



Université Claude Bernard



Lyon 1

LECTURE CRITIQUE D'ARTICLES

ICCA

La mesure de critère de jugement repose sur des instruments de mesure plus ou moins complexes. On doit connaître la validité de ces instruments ainsi que leur reproductibilité. La validité d'un instrument (ou son exactitude) est la capacité à bien mesurer ce qu'il est censé mesurer. La reproductibilité d'un instrument représente sa faculté à donner un même résultat lorsqu'il mesure le même phénomène de façon répétée. Par exemple, pour mesurer la reproductibilité du diagnostic de thrombose veineuse profonde par une phlébographie, on mesure la reproductibilité inter-observateurs en faisant lire les mêmes radios par plusieurs radiologues et en analysant la concordance de leurs résultats, ou la reproductibilité intra-observateur en faisant lire les mêmes radios plusieurs fois par le même radiologue et en analysant la concordance de ses résultats. Bien entendu, dans les deux cas ces lectures doivent être faites indépendamment des lectures précédentes.

LECTURE CRITIQUE D'ARTICLES

Ce livret a été réalisé par les enseignants de LCA de la faculté Lyon-Est
sous la direction du **Pr. Anne-Marie Schott**

AUTEURS

Anne-Marie Schott : PU-PH en Santé publique

Pauline Occelli : AHU en Santé Publique

Pascal Caillet : AHU en Santé Publique

Julie Haesebaert : AHU en Santé Publique

Muriel Rabilloud : MCU-PH en Biostatistiques

Jean-Pierre Fauvel : PU-PH en Néphrologie

Marie Viprey : AHU en Santé Publique

Zoé Boulot : étudiante en DFASM2

Nicolas Guibert : AHU en Médecine du travail

Barbara Charbotel : PU-PH en Médecine du Travail

CORRESPONDANCE

anne-marie.schott-pethelaz@chu-lyon.fr

brigitte.sebbane@chu-lyon.fr

marie.viprey@chu-lyon.fr

SOMMAIRE

1 / INTRODUCTION	1
1.1 À quoi vous sert la LCA ?	1
1.2 Quel format à l'ECNi ?	1
1.3 Comment débiter ?	1
1.4 Quel est l'objectif de la publication de ces études ?	2
1.5 Pour chaque article, comment savoir si vous devez prendre en compte ses résultats dans votre propre pratique ?	2
2 / LES DIFFÉRENTS TYPES D'ARTICLES	4
3 / STRUCTURE D'UN ARTICLE ORIGINAL : LA FORME	4
3.1 Le titre	4
3.2 L'introduction (appelée aussi justification, contexte...)	5
3.3 Matériel et Méthodes	5
3.4 Résultats	6
3.5 Discussion	8
3.6 Les références bibliographiques	8
3.7 Synthèse	9
4 / LES GRILLES DE LCA OU « CHECK-LIST »	11
4.1 La question de recherche (Le « PECO »)	12
4.2 Le type d'étude (synonymes : le design, le schéma, l'architecture, le dessin...)	21
4.3 La population/l'échantillon étudié(e) la population source (généralités)	24
4.4 Le critère de jugement (généralités)	27
4.5 L'élément évalué : l'intervention ou le facteur de risque/pronostique (généralités)	30
4.6 L'analyse statistique	30
4.7 Les résultats (généralités)	36
4.8 Les facteurs de confusion, les biais et les interactions (généralités)	37
4.9 La discussion - La conclusion	41

5 / LCA DES ESSAIS CLINIQUES, ÉVALUATION D'UNE INTERVENTION THÉRAPEUTIQUE	43
5.1 La question	43
5.2 Le type d'étude	44
5.3 La population / l'échantillon étudié	58
5.4 Le critère de jugement	63
5.5 L'intervention	67
5.6 Les biais et facteurs de confusion	68
5.7 Les analyses statistiques	70
5.8 Les résultats (ampleur de l'effet et signification statistique)	75
5.9 La conclusion	79
5.10 Les essais d'équivalence ou de non-infériorité	81
5.11 Évaluation d'une intervention de dépistage	84
6 / LCA DES ÉTUDES DE COHORTES	90
6.1 LCA des études de cohortes visant à mesurer un lien entre un facteur de risque supposé et la survenue d'une maladie	90
6.2 LCA des cohortes visant à analyser l'évolution d'une maladie et ses facteurs pronostiques	114
7 / LCA DES ÉTUDES CAS-TÉMOINS	134
7.1 La question	134
7.2 Le type d'étude	135
7.3 La population étudiée	136
7.4 Le critère de jugement : la maladie étudiée	137
7.5 Le facteur de risque	138
7.6 Analyses statistiques	140
7.7 Les résultats	140
7.8 Les biais et facteurs de confusion	142
7.9 La conclusion	145
7.10 Étude cas-témoins nichée dans une cohorte (« nested case-control study »)	146
7.11 Comparaison Étude cas-témoins et étude de cohorte	147

8 / LCA DES ÉTUDES DE TESTS DIAGNOSTIQUES	150
8.1 La question de recherche	152
8.2 Le type d'étude	153
8.3 La population étudiée	155
8.4 Le Test évalué	156
8.5 Le Gold Standard (ou Test diagnostique de Référence)	158
8.6 Analyse statistique	161
8.7 Les résultats : performances du test étudié	171
8.8 Les biais	173
8.9 La conclusion	174
9 / PRINCIPES GÉNÉRAUX DE LA LECTURE CRITIQUE	
D'ARTICLES ORIGINAUX	175
9.1 LES QUATRES POINTS CARDINAUX DE LA LCA	175
9.2 Conclusion	177
10 / STATISTIQUES	181
10.1 Organisation générale de l'analyse statistique	181
10.2 Différentes catégories de variables	181
10.3 Les différentes analyses statistiques possibles	182

1

INTRODUCTION

1.1 À QUOI VOUS SERT LA LCA ?

- Le médecin aujourd'hui ne peut plus s'appuyer uniquement sur son expérience personnelle ou sur celle de ses collègues, mais doit se tenir informé des évolutions scientifiques nécessaires à sa pratique médicale.
- La quantité croissante d'articles publiés et leur qualité inégale impose au médecin d'avoir la capacité de savoir quels articles lire, comment les lire et comment prendre en compte les résultats pertinents dans sa pratique quotidienne. Il s'agit donc de ne plus s'informer de façon passive mais active.
- La moitié des connaissances acquises à la faculté est obsolète 7 ans plus tard...
- Vous serez bac + 10... est-ce pour vous laisser manipuler par des visiteurs médicaux ou d'autres sources d'informations dont vous ne saurez pas juger la crédibilité et la pertinence ?
- Vous devez être capable de différencier les opinions des faits, de valider les informations que vous recevez et d'identifier celles qui vous manquent.
- ...Et pour l'instant cela fait partie des épreuves de l'ECN !

1.2 QUEL FORMAT À L'ECNI ?

- Une épreuve de **3 h**.
- **Deux articles, en anglais**, en format papier.
- **15 QCM** progressifs (= pas de retour en arrière possible) sur chaque article.
- **10 %** de la note totale des ECNi.

Remarque : Ce format est assez long, est peu habituel. Il est donc important de s'entraîner régulièrement, et d'adopter des méthodes de lecture permettant de bien gérer le temps pendant l'épreuve.

1.3 COMMENT DÉBUTER ?

Pour démarrer la lecture critique il faut avoir un minimum de connaissances sur :

- les règles de la rédaction médicale,
- les principes de l'épidémiologie et de la recherche clinique (et un peu de statistiques)...

1.4 QUEL EST L'OBJECTIF DE LA PUBLICATION DE CES ÉTUDES ?

Ces publications sont censées apporter de nouvelles informations scientifiquement valides pour quantifier la **fréquence d'une maladie** dans une population, identifier ses **causes** (biologiques, médicales, environnementales, socio-économiques...), identifier les **facteurs qui prédisent son évolution**, déterminer la façon la plus exacte de le **diagnostiquer**, ou identifier le **meilleur traitement**.

1.5 POUR CHAQUE ARTICLE, COMMENT SAVOIR SI VOUS DEVEZ PRENDRE EN COMPTE SES RÉSULTATS DANS VOTRE PROPRE PRATIQUE ?

Pour cela vous devez répondre à 4 questions de base :

- Quel **degré de confiance** puis-je donner à ces résultats (le résultat observé est-il réel ? Ne résulte-t-il pas seulement d'un biais de l'étude) ? (**validité interne**).
- Ce résultat apporte-t-il une **avancée importante** pour améliorer la prise en charge des patients ? (**pertinence**).
- Ce résultat est-il **concordant avec les autres connaissances** sur le sujet ? (**cohérence**).
- Ce résultat est-il **extrapolable à d'autres contextes** ? D'autres populations ? (**validité externe**).

La validité interne peut être définie comme la recherche des défauts méthodologiques de l'étude et des biais potentiels. Une étude avec une méthodologie adéquate, permettant de limiter le risque des principaux biais évoqués, aura une bonne validité interne. Les éléments de la LCA qui composent ce polycopié sont essentiellement destinés à évaluer la validité interne.

