle. cf tableau §3.A . *Voir glossaire (expressions comportant le terme essai)*

acronyme: ECHRAMS

Essai : Essai clinique thérapeutique, de prévention,

Contrôlé contre X (citer traitement du groupe contrôle),

Hypothèse testée : d'efficacité, (OU de non infériorité, -OU d'équivalence)

Randomisé,

Aveugle? en double aveugle, (OU en ouvert OU en simple aveugle),

Multicentrique, (OU monocentrique).

Schéma : à X (2, 3) bras parallèles, (OU croisé = en cross-over),

e. Le schéma d'étude des études d'évaluation d'une action de dépistage d'une maladie (application aux cancers)

cf tableau §3.A

Attention : l'évaluation d'un **test de dépistage** utilise les mêmes méthodes que l'évaluation diagnostique (seules les populations changent).

1. Evaluation d'une ***procédure*** **de dépistage** :

- Objectif : démontrer l'efficacité (et/ou la qualité) de la procédure de dépistage
- Design optimal :
  - essai contrôlé randomisé

    - 2 bras : 1 bras effectuant le dépistage, un bras n'effectuant aucun dépistage
- Biais spécifiques : avance au diagnostic (lead-time), biais de surdiagnostic (famille des biais de classement)
- Critère de jugement principal :
    - Si l'objectif est l'efficacité : le critère est la mortalité spécifique (car permet de s'affranchir du biais d'avance au diagnostic : ex dépistage des cancers), ou l'incidence du cancer si l'objectif est de dépister des lésions pré-cancéreuses
    - si l'objectif est l'évaluation de la qualité : les critères de jugement sont la sensibilité, spécificité et valeurs prédictives de la *procédure* (test+bilan diagnostique)
    - Sensibilité de la procédure : impose l'existence d'une structure qui enregistre de façon exhaustive les cas de maladie (cancers de l'intervalle)
    - Spécificité de la procédure: impose que l'on dispose d'une organisation qui permette de suivre jusqu'au terme du bilan diagnostique tous les sujets ayant un test de dépistage « positif »
- Analyse en intention de dépister
- Si suivi long : problème de la contamination des groupes

2. Evaluation d'un ***programme*** **de dépistage** :

- Objectif : vérifier que, lors de la généralisation du dépistage à une population ou un sous-groupe de population (dépistage organisé), le programme de dépistage respecte les conditions dans lesquelles on peut attendre des résultats similaires à ceux des essais d'efficacité.
- Evaluation le plus souvent selon un cadre réglementaire (ex : dépistages des cancers).
- Critères d'évaluation calculés sur le sous-groupe de population concerné par le dépistage (participants ou non)
- Critères d'évaluation principaux : participation, qualité, efficacité du programme, respect des délais de prise en charge, parfois évaluation des conséquences psychologiques et sociales du dépistage.
- Dans le cas particulier des programmes de dépistage des cancers, les critères d'efficacité sont d'abord des critères *intermédiaires* (le calcul du taux de mortalité demandant le plus souvent un long délai) portant essentiellement sur la taille et le degré d'envahissement (facteurs pronostiques majeurs) de la maladie.

C. **Exemples:**

Les exemples suivants (questions de recherche et les plans expérimentaux appropriés) permettent de comprendre les réponses à donner à cette question.

1) ***Objet de l'article****:* Evaluation de l'efficacité thérapeutique d'une intervention médicale

- Exemple d'objectif : Mesurer (Evaluer) l'efficacité d'un médicament antivitamine K pour prévenir la survenue d'accident thrombo-embolique chez des sujets de plus de 65 ans, en arythmie cardiaque par fibrillation atriale.

- Hypothèse : antivitamine K réduit le risque de survenue d'accident thrombo-embolique chez des sujets de plus de 65 ans, en arythmie cardiaque par fibrillation atriale.

- Type d'étude: essai thérapeutique d'efficacité multicentrique, contrôlé, randomisé, en double aveugle, à deux groupes parallèles : Antivitamine K versus placebo,

2) ***Objet de l'article :*** **Evaluation d'une procédure de dépistage**

- Exemple d'objectif : Déterminer si la prescription systématique d'un test PSA à une population masculine de 50 à 75 ans permet de réduire la mortalité par cancer de la prostate.

- Hypothèse : la prescription systématique d'un test PSA à une population masculine de 50 à 75 ans réduit la mortalité par cancer de la prostate.

- Type d'étude: essai contrôlé randomisé prospectif à deux bras parallèles avec prescription d'un dépistage systématique versus non prescription de dépistage

3) ***Objet de l'article :*** **Evaluation d'une intervention**

- Exemple d'objectif: La mise en place de séances d'information sur les risques liés au tabagisme dans les écoles primaires permet-elle de diminuer le nombre de fumeurs chez les adolescents ?

- Hypothèse : Une information sur les risques liés au tabagisme dans les écoles primaires diminue le nombre de fumeurs chez les adolescents.

- Type d'étude : essai randomisé entre les diverses écoles d'une même ville : essai en cluster contrôlé randomisé, séances d'information versus pas de séance

4) ***Objet de l'article :*** **Enquête épidémiologique**

Descriptive

- Exemple d'objectif : Déterminer la prévalence des infections nosocomiales dans les établissements de courts séjours publics et privés français.

- Hypothèse : la prévalence des infections nosocomiales dans les établissements de courts séjours publics et privés français est de x%.

- Type d'étude: étude épidémiologique d'observation transversale.

Recherche de facteurs de risque

- Exemple d'objectif : Déterminer le risque de sclérose en plaques lié à la vaccination contre l'hépatite B.

- Hypothèse : le vaccin contre l'hépatite B est associé à un risque accru de sclérose en plaque.

- Type d'étude: étude épidémiologique d'observation rétrospective (évènement rare) à visée analytique de type cas-témoin

5) ***Objet de l'article :*** **Evaluation d'une procédure diagnostique**

- Exemple d'objectif : Evaluer la valeur diagnostique de la scintigraphie pulmonaire ventilation/perfusion, chez un patient avec suspicion d'embolie pulmonaire

Hypothèse : la scintigraphie pulmonaire ventilation/perfusion permet de détecter une embolie pulmonaire.

- Type d'étude : étude transversale avec sujet propre témoin

6) ***Objet de l'article :*** **Estimation d'un pronostic : identifier des critères pronostiques**

- Exemple d'objectif : Identifier les facteurs qui augmentent le risque de séquelles chez les enfants atteints de méningite bactérienne

- Hypothèse : Les facteurs X, Y, Z augmentent le risques de séquelles chez les enfants atteints de méningite bactérienne-

- Type d'étude : étude observationnelle de cohorte prospective à visée analytique.

OBJECTIF 9 : VERIFIER QUE LES ANALYSES STATISTIQUES (EN FONCTION DE NOTIONS ELEMENTAIRES) SONT COHERENTES AVEC LE PROJET DU TRAVAIL ; CONNAITRE LES LIMITES DE L'ANALYSE PAR SOUS GROUPE ; CONNAITRE LA NOTION DE PERDUS DE VUE.

1. LES QUESTIONS A SE POSER

Certains aspects sont un peu complexes ; nous nous contenterons de lister quelques points-clés simples à vérifier :

a. La stratégie d'analyse est-elle cohérente avec le projet du travail ?
   i. Si l'étude est un essai clinique d'efficacité (de supériorité ou de différence), l'analyse est-elle réalisée en intention de traiter ?
   ii. Si l'étude est un essai clinique de non infériorité, deux analyses sont-elles réalisées, l'une en intention de traiter, l'autre en per protocole, et leurs résultats ont-ils été confrontés ?
   iii. Si l'étude est observationnelle à visée analytique, une analyse multivariée avec ajustement sur les facteurs de confusion ou sur les facteurs pronostiques a t-elle été effectuée ?
b. Les tests choisis sont-ils adaptés au type de données analysées ?
c. Les intervalles de confiance des résultats sont-ils précisés ?
d. Si des analyses en sous-groupes ont été effectuées, étaient-elles prévues dans le protocole, argumentées, avec un but précisé (exploratoire, OU confirmatoire)? La méthode utilisée permet-elle de fournir une conclusion valide pour ces analyses ? (La randomisation a t-elle été stratifiée ? L'interaction a t-elle été testée ? A t-on corrigé le risque alpha ? L'interprétation repose t-elle sur l'interaction ? Le calcul du NSN a-t-il tenu compte des analyses en sous-groupes ?)
e. S'il y a des perdus de vue, engendrent-ils une perte de puissance importante ? Engendrent-ils un biais majeur ?
f. S'il y a une (des) analyses intermédiaires,
   i. Ont-elles été prévues dans le protocole ?
   ii. A t-on tenu compte de ces analyses intermédiaires dans le calcul du NSN ?
   iii. A t-on précisé un seuil de signification pour chacune d'elles pour fonder les conclusions statistiques ?

Pour les notions de causalité se reporter au point D et à l'objectif 17

2. OU CHERCHER

Stratégie d'analyse et analyses statistiques : partie méthode: analyse, partie résultats, diagramme de flux (stratégie d'analyse)

Pour les analyses en sous-groupes : planification et argumentation : partie introduction

Planification: partie méthode : randomisation et analyses

Interprétation : partie résultats et discussion.

Pour les perdus de vue : partie méthode (NSN) et partie résultats et discussion (biais, puissance).

Pour les analyses intermédiaires : parties méthode et résultats

3. RAPPELS THEORIQUES

A. Intention de traiter

Définition: Analyse en intention de traiter = Méthode qui consiste à analyser les données de tout patient inclus, et ce dans le « bras » (groupe de tirage au sort) dans lequel il a été randomisé au début de l'étude.

Cette analyse permet:

- De ne pas détruire la randomisation
- D'analyser les données de façon pragmatique, comme dans la réalité clinique (le traitement pouvant être non actif chez certains patients et donc changé par le praticien). On compare donc les deux interventions (intervention à tester et intervention contrôle) sur leurs effets et ce dans les conditions les plus proches de "la pratique courante".
- Donc d'analyser les sujets randomisés dans leur groupe de randomisation:
  - ✓ Quelle que soit leur observance au traitement
  - ✓ Quel que soit le traitement réellement reçu
  - ✓ Quel que soit l'éventuel retrait du patient de l'étude ou l'éventuelle déviation au protocole.

Vérification

- Les auteurs auront dans la plupart des cas noté, en partie méthodologie - analyse -, que leur analyse a été faite "en intention de traiter".
- La vérification s'appuie sur le tableau du flow-chart ainsi que sur les tableaux de résultats en comparant le nombre de patients randomisé et le nombre de patients analysés dans chaque groupe de traitement.

Cette méthode se distingue de

- "l'analyse per protocole" où seuls les sujets ayant suivi jusqu'au bout le protocole sont analysés (des sujets seront exclus de l'analyse sur la base d'informations observées après leur randomisation : non observants, changement de posologie ou arrêt de traitement par le médecin, perdus de vue, permutation de traitement…)
- et de « l'analyse en traitement réellement reçu » où les sujets sont analysés selon le traitement effectivement reçu, même s'il ne s'agit pas du traitement qui leur avait été attribué initialement par la randomisation.

Comparaison des résultats et conclusions issus des analyses per protocole ou en traitement réellement reçu et en ITT.

En général

- Différence estimée avec l'analyse per protocole plus grande qu'avec ITT.
- Puissance moins élevée avec analyse per protocole qu'avec ITT
- Biais de sélection potentiels plus élevés avec analyse per protocole qu'avec ITT.
- Tendance de l'ITT à être conservatrice (ne pas mettre en évidence une différence qui existe vraiment).

Remarque : Moins il y a d'écarts au protocole, de perdus de vue, d'arrêts prématurés des traitements, de données manquantes, plus les résultats des deux types d'analyses sont proches. En l'absence d'écarts au protocole, de sorties d'étude, d'arrêts prématurés des traitements, de données manquantes, les résultats des deux analyses sont identiques.

B. Les tests choisis sont-ils adaptés à la question et au type de données analysées

On vérifie que les tests classiques pour données qualitatives ou quantitatives, décrits dans la partie méthode, et référencés, ont été utilisés, de même que les tests spécifiques en cas d'écarts aux conditions de validité des tests classiques (ex : faible nombre de sujets (<30), distribution non gaussienne), de séries appariées, de type d'étude particulier.

Les tests doivent être appropriés aux données, et il est donc indispensable de vérifier l'adéquation du test statistique avec :

- Le type de la variable,
- La distribution des variables. Les variables non gaussiennes doivent le plus souvent être analysées avec des tests non paramétriques, (c'est le cas des petits effectifs),
- Le type d'expérience : groupes distincts (parallèles) ou mesures répétées chez un même sujet.

(Le détail des tests à utiliser en fonction de toutes ces données est joint en annexe 1. Une liste de références bibliographiques adaptées est jointe en annexe 2).

Si l'objectif de l'étude était une comparaison de groupes, vérifier qu'un test statistique a bien été effectué pour comparer ces groupes.

Si l'étude est observationnelle avec une recherche de facteurs de risque ou de facteurs pronostiques, vérifier qu'un ajustement sur les facteurs de confusion ou les autres facteurs pronostiques a été réalisé.