La validité externe peut se définir comme la transposabilité (extrapolabilité, applicabilité, généralisabilité) des résultats à la pratique courante. Son évaluation repose sur les éléments suivants :

- **Caractéristiques des patients inclus dans l'étude** : la population de patients inclus dans l'étude correspond-elle à la population de patients à laquelle on appliquera les résultats ? Autrement dit, la population incluse dans l'étude est-elle représentative de la population cible (âge, sévérité...) ?

- **Caractéristiques des centres** : les résultats des études monocentriques sont moins généralisables que les résultats des études multicentriques. Le niveau d'expertise des centres est également à prendre en compte pour évaluer la transposabilité des résultats. Par exemple, si un essai est réalisé dans un centre de référence universitaire avec un volume d'activité important, les résultats ne seront pas forcément généralisables à des centres pratiquant moins d'interventions ou avec un plateau technique différent.
- **Caractéristiques des pays** : les résultats de certaines études réalisées en Chine sont-ils généralisables en France ?
- **Caractéristiques des interventions dans les essais** : est-il possible de reproduire les interventions évaluées dans l'essai dans notre contexte de soins ? Les interventions sont-elles suffisamment bien décrites pour être reproduites ?

Au final la LCA est un outil indispensable pour pratiquer la médecine factuelle (evidence based medicine ou EBM)

Médecine factuelle : chaque fois que possible, une décision clinique doit reposer sur des arguments objectifs et prouvés. Je prescris le traitement X chez ce patient parce qu'il y a des preuves (« evidence » en anglais) que ce traitement est l'option la plus efficace pour ce type de patient et dans cette situation. Cela nécessite trois étapes :

- Rechercher la littérature adéquate (recherche bibliographique),
- Évaluer quels résultats sont pertinents et crédibles (LCA ECN !),
- Appliquer ces résultats dans sa pratique (pour plus tard !).

La difficulté particulière de la lecture critique d'article réside dans la nécessité de développer une gymnastique d'esprit, afin d'apprendre à jongler avec les connaissances théoriques tout en restant pragmatique. Il est donc nécessaire d'assimiler ces connaissances petit à petit, donc de commencer tôt !

Ce polycopié a été fait avec l'aide des enseignants de LCA de la Faculté Lyon-Est et nous espérons qu'il vous sera utile.

Une grande partie des notions présentées dans ce polycopié représente le socle de base pour la LCA, qu'on peut appeler niveau 1, c'est la cible du DFGSM3. Le niveau 2 correspond à des notions plus poussées qu'il faudra aborder une fois le niveau 1 maîtrisé. Il correspond au niveau de l'ECN et aux examens de LCA à partir du DFASM1.

2

LES DIFFÉRENTS TYPES D'ARTICLES

Un « article original » est la publication d'un travail de recherche clinique ou épidémiologique.

Seuls les articles originaux peuvent faire l'objet de la LCA à l'ECN.

Les autres types d'articles sortent du champ de la LCA ECN : les avis d'auteurs, éditorial (l'avis d'un expert sur un sujet) et mise au point ; les revues de la littérature (revues de plusieurs articles et méta-analyses) ; la description clinique d'un ou de plusieurs cas (cas clinique – « case report » – ou série de cas), les articles didactiques à visée pédagogique (formation, enseignement), les évaluations économiques, les lettres de réponse à des articles originaux.

3

STRUCTURE D'UN ARTICLE ORIGINAL : LA FORME

Un article original est en général structuré de façon standardisée : la structure **IMRAD** (de l'anglais Introduction, Methods, Results And Discussion).

3.1 LE TITRE

Il doit être concis mais informatif, refléter le contenu scientifique de l'article et rappeler les éléments clefs de l'objectif principal de l'étude.

Le lecteur **devrait pouvoir comprendre** instantanément de quoi traite l'article en ne lisant que le titre.

Exemples :

Risque de fracture pelvienne après irradiation pelvienne chez la femme âgée.

Radiothérapie conventionnelle versus à dose élevée dans les adénocarcinomes localisés de la prostate : essai comparatif randomisé.

Vaccination contre l'hépatite B et premier épisode démyélinisant du système nerveux central : une étude cas-témoins.

3.2 L'INTRODUCTION (APPELÉE AUSSI JUSTIFICATION, CONTEXTE...)

Cette partie est destinée à replacer l'étude dans son contexte :

- Ce que l'on sait déjà ou « état de l'art » : **synthèse des connaissances** objectives sur le sujet, la nature et l'**importance du problème** avec des **données objectives chiffrées** et des références bibliographiques les plus récentes possibles.
- Les **lacunes et interrogations** justifiant l'étude proposée.
- **La question de recherche doit figurer en fin d'introduction** sous forme d'un **objectif principal** qui doit être le plus précis possible, plus accessoirement apparaissent les objectifs secondaires.

3.3 MATÉRIEL ET MÉTHODES

Cette partie doit contenir la **description de la méthode utilisée pour concevoir et conduire l'étude**. Elle comporte les éléments suivants :

- **Population étudiée** – Méthode de sélection :
 - Population source : critères d'inclusion et de non inclusion.
 - Lieu et période d'inclusion.
 - Modalités de sélection (ex. : patients consécutifs sur une période donnée, ou tirés au sort, ou appariement des témoins et des cas...).
- **Type d'étude** :
 - Essai clinique.
 - Étude de cohorte.
 - Étude cas-témoin.
 - Étude transversale.
- **Plan expérimental** (avec des précisions sur le schéma de l'étude).
- **Description de l'élément évalué** (terme générique qui désigne l'élément censé avoir un impact sur le critère de jugement), il s'agit soit d'un facteur de risque, soit d'une intervention en fonction du type d'étude :
 - **un facteur de risque**, s'il s'agit d'une étude étiologique (ex. : consommation de tabac en nombre d'années, ou en nombre de paquets-années, alcool, etc.).

S'il s'agit d'un **facteur de risque**, sa **méthode de mesure** et les **procédures de recueil doivent être décrits** :

 - > qui (patient lui-même par auto-questionnaire, enquêteurs...),
 - > où (hôpital, domicile...),
 - > quand (dates, durée...),
 - > comment (courrier, téléphone, face à face...),

- > outil de mesure (questionnaire, balance, examens biologiques...).
- **une intervention** s'il s'agit d'une étude d'intervention (ex. : si médicament : posologie, durée et voie d'administration..., si chirurgie : type de chirurgie, voie d'abord, etc.).
- **Description du critère de jugement principal** : c'est l'évènement ou l'état de santé étudié (ex. d'évènement : infarctus, AVC, diabète... ; ex. d'état : qualité de vie, douleur, anxiété...) :
 - **sa méthode de mesure** (ex. infarctus : transaminases X par 2 + douleur thoracique + sus-décalage de ST).
 - **les procédures de recueil** :
 - > **qui** (patient lui-même par auto-questionnaire, médecins, enquêteurs...),
 - > **où** (hôpital, domicile...),
 - > **quand** (dates, période de l'année, heures de la journée...),
 - > **comment** (courrier, téléphone, examen clinique...),
 - > **outil de mesure** (questionnaire, score, radiographies, examens biologiques...).
- **Éléments mis en place pour limiter les biais** (ex. : double insu, randomisation...).
- Description des données recueillies et de la méthode de recueil, au moment de **l'inclusion et pendant le suivi des patients**.
- **Statistiques** :
 - Calcul du nombre de sujets nécessaire,
 - Plan d'analyse statistique,
 - Tests statistiques utilisés et seuils de significativité retenus,
 - Méthode de calcul de la précision des estimations (intervalle de confiance à 95% (IC 95%) et/ou petit p),
 - Méthode de calcul de la mesure d'association retenue (Risque Relatif, Odds Ratio, Différence de Risque, Nombre de patients à Traiter pour éviter un évènement...),
 - Méthode de gestion des données manquantes.

3.4 RÉSULTATS

Cette partie doit décrire tous les résultats et rien que les résultats.

PRÉSENTATION DES RÉSULTATS

1. Description de la population recrutée et effectivement étudiée

On doit pouvoir suivre le nombre des patients aux principales étapes de l'étude : patients identifiés (« screened »), patients inclus, patients analysés, l'idéal est une représentation sous forme d'un **graphique de flux** (« flow chart »)

(Voir chapitre 5.3 : Comparabilité au cours de l'étude)

Les **caractéristiques initiales** (à l'inclusion, « baseline characteristics ») de l'échantillon étudié figurent en général dans le(s) premier(s) tableau(x) de l'article, s'il y a plusieurs groupes, figurent également des **informations sur la comparabilité initiale des groupes**.

2. Résultats pour l'objectif principal

Les résultats de l'**analyse principale** ++ (répondant à la question posée et à l'**objectif principal (primary objective)**) sont en général présentés avant ceux des analyses secondaires.

Ces résultats figurent dans le texte et/ou, tableaux et/ou figures suivants le(s) tableau(x) de description de l'échantillon étudié.

3. Résultats pour les objectifs secondaires

Les résultats des **analyses secondaires** (répondant aux objectifs secondaires) sont présentés ensuite dans le texte et/ou, tableaux et/ou figures suivants.