C. Les intervalles de confiance des résultats sont-ils précisés ?

L'intervalle de confiance d'un résultat indique les limites à l'intérieur desquelles la "vraie" valeur estimée (différence entre traitements, risque relatif, sensibilité etc…) est susceptible de se trouver. Qu'il s'agisse de pourcentage, de moyenne, de risque, il est indispensable d'assortir les résultats de leur intervalle de confiance (en général à 95%, soit alpha = 5%), ce qui permet de juger de la précision des résultats et des valeurs auxquelles on peut s'attendre lorsque l'on appliquera les résultats dans la pratique.

D. Analyses en sous-groupes

1. **Il existe trois cas**
   - Analyses en sous-groupes planifiées (prévues dans le protocole)
     - A visée de confirmation
     - A visée exploratoire
   - Analyses post-hoc (non planifiées, non prévues dans le protocole)

   Si elles n'ont pas été planifiées comme analyses à visée de confirmation, les analyses en sous-groupe sont toujours exploratoires. Aucune conclusion clinique ne peut être fondée sur les résultats des ces analyses non planifiées ou à visée exploratoire.

2. **Conditions nécessaires pour fonder une conclusion clinique.** Les analyses en sous-groupe doivent être
   - planifiées comme analyse à visée de confirmation,
   - justifiées (cf arguments § E.5),

- limitées en nombre
- la randomisation doit avoir été stratifiée sur la variable qui définit les sous-groupes *(exemple genre).*
- l'essai doit avoir la puissance suffisante pour détecter les interactions (remarque : généralement le NSN n'est pas calculé pour garantir une puissance pour détecter une certaine interaction pour des raisons de faisabilité et de coût).
- l'interaction doit avoir été testée.
  - **Définition interaction** (glossaire) **:** Mesure dans laquelle l'effet d'un facteur est modifié en fonction de l'action d'un ou de plusieurs autres facteurs. *Ex : interaction du genre sur l'effet d'un traitement*
  - **Interaction quantitative** : le traitement est bénéfique (ou délétère) dans tous les sous-groupes définis par une variable mais a une taille d'effet différente. *EX: effet bénéfique du traitement A deux fois plus important chez les femmes que chez les hommes.*
  - **Interaction qualitative** : le traitement est bénéfique dans un sous-groupe mais est délétère dans un autre. *Ex: effet du traitement B bénéfique chez les femmes mais délétère chez les hommes.*

3. **Conclusions d'une analyse en sous-groupe**

- **Si l'essai est concluant** (différence significative quand on analyse tous les sujets globalement), si **l'interaction est très significative, et si elle est justifiée par un argument fort** (voir ci-dessous), **les résultats des analyses en sous-groupes peuvent être pris en compte dans la conclusion clinique.**) On peut alors conclure que l'effet du traitement est différent entre les sous-groupes : *exemple - traitement efficace chez les femmes mais délétère chez les hommes.*

  Il faudra cependant **corriger le risque alpha** en raison de la multiplicité des tests réalisés (inflation du risque alpha) si l'on souhaite conclure dans des analyses à visée de confirmation.

- **Attention** aux interactions significatives par le fait du hasard.
- **Si l'essai est non concluant** (pas de différence significative quand on analyse tous les sujets), **les analyses en sous-groupes devraient être exploratoires.**
- **En règle générale, les résultats d'une analyse en sous-groupe nécessitent d'être confirmés dans une autre étude.**

4. **Arguments justifiant les analyses en sous-groupe** : ils peuvent être de nature diverse
   - Exemples : Variation génétique
     - *Exemple en évaluation thérapeutique : Dans le cancer du colon, la réponse à la chimiothérapie dépend de l'expression de gènes. La conservation de l'hétérozygotie au niveau des loci des chromosomes 17p et 18q est un facteur prédictif de l'effet d'une chimiothérapie avec du fluorouracile. Le traitement augmente la survie chez les patients n'ayant pas perdu l'hétérozygotie. En revanche, il ne présente pas de bénéfice chez les patients ayant complètement perdu l'hétérozygotie.*
     - *Exemple en recherche épidémiologique: Infection au VIH et récepteur CCR5 des lymphocytes T CD4+. Le corécepteur CCR5 est indispensable à l'entrée du VIH dans les lymphocytes T CD4+. Certains sujets homozygotes pour la délétion de la séquence génétique qui rend CCR5 non fonctionnel sont résistants à l'infection par le VIH.*

E. Analyses intermédiaires

- **Définition (glossaire)** : Analyse effectuée avant l'inclusion de tous les sujets prévus. Elle est réalisée le plus souvent lorsque l'étude est longue. **Elle doit être prévue dans le protocole (argumentée), et le nombre de sujets nécessaires prend en compte le nombre d'analyses intermédiaires qui sont prévues. Pour chaque analyse intermédiaire, un seuil de signification doit être déterminé (plusieurs méthodes sont possibles).**
- Arguments pour justifier une analyse intermédiaire:
    - Les hypothèses de taille d'effet ne sont pas très précises.
        i. Dans le cas où on prévoit un bénéfice net précoce : il n'est pas éthique de priver des patients d'un traitement qui s'avérerait plus efficace que prévu. le principe d'équipoise entre les groupes de traitements n'est plus respecté, autrement dit, il n'y a plus de réelle incertitude à l'intérieur de la communauté scientifique quant au meilleur traitement. : arrêt de l'essai pour efficacité.
        ii. De la même façon, il n'est pas éthique de soumettre des patients à un traitement qui s'avérerait moins efficace que prévu (arrêt de l'essai pour futilité)
    - Les effets secondaires attendus sont potentiellement graves pouvant remettre en question la balance bénéfice-risque : il n'est pas éthique de poursuivre un essai dans ces conditions : arrêt de l'essai pour sécurité.

- Pour chaque analyse intermédiaire, le seuil de signification doit avoir été corrigé pour garantir un risque α global. *Ex : si α global fixé à 5%, $alpha_{int}$ << 5%.*
- Dans tous les cas, les règles d'arrêt ou de poursuite de l'essai doivent être définies explicitement a priori.

F. Perdus de vue

- **Définition** (glossaire) : Patient qui n'est pas suivi sur la totalité de la période prévue par le protocole d'un essai ou d'une étude épidémiologique. On ne sait pas si le patient a guéri, s'il a eu une complication ou des effets secondaires, et pourquoi il n'est pas revenu.
- **Conséquences des perdus de vue**
    - Perte de précision et/ou perte de puissance dans une comparaison
    - Biais de sélection si les perdus de vue sont différents (pour les facteurs pronostiques) des patients restés dans l'étude. Biais négligeable si taux < 10%, acceptable si 10% ≤ taux < 20%, important si taux ≥ 20% (taux empiriques). Le biais est d'autant plus important que les taux de perdus de vue sont déséquilibrés entre les groupes comparés. *Ex: 30% vs 10%*

OBJECTIF 10 : VERIFIER LE RESPECT DES REGLES D'ETHIQUE

1. LES QUESTIONS A SE POSER

- Le protocole de l'étude a-t-il été soumis à un comité d'éthique indépendant avant l'inclusion du premier patient ? A-t-il reçu l'approbation du comité?
- Le consentement, libre et éclairé, de chaque patient a-t-il été recueilli ?
- Les données sont elles anonymes ? Sinon ont-elles reçu une habilitation de stockage?

Les règles de confidentialité des données recueillies ont-elles été énoncées et respectées ?

- Si, dans un essai clinique, on utilise un placebo, cette utilisation est-elle justifiée ?

2. OU CHERCHER

Dans la partie méthode

Pour l'utilisation du placebo, parties introduction, méthodes, discussion.

3. RAPPELS THEORIQUES :

A. Les trois principes fondamentaux

Ils découlent de la **déclaration d'Helsinki**, adoptée au niveau international et donnant des recommandations concernant tous les participants à une recherche biomédicale et **des bonnes pratiques cliniques** (conférence internationale d'harmonisation (ICH : international conference of harmonization).

Respect de la personne humaine:

Toute recherche implique une information juste, complète et accessible au sujet quant aux finalités et aux modalités de l'étude, ainsi qu'aux risques auxquels le sujet pourrait être exposé.

Le sujet doit donner son consentement libre et éclairé à sa participation à l'étude

Principe d'utilité

Toute recherche doit amener un "bien".

La balance bénéfice-risque doit être en faveur du bénéfice pour le participant.

Principe de justice

Les sujets, égaux en dignité et en droits, doivent être traités de façon équitable.

Les populations vulnérables ont besoin d'une protection spéciale. 'La recherche médicale impliquant une population ou une communauté défavorisée ou vulnérable se justifie uniquement si la recherche répond aux besoins et priorités sanitaires de cette population ou communauté et si, selon toute vraisemblance, les résultats de la recherche seront bénéfiques à cette population ou communauté (déclaration d'Helsinki).

B. Conséquences

1. Toute recherche bio-médicale sur un individu, son dossier médical ou des produits issus du corps humain (produits sanguins, tissus…) est soumise à une réglementation stricte (*en France, loi du 20 décembre 1988 modifiée en 2004, dite loi "Huriet-Sérusclat", et en mars 2012 (dite loi "Jardé") lois de bioéthique et loi "informatique et libertés"*) et donc à l'approbation de comités et d'instances légales.

2. Le projet doit avoir été soumis par l'investigateur à un **Comité d'Ethique Indépendant** (*CPP ou Comité de Protection des Personnes, en France, ethical committee, ailleurs*. Parfois l'accord du comité institutionnel (*Institutional Review Board, IRB*) est mentionné. C'est une instance locale qui ne garantit pas l'indépendance par rapport aux investigateurs et/ou au promoteur. Dans le cadre d'un projet international: soumission à un comité par pays. L'approbation doit avoir été donnée avant toute inclusion dans l'étude.
3. Corollaire:
   - ✓ Dans le cas où l'accord du Comité d'éthique n'a pas été obtenu avant le début de l'étude, les auteurs doivent le noter en en donner les raisons.
   - ✓ Dans le cas où la recherche ne nécessiterait pas l'accord du Comité, les auteurs doivent en spécifier les raisons et spécifier comment ils se sont assurés que leur travail sortait bien de ce cadre.
   - ✓ Toute modification du protocole nécessite un accord complémentaire du (des) comité(s)
4. Tout patient doit donner son **consentement libre et éclairé** à sa participation à l'étude. En cas contraire, le patient ne peut être inclus. Sa non inclusion doit être notée dans le tableau de flux. Un patient peut se retirer à tout moment de l'étude.
5. Le promoteur de l'étude doit avoir souscrit une assurance couvrant les risques pour les participants, le cas échéant.
6. Toute collecte automatisée de données permettant l'identification, même indirecte, des participants, doit avoir été soumise au Comité d'Ethique (*pour la France, à la CNIL, Commission Informatique et Libertés*) et reçu un avis favorable préalablement à toute inclusion.
7. En cas de recherche impliquant la personne humaine de type interventionnel , l'avis de l'autorité compétente (*en France AFSSAPS, Agence française de sécurité sanitaire des produits de santé*) est obligatoire.

Remarque: Ces trois derniers points ne sont généralement pas mentionnés dans les articles publiés dans les revues internationales

C. **Conditions d'utilisation d'un placebo dans les essais cliniques**

- **Déclaration d'Helsinki**

Les bénéfices, risques, dangers, efficacité d'une nouvelle intervention doivent être comparés à ceux de la meilleure intervention actuelle reconnue (preuve scientifique établie de son efficacité), sauf dans les cas suivants:

i. L'utilisation d'un placebo ou l'absence de traitement est acceptable dans les études où aucune intervention reconnue n'existe
ii. Quand pour des raisons irréfutables et scientifiquement valables, l'utilisation d'un placebo est justifiée pour évaluer l'efficacité et la tolérance d'une intervention, et les patients qui reçoivent le placebo
iii. d ne seront pas exposés à un risque/danger grave ou irréversible.

- Conséquences
  i. Comité d'éthique juge si, au vu des connaissances du moment, l'utilisation d'un placebo est éthiquement acceptable => rechercher si l'étude a été approuvée par un comité d'éthique
  ii. Rechercher si consentement libre et éclairé a été signé par les patients

OBJECTIF 11°: ANALYSER LA PRESENTATION, LA PRECISION ET LA LISIBILITE DES TABLEAUX ET DES FIGURES, LEUR COHERENCE AVEC LE TEXTE ET LEUR UTILITE.

1. LES QUESTIONS A SE POSER

A. Quels sont les critères de qualité d'un tableau ?
B. Quels sont les critères de qualité d'une figure ?
C. Où chercher les informations ?

2. OU CHERCHER

Tout ce qui figure dans un tableau ou une figure concerne généralement les résultats de l'étude. C'est donc dans la section « Résultats » que se situeront les tableaux et les figures, ainsi que le texte s'y rapportant.

3. RAPPELS THEORIQUES :

1. TABLEAUX

Les tableaux sont indiqués pour présenter soit des données répétitives sous forme synthétique soit des données utiles pour vérifier des résultats importants. Les données répétitives sont par exemple les données servant à vérifier la comparabilité des groupes dans un essai randomisé. Ils servent à gagner de la place et à gagner en clarté. Les critères de qualité d'un tableau à vérifier concernent les points suivants :

1. Sa taille doit être raisonnable. Un tableau trop grand présentant des types d'informations trop différentes est difficile à lire.
2. Il doit comporter un titre informatif situé au dessus du tableau et doit indiquer le contenu du tableau et la population concernée (individus, lieu, temps).
3. Les variables (ou facteurs) qu'il décrit doivent être identifiées clairement (les abréviations doivent être définies dans une note de bas de tableau), avec leur unité pour les variables quantitatives. Les effectifs de sujets sur lesquels sont donnés les résultats doivent être indiqués systématiquement (et pas seulement les pourcentages).
4. Les nombres figurant dans les cellules du tableau représentent les résultats : ceux-ci doivent être présentés de façon identique et regroupés pour toutes les variables du même type (qualitative versus quantitative)
5. Les indices de dispersion doivent être reportés. Il n'est pas informatif de donner une moyenne sans son écart type, ou une médiane sans les quartiles ou les valeurs extrêmes. Lorsqu'on donne les résultats d'une estimation telle que prévalence, incidence, risque relatif, odds ratio, sensibilité…, l'intervalle de confiance et son niveau de confiance doivent être donnés *(ex : sensibilité 97%; intervalle de confiance à 95% = [87,9 ; 99,6]).*
6. Les valeurs statistiquement significatives doivent être clairement identifiables.
7. Il doit pouvoir se lire de façon autonome (auto-suffisant : il n'est pas nécessaire de s'aider du texte pour en comprendre le sens).
8. Il doit être référencé dans le texte.
9. Les informations qu'il contient ne doivent pas être redondantes avec le texte. Cependant, pour les données concernant les résultats principaux de l'étude, il est normal de les énoncer de façon synthétique dans le texte et de les faire figurer également avec tous les détails nécessaires dans le tableau.

2. FIGURES

Elles doivent être claires, compréhensibles et crédibles. Elles servent à représenter des résultats importants. Les critères de qualité d'une figure à vérifier concernent les points suivants :

1. Y a-t-il un titre informatif à la figure ? Une légende explicative ?
2. La figure est-elle bien référencée dans le texte ?
3. Les axes sont-ils sur la figure ? Sont-ils bien référencés ? Les échelles et les divisions sont-elles bien marquées ? Les unités sont-elles précisées ? Les minimums et maximums choisis sur chaque axe sont-ils pertinents ?
4. Le nombre de figures est-il raisonnable ? N'y a t-il pas de redondances d'information avec le texte?
5. Les figures portant sur les mêmes grandeurs sont-elles toutes à la même échelle pour éviter les distorsions de lecture?
6. Les effectifs sont-ils mentionnés ? La variabilité est-elle représentée (écart-type, intervalle de confiance…). Les valeurs statistiquement significatives sont elles clairement identifiables ?

OBJECTIF 12°: VERIFIER LA PRESENCE DES INDICES DE DISPERSION PERMETTANT D'EVALUER LA VARIABILITE DES MESURES ET LEURS ESTIMATEURS

1. LES QUESTIONS A SE POSER

A. Tous les résultats concernant les critères de jugement sont-ils exprimés avec les intervalles de confiance voire le paramètre de dispersion approprié ?
B. Pour la description initiale de l'échantillon concernant les variables quantitatives, a-t-on choisi correctement la présentation de la valeur centrale de la distribution (moyenne ou médiane) et de sa dispersion (écart type, quartiles..) ?
C. Le choix de la représentation des résultats par des figures ou des tableaux est-il judicieux ?

2. OU CHERCHER LA PRESENCE DES INDICES DE DISPERSION ?

Dans la section résultats sans oublier les tableaux et figures.

3. RAPPELS THEORIQUES : INDICES DE DISPERSION ET INTERVALLE DE CONFIANCE

Un indice de dispersion renseigne sur l'étalement de la distribution d'une variable. Il complète le paramètre de position et dépend de la nature de la variable.

Quand on estime un critère de jugement, on doit renseigner la précision de l'estimation par un indice de dispersion ou par l'intervalle de confiance à 95%. Pour chaque type de variable, l'estimateur et l'indice renseignant la précision sont reportés dans le tableau ci-dessous.

Variables, estimateurs et indices renseignant la précision

Variable	***Estimateur***	***Indice de dispersion /intervalle de confiance***	***Exemples***
Qualitative	pourcentage	IC 95%	
Quantitative	Moyenne (distribution gaussienne)	Ecart-type (= SD) (distribution gaussienne) IC 95%	*Age: moyenne, (écart-type) : 61,3 (11,3) (N=299)*
	Médiane (distribution non gaussienne)	Interquartile, (Q1-Q3)=P25 - P75 min - max	*Age: médiane (Q1 - Q3) : 25,7 [15 -47]*
Prévalence incidence	Estimation ponctuelle	IC 95%	*Prévalence de l'asthme induit par l'effort: 9,9% IC95%: [8,2% – 11,7%]*
Sensibilité Spécificité VPP, VPN	Estimation ponctuelle	IC 95%	*Sensibilité du scanner multicoupe vs coronarographie, Nb/Tot (% [IC95%]): 149/157 (IC95%, [8,2% – 11,7%])*
OR RR, RRI HR	Estimation ponctuelle	IC 95%	*RRI de Tachyarythmie auriculaire de l'amiodarone versus placebo RRI= 0,52 IC95% [0,34 - 0,69]*

Abréviations : IC 95% : intervalle de confiance à 95%; VPP : valeur prédictive positive; VPN : valeur prédictive négative; OR : odds ratio ou rapport de cotes ; RR : risque relatif; RRI ou HR : rapport de risques instantanés ou hazard ratio; SD : déviation standard ; Min : minimum; Max : maximum; Interquartile : les valeurs du premier quartile (Q1) ou du 25ème percentile (P25) et du troisième quartile (Q3) ou du 75ème percentile (P75) sont données

Remarque : Les indices reportés sur même ligne doivent être associés.

Exemple : un interquartile ne doit pas accompagner une moyenne.

Ecart-type de la moyenne (SEM) - Attention: ne pas confondre l'écart-type de la moyenne (SEM) et l'écart-type (SD). $SEM = SD / \sqrt{N}$

où N est l'effectif et SD la déviation standard ou écart-type.

Si N=100, SEM est 10 fois plus petit que l'écart-type. La dispersion de la variable est pourtant toujours la même car elle est donnée par l'écart-type.

OBJECTIF 13°: DISCUTER LA NATURE ET LA PRECISION DES CRITERES DE JUGEMENT DES RESULTATS

1. LES QUESTIONS A SE POSER : CRITERE DE JUGEMENT

A. Sa définition est-elle précise ?

Le moment où le critère de jugement (CJ) est recueilli et sa méthode d'évaluation sont-ils précisés ?

B. Est-il pertinent pour répondre à l'objectif de l'étude ?

- Est-il clinique, intermédiaire ou de substitution ?
- Le moment de son recueil est-il pertinent ?
- Est-il unique simple, unique composite ou multiple ?
- Le critère de jugement principal (CJP) est-il unique et pertinent ? (à défaut : consensuel ?)
- S'il est unique composite, ses composantes ont-elles toutes la même pertinence clinique ?
- Si le CJP est multiple, les critères ont-ils tous la même pertinence clinique ?
- S'il y a plusieurs CJ, sont-ils hiérarchisés ?

C. Est-il sujet à des biais de classement/information ?

- Sa mesure est-elle standardisée ? (identique pour tous les sujets)
- Est-il objectif ou subjectif ?
- S'il est subjectif, est-il validé ? (référence dans le texte à une étude de validité)
- Est-il fiable et reproductible ?
- Le recueil du critère de jugement est-il fait en insu du groupe d'étude ?

D. Sa précision est-elle adéquate ?

- Le nombre de sujets nécessaire a t-il été calculé pour le CJP ?
- L'intervalle de confiance (IC) au risque alpha (alpha est défini pour le calcul du NSN dans la section méthodes, généralement alpha = 5% et IC à 95%) est-il donné pour le CJP ?
- Si les critères de jugement ne sont pas hiérarchisés ou si le CJP est multiple, a t-on précisé une règle de décision pour conclure au niveau statistique ?

2. OU CHERCHER LES INFORMATIONS: RELATIVES A LA NATURE ET A LA PRECISION DES CRITERES DE JUGEMENT DES RESULTATS ?

- Pour la nature des critères de jugement, dans la section méthodes et parfois dans la section introduction (objectif)
- Pour la précision des critères de jugement, dans les sections méthodes et résultats

3. RAPPELS THEORIQUES :

3.1. DEFINITION (GLOSSAIRE CNCI)

Critère de jugement : variable observée et/ou mesurée dont l'interprétation va permettre de répondre à la question posée dans l'objectif.

Dans un essai clinique, il s'agit du critère qui permet d'évaluer l'effet du traitement ou d'une intervention.

Dans une enquête épidémiologique, il s'agit d'un événement ou de la survenue d'un événement. (enquête exposés/non exposés), ou d'une exposition (enquête cas-témoins), d'un facteur pronostique.

Toujours discuter la nature d'un critère de jugement en fonction des objectifs de l'étude.

3.2. CARACTERISTIQUES

Nature

❶	Unique simple	*survenue de décès dans les 6 mois post randomisation*
	Unique composite	*survenue de décès ou d'infarctus du myocarde dans les 6 mois post randomisation*
	Multiple	*score d'incapacité à l'échelle AVQ-B et score à l'échelle AVQ-D*
❷	Objectif	*décès, poids*
	Subjectif	*échelle de qualité de vie SF36*
❸	Clinique	critère de jugement correspondant à des objectifs thérapeutiques (objectif de guérison (maladie bénigne), de prévention primaire ou secondaire, symptomatologique, amélioration de la qualité de vie). *Ex: survenue d'infarctus du myocarde, décès*
	Intermédiaire	critère de jugement documentant les mécanismes d'action du traitement *Ex: paramètres biologiques ou physiologiques, pression artérielle* *Ex taille des cancers dépistés dans l'évaluation d'un dépistage*
	De substitution	examen de laboratoire ou signe physique utilisé à la place d'un critère clinique. Il doit être très bien « corrélé » au critère clinique ou prédictif du critère clinique. *Ex: volume d'expiration forcée au lieu de brachypnée*

Exemples

Pathologie	*Critère intermédiaire*	*Critère clinique*
Hypertension artérielle	*Pression artérielle*	*Evènements coronariens, AVC*
Dépression	*Score de dépression*	*Normalisation du score de dépression*
Rhumatisme inflammatoire	*Vitesse de sédimentation*	*Handicap fonctionnel, déformation*
Ostéoporose	*Densité osseuse*	*Fracture du col fémoral*
Infarctus du myocarde	*Taux de reperméabilisation coronaire*	*Décès, évènements coronariens*
Diabète non insulino-dépendant	*Glycémie, hémoglobine glycosylée*	*Evènements cardiovasculaires, rétinopathie, insuffisance rénale ..*

Hiérarchisation des critères de jugement

On distingue le critère de jugement principal ou primaire (CJP) des critères de jugement secondaires (CJS).

L'étude vise à répondre à l'objectif principal exprimé avec le critère de jugement principal.

Le calcul du nombre de sujets nécessaire est basé sur le critère de jugement principal.

Pertinence d'un critère de jugement

Un critère est pertinent s'il permet de répondre à l'objectif de l'étude.

Ce tableau regroupe les critères de jugement attendus et les paramètres statistiques associés en fonction du type de question

Question	***Critère de jugement & paramètre statistique associé***
Epidémiologie descriptive	Pathologie : Incidence d'une pathologie
	Pathologie : Prévalence d'une pathologie
Evaluation de traitement	Efficacité: survenue d'un événement mesurant l'efficacité; incidence de l'événement, qualité de vie : score sur échelle spécifique
	Sécurité: effets secondaires; incidence des effets secondaires.
Evaluation de procédure diagnostique	Fiabilité:Signe, pathologie; indice de concordance (ex: kappa)
	Validité: Signe, pathologie; sensibilité, spécificité, rapport de vraisemblance, valeurs prédictives, valeur seuil pour les tests quantitatifs / gold standard
Evaluation d'une procédure de dépistage	Qualité : sensibilité, spécificité, valeurs prédictives de la *procédure* (test+bilan diagnostique) Efficacité : critère indépendant de l'avance au diagnostic (ex : mortalité et non survie dans le dépistage des cancers)
Epidémiologie analytique, Recherche de facteur de risque	Survenue de maladie; Incidence maladie chez exposés et non exposés (cohorte)
Estimation d'un pronostic	Survenue (incidence) de décès, complication, rechute; (Variable dépendant du temps)

Si on recherche une pertinence clinique (recherche clinique à visée pragmatique, épidémiologie) le CJP doit être clinique ou correspondre à des marqueurs biologiques de l'évolution de la maladie *(ex taux de CD4 dans le VIH, …)*

En recherche clinique à visée explicative, le CJP peut être intermédiaire. Il faut alors s'assurer de la relation de transitivité entre ce critère intermédiaire et un critère clinique.

Si le CJP est unique composite, il faut que ses composantes aient la même pertinence clinique

Contre exemple: décès, infarctus du myocarde, hospitalisation

Critères de jugement et biais

L'évaluation d'un critère de jugement est sujet à des biais de classement (cf objectif 14 - biais).

Un CJ objectif est fiable et reproductible. Il est très peu sujet à des biais de classement. Un CJ subjectif est sujet à des biais de classement.

Pour limiter le biais de classement il est nécessaire que le CJ soit validé (c'est à dire une étude a évalué sa bonne fiabilité ou sa bonne reproductibilité), le biais de classement est alors limité.