Complémentarité entre le texte, les tableaux et les figures

Le texte doit être cohérent avec les résultats présentés dans les tableaux et figures, néanmoins les résultats rapportés dans le texte ne figurent pas forcément dans les tableaux ou figures et vice-versa. Les résultats éventuellement négatifs doivent être aussi présentés.

Il ne doit en principe PAS y figurer de commentaires, ni d'interprétations, ni de comparaisons avec les données de la littérature.

LES VALEURS QUE VOUS DEVEZ POUVOIR TROUVER :

- Pour un **critère de jugement binaire (Oui/Non)** (ex. : survenue ou non d'une fracture, d'un infarctus, d'un AVC...) :
 - Le **nombre** d'évènements dans chaque groupe et leur **fréquence** (%) (risque de survenue de l'évènement dans chaque groupe),
 - S'il s'agit d'un traitement (essai thérapeutique) : les **indices d'efficacité du traitement** (ex. : risque relatif, Odds ratio, hazard ratio...) avec leur intervalle de confiance à 95 %,

- S'il s'agit d'une étude étiologique : les mesures de l'**association entre facteur de risque et maladie** (ex. : risque relatif, Odds ratio...) avec leur intervalle de confiance à 95 %.
- Si le **critère de jugement est une variable continue** (ex. : poids, taille, glycémie...) :
 - La **moyenne** et l'**écart-type** (synonyme : déviation standard),
 - Ou La **médiane** et l'étendue (min-max) ou les quartiles.

3.5 DISCUSSION

Ce que l'on doit trouver dans une discussion :

- Synthèse brève des principaux résultats de l'étude (aspects nouveaux et importants),
- Réponse à la question posée,
- Limites de l'étude (biais possibles et leurs conséquences possibles sur l'étude),
- Comparaison avec la littérature existante (**données objectives, publiées, référencées**),
- Portée des résultats : sont-ils extrapolables ?,
- Conclusion : implication pour la pratique médicale,
- Formulation de nouvelles hypothèses ou perspective de recherche.

Ce qui ne doit pas figurer dans la discussion : la méthodologie mise en œuvre ni une discussion sur des résultats qui n'auraient pas été présentés dans la section résultats.

3.6 LES RÉFÉRENCES BIBLIOGRAPHIQUES

LE CONTENU

- Dans l'**introduction et la discussion**, elles doivent justifier tous les faits énoncés.
- Elles doivent permettre aux lecteurs de retrouver facilement les articles concernés.
- Seules les références effectivement numérotées dans le texte doivent être citées.
- La bibliographie doit être à jour c'est-à-dire qu'elle **peut comporter des articles anciens si ce sont des articles princeps mais il doit y avoir un certain nombre de références très récentes.**

- Il faut éviter de référencer trop d'abstracts, de lettres, d'actes de congrès non publiés, de communications personnelles de travaux non évalués par des comités de lecture (thèse, mémoire).

LA FORME

- Le plus souvent, dans le texte, on trouve un numéro entre crochet qui renvoie à la citation complète à la fin de l'article, numérotation dans l'ordre d'apparition.
- La plus répandue – système numérique séquentiel de Vancouver : premier auteur (nom en entier et première lettre du prénom), autres auteurs (si plus de 6 auteurs on rajoute « *et al* »), titre, journal, année, volume, page.
- Le titre des journaux doit être abrégé selon l'Index Medicus.
- Les articles en cours de publication peuvent être cités : article « soumis », article « sous presse » ou « accepté ».

3.7 SYNTHÈSE

Quelles informations doit on retrouver dans chaque partie de l'article ?

SECTIONS

INTRODUCTION

I

QUELLE INFORMATION RECHERCHER EN PRIORITÉ ?

État de l'art sur le sujet :

Ce qu'on sait.

Les informations qui manquent :

- pas d'études,
- études non concluantes,
- études discordantes.

Pourquoi il est important de répondre à la question

(gravité, coût, fréquence, traitements efficaces ou prévention possible...).

Objectif (en général dernière phrase de l'introduction) : « PECO »

(plus ou moins précis) :

- Population,
 - Élément évalué : Intervention (traitement...) ou Facteur de risque/Facteur pronostique étudié,
 - Comparaison,
 - Outcome (Critère de jugement, critère d'évaluation).
-

MÉTHODE

M

Population source (a priori).

Critères d'inclusion et de non inclusion.

Modalités de recrutement (hôpital, population générale, cabinets, mutuelles, screening systématique ou non, patients consécutifs ou non...).

Lieux et date.

Critère de jugement Principal (CJP) et Critères de jugement secondaires :

- Définition,
- Méthodes de mesures (instrument de mesure, insu...),
- Modalités de recueil (par téléphone, face à face...),
- Temps auquel (auxquels) les mesures sont faites pendant le suivi.

Intervention/Facteur de risque :

- si étude d'efficacité : Description de l'intervention dans le groupe traité et de ce qui est fait dans le groupe contrôle (placebo, rien...),
 - si étude étiologique : Description du Facteur de risque,
 - Définition,
 - Méthodes de mesures (instrument de mesure, conditions...),
 - Modalités de recueil (visite systématique, visite si symptômes, faces à face ou téléphone...).
-

RÉSULTATS

R

Tous les résultats mais rien que les résultats

(pas d'interprétation, pas de méthode ni de commentaires).

Les caractéristiques de l'échantillon étudié dans son ensemble et des deux groupes comparés :

- En général données présentées dans le tableau 1,
- Permet de comparer les caractéristiques des patients de l'étude avec ceux d'autres études et permet de voir la comparabilité des groupes.

L'ampleur d'effet :

- Importance de l'efficacité (si on est dans l'efficacité d'un traitement),
- Force de l'association entre facteur de risque et maladie (si on est dans l'étiologie),
- Sensibilité, spécificité, valeurs prédictives positives et négatives si l'on est dans la validation d'un test diagnostique.

La significativité statistique (p-value et intervalle de confiance : l'association observée peut-elle être seulement le fruit du hasard ou est-elle trop forte pour être seulement le fruit du hasard?).

DISCUSSION

And D

Ce qu'apporte cette étude par rapport aux autres études publiées.

Les limites de l'étude (discussion des biais possibles et de leurs conséquences sur la validité des résultats).

LES GRILLES DE LCA OU « CHECK-LIST »

GÉNÉRALITÉS

Elles permettent de ne pas oublier les points essentiels à vérifier en fonction du type d'étude, il en existe de nombreuses.

Check-list généraliste :

1. L'hypothèse de recherche (ou question de recherche)
2. Le type (design, architecture) de l'étude
3. La population étudiée
4. Le critère de jugement
5. L'intervention ou le facteur de risque
6. Les analyses statistiques
7. Les résultats (ampleur de l'effet et significativité statistique)
8. Les biais et facteurs de confusion
9. La conclusion

Cette check-list est applicable pour les études de cohortes et les essais randomisés. C'est donc le plan que nous adopterons pour ces deux chapitres.

Nous avons modifié cette check-list pour l'adapter aux études cas-témoins et aux études sur la performance d'un test diagnostique.

Check-list pour les études cas-témoin

1. L'hypothèse de recherche (ou question de recherche)
2. Le type (design, architecture) de l'étude
3. La population étudiée et la population source
4. Des cas
5. Des témoins
6. Le critère de jugement
7. Définition des cas
8. Définition des témoins
9. Le facteur de risque étudié
10. Les analyses statistiques
11. Les résultats (ampleur de l'effet et significativité statistique)
12. Les biais et facteurs de confusion
13. La conclusion

Nous vous proposons ci-dessous une **check-list modifiée pour les études diagnostiques** :

1. L'hypothèse de recherche (ou question de recherche)
2. Le type (design, architecture) de l'étude
3. La population étudiée – la population source
4. Le gold standard (= méthode diagnostique/test de référence)
5. La méthode diagnostique/test à évaluer
6. Les analyses statistiques
7. Les résultats (sensibilité, spécificité, valeurs prédictives, ratios de vraisemblance)
8. Les biais et facteurs de confusion
9. La conclusion

Dans ce chapitre sont traitées les généralités de chaque point de la checklist. Chaque point est ensuite analysé de façon spécifique dans le contexte de chaque type d'étude.

4.1 LA QUESTION DE RECHERCHE (LE « PECO »)

Un article original tente d'apporter une réponse à une question de recherche.

OÙ TROUVER CETTE QUESTION DANS L'ARTICLE ?

- À la **fin de l'introduction** (c'est là qu'elle devrait être) rechercher l'objectif principal (puisque l'objectif est de répondre à la question posée).
- Parfois dans le titre.
- Si elle est difficile à trouver, il faut repérer le **critère de jugement principal** ou le paragraphe **calcul de la taille de l'échantillon** (analyse statistique) dans lequel doit figurer théoriquement la question de recherche.

Les composantes de la question correspondent à celles de l'objectif principal.

La question de recherche doit être la plus précise possible :

Question peu précise :
La prévention permet-elle de réduire les AVC ?



Question un peu plus précise :
La prévention permet-elle de réduire les AVC ischémiques ?



Question plus précise :
Le traitement des patients atteints d'arythmie cardiaque permet-il de réduire les risques d'AVC ischémique ?



Question de plus en plus précise :
Le traitement par AVK des patients âgés de plus de 50 ans atteints d'arythmie cardiaque permet-il de réduire le risque d'AVC ischémique ?