Le biais de classement est différentiel si l'évaluation n'est pas faite en insu du groupe étudié.

Le biais de classement est non différentiel si l'évaluation est faite en insu du groupe étudié.

Critères de jugement et précision

A. Précision de la définition d'un critère de jugement
   - La définition d'un critère de jugement doit être précise : elle contient une variable, sa méthode de mesure et le moment où elle est évaluée
   - *Exemple : acquisition du VIH (évaluée par PCR ADN) entre les âges de 7 jours et 38 semaines*
   - *Si le CJP est la résultante d'un calcul particulier, celui-ci doit être fourni en annexe.*
B. Précision d'un critère de jugement
   - Le nombre de sujets nécessaire est calculé pour le critère de jugement principal. Vérifier que le nombre de sujets recrutés (ou d'évènements) correspond aux hypothèses de calcul du NSN. Remarque : Un critère de jugement fréquent permet de réduire le NSN.
   - Si l'analyse du CJ est à visée de confirmation, on doit préciser la règle de décision relative au risque alpha pour conclure au niveau statistique (Le risque alpha est précisé). On ne corrige pas alpha si l'analyse est à visée exploratoire. Généralement, l'analyse principale portant sur le CJP est à visée de confirmation.
   - Le résultat doit être exprimé avec l'estimation ponctuelle accompagnée de l'intervalle de confiance au seuil alpha (généralement à IC 95%)
   - Dans le cas d'un CJP multiple, la règle de décision relative au risque alpha pour conclure au niveau statistique doit être précisée. Exemple de règle : dans le cadre d'un essai thérapeutique avec plusieurs CJP, on ne peut conclure à l'efficacité du traitement que si tous les petits $p < 0,05$. Ainsi, si un seul petit p est $> 0,05$, on ne peut pas conclure à l'efficacité du traitement.
   - Les analyses des CJS sont exploratoires sauf mention contraire.

La conclusion ne porte que sur le CJP. Les analyses des CJS sont exploratoires sauf mention contraire. Elles viennent en complément de l'analyse principale du CJP.

Le CJP doit être le plus pertinent possible. Dans un essai clinique, c'est l'intérêt du patient qui prime et non l'intérêt du traitement.

OBJECTIF 14°: RELEVER LES BIAIS QUI ONT ETE DISCUTES, RECHERCHER D'AUTRES BIAIS D'INFORMATION ET DE SELECTION EVENTUELS NON PRIS EN COMPTE DANS LA DISCUSSION ET RELEVER LEURS CONSEQUENCES DANS L'ANALYSE DES RESULTATS

1. LES QUESTIONS A SE POSER

a. Les auteurs soulignent-ils dans la discussion un ou des biais à prendre en compte dans l'interprétation des résultats ? ont-ils cherché à en déterminer le sens (sur/sous-estimation du paramètre) ?
b. Ont-ils essayé de les contrôler ou d'en atténuer l'effet par le protocole mis en place et/ou les ajustements à l'analyse?
c. Existe t-il d'autres biais omis par les auteurs ?
d. L'ensemble des biais identifiés (atténués ou non, explicités ou non) invalide t-il les résultats de l'étude? Donner leur nature, le sens du biais et quantifier le biais si possible

2. OU CHERCHER LES INFORMATIONS RELATIVES AUX BIAIS ?

e. Pour relever les biais qui ont été discutés, dans la section discussion (souvent paragraphe limites de l'étude)
f. Pour rechercher d'autres biais éventuels non pris en compte dans la discussion et relever leurs conséquences dans l'analyse des résultats, dans les sections méthodes et résultats

3. RAPPELS THEORIQUES

3.1. GENERALITES SUR LES BIAIS

Définition (glossaire CNCI)

Un biais est une erreur systématique qui fausse les résultats dans un sens donné .

Attention : ne pas confondre un biais avec une erreur aléatoire (fluctuation d'échantillonnage qui débouche sur une imprécision) ou avec une erreur dans une conclusion statistique (risque alpha et risque bêta ou manque de puissance).

Pour tout biais, donner

- le nom de sa famille,
- son sens (surestimation ou sous-estimation, si possible) et
- le quantifier (biais négligeable, faible, modéré, important, si possible).

On classe les biais en 3 familles (glossaire CNCI) selon leur cause

- Biais de sélection
- Biais de classement
- Biais de confusion

Deux types d'erreurs de classement: biais différentiel et biais non différentiel

- **Un biais est différentiel** si les erreurs sur les informations recueillies affectent différemment les groupes soumis à comparaison. Le sens du biais est imprévisible a priori : on a soit une surestimation soit une sous-estimation de l'association.

  Exemple : l'évaluation de l'exposition des sujets dans une enquête cas-témoin (non en aveugle) faite en connaissant le statut de la maladie

- **Un biais est non différentiel** si les erreurs sur les informations recueillies affectent indifféremment les groupes soumis à comparaison. Ces erreurs sont statistiquement indépendantes du groupe de comparaison. Le sens du biais est prévisible : on a toujours une diminution de l'association entre le facteur étudié et le critère de jugement : OR tend vers 1; RR vers 1; HR vers 1

  Exemple : évaluation de la consommation d'alcool en insu du traitement reçu dans un essai randomisé - évaluation erronée de l'exposition à un facteur pour des questions liées à l'outil de mesure (se produit quels que soient les sujets)

3.2. FAMILLE DES BIAIS DE SELECTION

Définition (glossaire CNCI)

Biais dans la constitution de l'échantillon, qui va se retrouver non représentatif de la population cible pour des facteurs liés au problème étudié.

Attention : il s'agit de l'échantillon analysé.

Processus

Un biais de sélection peut survenir à 2 niveaux :

- **Au niveau de la constitution de l'échantillon**. Il s'agit d'un échantillon non représentatif de la population source ou de la population à laquelle on veut extrapoler les résultats

  Exemple : sujets âgés de 70 ans ou plus sans co-morbidité ne sont pas représentatifs de la population des sujets âgés

- **Au niveau des groupes analysés** : les groupes analysés ne diffèrent pas seulement par le facteur étudié mais aussi par un autre facteur (facteur pronostique) qui peut modifier les résultats.

  Exemple : essai contrôlé randomisé avec 10% de sorties d'étude liées à des effets indésirables dans un groupe de traitement => résultats d'efficacité biaisés

Biais inclus dans cette famille (noms des biais non au programme des ECN)

- **Biais d'attribution** (allocation bias) : biais survenant quand les sujets ne sont pas répartis aléatoirement dans les groupes étudiés

  Exemple : procédure de randomisation détournée

- **Biais d'attrition** : biais survenant quand on exclut de la population d'analyse des sujets initialement inclus dans l'étude et que ces exclusions sont liées au résultat

Exemple : dans un essai randomisé, analyse non en intention de traiter et motifs des exclusions liées au résultat

- **Biais de perdus de vue** : biais survenant quand les motifs des pertes de vue sont liés au résultat.

Comment contrôler les biais de sélection

On peut les contrôler à deux niveaux : lors de la planification de l'étude ou lors de l'analyse

- **Planification**

Recrutement par tirage au sort	concerne surtout les enquêtes d'observation *Exemples : enquête descriptive : sondage aléatoire; enquête cas-témoin : sélection par tirage au sort des témoins dans population d'où sont issus les cas*
Recrutement exhaustif	*Exemple : enquête cas-témoin : recrutement systématique de tous les cas définis par des critères diagnostiques valides et appliqués à tous les sujets* *Exemple: enquête de cohorte, essai clinique, évaluation procédure diagnostique : recrutement systématique de tous les patients éligibles sur une période donnée*
Randomisation du facteur étudié	si l'objectif de l'étude le permet. S'assurer que la randomisation est bien faite (cf objectif 5)
Suivi complet et/ou relances	concerne toutes les études. Pour éviter les perdus de vue dans études prospectives et les données manquantes dans toutes les études

- **Analyse (limite le biais mais ne parvient pas à le neutraliser totalement)**

Comparaison initiale des groupes	Les facteurs pronostiques doivent être similaires (comparabilité clinique et non statistique). Sinon, ajustement sur les facteurs pronostiques déséquilibrés entre les groupes.
Analyse en intention de traiter (ITT) /analyse de tous les sujets inclus dans leur groupe initial de randomisation	L'analyse en ITT concerne uniquement les essais cliniques. Elle n'élimine pas le biais si le suivi des patients est incomplet **et si les** motifs de sortie d'étude ou de perdus de vue sont liés aux résultats
Analyse en intention de dépister	Même commentaire que pour l'analyse en ITT
Ajustement sur facteurs	Il s'agit des facteurs pronostiques déséquilibrés, des facteurs sur lesquels la randomisation a été stratifiée (centre …) et les facteurs d'appariement (enquêtes cas-témoins, enquêtes de cohorte) Remarque : si les groupes analysés sont très différents, un ajustement ne suffit pas à neutraliser le biais car on ne peut pas ajuster sur tous les facteurs pronostiques (tous ne sont pas recueillis)
Analyse de sensibilité	Elle permet de déterminer le sens du biais et de le quantifier dans la mesure du possible *Exemple : analyse de la robustesse des résultats si les perdus de vue sont liés au facteur étudié ou à la réponse.*

3.3. LA FAMILLE DES BIAIS DE CLASSEMENT

Définition (glossaire CNCI)

Biais dans la mesure du facteur de risque ou dans la certitude de la maladie. Cette erreur est quasi inévitable puisque aucun outil de mesure (interrogatoire, examen, test) n'est parfait.

Ce biais peut toucher aussi bien le facteur étudié (par exemple l'exposition dans une enquête de cohorte exposés-non exposés) que les critères de jugement (exemple : score d'un questionnaire, décès pour cause spécifique).

Processus

Un biais de classement peut résulter de

- la subjectivité de l'enquêteur : variabilité
- la subjectivité des enquêtés : mémorisation, refus de répondre, déni
- la méthode de mesure : rythme de suivi différent, temps de mesure différent, outil non valide

Biais inclus dans cette famille (noms des biais non au programme des ECN)

- **Biais d'information/mesure** : biais survenant quand le protocole de mesure n'est pas standardisé (le même pour tous) et/ou est sujet à interprétation.
  Exemple : questionnaire de qualité de vie non validé

- **Biais de mémorisation (recall bias)** : biais survenant lors du rappel de l'exposition ou du critère de jugement dans la vie passée.
  Exemple : enquête cas-témoin : les cas peuvent faire plus d'effort pour se souvenir d'une exposition passée que les témoins

- **Biais d'évaluation :** biais survenant quand des facteurs subjectifs externes influencent l'évaluation du critère de jugement ou de l'exposition.
  Exemple : dans toute étude, ce biais survient quand l'évaluation du critère de jugement ou de l'exposition n'est pas faite en aveugle. Il est moindre quand le critère est objectif.

Comment contrôler les biais de classement

Les moyens à disposition pour contrôler les biais de classement doivent être prévus lors de la planification de l'étude

Les règles suivantes s'appliquent au facteur étudié et au critère de jugement et à toutes les études

- **validité et fiabilité des mesures :**
  - ✓ la mesure doit être précise, exacte et reproductible (reproductibilité intra-observateur et inter-observateur).
  - ✓ Pour les critères subjectifs, il faut une référence dans le texte à une étude de validation; à défaut, des résultats de reproductibilité
    Exemple : questionnaire de qualité de vie (subjectif) doit avoir été validé. Vérifier qu'il y a une référence dans le texte à une étude de validation.

- **Standardisation des procédures**
  - ✓ Les mêmes procédures doivent être appliquées à tous les sujets de l'étude: même procédé de recueil de mesure, même rythme de mesure (si prospectif), même définition du facteur étudié et des critères de jugement pour tous les sujets.
    Exemples : mêmes critères diagnostiques appliqués à tous les sujets dans une enquête cas-témoin, à chaque fois que c'est possible même contenu et rythme de suivi dans les enquêtes exposés/non exposés

- **Evaluation/recueil des données en aveugle/en insu.**
  - ✓ Essais cliniques : évaluation des critères de jugement en insu du traitement
  - ✓ Résultat d'un examen diagnostique en insu du résultat du gold standard et vice versa
  - ✓ Enquête cas-témoin : évaluation de l'exposition en insu du statut malade/non malade du sujet
  - ✓ Enquête de cohorte exposés non exposés : évaluation de la maladie en insu de l'exposition
- **Formation des enquêteurs**
  Pour avoir une standardisation des procédures

3.4. LA FAMILLE DES BIAIS DE CONFUSION

Définition (glossaire CNCI)

Biais provoqué par un facteur de confusion interagissant avec le facteur de risque étudié dans l'étude du lien entre ce facteur et la maladie.

Un facteur F joue le rôle de facteur de confusion dans la relation entre E et M si dans la population cible

- la relation brute entre E et M n'est pas la même que celle obtenue en tenant compte de F (relation ajustée).
- et F est lié à M et F est lié à E mais F n'est pas une conséquence de E

Un facteur de confusion existe au niveau de la population avant même de constituer l'échantillon de l'étude.

Comment neutraliser les facteurs de confusion

On peut neutraliser les facteurs de confusion lors de la planification de l'étude et/ou lors de l'analyse, mais cette neutralisation n'est jamais complète.

- **Planification**

tirage au sort du facteur étudié (randomisation)	=> comparabilité des groupes *Exemple : essais thérapeutiques.* Cette règle n'est pas applicable à toutes les études.
restriction de la population d'étude à certaines catégories	*Exemple : exclusion de patients atteints de bilharziose (facteur de risque connu de cancer de la vessie) dans une enquête étiologique sur le cancer de la vessie en Europe. Les cas de bilharziose sont rares en Europe. Leur exclusion ne met pas en péril l'extrapolation des résultats.*
appariement sur les facteurs de **confusion** voire de risque de la maladie déjà connus	*En général âge, sexe …* enquêtes étiologiques Attention: ne pas confondre facteur de risque et facteur de confusion un sur-appariement peut masquer une liaison

- **Analyse**

ajustement sur les facteurs de confusion	On utilise des modèles multivariés (voir cours de statistiques) qui donnent l'effet propre du facteur étudié indépendamment des autres co-variables entrées dans le modèle. On obtient un OR, RR ou HR ajusté sur les facteurs de confusion entrés dans le modèle multivarié. Attention: ceci nécessite d'avoir prévu dans le protocole de recueillir les facteurs de confusion.

Remarque :

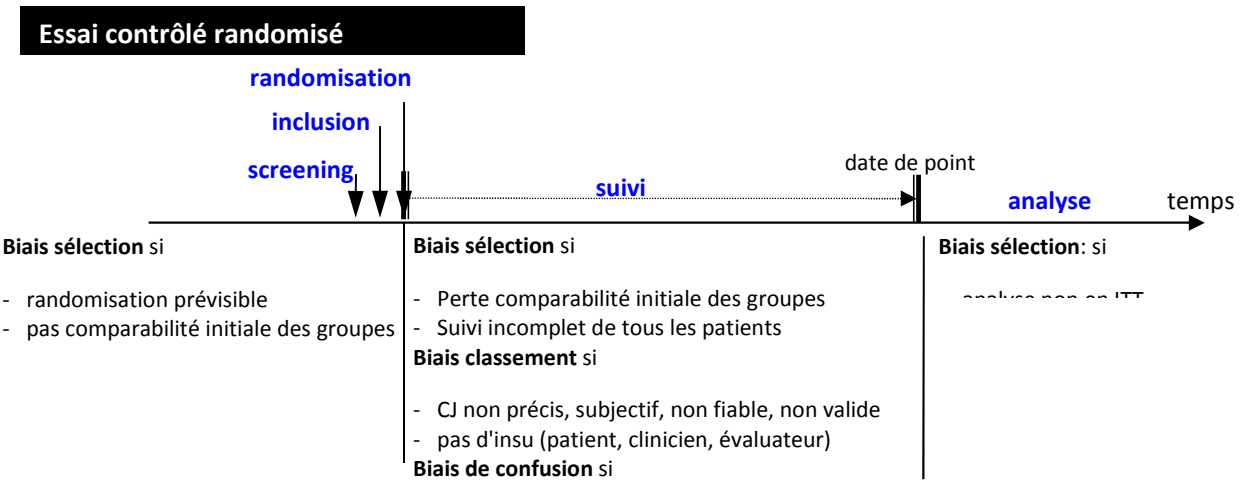
Certaines situations relèvent de plusieurs familles de biais

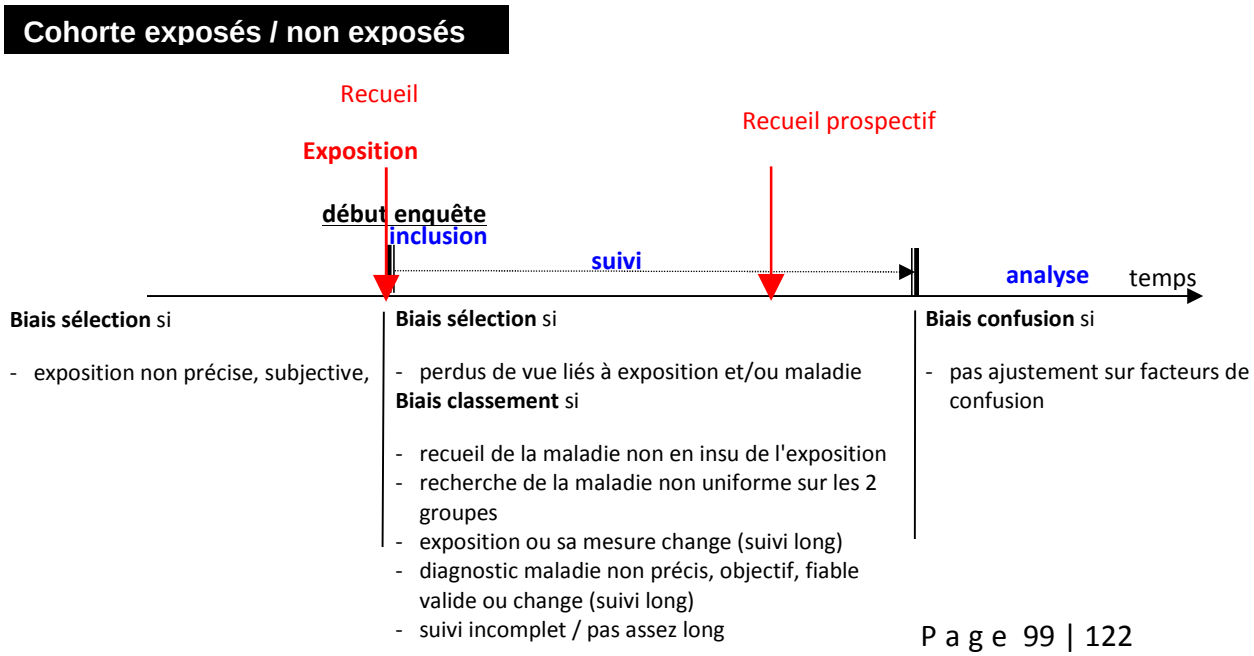
Exemples : dans une enquête cas-témoins où le recrutement se déroule sur une assez longue période avec une évolution des critères diagnostiques, on a à la fois un biais de classement (un sujet classé au départ non malade aurait été classé ensuite malade) et un biais de sélection car on est à la phase de constitution de l'échantillon

Les moyens même mis en place pour prévenir les biais ne garantissent pas l'absence de biais et ne dispensent pas les auteurs d'en discuter l'existence.

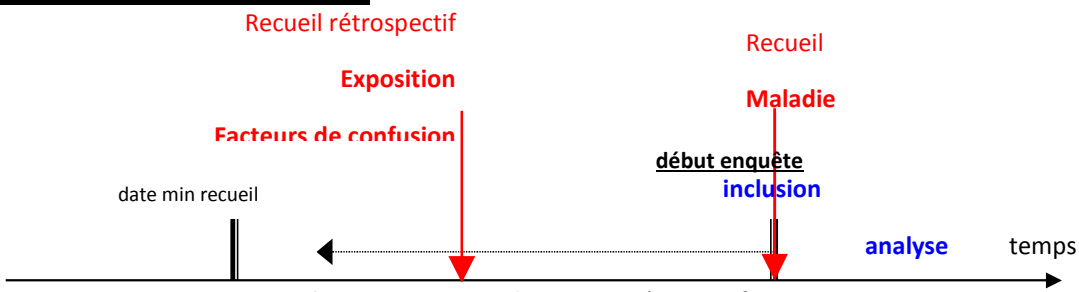
3.5. PLACE DES BIAIS DANS LES ETUDES

Les biais peuvent se situer à chaque stade de l'étude : lors de la planification, lors de la réalisation et lors de l'analyse des données





Cas - Témoins



Biais sélection si

- ✓ Cas : diagnostic maladie non précis, subjectif, non fiable non valide
- ✓ Cas :mêmes critères diagnostiques non appliqués à tous les sujets
- ✓ Ccas recrutés dans services non représentatifs de la population des cas
- ✓ Témoins : témoins non représentatifs de la population des cas

Biais classement car recueil exposition rétrospectif (mémorisation)

- Si recueil de l'exposition non en insu de la maladie
- Si recueil exposition non uniforme sur les 2 groupes
- Si exposition non précise, objective, fiable, valide

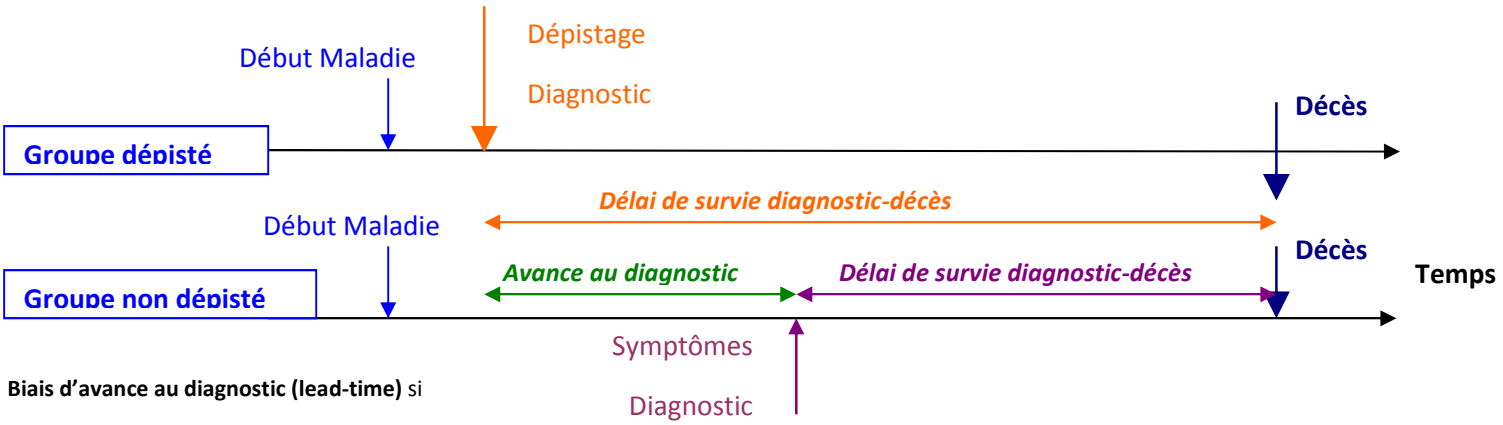
Biais confusion si

- pas ajustement sur facteurs de confusion

Evaluation de procédure diagnostique - Valeur diagnostique
Grille de lecture - Validité des résultats – 8 points clés

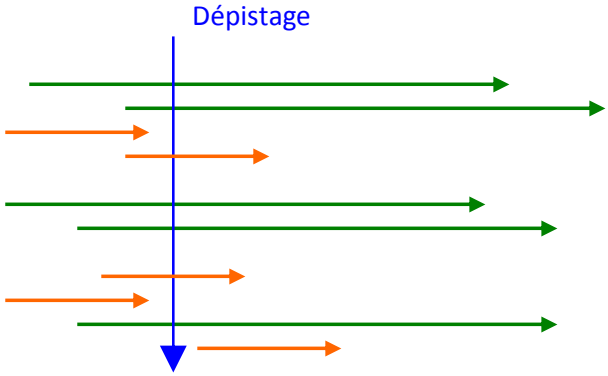
	Oui	Non
• Étude prospective voire transversale ? Les deux examens évaluent le même état du sujet		Biais de classement
• Échantillon de patients représentatif de la population cible (prévalence et sévérité de la pathologie) ?		Biais de sélection
• Définition parfaite de la maladie ?		Biais de sélection
• Test référence (gold standard) valide et reconnu ?		Biais de classement
• Test de référence déterminé *a priori* (indépendant du test étudié) ?		Biais de classement
•Test étudié fiable et ses conditions d'utilisation décrites avec précision (positivité définie) ?		Biais de classement
• Insu des deux tests ?		Biais de classement
• Tous les sujets subissent les deux tests ?		Biais de sélection

Essai randomisé prospectif appliqué au dépistage : biais spécifiques



Biais d'avance au diagnostic (lead-time) si

- Comparaison de la survie diagnostic-décès : toujours plus élevée dans le groupe dépistage même s'il est inefficace (date du décès inchangée)



Biais lié à la vitesse d'évolutivité (length-time)

- tumeurs lentes __: 5/5 des sont identifiées par le dépistage

- tumeurs rapides __: 2/5 des sont identifiées par le dépistage

⇒

- Taux plus élevé dans le groupe dépisté de formes à

Autres biais survenant dans les évaluations de dépistages

- biais de « sur diagnostic »
- biais de sélection liés aux sujets par définition volontaires
- biais d'indication : certains sujets sont adressés au

OBJECTIF 15 : VERIFIER LA LOGIQUE DE LA DISCUSSION ET SA STRUCTURE, RECONNAITRE CE QUI RELEVE DES DONNEES DE LA LITTERATURE ET CE QUI EST L'OPINION PERSONNELLE DE L'AUTEUR

1. LES QUESTIONS A SE POSER :

a. Les auteurs apportent-ils une réponse à la question posée ?
b. L'interprétation des résultats observés dans l'étude est-elle rationnelle ?
c. Les points forts de l'étude sont-ils exposés ?
d. Les points faibles et les limites de l'étude sont-ils répertoriés ? Sont-ils appropriés ?
e. Les auteurs ont-ils confronté leurs résultats à la littérature récente ?
f. Ce qui relève de la littérature a t-il été référencé ?
g. Le caractère généralisable des résultats de l'étude a t-il été discuté ? Si les auteurs concluent qu'on peut extrapoler les résultats à telle population, est-ce justifié ?
h. La conclusion des auteurs est-elle étayée par les résultats de l'étude ?
i. Y a t-il des pistes de recherche complémentaire proposées ?
j. N'y a-t-il pas des erreurs de forme ?
   - ✓ Discussion trop longue, revue de la littérature injustifiée (non en rapport direct avec les résultats de l'étude)
   - ✓ présentation de nouveaux résultats non présentés dans le chapitre résultats (par exemple résultats d'analyse en sous-groupes)…

2. OU CHERCHER LES INFORMATIONS RELATIVES A LA DISCUSSION ?

- dans la section discussion (la section des résultats ne doit pas faire l'objet d'interprétation/ discussion)

3. RAPPELS THEORIQUES

Une discussion doit présenter les 6 points suivants. Généralement elle suit la logique suivante mais ce n'est pas une règle (cf objectif 22).