Question finale
Le traitement par AVK au long cours des patients âgés de plus de 50 ans atteints d'arythmie cardiaque permet-il de réduire le risque d'AVC ischémique de 30% sur un suivi de 5 ans ?

MNÉMOTECHNIQUE DES COMPOSANTS DE LA QUESTION DE RECHERCHE :

le **PECO**

P

POPULATION ÉTUDIÉE

E

E : ÉLÉMENT ÉVALUÉ

I : INTERVENTION, ou F : FACTEUR DE RISQUE/PRONOSTIQUE

- Intervention s'il s'agit d'une étude d'intervention,
- ou Facteur de risque si étude étiologique,
- ou Facteur pronostique si étude pronostique.

C

COMPARAISON (groupe « contrôle » = groupe de référence)

O

« OUTCOME » c'est l'état de santé auquel on s'intéresse, appelé dans le jargon de la LCA : « Critère de jugement principal ». Ex. : l'infarctus, la fracture du col, la démence, la dépression, la qualité de vie, la douleur, etc.

EXEMPLE DE PECO :

LA QUESTION :

Un **traitement par antivitamine K** est-il plus efficace qu'un **traitement par aspirine** pour **réduire le risque de survenue d'accident vasculaire cérébral (AVC)** chez des **sujets âgés de 65 à 85 ans en arythmie cardiaque par fibrillation auriculaire** ?

Lien entre question de recherche – Objectif – hypothèse

L'OBJECTIF :

Mesurer l'**efficacité d'un médicament antivitamine K** pour prévenir la survenue d'**accident vasculaire cérébral (AVC)** chez des **sujets de 65 à 85 ans en arythmie cardiaque par fibrillation auriculaire** par rapport à un **traitement par aspirine**.

L'HYPOTHÈSE :

Un **traitement par antivitamine K** permet de réduire le risque de survenue d'**un accident vasculaire cérébral (AVC)** chez des **sujets de 65 à 85 ans en arythmie cardiaque par fibrillation auriculaire** par rapport à un **traitement par aspirine**.

La question de recherche peut TOUJOURS être classée dans une des catégories suivantes (Voir chapitre 4.2 : Le type d'étude) :

- EFFICACITÉ D'UNE INTERVENTION
- ÉTIOLOGIE / CAUSALITÉ
- PRONOSTIC
- PERFORMANCE TEST DIAGNOSTIQUE
- INCIDENCE / PRÉVALENCE

EFFICACITÉ D'UNE INTERVENTION

sur une maladie / un état de santé

Le chercheur maîtrise l'intervention

L'intervention en santé est ici envisagée au sens large : traitements médicamenteux, autres types de traitements (Kinésithérapie, chirurgie, etc.), autres interventions (dépistage, formation de professionnels de santé, etc.).

ÉVALUATION DE L'EFFICACITÉ D'UN TRAITEMENT :

- sur la survenue d'un évènement de santé (survenue d'un infarctus, d'une chute, d'une infection, d'une insuffisance rénale, récurrence d'un cancer...),
 - sur l'évolution d'un état de santé ou d'un symptôme (évolution de la fatigue, du poids, de la tension artérielle, de l'anxiété, de la qualité de vie, des douleurs...),
- ... tout en évaluant également les **effets potentiellement néfastes** (c'est le rapport-bénéfice/ risque qui est important pour le patient).

Au final il s'agit de répondre à la question : Cette thérapeutique est-elle utile, inutile ou nuisible pour mes patients ?

Exemple :

Un **traitement de 4 ans par risédronate et calcium-vitamine D** permet-il de **réduire la fréquence des fractures ostéoporotiques** chez des **femmes âgées de 70 à 90 ans ostéoporotiques** par rapport au **calcium-vitamine D seul** ?

CAS PARTICULIER : ÉVALUATION DE L'EFFICACITÉ D'UN PROGRAMME DE DÉPISTAGE

Dans ce cas, la mise en place du programme de dépistage est une intervention de santé maîtrisée par l'investigateur de l'étude. On est donc dans la même catégorie que précédemment. L'intervention pourrait également être un programme de formation des professionnels, la mise en place de la télémédecine, l'utilisation de la checklist au bloc opératoire, etc.

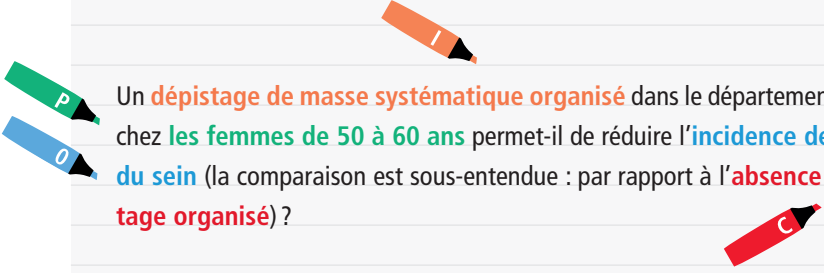
Comme un traitement, un programme de dépistage peut avoir des effets néfastes liés aux faux positifs (entraînant des examens invasifs ou des traitements inutiles) ou aux faux négatifs (responsables de fausse réassurance des individus) qu'il faut évaluer.

Au final, il s'agit de répondre à la question : ce programme de dépistage est-il utile, inutile ou nuisible pour les personnes concernées par ce programme ?

Le principe est donc proche de l'évaluation d'un traitement.

La principale différence : un programme de dépistage est une intervention beaucoup plus complexe que la simple prescription de traitements, son résultat dépend de la **qualité du test utilisé pour le dépistage ET de l'efficacité des traitements** entrepris pour les malades dépistés, **ET de la participation de la population**.

Exemple :



Un **dépistage de masse systématique organisé** dans le département de l'Isère chez **les femmes de 50 à 60 ans** permet-il de réduire l'**incidence des cancers du sein** (la comparaison est sous-entendue : par rapport à l'**absence de dépistage organisé**) ?

La question de recherche porte sur

L'ÉTILOGIE

« ASSOCIATION » entre Un Facteur de Risque ET Survenue d'une maladie ?
Le CHERCHEUR OBSERVE mais n'intervient pas sur le facteur de risque

Parler d'« ASSOCIATION » ou « lien » ou « relation » sous-entend qu'on observe un lien statistique au-delà du hasard entre ces deux éléments mais qu'on ne peut pas aller plus loin sur l'existence d'une relation causale.

Bien sûr, ce qui nous intéresse au fond est de savoir si un facteur a ou non **une relation CAUSALE avec la maladie**.

Mais il est **très difficile d'établir formellement une relation de causalité**.

(Voir chapitre 6.1 : Les résultats, études de cohorte)

ESTIMATION DE L'EFFET D'UN FACTEUR DE RISQUE SUR LA SURVENUE D'UNE MALADIE :

- Se fait au moyen d'études analytiques (encore appelées étiologiques).
- Dans lesquelles l'investigateur ne maîtrise pas le facteur étudié, c'est à dire le facteur qui est supposé avoir un effet sur la survenue de la maladie ou l'évènement de santé étudié.

- Cet effet est le plus souvent délétère (facteur de risque, par exemple tabac et cancer du poumon), mais peut-être (très rarement) protecteur (facteur protecteur, par exemple tétine et risque de mort subite).

Au final, Ce facteur de risque est-il associé à une augmentation du risque de survenue de la maladie ?

Exemple :

La **position de couchage ventrale** augmente-t-elle le **risque de mort subite** chez des **nourrissons en bonne santé âgés de 3 à 24 mois** par rapport à des **nourrissons couchés sur le dos ou sur le côté** ?



La question de recherche porte sur le

PRONOSTIC

« ASSOCIATION » entre Facteur Pronostique

ET Évolution (aggravation ou amélioration) d'une maladie ?

LE CHERCHEUR OBSERVE mais n'intervient pas sur le facteur pronostique

ÉVOLUTION D'UNE MALADIE EN FONCTION DE LA PRÉSENCE D'UN FACTEUR PRONOSTIQUE :

- Se fait au moyen d'études pronostiques.
- L'investigateur observe le(s) facteur(s) supposé(s) être positivement ou négativement associé(s) au pronostic (à l'évolution) de la maladie ou de l'évènement de santé (complications ou décès).
- Parfois assez proche d'une étude étiologique, la différence principale est qu'elle est en général menée dans une cohorte de malades.
- Nécessite en théorie une étude avec un suivi longitudinal (étude de cohorte) seule permettant de compter les cas de complications et les décès.

Au final, ce facteur est-il associé à l'évolution (au pronostic) de la maladie ?

Exemple 1 :

Chez des **patientes atteintes d'un premier cancer du sein invasif non métastatique**, quelle est l'**incidence des récives métastatiques à 5 ans en fonction de la corpulence** ?

Exemple 2 :

Un **score de risque clinique** permet-il de cibler un groupe à haut risque de **récidive de lymphome dans les 5 ans** dans une population de **patients âgés de 18 à 65 ans ayant eu un lymphome après l'âge de 18 ans** ?


La question de recherche porte sur la
PERFORMANCE TEST DIAGNOSTIQUE

ÉVALUATION DE LA PERFORMANCE D'UN TEST DIAGNOSTIQUE

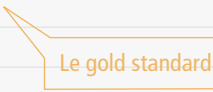
Un examen diagnostique vise à classer les patients en « malades » ou « non malades » par rapport à une maladie donnée. Pour évaluer sa performance on mesure sa sensibilité, sa spécificité et ses valeurs prédictives par rapport à un examen diagnostique de référence censé être parfait : le « gold standard » (ou « étalon or »).

Exemple :

 Test diagnostique à évaluer



Le **nouveau test de diagnostic rapide de l'angine bactérienne** permet-il de diagnostiquer de façon performante l'étiologie de l'angine (bactérienne ou virale) chez **les enfants suivis en médecine de ville consultant pour une angine banale** en utilisant **la mise en culture des prélèvements de gorge** comme gold standard ?



 Le gold standard

REMARQUE : Performance d'un test diagnostique = capacité du test à classer correctement les personnes en malades et non-malades (examen POSITIF/examen NEGATIF), on ne parle en général pas de l'efficacité d'un test diagnostique.




NIVEAU 2

L'évaluation de la performance d'un test diagnostique doit être distinguée d'une situation totalement différente : l'évaluation de L'EFFICACITÉ d'une STRATÉGIE diagnostique car on se situe alors dans le champ de l'évaluation de l'efficacité d'une « intervention » (en l'occurrence « l'intervention » est la réalisation du test diagnostique).

Exemple : évaluer L'EFFICACITÉ d'une stratégie diagnostique consistant à utiliser le test de diagnostic rapide de l'angine bactérienne :



L'utilisation systématique du nouveau test de diagnostic rapide de l'angine bactérienne chez **l'enfant en médecine de ville** permet-il de **réduire l'utilisation d'antibiotiques sans augmenter les complications cliniques** des angines par rapport à la **prise en charge habituelle**.

INCIDENCE / PRÉVALENCE

MESURE d'une INCIDENCE, MESURE d'une PRÉVALENCE,
LE CHERCHEUR DECRIT un phénomène (pas de comparaison)

ESTIMATION D'UNE INCIDENCE

Nécessite une étude avec un suivi longitudinal (étude de cohorte), seul type d'étude permettant de compter les nouveaux cas pour calculer l'incidence.

Exemple :



Quelle est l'incidence du cancer du sein chez les femmes âgées de 40 à 75 ans en Isère en 2004 ?

Réponse : l'incidence du cancer du sein était de 15 pour 1000 personnes-année chez les femmes de 40 à 75 ans résidant en Isère sur l'année 2004.

ESTIMATION D'UNE PRÉVALENCE

Exemple :



Quelle est la prévalence de l'insuffisance rénale chronique terminale dialysée en France en 2004 chez les personnes âgées de plus de 50 ans ?



Réponse : 5 % des personnes de plus de 20 ans étaient en dialyse pour une insuffisance rénale chronique terminale en France au 31 décembre 2004.

REMARQUE, études mesurant l'INCIDENCE ou la PRÉVALENCE d'une maladie :
Estimation d'un chiffre, pas d'hypothèse ni de comparaison.
DANS LES AUTRES CAS, on cherche une ASSOCIATION entre deux facteurs en
COMPARANT deux GROUPEs (au moins), l'un étant considéré comme RÉFÉRENCE.

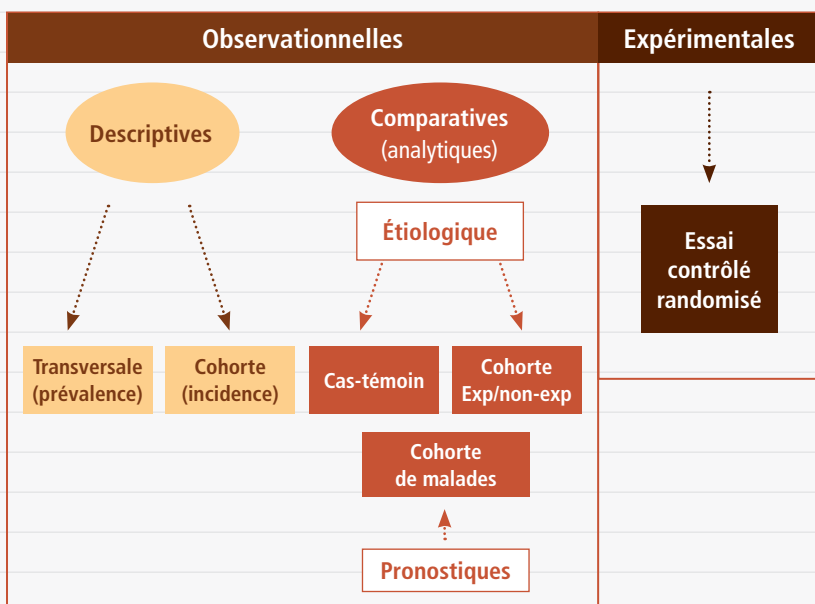
4.2 LE TYPE D'ÉTUDE (SYNONYMES : LE DESIGN, LE SCHÉMA, L'ARCHITECTURE, LE DESSIN...)

On classe les types d'études selon différentes dimensions :

- En fonction de l'objectif :
 - Les études descriptives,
 - Les études analytiques,
 - Les études diagnostiques,
 - Les études d'efficacité,
 - Les études pronostiques.
- En fonction du schéma d'étude :
 - Les études transversales,
 - Les études longitudinales (cohortes),
 - Les études cas-témoins,
 - Les essais randomisés.
- En fonction de l'intervention :
 - Les études observationnelles,
 - Les études interventionnelles.

L'important est de connaître le type d'étude LE MIEUX adapté à chaque catégorie de question de recherche. Il faut aussi vérifier que le niveau de preuve du type d'étude choisi est le meilleur possible. Cependant, le choix du type d'étude peut être limité par des questions éthiques, financières, ou de possibilité de réalisation.

LES TYPES D'ÉTUDES



RELATIONS ENTRE QUESTION POSÉE - TYPE D'ÉTUDE

TABLEAU À CONNAITRE PARFAITEMENT

QUESTION POSÉE TYPE D'ÉTUDE

Efficacité d'une intervention Essai Clinique Randomisé
(Traitement / Dépistage / formation...)

Étiologie / Causalité Cohorte
(meilleur niveau de preuve que cas-témoins)
Cas-Témoin

Pronostic Cohorte

Test diagnostique

Performance d'un test diagnostique Étude Transversale
(comparaison avec un gold standard)

Efficacité d'une stratégie diagnostique Essai Randomisé
(idem efficacité d'une intervention) (nouvelle stratégie diagnostique vs ancienne)

Efficacité Essai Randomisé
(idem efficacité d'une intervention) (nouvelle stratégie diagnostique vs ancienne)

Prévalence Étude Transversale
(possible évidemment à partir d'une étude de cohorte)

Incidence Étude Longitudinale
(suivi de cohorte descriptive, ou registre)

LE NIVEAU DE PREUVE DES DIFFÉRENTS TYPES D'ÉTUDES

En fonction du risque de biais, les différents types d'étude n'apportent pas le même degré de certitude et de confiance vis-à-vis de leurs résultats. Les études observationnelles qui ne permettent pas à l'investigateur de maîtriser l'intervention apportent par exemple un niveau de preuve inférieur à celui des études randomisées en double insu.

L'essai randomisé en double insu apporte le meilleur niveau de preuve car il réduit le risque de biais et donc de conclusion erronée. Aussi pour valider l'efficacité d'un traitement il est nécessaire de disposer d'essais cliniques randomisés.

MAIS on ne peut pas obliger un groupe de patients à fumer pour vérifier que le tabac augmente le risque de cancer du poumon : **pour explorer un lien entre facteur de risque et maladie, cette méthode n'est PAS UTILISABLE** pour des raisons éthiques. **Donc pour ce qui concerne l'étude des facteurs de risque on devra se contenter d'études d'observation qui devront être réalisées avec la meilleure qualité possible.**

La Haute autorité de santé (HAS) a défini les NIVEAUX (1 à 4) de preuve scientifique de chaque type d'étude. Ainsi lorsqu'un groupe d'expert de la HAS émet des recommandations de pratiques, il doit explicitement dire quel est le GRADE (A, B ou C) de chaque composant de cette recommandation, et ce grade est directement lié au niveau de preuve des études sur lesquelles la recommandation a été établie. Parfois, il n'y a aucune étude probante permettant d'établir une recommandation alors en attendant de disposer d'une étude la recommandation est basée sur un « accord professionnel d'experts », ce qui représente le plus faible grade d'une recommandation mais c'est mieux que rien...

Grade des recommandations	Niveau de preuve scientifique fourni par la littérature
A Preuve scientifique établie	Niveau 1 - essai comparatifs randomisés de forte puissance, - méta-analyse d'essais comparatifs randomisés, - analyse de décision fondée sur des études bien menées.
B Présomption scientifique	Niveau 2 - essai comparatifs randomisés de faible puissance, - études comparatives non randomisées bien menées, - études de cohorte.
C Faible niveau de preuve scientifique	Niveau 3 - études de cas témoins.
	Niveau 4 - études comparatives comportant des biais importants, - études rétrospectives, - séries de cas, - études épidémiologiques descriptives (transversale, longitudinale).