1.Résumé des principaux résultats

2.Discussion de leur validité : points forts et points faibles

3.Discussion de leur caractère généralisable

4. Mise en perspective de cette réponse : confrontation des résultats à la littérature

5.Conclusion sur la réponse à la question posée : conclusion clinique

6.Implications et ouverture, perspectives

OBJECTIF 16°: DISCUTER LA SIGNIFICATION STATISTIQUE DES RESULTATS.

1. LES QUESTIONS A SE POSER

a. Les tests statistiques sont-ils adaptés à la question posée ?
b. L'interprétation de leurs résultats est-elle adéquate ?
c. La valeur du « p » est-elle interprétée correctement ?
d. Y a t-il une signification clinique associée à la signification statistique ?
e. Le nombre de tests réalisés est-il justifié ?
f. En cas d'analyses en sous-groupes, sont-elles licites et leur interprétation est-elle adéquate? (cf objectif 9)

2. OU CHERCHER LES INFORMATIONS

- dans la partie méthodes, paragraphe analyse statistique pour vérifier leur adéquation à la question posée et la justification des études de sous-groupes
- dans la partie résultats, souvent dans les tableaux pour juger de la signification statistique des résultats.
- Parfois en annexe (indices ou tests particuliers)

3. RAPPELS THEORIQUES

A. Risque alpha et risque beta

La signification statistique exprime une probabilité. Elle varie en fonction du type de test utilisé et est dépendante de l'effectif sur lequel est réalisé le test. Ainsi toute différence, aussi petite soit-elle, peut devenir statistiquement significative, si l'effectif sur lequel elle est calculée est suffisamment grand.

Le risque alpha ou risque de 1ère espèce exprime le risque de déclarer une différence significative (sous l'effet d'un traitement ou d'autres facteurs extérieurs) alors qu'elle n'est que le résultat de fluctuations aléatoires. Ce risque est jugé acceptable s'il est inférieur à 5%.*(ce qui revient à un taux maximal de faux positifs toléré : 5%*.) Autrement dit, c'est le risque de rejeter à tort l'hypothèse nulle et de conclure par exemple qu'un traitement est efficace alors qu'il ne l'est pas.

Le risque bêta ou risque de 2ème espèce exprime le risque de déclarer une différence non significative alors qu'elle traduit en fait l'effet d'un traitement ou d'autres facteurs extérieurs. Il est encore appelé manque de puissance et est jugé acceptable s'il est inférieur à 20% (ce qui revient à un taux maximal de faux négatifs toléré : 20%), voire plus souvent 10% dans les essais thérapeutiques de phase III). Autrement dit, c'est le risque de ne pas rejeter l'hypothèse nulle alors qu'elle est vraie et de conclure par exemple qu'un traitement n'est pas efficace alors qu'il l'est.

Un même ensemble de données peut être analysé de plusieurs manières, dont certaines peuvent ne pas être strictement appropriées. Les points suivants sont donc essentiels :

- Les auteurs doivent indiquer quelle méthode statistique a été employée pour chaque analyse (présence d'un chapitre « considérations statistiques » dans la partie « méthodes »)
- Quand on décrit les caractéristiques de chaque groupe soumis à comparaison, on recherche une comparabilité clinique et non statistique car les effectifs ont été calculés pour détecter

une différence sur le critère de jugement principal et non sur les caractéristiques initiales. Faire des tests statistiques n'est pas recommandé dans ce cas. L'absence de significativité du test ne doit pas conduire à négliger une différence si elle est jugée importante sur le plan clinique

- Une valeur de p très petite ne signifie pas forcément qu'il existe une grande différence entre les groupes de comparaison : ne pas confondre degré de signification statistique (p) et quantité d'effet (ampleur de la différence entre les groupes : modification de la moyenne ou du taux d'évènements). Une valeur de « p » > 5% ne signifie pas qu'il n'existe pas de différence, mais seulement qu'on n'a pas mis en évidence de différence significative (= non due au hasard). Plusieurs raisons peuvent être à l'origine de cette non signification : l'absence réelle de différence, l'existence d'une différence plus petite que celle qu'on désirait détecter *a priori*, le manque de puissance, un biais…
- Toute différence, aussi petite soit-elle, peut devenir statistiquement significative, si l'effectif sur lequel elle est calculée est suffisamment grand mais cette différence statistiquement significative peut ne pas être cliniquement significative, c'est-à-dire n'avoir aucun intérêt pour la santé des patients (cf objectif 17).
- Plus on multiplie les tests statistiques, plus on a de chances d'avoir un test significatif uniquement par hasard. Donc vérifier que les auteurs n'ont pas réalisé des tests tous azimuts mais sur les critères de jugement annoncés et qu'ils ont utilisé des corrections *ad hoc pour garantir un risque α global de 5%*.
- Il est toujours très tentant d'analyser le résultat dans des sous groupes d'intérêt (hommes/femmes, formes graves/bénignes de la maladie etc.). On court alors trois risques : le manque de puissance par diminution des effectifs, l'observation d'un résultat significatif du seul fait du hasard, en raison de la multiplication des tests et la destruction de la comparabilité initiale des groupes. Les analyses en sous-groupes doivent donc être justifiées et prévues AVANT le début de l'étude de façon à ce que la planification de l'étude en tienne compte (justification des analyses, hypothèse à vérifier, nombre de sujets à inclure, stratification de la randomisation, test de l'interaction, correction du risque alpha). Dans le cas où elles seraient décidées a posteriori ou sur un effectif trop faible, leurs résultats ne doivent être donnés qu'à titre exploratoire et doivent être vérifiés avec d'autres études. Elles doivent de toute façon être en nombre limité.

B. LA VALEUR DU « P » EST-ELLE INTERPRETEE CORRECTEMENT ?

Les résultats peuvent être évalués en termes de signification statistique. La valeur de « p » représente la probabilité que la différence observée le soit par hasard, quand en réalité il n'y a pas de différence entre les groupes. Ainsi, une valeur de p de 0,01 signifie que l'on a seulement 1 chance sur 100 d'observer au moins cette différence sous le seul effet du hasard, quand il n'y a pas de différence entre les groupes.

Le seuil de signification (risque α) est fixé à 5% par consensus; on conclut, lorsque $p<5\%$, qu'il y a une différence statistiquement significative : on considère peu probable que la différence soit due au hasard.

Remarque importante : une valeur de p très petite ne signifie pas forcément qu'il existe une grande différence entre les groupes de comparaison. Cela signifie simplement que l'hypothèse nulle (pas de différence) est très peu vraisemblable. Toutes choses égales par ailleurs, la valeur de p sera d'autant plus petite que l'effectif de sujets est grand.

En conclusion :

- **p<5% = test significati**f, la différence observée n'est probablement pas due au hasard,

- **p>5% = test non significatif**, la différence observée pourrait être due au hasard :.

- **p compris entre 5% (risque α) et 10% : une tendance statistique** est mise en évidence**.** Pour infirmer ou confirmer une tendance statistique, il aurait fallu recruter plus de sujets et/ou allonger la durée de suivi de l'étude pour avoir plus d'événements.

Mais un résultat statistiquement significatif peut ne pas être cliniquement significatif.

Nota bene : ne pas confondre tendance statistique et test de tendance (trend test).

Pour savoir s'il y a une relation monotone entre un facteur d'exposition analysé en catégories et le risque d'un phénomène (maladie), c'est à dire si la relation se renforce (ou au contraire diminue) avec des catégories du facteur d'exposition plus élevées, on fait un test de tendance.

Exemple : on étudie la relation entre activité physique et incidence de diabète. L'activité physique est classée en 5 classes selon la fréquence. Les rapports de risque instantanés de survenue de diabète (hazard ratio=HR) sont reportés dans la table ci-dessous (The American journal of medicine, 122: 1115-21).

	Rarely/Never	1-3/mo	Once/wk	2-4×/wk	≥5 times/wk	*P*, Trend
Age-adjusted	1.00	0.79 (0.68-0.93)	0.74 (0.64-0.86)	0.57 (0.50-0.65)	0.43 (0.36-0.52)	<.001

On observe que HR diminue avec les catégories de fréquence d'activité physique plus élevées. La colonne "p, trend" donne le résultat du test de tendance. Il est significatif (p<0.05). On met en évidence une relation inverse entre fréquence de l'activité physique et risque de diabète. Plus l'activité physique est fréquente, plus le risque de diabète diminue.

OBJECTIF 17° DISCUTER LA PERTINENCE CLINIQUE DES RESULTATS

1. LES QUESTIONS A SE POSER

a. La signification clinique est-elle analysée après avoir évalué les biais ?
b. La signification clinique est-elle analysée en référence à la « taille d'effet » et non à la signification statistique ?
c. La variable choisie pour cette analyse est-elle pertinente : s'agit-il bien de l'une des variables traduisant les objectifs de l'étude (critères de jugement définis) ?
d. Le jugement sur la taille de l'effet utilise-t-il l'intervalle de confiance ?
e. Le jugement sur la taille de l'effet fait-il référence à des données externes à l'étude (si oui, sont-elles référencées ?)
f. Dans les enquêtes observationnelles, peut-on conclure à la causalité ?

***Remarque**: privilégier le point de vue clinique.*

2. OU CHERCHER LES INFORMATIONS

- dans les résultats pour juger de la « taille » de l'effet.
- dans l'introduction, les méthodes (calcul du nombre de sujets nécessaire) et la discussion pour trouver les éléments permettant de juger de l'intérêt de la taille de l'effet.

3. RAPPELS THEORIQUES

A. Signification clinique

La signification clinique exprime un jugement porté sur l'importance au sens de « taille » du résultat obtenu dans l'étude. Elle justifie la décision clinique, diagnostique, pronostique ou thérapeutique. La taille de l'effet fait référence à l'une des quantités résumant l'étude qui quantifie un écart d'effet entre les groupes comparés.

La taille de l'effet est exprimée en bénéfice relatif et/ou bénéfice absolu.

Le bénéfice relatif caractérise le facteur étudié. Pour des résultats binaires (ex : décès oui/non), il est exprimé par le risque relatif, ou le rapport des cotes (odds ratio, OR) selon l'étude. Pour des données censurées (ou de survie), il est exprimé par le rapport de risques instantanés (hazard ratio, HR).

Le bénéfice absolu caractérise un facteur étudié dans un contexte. Pour des résultats binaires ou les données censurées (ex : décès oui/non), il est exprimé par la réduction absolue de risque ($RAR=R_E-R_C$) et/ou par le nombre de sujets à traiter (NNT) pour éviter un accident (1/RAR). Pour des variables quantitatives (taux de cholestérol), il est exprimé par la différence des moyennes [M_E-M_C] et/ou par le d de Cohen ou le g de Hodge ([M_E-M_C]/écart-type combiné). Le d de Cohen ou le g de Hodge sont les équivalents du nombre de sujets à traiter pour les variables qualitatives ou censurées.

L'effet du traitement est d'autant plus grand que la RAR est élevée ou que le NNT est faible.

Exemple : un traitement X réduit le risque de décès de 50% par rapport au placebo (HR=0,5). C'est le bénéfice relatif.

Si on applique ce traitement X à une population à bas risque (risque de décès 4% dans le groupe placebo) on s'attend à avoir en moyenne un risque de décès de 2%. Le bénéfice absolu est de 2% (4%-2%). Ceci signifie que si on traite 100 patients à bas risque, on évite en moyenne 2 décès par rapport à un traitement par placebo. Autrement dit, il faut traiter en moyenne 50 patients à bas risque pour éviter un décès.

Si on applique ce traitement X à une population à haut risque (risque de décès 40% dans le groupe placebo) on s'attend à avoir en moyenne un risque de décès de 20%. Le bénéfice absolu est de 20% (40%-20%). Ceci signifie que si on traite 100 patients à haut risque, on évite en moyenne 20 décès par rapport à un traitement par placebo. Autrement dit, il faut traiter en moyenne 5 patients à haut risque pour éviter un décès.

Pour chaque critère de jugement, les résultats de l'étude doivent au préalable être rapportés sous la forme de la mesure appropriée dans chaque groupe (par exemple, la proportion de participants avec ou sans l'événement, ou la moyenne et l'écart type des mesures).