4.3 LA POPULATION/L'ÉCHANTILLON ÉTUDIÉ(E) LA POPULATION SOURCE (GÉNÉRALITÉS)

Quelle est la population source de laquelle sont issus les individus de l'échantillon étudié (sélection *a priori*) ?

- Sélection *a priori* des personnes selon des **critères d'inclusion et de non inclusion**.

Ces informations doivent figurer dans la section matériel et méthodes.

Quelle est la population étudiée (ou plutôt l'échantillon étudié) ?

Finalement, *a posteriori*, qui sont les personnes qui ont réellement participé à l'étude ?

- Sélection de fait liée au **contexte** (centres, période d'étude, modalités de sélection des sujets...) dans lequel les patients sont recrutés pour l'étude => ces informations de situent dans la section matériels et méthodes.
- Description de leurs **caractéristiques** => ces informations se trouvent dans la **section résultats** (en particulier dans le tableau 1).

Ces questions concernent la **validité externe** et permettent de réfléchir à quels types de patients les résultats sont extrapolables.

La description de la population sélectionnée *a priori* commence par la liste des critères d'inclusion et de non inclusion (souvent appelés à tort critères d'exclusion).

- Le groupe est-il **très sélectionné** (multiples critères de non inclusion) ? Si oui, alors l'extrapolabilité des résultats est limitée.
- À l'inverse le groupe est-il **peu sélectionné** ? Si oui, on a alors une meilleure extrapolabilité mais au prix d'une plus grande variabilité (d'un « bruit de fond » plus important) donc d'une plus grande difficulté à mettre en évidence un effet s'il existe.

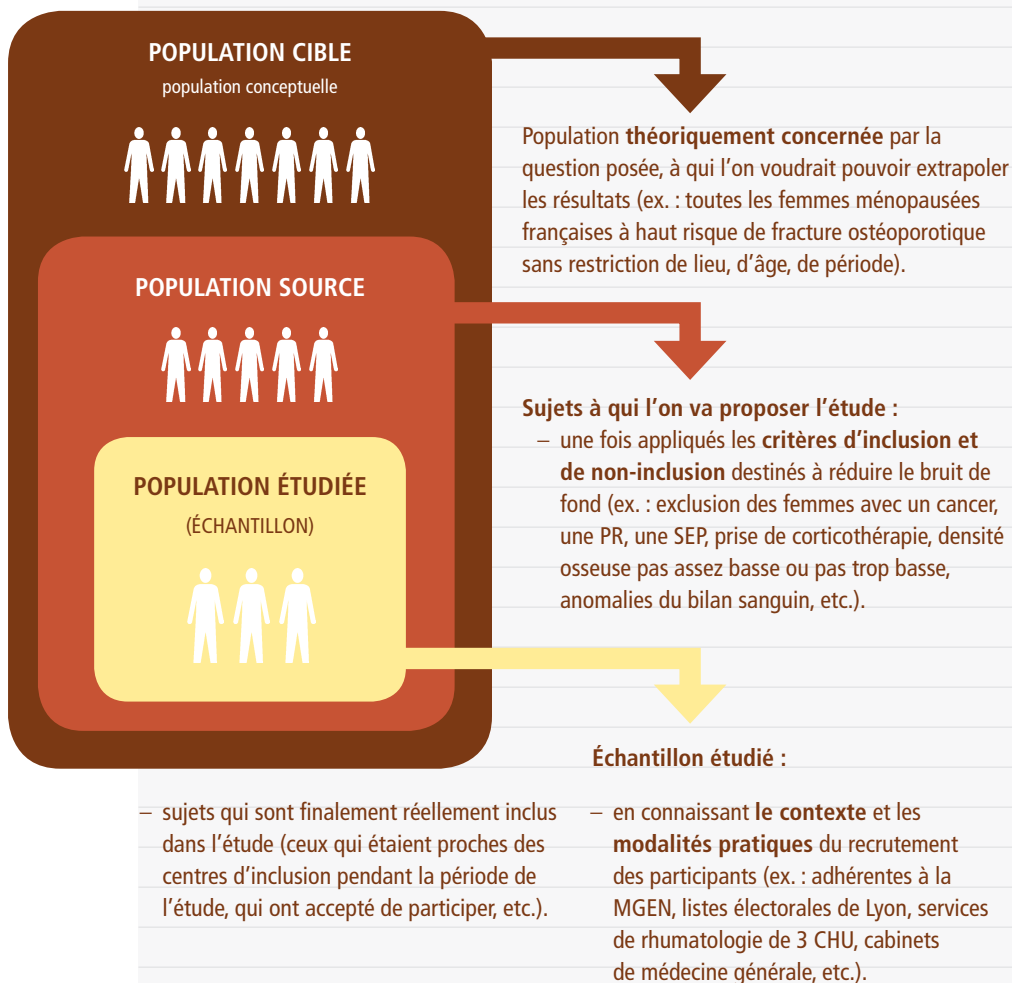
Exemple : on inclut des patients âgés de 20 à 50 ans consultant en médecine générale pour une fatigue générale et/ou des douleurs diffuses, sans diabète ni hypertension artérielle ni cancer évolutif. Cet échantillon n'est pas trop sélectionné il est beaucoup plus extrapolable que si les patients étaient recrutés à partir de centres universitaires très spécialisés.

Ces critères d'inclusion et de non inclusion doivent avoir une définition précise.

Exemple : « hypertension artérielle » sans précision n'est pas suffisant, il faut la définition chiffrée : tension artérielle strictement supérieure à 140/90 mm Hg, mesurée au repos et 2 fois à quelques minutes d'intervalle afin de s'assurer que pour tous les patients, la même définition est appliquée.

Les caractéristiques de l'échantillon étudié *a posteriori* (échantillon qui participe réellement à l'étude et sur lequel vont porter les résultats) permettent d'avoir une deuxième indication sur l'échantillon. Dans l'article, ces données figurent dans le tableau descriptif de la population, en général le tableau n°1.

Exemple : 126 patients ont participé, 80 hommes et 46 femmes, l'âge moyen était de 46 ans (SD 4 ans).



Comment est définie la population source ?

Le lieu et la période d'étude :

- Début et fin de période d'inclusion,
- Sujets sains ou malades,
- Médecin de ville ou hôpital,
- Ambulatoire, court, moyen séjour,
- CH général, privé ou service spécialisé d'un CHU,
- (Dans les services hyperspécialisés dits référents, il y a souvent un plus fort pourcentage de patients avec des pathologies plus avancées ou plus compliquées).

Les critères d'inclusion et de non-inclusion :

- Âge, sexe,
- Autres caractéristiques sociodémographiques,
- Co-morbidités, facteurs de risque ou facteurs pronostiques,
- Stades de la maladie, formes particulières, prises en charge antérieures ou traitements concomitants.

Comment est sélectionné l'échantillon ?

L'idéal : inclure toute la population...

MAIS impossible (faisabilité et coût).

Méthodes de sélection de la - à la + biaisée :

- Recrutement aléatoire (tirage au sort pour décider devant chaque patient éligible si on l'inclut dans l'étude ou non),
- Recrutement de tous les patients consécutifs sur une période,
- Recrutement séquentiel (1 patient tous les X patients),
- Volontariat (biais +++ : patients les plus motivés, ou les moins graves, ou les plus graves, ou les moins satisfaits ?, etc.).

Les conséquences sur les résultats dépendent du type d'étude.

FLUCTUATIONS D'ÉCHANTILLONNAGE

Si on considère différents échantillons issus de la population source, l'estimation de paramètres statistiques (ex. : moyenne) pourra varier d'un échantillon à l'autre, simplement du fait du hasard. Cette fluctuation est inévitable, mais diminue à mesure que la taille de l'échantillon augmente.

Pour prendre en compte cette fluctuation, on utilise souvent l'intervalle de confiance (en général à 95 %), qui représente la fourchette des valeurs dans laquelle le paramètre mesuré se situerait dans 95 % des cas, que quel que soit l'échantillon dans la population source.

Ceci explique la différence entre population source et échantillon (ou population) étudié.

Par exemple, imaginons un essai réalisé pour tester un traitement contre l'ostéoporose chez des femmes ménopausées. Pour une même population source de femmes âgées de plus de 50 ans, avec des critères d'inclusion identiques, si on fait plusieurs études différentes on pourra avoir dans une étude un échantillon d'âge moyen 60 ans et dans une autre étude un échantillon d'âge moyen 70 ans...

4.4 LE CRITÈRE DE JUGEMENT (GÉNÉRALITÉS)

C'est la façon de mesurer la maladie / l'évènement de santé / l'issue clinique (décès, infarctus, récurrence de cancer, ulcère duodéal...) ou l'état de santé (fatigue, dépression, qualité de vie...) sur lequel le facteur de risque (dans les études observationnelles) ou le traitement (dans les essais) est censé avoir un effet.

C'est la traduction imparfaite de l'anglais « **OUTCOME** » qui signifie conséquence, issue, aboutissement. Dans les essais on parle volontiers de **critère d'efficacité**.