B. Taille de l'effet

L'estimation de la taille d'effet doit être suffisamment précise (importance de donner son intervalle de confiance) pour pouvoir éliminer le fait que l'effet puisse être petit (borne inférieure de l'intervalle) donc sans intérêt en pratique. La définition du « plus petit effet intéressant » en pratique est arbitraire et difficile. Elle doit tenir compte de plusieurs paramètres qui sont extérieurs à l'étude et qui doivent être référencés dans la discussion:

- La gravité de la pathologie
  - ✓ Traiter 100 malades pendant 2 ans pour éviter une hospitalisation n'a pas la même « importance » que traiter 100 malades pour éviter 1 décès.
- De la fréquence de la maladie : plus une maladie est fréquente plus un petit bénéfice peut correspondre à un bénéfice substantiel en terme de nombre de sujets concernés par le bénéfice dans la population.
  - ✓ Ainsi, un risque relatif de 1.3 peut être vu comme négligeable au niveau de l'individu, mais important au niveau collectif si ce risque concerne un problème fréquent dans la population.
  - ✓ Une augmentation relative de risque de 30% peut paraître importante. Si elle s'applique à une maladie dont la fréquence est de $1/10^6$, le risque de maladie passe de 1 à $1{,}3x10^{-6}$, ce qui parait négligeable…

Ce problème ne concerne pas seulement l'essai contrôlé ou les études de cohorte. On doit aussi se poser cette question pour d'autres études par exemple les études diagnostiques (un test ayant une sensibilité de 90% est-il « un bon test » ?) ou les études analysant la relation entre deux variables (un coefficient de corrélation de 0.90 traduit-il une « forte relation » entre ces 2 variables ?). En résumé, l'adjectif « bon », « mauvais », « fort » ou « faible » qualifiant un résultat doit être justifié.

C. Peut-on conclure à la causalité ?

La causalité ne s'affirme pas du seul fait d'un test « statistiquement » significatif, la causalité peut s'affirmer si, et seulement si :

- il s'agit d'un essai randomisé,
- il n'y a pas de biais dans la randomisation,
- la randomisation n'est pas détruite par un trop grand nombre de perdus de vue, de données manquantes ou de modifications du protocole,
- l'analyse est réalisée en intention de traiter.
- Dans une étude observationnelle, un faisceau d'arguments doit être réuni point D objectif 17.

D. Critères permettant de conclure à la causalité

La causalité ne s'affirme pas du seul fait d'un test « statistiquement » significatif, la causalité peut s'affirmer si, et seulement si :

- ✓ il s'agit d'un essai randomisé,
- ✓ il n'y a pas de biais dans la randomisation ou les groupes sont comparables au départ pour les facteurs pronostiques,
- ✓ la randomisation n'est pas détruite par un trop grand nombre de perdus de vue, de données manquantes ou de modifications du protocole (traitement efficace concomitant),
- ✓ l'analyse est réalisée en intention de traiter.

Dans une enquête observationnelle, la causalité entre le facteur étudié et la maladie est suggérée par un faisceau d'arguments.

Critères de Hill

Critères internes à l'étude

- **Force de l'association :** quand le degré d'association (RR ou OR) est très grand il faudrait un fort effet de confusion pour faire disparaitre l'association
- **Cohérence chronologique**: Exposition au facteur doit précéder l'apparition de la maladie : permet d'exclure la relation inverse
- **Relation dose-effet** : si l'exposition augmente, le risque augmente : quand ce critère est présent, il est très spécifique mais il n'est pas vrai dans tous les cas

Spécificité de l'association : si un facteur de risque est constamment relié uniquement à la maladie étudiée, il apparait vraisemblable qu'une relation causale existe : ce critère découle d'une analogie avec les maladies infectieuses mais son absence ne remet pas en question une relation causale

•Critères externes à l'étude

– **Reproductibilité** : Constance de l'association et reproductibilité des résultats de l'étude avec d'autres équipes, en différents lieux, circonstances et temps : c'est un bon critère.
– **Cohérence des variations du facteur et de la maladie dans le temps et dans l'espace :** Étudiée à partir des corrélations « écologiques »
– **Plausibilité** ou **Cohérence avec les connaissances actuelles** (sur l'histoire naturelle de la maladie, la physiopathologie, la biologie
– **Analogie avec d'autres facteurs de risque démontrés** : une hypothèse causale est renforcée par le fait qu'une relation causale analogue est connue, rendant l'hypothèse plus crédible ; *ex: l'hypothèse que l'inhalation d'amiante cause le cancer du poumon est rendue plus plausible par la connaissance qu'on a du rôle causal du tabac dans le cancer du poumon.*

•Conclusion : une seule étude épidémiologique ne permet pas d'affirmer un lien de cause à effet.

OBJECTIF 18°: VERIFIER QUE LES RESULTATS OFFRENT UNE REPONSE A LA QUESTION ENONCEE ET OBJECTIF 19°: VERIFIER QUE LES CONCLUSIONS SONT JUSTIFIEES PAR LES RESULTATS.

Ces deux objectifs seront traités ensemble car ils sont très liés : en effet pour vérifier que les résultats offrent une réponse à la question posée, il ne suffit pas de vérifier l'adéquation entre l'objectif principal de l'étude et le résultat principal de l'étude mais également que la méthodologie employée permet bien de répondre à la question posée ; ceci revient à évaluer la validité interne de l'étude. Cette évaluation sert aussi l'objectif 19 car les auteurs peuvent avoir tendance à sous évaluer les biais.

1. LES QUESTIONS A SE POSER

a. Les auteurs ont ils répondu à la question posée (objectif) dans l'introduction ?
b. Ont-ils donné une explication rationnelle des résultats observés et de leur interprétation ?
c. Les auteurs ont-ils évalué la validité interne de leur travail, c'est-à-dire la fiabilité de leurs résultats? En d'autres termes leurs résultats sont-ils non biaisés et suffisamment puissants?
d. Les auteurs ont-ils discuté le caractère généralisable (validité externe) de leurs résultats ?
e. Dans un essai clinique, si la conclusion est basée sur les résultats d'une analyse intermédiaire,
    - La ou les analyse(s) intermédiaire(s) ont-elles été prévues dans le protocole ?
    - A t-on tenu compte de ces analyses intermédiaires dans le calcul du NSN ?
    - A t-on précisé un seuil de signification pour chacune d'elles pour fonder les conclusions statistiques ? (**cf objectif 9)**
f. Si des analyses en sous-groupes ont été effectuées, étaient-elles prévues dans le protocole, argumentées, avec un but précisé (exploratoire, OU confirmatoire)? La méthode utilisée permet-elle de fournir une conclusion valide pour ces analyses ? (La randomisation a t-elle été stratifiée ? L'interaction a t-elle été testée ? A t-on corrigé le risque alpha ? L'interprétation repose t-elle sur l'interaction ?) (**cf objectif 9)**

2. OU CHERCHER LES INFORMATIONS

Dans toutes les sections : introduction pour vérifier l'objectif, méthodes & résultats pour évaluer la validité interne et la représentativité, discussion pour évaluer l'interprétation que les auteurs ont faite de leur travail.

3. RAPPELS THEORIQUES

1. VALIDITE INTERNE : elle permet de s'assurer que le résultat obtenu n'est pas dû à un biais, au hasard ou à un manque de puissance donc que la méthode utilisée pour l'obtenir est adéquate pour répondre à la question.

- Absence de biais : cf objectif 14.
- Réalité statistique du résultat : cf objectif 16.
- Calcul du nombre de sujets réalisé pour garantir une puissance suffisante

2. VALIDITE EXTERNE : elle regroupe les notions de cohérence externe et de représentativité.

- **Cohérence externe** : permet de s'assurer que le résultat obtenu n'est pas isolé mais s'intègre dans un contexte logique (cohérence avec les connaissances expérimentales, épidémiologiques, les études de même nature). Pour les enquêtes épidémiologiques

observationnelles, il s'agit de s'assurer que les arguments en faveur d'une relation de cause à effet entre l'exposition et la maladie sont rassemblés.

- **Représentativité :** les résultats sont-ils extrapolables aux patients rencontrés couramment dans la pratique médicale ? En d'autres termes, les patients de l'étude sont ils représentatifs des patients vus en pratique médicale courante (même définition de la maladie, pas de sélection excessive (attention aux critères d'inclusion trop « durs ») ? L'objectif 21 détaille ce point.

OBJECTIF 20 : INDIQUER LE NIVEAU DE PREUVE DE L'ETUDE (GRILLE DE L'HAS / ANAES)

1. LES QUESTIONS A SE POSER :

a. Quel est l'objectif de l'étude?
b. Quel est le type d'étude? Est-il le plus approprié pour répondre à l'objectif?
c. Les biais sont-ils importants dans cette étude ?
d. La puissance de l'étude est-elle satisfaisante ? L'effectif repose t-il sur un calcul du nombre de sujets nécessaire ? Quel est le taux de sorties d'étude ou de données manquantes pour l'analyse principale ?

2. OU CHERCHER LES INFORMATIONS ?

Dans toutes les sections : introduction (objectif); méthodes; résultats; discussion (limites de l'étude)

3. RAPPELS THEORIQUES

DEFINITION DU NIVEAU DE PREUVE (ANAES/HAS)(4)

Le niveau de preuve d'une étude caractérise la capacité de l'étude à répondre à la question posée.

Cette capacité se juge, d'une part, par la correspondance de l'étude au cadre du travail (sujet, population, paramètres de jugement pris en compte), et d'autre part par les caractéristiques suivantes :

- l'adéquation du protocole d'étude à la question posée,
- l'existence ou non de biais importants dans la réalisation,
- l'adaptation de l'analyse statistique aux objectifs de l'étude,
- la puissance de l'étude et en particulier la taille de l'échantillon.

Une classification générale du niveau de preuve d'une étude peut être proposée à partir des classifications de la littérature et des composantes vues ci-dessus :

- **un fort niveau de preuve** correspond à une étude dont :
  - ✓ le protocole est adapté pour répondre au mieux à la question posée,
  - ✓ la réalisation est effectuée sans biais majeur,
  - ✓ l'analyse statistique est adaptée aux objectifs,
  - ✓ la puissance est suffisante ;
- **un niveau intermédiaire** est donné à une étude de protocole similaire, mais présentant une puissance nettement insuffisante (effectif insuffisant ou puissance a posteriori insuffisante) et/ou des anomalies mineures ;
- **un faible niveau de preuv**e peut être attribué aux autres études.

Des distinctions plus fines ont été proposées par certains auteurs. Elles ne concernent que les études thérapeutiques et ne sont pas utilisables pour d'autres types d'études (diagnostic, causalité, cohorte).

GRILLE DE L'HAS/ANAES

GRILLE A INTERPRETER DANS LE CAS OU ON RECHERCHE UN LIEN DE CAUSALITE ENTRE UNE ACTION ET UN EFFET

Niveau 1	Essais comparatifs randomisés de forte puissance	*Recommandation* *Preuve scientifique établie*
	Méta-analyse d'essais comparatifs de forte puissance	
	Analyse de décision basée sur des études bien menées	
Niveau 2	Essais comparatifs randomisés de faible puissance	*Recommandation* *Présomption scientifique*
	Etudes comparatives non randomisées bien menées	
	Etudes de cohorte	
Niveau 3	Etudes cas-témoins	*Recommandation* *Faible niveau de preuve scientifique*
Niveau 4	Etudes comparatives comportant des biais importants	
	Etudes rétrospectives	
	Séries de cas	
	Etudes épidémiologiques descriptives (transversale, longitudinale)	

OBJECTIF 21°: DISCUTER LA OU LES APPLICATIONS POTENTIELLES PROPOSEES PAR L'ETUDE

1. LES QUESTIONS GENERALES A SE POSER :

a. Quel est le niveau de preuve ?
b. A quels patients les résultats sont-ils réellement applicables ? (cf objectif 3)
c. Tous les critères cliniquement importants ont-ils été pris en compte dans la conclusion ?
d. La balance bénéfice-risque est-elle favorable ?
e. Les applications proposées sont-elles une conséquence directe des objectifs de l'étude?
f. L'étude permet-elle de faire cette recommandation?
g.

2 . LES QUESTIONS A SE POSER

POUR UN ESSAI CLINIQUE

- La population de l'étude correspond-elle à la population habituellement traitée ?
  Si non : existe t-il une cause de force majeure pour ne pas appliquer les résultats à mon patient ?

- Toutes les variables cliniquement pertinentes ont-elles été évaluées ? Le critère de jugement est-il clinique ? *Par exemple : A t-on dans un traitement pour le cancer pris en compte en plus de la mortalité, la qualité de vie ?*
- Les bénéfices dus au traitement dépassent-ils les risques et le coût encourus ?
- Si analyse en sous-groupe, **(cf objectif 9)**
  - ✓ cette analyse a t-elle été prévue (visée confirmatoire, justification, la randomisation a-t-elle été stratifiée sur la variable définissant les sous-groupes *(exemple : sexe pour analyse chez les hommes, et les femmes)* et le risque alpha a-t-il été corrigé ?
  - ✓ l'effet est-il significatif et important ? (test de l'interaction)
  - ✓ l'effet est-il retrouvé dans d'autres études ?
- Si arrêt prématuré de l'essai, des règles d'arrêt avaient-elles été définies à l'avance dans le protocole ? Ces règles ont-elles été respectées ? **(cf objectif 9, analyse intermédiaire)**
- **L'essai est-il pragmatique ou explicatif? (cf addendum 2 : types d'études)**

POUR UNE EVALUATION DE PROCEDURE DIAGNOSTIQUE

- Dans le cadre qui nous intéresse, les résultats du test diagnostique sont-ils reproductibles et leur interprétation satisfaisante ?
  - ✓ A t-on des informations sur la reproductibilité du test diagnostique ? Qui étaient les évaluateurs ?
  - ✓ Les réponses « douteuses » lorsqu'elles existent ont-elles été prises en compte dans l'analyse ? classées dans quel groupe (positif ou négatif) ?
  - ✓ Un calcul NSN a–t-il été effectué pour garantir une précision correcte dans l'estimation des paramètres (cf les intervalles de confiance des Se, Sp, VPP, VPN.)
  - ✓ Le test diagnostique est-il utile compte tenu de sa valeur informative et de sa reproductibilité dans le cadre qui nous intéresse ?
- Les résultats du test sont-ils applicables à la population à laquelle on veut appliquer le test diagnostique ?
  - ✓ Les conditions d'application du test diagnostique sont-elles les mêmes que dans l'étude ?
  - ✓ La population cible vérifie t-elle les critères d'inclusion et ne viole t-elle pas les critères de non inclusion ?

- Les résultats du test changent-ils la prise en charge ? Le test est-il informatif : permet-il d'éliminer le diagnostic ou au contraire d'affirmer le diagnostic au vu du résultat et compte tenu de la probabilité pré-test ? quel est le « gain » diagnostique d'un test positif (probabilité post-test) ou quel est « la perte » diagnostique d'un test négatif (probabilité post-test) ?
- Les patients auxquels on veut appliquer le test seront-ils mieux après la conclusion du test ?
  Les bénéfices dus au résultat du test dépassent-ils les risques et le coût encourus par le test diagnostique ?

POUR UNE ENQUETE EPIDEMIOLOGIQUE ANALYTIQUE OU A VISEE ETIOLOGIQUE

- La population à laquelle on veut extrapoler les résultats est-elle comparable aux sujets recrutés dans l'étude, pour l'exposition, la maladie, l'âge, l'ethnie, et les autres caractéristiques importantes ?
- Quel est le niveau de preuve de l'étude ?
- Le faisceau d'arguments en faveur d'une relation de cause à effet entre l'exposition et la maladie est-il rassemblé ?
- Le risque est-il élevé si l'exposition est poursuivie ?
- Quelles sont les conséquences si l'exposition est supprimée ou réduite ?

POUR UNE ETUDE PRONOSTIQUE

- Les patients de l'étude sont-ils similaires aux patients auxquels on veut extrapoler les résultats de l'étude ?
- la population de l'étude est-elle décrite avec précision ?
- Les résultats de l'étude impliquent-ils une prise en charge spécifique : un arrêt de traitement ou une mise sous traitement ?
- Les résultats de l'étude permettent-ils de rassurer ou de conseiller mes patients?
  - ✓ Le pronostic est-il valide (sans biais) ?
  - ✓ Le pronostic est-il précis ?
  - ✓ Le pronostic est-il généralisable ?

3. OU CHERCHER LES INFORMATIONS ?

Dans toutes les sections : introduction (objectif); méthodes; résultats; discussion (limites de l'étude)

4. NOTIONS THEORIQUES

4.1. ETAPES DU GUIDE DE LECTURE CRITIQUE D'UN ARTICLE

Lire de façon critique un article comporte 3 étapes

1. Les résultats de l'étude sont-ils valides (non biaisés) ?
2. Quels sont les résultats de l'étude ? Autrement dit, la taille de l'effet est-elle importante ?
3. Quelles sont les implications des résultats pour ma pratique ?

La question des applications potentielles des résultats d'une étude est la troisième étape qui touche la validité externe de l'étude.

Elle concerne l'application au niveau individuel, et au niveau de la population.

4.2. LES GRILLES DE LECTURE POUR CHAQUE TYPE D'ETUDE

Le groupe de l'Evidence Based Medecine (EBM) (5-11) a publié des grilles de lecture qui regroupent un ensemble de points à passer en revue afin d'analyser dans quelle mesure les résultats sont applicables à un type de patient donné.

On peut transposer ces grilles au niveau des populations.

Ces grilles sont détaillées dans les paragraphes 2, 3,4 et 5 de cet objectif 21.

OBJECTIF 22 : IDENTIFIER LA STRUCTURE IMRAD (INTRODUCTION, MATERIEL ET METHODE, RESULTATS, DISCUSSION) ET S'ASSURER QUE LES DIVERS CHAPITRES REPONDENT A LEURS OBJECTIFS RESPECTIFS.

1. LES QUESTIONS A SE POSER :

- L'article contient-il quatre chapitres (Introduction, Matériel et méthode, Résultats, Discussion)?
- Chaque chapitre répond-il à ses objectifs ? Quels éléments ont été omis ? Quels éléments ne devraient pas y figurer ?

2. OU CHERCHER LES INFORMATIONS ?

DANS TOUT L'ARTICLE, EN SUIVANT L'ORDRE DE L'ARTICLE.

3. RAPPELS THEORIQUES

A. STRUCTURE IMRAD

Section	*Rôles et buts*	*Plan*
Introduction	Raisons de l'étude	Santé publique Clinique/physiopathologie Lacunes dans les connaissances
	Enoncer la question étudiée	Population Intervention/facteur étudié Comparateur Observé (outcome)
Méthodes	Décrire ce qui a été fait Montrer que les résultats serontvalides Permettre la réplication de l'étude	**S**chéma d'étude **P**opulation **L**ieu(x) **I**ntervention(s)/facteur étudié **C**ritère de jugement **A**nalyse
Résultats	Donner les éléments de réponse à la question	Description de l'échantillon étudié et date. Résultats sur CJP Résultats sur CJS
Discussion	Interpréter et critiquer les résultats Proposer une implication clinique	Résumé des principaux résultats Points forts et limites de l'étude Discussion de leur validité Généralisation et confrontation à la littérature Conclusion sur la question Implications de la conclusion

B. La partie Méthodes

La partie Méthodes est la partie de l'article détaillant quels "matériels & méthodes" ont été à la base des résultats de la publication et des conclusions.

Cette partie Méthodes suit, normalement, le plan suivant (SPLICA)

- **S**chéma d'étude : ses paramètres (variables selon le type d'étude).
- **P**opulation étudiée : critères d'inclusion, de non inclusion, nombre d'échantillons, mode de randomisation
- **L**ieu de l'étude : structure hospitalière ou structure ambulatoire, nombre de structures (étude multi-centrique)
- **I**ntervention(s) Facteur étudié: caractéristiques des interventions (chirurgicales, médicamenteuses …) et leur durée, intervalle éventuel entre elles si plusieurs
- **C**ritère(s) d'évaluation : bien noter le critère d'évaluation principal (qui doit avoir été la base du calcul du nombre de sujets nécessaires et dont les résultats sont la base de la conclusion). Mode de recueil et d'évaluation (Notion d'aveugle) …Noter si ce critère a été choisi après collecte des données de l'étude et pourquoi.
- **A**nalyse: calcul du nombre de sujets nécessaire (vérifier si les paramètres de ce calcul sont précisés), plan d'analyse (par exemple : analyse en intention de traiter pour un essai clinique, analyses principales, secondaires, analyses intermédiaires, analyses en sous-groupes, principe de construction des modèles multivariés), outils statistiques utilisés, valeur du risque α choisi.

- A ces points s'ajoutent les considérations réglementaires, éthiques.

Remarque : Cette question concerne la forme (structure) de l'article. => **Ne pas discuter** le contenu de l'article, la pertinence des méthodes, leur validité ….

OBJECTIF 23 : FAIRE UNE ANALYSE CRITIQUE DE LA PRESENTATION DES REFERENCES

1. LES QUESTIONS A SE POSER :

- Quel type de classement a été utilisé ?
- Les références sont-elles présentées de façon complète ?

2. OU CHERCHER LES INFORMATIONS ?

A LA FIN DE L'ARTICLE ET DANS L'ARTICLE POUR LES APPELS DANS LE TEXTE.

3. RAPPELS THEORIQUES

A. FORME D'UNE REFERENCE

- Auteurs : nom, initiales du prénom. Si plus de 6 auteurs: mettre « et al. », séparés par des virgules. (Remarque : parfois « et al. » est mis après les 3 premiers auteurs)
- Titre de la publication
- Support :
    - Revue : titre abrégé international (Index Medicus) ex: N Engl J Med, JAMA
    - Livre : nom de l'éditeur et lieu de publication
- Année de publication
- Volume et numéro de publication
- Première et dernière pages

Remarque: Si l'article est dans un ouvrage collectif : références de l'article et références de l'ouvrage (auteurs et titre de l'ouvrage en plus du nom de l'éditeur et du lieu de publication)

EXEMPLES DE REFERENCES

- ***Article de revue :*** *Santen SA, Holt DB, Kemp JD, Hemphill RR. Burnout in medical students: examining the prevalence and associated factors. South Med J 2010; 103(8): 758-63.*
- ***Ouvrage :*** *Caplan LR. Posterior circulation disease: clinical findings, diagnosis and management. Cambridge, Mass: Blackwell Science, 1996: 555-556*
- ***Article dans un ouvrage collectif :*** *Goadsby PJ, Silberstein, SD, eds. Pathophysiology of migraine: a disease of the brain. In: Goadsby PJ, et al. Headache. Boston, Mass: Butterworth-Heinemann; 1997: 5-25.*
- ***Rapport :*** *Stoebner A, Lehmann M, Sancho-Garnier H. Evaluation d'une action de prévention du tabagisme en entreprise. Rapport février 1996*

B. CLASSEMENT DES REFERENCES :

- Auteur-année
    - Dans le texte : [nom auteur, année]
    - Liste à la fin de l'article: par ordre alphabétique selon le nom de famille du premier auteur
- Alphabétique-numérique
    - Dans le texte : [numéro]

    - Liste à la fin de l'article: numéro par ordre alphabétique selon le nom de famille du premier auteur
- Numérique-séquentiel (Vancouver)
    - Dans le texte : [numéro]
    - Liste à la fin de l'article: numéro par ordre d'apparition dans le texte

OBJECTIF 24 : FAIRE UNE ANALYSE CRITIQUE DU TITRE

1. LES QUESTIONS A SE POSER :

- Le titre est-il informatif ou indicatif ?
- Le titre est-il attractif ?

2. OU CHERCHER LES INFORMATIONS ?

- Dans le titre, l'introduction et les méthodes

3. RAPPELS THEORIQUES

A BUT DU TITRE : ANNONCER LE CONTENU SIGNIFICATIF DE L'ARTICLE

B.PRINCIPES :

- Concision
- Précision
- Mots les plus informatifs au début (voire à la fin)
- Éviter jargon et abréviations

C. Contenu : PICOS

- Population
- Intervention (Facteur étudié)
- Comparaison (le cas échéant)
- Outcome (ce qui est observé, résultat)
- Schéma d'étude

D. Types :

- • Indicatif : orientation générale du contenu de l'article
- • Informatif : fournit des éléments précis sur le contenu significatif de l'article

E. Forme :

- Syntaxe souple
- Attractif

F. Exemples

Facteurs de risque environnementaux (Facteurs étudiés) des Lymphomes Malins Non Hodgkiniens (Outcome): une étude cas-témoins (Schéma d'étude) de population en Languedoc-Roussillon, France (Population/Lieu) =>titre informatif

• Les troubles (Outcome) liés à l'usage de drogues illicites (Facteurs étudiés) chez les conducteurs (Population) par comparaison avec les étudiants (Comparaison) = >titre indicatif car aucune des composantes n'est précise.

REFERENCES BIBLIOGRAPHIQUES

1. Mitchell LB, Exner DV, Wyse DG, Connolly CJ, Prystai GD, Bayes AJ, et al. Prophylactic Oral Amiodarone for the Prevention of Arrhythmias that Begin Early After Revascularization, Valve Replacement, or Repair: PAPABEAR: a randomized controlled trial. Jama; 2005. p. 3093-100.

2. Debrock C, Menetrey C, Bonavent M, Antonini MT, Preux PM, Bonnaud F, et al. [Prevalence of exercise-induced asthma in school children]. Rev Epidemiol Sante Publique; 2002. p. 519-29.

3. Hoffmann MH, Shi H, Schmitz BL, Schmid FT, Lieberknecht M, Schulze R, et al. Noninvasive coronary angiography with multislice computed tomography. Jama. [DIAGNOSTIC]. 2005 May 25;293(20):2471-8.

4. ANAES. Guide d'analyse de la litterature et gradation des recommandations. In: Santé ANdAedEeSHAe, editor.; 2000.

5. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993 Dec 1;270(21):2598-601.

6. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA. 1994 Jan 5;271(1):59-63.

7. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994 Feb 2;271(5):389-91.

8. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994 Mar 2;271(9):703-7.

9. Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. JAMA. 1994 Jul 20;272(3):234-7.

10. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. Evidence-Based Medicine Working Group. JAMA. 1994 May 25;271(20):1615-9.

11. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA. 1993 Nov 3;270(17):2093-5.

12. Beuscart R., Bénichou J., Roy P., Quantin C. : Biostatistique: Montreuil, Omniscience, 2009.