Parmi les événements / état de santé auxquels on s'intéresse, on distingue classiquement :

- ceux qui sont facilement **mesurables sans biais** (ex. : décès), critères de jugement « durs »,
- ceux qui nécessitent un jugement **clinique/biologique/radiologique** (ex. : infarctus du myocarde, AVC),
- ceux qui nécessitent un jugement **subjectif** (ex. : qualité de vie, handicap), critères dit « mous » ou « subjectifs » car plus influençables dans leur mesure.

Quelles que soient les études, il doit être mesuré de façon standardisée et reposer sur des critères explicites et bien définis.

La mesure de critère de jugement repose sur des instruments de mesure plus ou moins complexes.

La **validité** d'un instrument (= **son exactitude**) est sa capacité à bien mesurer ce qu'il est censé mesurer. La **reproductibilité** d'un instrument représente sa capacité à produire un même résultat lorsqu'il mesure le même phénomène de façon répétée.

Exemple : pour mesurer la reproductibilité du diagnostic de thrombose veineuse profonde à partir d'une phlébographie, on mesure la reproductibilité inter-observateur en faisant lire les mêmes radios par plusieurs radiologues et en analysant la concordance de leurs résultats, et la reproductibilité intra-observateur en faisant lire les mêmes radios plusieurs fois par le même radiologue et en analysant la concordance de ces résultats. Bien entendu dans les deux cas ces lectures doivent être faites en insu des lectures précédentes.

Les événements/états de santé auxquels on s'intéresse sont parfois très subjectifs, pour autant ils peuvent être tout à fait pertinents s'ils sont importants pour le patient. C'est le cas de la mesure de la fatigue, de la douleur, de la dépression, de l'anxiété, de la qualité de vie. Les instruments de mesure (questionnaire ou échelle) doivent avoir été validés notamment pour ce qui concerne leur exactitude et leur reproductibilité.

La définition des **critères de jugement** doit être établie de façon **claire et précise** avant le début de l'étude.

Exemple : si l'on étudie l'efficacité d'un traitement sur le risque de fracture dans un essai thérapeutique ou l'impact d'un facteur de risque sur la survenue de fractures, il faut définir avant de débiter l'étude de quelles fractures il s'agit : faut-il considérer uniquement les fractures sans traumatisme majeur ou faut-il également considérer des fractures survenant lors de traumatisme violent ? Faut-il considérer uniquement les fractures importantes telles que fracture du col du fémur, du poignet, de l'épaule, des vertèbres ou doit-on considérer également des fractures des orteils ou des métacarpiens ? Pour les tassements vertébraux, il est parfois difficile d'affirmer la survenue d'un nouveau tassement surtout s'il est mineur, si les deux radiologues sont en désaccord faut-il demander à un troisième radiologue où exiger un consensus des deux premiers radiologues ?

Pour minimiser le risque de **biais de mesure**, la mesure du critère de jugement doit se faire :

- **en aveugle (insu) des facteurs de risques / ou du traitement étudiés** : celui qui mesure le critère de jugement ne sait pas si le sujet est exposé au facteur de risque ou s'il prend le traitement ou le placebo,
- **avec des outils (ou instruments de mesure) standardisés et validés** : questionnaires de qualité de vie SF-36, questionnaire sur la dépression, échelles de mesure de la douleur, de l'incapacité, de l'asthénie...

Ceci est d'autant plus important que le critère de jugement est subjectif.

- **tous les patients/individus de l'étude doivent être soumis aux mêmes procédures diagnostiques, mêmes questionnaires.**

On distingue les biais de mesure différentiels et non différentiels :

- biais de mesure **non différentiel** : biais aléatoire qui peut survenir dans les deux groupes avec le même risque,
- biais de mesure **différentiel** : biais qui survient préférentiellement dans un groupe et qui peut influencer les résultats de l'étude.

Exemple : considérons une étude dans laquelle on analyse le lien entre exposition à des ondes électromagnétiques et la survenue de troubles du sommeil. Si tous les individus de l'échantillon étudié ont tendance à surestimer leurs troubles, qu'ils soient exposés ou non au facteur de risque (ondes électromagnétiques), c'est un biais non différentiel : il ne risque pas de faire croire de façon erronée à un lien entre ondes électromagnétiques et troubles du sommeil. En revanche, si les individus se savent exposés aux ondes magnétiques et qu'ils pensent que c'est cela qui entraîne leurs troubles, ils peuvent avoir tendance à surestimer systématiquement ces troubles, alors que les personnes non exposées ne surestiment pas leurs troubles du sommeil de façon systématique. Ce biais peut faire croire de façon erronée à un lien entre ondes électromagnétiques et troubles du sommeil.

Critères de qualité d'un critère de jugement :

- **critère simple > critère composite.**

Exemple, critère simple : survenue d'un AVC ischémique ; critère composite : survenue d'un événement cardio vasculaire comportant soit un AVC, soit un syndrome coronarien aigu, soit un décès par AVC ou syndrome coronarien aigu.

- **validité** = sensibilité et spécificité élevées (le plus proche possible de 100%).
- **reproductibilité et fiabilité** = concordance élevée entre mesures répétées.
- **critère objectif > critère subjectif.**
- **critère consensuel > critère improvisé.**

Exemple : critère décidé par un groupe d'experts reconnus, si possible avec une méthodologie reconnue d'établissement de consensus, avant ou au début de l'étude, versus critère décidé par l'auteur de la publication ne reposant pas sur une méthode de choix explicite.

- **critère pertinent d'un point de vue clinique.**

Exemple : un critère pertinent est important pour le patient, il peut être objectif (diabète, AVC..) ou subjectif (douleurs, fatigue, dépression...), un critère intermédiaire est en général considéré comme non pertinent (paramètre biologique ou radiologique per exemple) sauf dans quelques cas particuliers dans lesquels les critères intermédiaires ont été validés.

4.5 L'ÉLÉMENT ÉVALUÉ : L'INTERVENTION OU LE FACTEUR DE RISQUE/PRONOSTIQUE (GÉNÉRALITÉS)

C'est l'élément (facteur de risque, facteur pronostique, traitement) qui est censé avoir un impact sur le critère de jugement :

- Si c'est un facteur de risque, il est associé au risque de la maladie étudiée (ou aggrave l'état de santé), en général il augmente le risque, plus rarement certains facteurs sont protecteurs, on les désigne néanmoins sous le nom générique de facteur de risque.
- Si c'est un facteur pronostique, il est censé être associé avec l'évolution de la maladie étudiée.
- Si c'est un traitement, il est censé avoir un effet qui diminue le risque de la maladie (ou améliore l'état de santé).

4.6 L'ANALYSE STATISTIQUE

LA TAILLE DE L'ÉCHANTILLON MINIMUM NÉCESSAIRE

(= CALCUL DU NOMBRE DE SUJETS NÉCESSAIRES = CALIBRAGE DE L'ÉTUDE)

S'il existe réellement un lien significatif entre un facteur de risque et la survenue d'une maladie/ou si un traitement est réellement efficace sur une maladie, alors on veut être en capacité de le démontrer.

Pour cela l'étude doit avoir une puissance statistique suffisante c'est-à-dire un effectif suffisant.

Le calcul de la taille d'échantillon repose sur la **différence attendue** entre les groupes, et les **risques alfa et beta choisis** par l'investigateur de l'étude.

- la **différence** que l'on pense raisonnablement pouvoir mettre en évidence entre les groupes étudiés est estimée par l'investigateur à partir des données de la littérature, ou des résultats d'études pilotes ou préliminaires.

Exemple, l'hypothèse est la suivante :

Chez des sujets de 65 à 85 ans en arythmie cardiaque par fibrillation auriculaire, d'après les études publiées, **le risque d'AVC dans les 10 ans** des

patients sous aspirine serait de 5% et un traitement par antivitamine K permettrait de réduire ce risque à 3% (soit une baisse relative de 40%) par rapport à un traitement par aspirine. L'objectif sera d'essayer de montrer que cette hypothèse est vraie.

- **Le risque beta et la puissance statistique** de l'étude varient avec la taille de l'échantillon. **Une puissance de 80 % signifie que si notre hypothèse est juste, on a 80 chances sur 100 de le confirmer** avec cette étude, **et donc 20 % de risque de conclure à tort que ce n'est pas un facteur de risque : c'est le risque beta = 1 - puissance.**

Ce calcul doit avoir été réalisé *a priori* à partir d'une hypothèse qui doit être clairement exprimée et justifiée. C'est l'investigateur de l'étude qui choisit la puissance statistique. Elle doit être au moins égale à 80% (si possible 90%), un risque beta de plus de 20% étant peu acceptable.

Parfois on ne trouve pas ce calcul dans l'article ce qui est un défaut GRAVE pour la LCA.

- **Le risque alpha doit aussi être déterminé par l'investigateur**, c'est le risque de conclure à une association alors qu'il n'y en a pas, il est le plus souvent **fixé à 0,05.**

Exemple de calcul de taille d'échantillon dans un essai : dans l'essai précédent, mesurer la véritable efficacité nécessiterait d'étudier la totalité des patients en arythmie cardiaque par FA, bien sûr ceci n'est pas possible. On doit donc, à partir d'un essai mené dans un échantillon de patients, estimer les vrais résultats (ceux que l'on obtiendrait si on pouvait étudier tout le monde). Cette estimation, même si l'étude est très bien faite, est soumise à la possibilité d'erreur du simple fait de la variabilité des personnes et des phénomènes (erreur aléatoire). Ces risques d'erreur sont directement liés au nombre de patients inclus dans l'essai, ce sont le risque alpha et le risque beta. Avant de débiter une étude on doit choisir la valeur de ces deux risques.

Dans un essai thérapeutique, le risque alpha est celui de conclure à l'efficacité du traitement alors qu'en réalité il n'est pas efficace, et le risque beta celui de conclure à l'inefficacité du traitement alors qu'il est efficace. En général on choisit une taille d'échantillon suffisante pour que ce risque de conclure à un tort à l'efficacité du traitement soit au maximum de 5% ($\alpha=0.05$). S'il s'agit d'un

traitement lourd avec des effets secondaires importants on peut même réduire ce risque, par exemple à 1 %.

Interprétation des risques alpha et bêta

		Existence d'une vraie différence entre les groupes (inconnue)	
		Présente	Absente
Conclusion des tests statistiques (connue)	Différence observée entre les groupes	Correct	Erreur de type I (risque α)
	Pas de différence observée entre les groupes	Erreur de type II (risque β)	Correct

REMARQUE : le raisonnement pour le calcul de la taille d'échantillon est identique pour les études étiologiques dans lesquelles les groupes comparés ne sont pas différents en termes de traitement mais en termes d'exposition à un facteur de risque (étude de cohorte) ou d'existence ou non d'une maladie (étude cas-témoins).

LES ANALYSES STATISTIQUES (GÉNÉRALITÉS)

Si l'on est dans le cadre d'un essai randomisé (comparaison de deux ou plusieurs groupes de traitements) ou dans celui d'une étude étiologique, l'analyse statistique principale consiste en général à **comparer des groupes**.

Dans les essais randomisés

Exemple : Dans un essai thérapeutique, 10 000 femmes sont réparties de façon aléatoire dans 2 groupes, 5 000 femmes sous placebo et 5 000 femmes sous bisphosphonate censé réduire le risque de fracture ostéoporotique (c'est le critère de jugement). On va donc comparer le taux de nouvelles fractures entre les deux groupes. Si, au cours des 2 ans de suivi, 3 % des femmes sous placebo ont une nouvelle fracture contre seulement 1 % dans le groupe bisphosphonates.

La différence entre les groupes traduit l'efficacité du traitement. Cette réduction du risque peut être quantifiée au moyen de différents indicateurs appelés « mesures d'association ».

Mesures d'association

- **Différence absolue ou, plus fréquemment, différence relative.**

Dans cet exemple la **différence relative est de 66%** $(0,03 - 0,01)/0,03$ et la **différence absolue est de 2%** $(0,03 - 0,01)$.

REMARQUE : on voit que la différence relative donne une impression de très grande réduction du risque même lorsque l'impact en termes de nombre absolu de fractures évitées est très faible. **ATTENTION :** quand une réduction de 66% du risque est annoncée, vous devez être capable de distinguer si on parle de différence relative ou absolue.

- **Rapports de risques** (au lieu des différences de risques).

Le rapport du risque de fracture du groupe bisphosphonates (1%) sur le risque de fracture du groupe placebo (3%) représente le **risque relatif** de nouvelle fracture lié à la prise de bisphosphonates vs placebo.

$$RR = \frac{R \text{ fracture}_{\text{bisphosphonate}}}{R \text{ fracture}_{\text{placebo}}} = \frac{1\%}{3\%}$$

En général au numérateur figure le risque du groupe traitement évalué (ou nouveau traitement) et au dénominateur le risque du groupe placebo (ou traitement de référence). **Si le traitement est efficace le risque relatif varie entre 0 et 1.** Plus le risque relatif est proche de 1, moins l'effet du traitement est important (un rapport des risques = à 1 signifie que le risque du groupe traitement (numérateur) est identique au risque du groupe placebo (dénominateur)).

Dans cet exemple **le risque relatif est de 0,33**, le risque sera donc réduit de $1 - 0,33 = 0,67$ soit 67%.

- **Nombre de sujets à traiter** pour éviter un événement (NNT = number needed to treat).

Dans notre exemple la différence absolue de nouvelles fractures entre les deux groupes est de 2%, ce qui signifie qu'il faut traiter par bisphosphonates 100

patients pour éviter 2 fractures, donc le NNT est de 50 (pour éviter une fracture il faut traiter 50 patients).

On peut aussi comparer des temps, par exemple comparer la durée de suivi sans nouvelle fracture.

Dans cet exemple le critère de jugement est mesuré par une **variable binaire ou dichotomique : présence / absence**. Ce sont donc des pourcentages de patients que l'on compare. Si en revanche le critère de jugement est estimé par une **variable quantitative continue** comme le poids ou la tension artérielle, l'effet du traitement pourra être exprimé sous la forme d'une **différence de moyenne** entre les deux groupes (par exemple, tension artérielle systolique moyenne de 17,2 +/- 3,2 dans le groupe placebo est de 15,6 +/- 2,6 dans le groupe traité par antihypertenseur).

Dans les études étiologiques de type cohorte visant à mesurer le lien entre facteur de risque et survenue d'une maladie, le groupe exposé au facteur de risque est comparé au groupe non exposé. L'effet du facteur de risque sur la survenue de la maladie est souvent estimé par le **risque relatif (RR)**.

Exemple : le risque de mort subite dans le groupe des nourrissons qui dorment sur le dos est de 0,1 pour 1000 par an, ce risque est de 0,3 pour 1 000 par an chez les nourrissons qui dorment sur le ventre soit un risque relatif $0,003/0,001 = 3$.

En général le risque de maladie chez les exposés figure au numérateur et le risque chez les non exposés au dénominateur, ainsi si le **facteur étudié augmente réellement le risque** de la maladie, le **risque relatif** lié à ce facteur est **supérieur à 1**. Parfois le **facteur étudié est protecteur** vis-à-vis du risque, dans ce cas le **RR est compris entre 0 et 1**. (Par exemple, l'utilisation d'une tétine s'accompagne d'une réduction du risque de mort subite du nourrisson).

$$RR_{\text{cohorte}} = \frac{R_{\text{Malade}}_{\text{exposés}}}{R_{\text{Malade}}_{\text{non exposés}}}$$

RR et ICC 95 % > 1 le facteur étudié augmente le R de la maladie.

0 < RR et ICC 95 % < 1 le facteur étudié est protecteur (beaucoup plus rare).

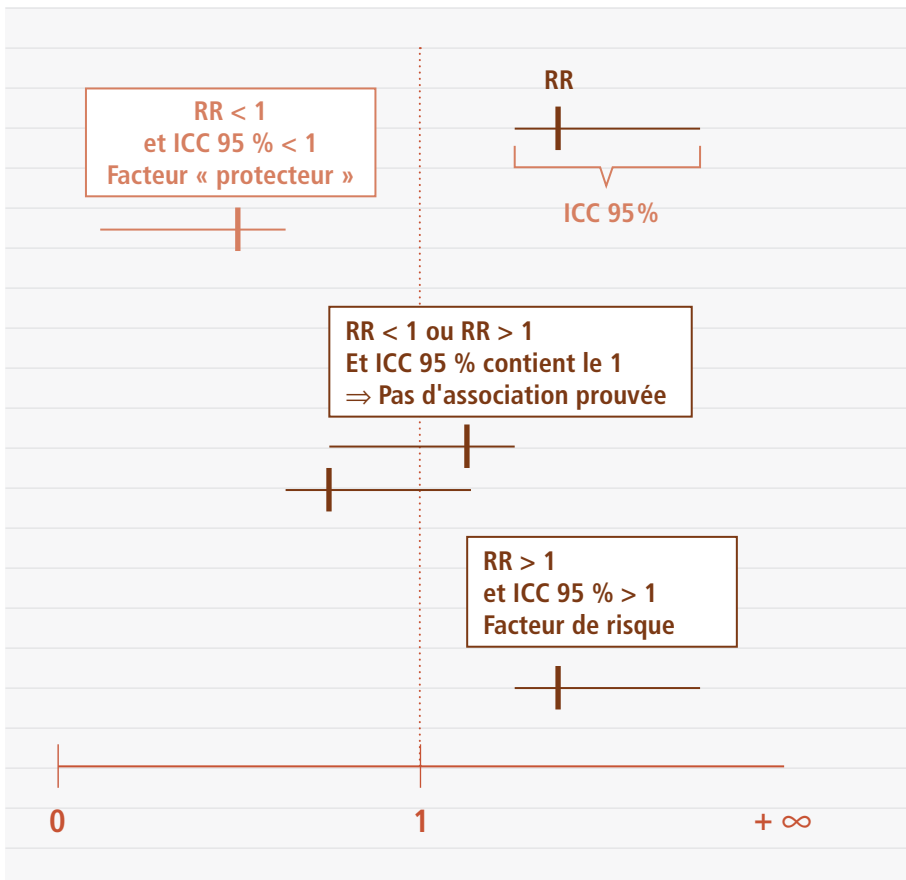


Figure inspirée du document de J Labarère (faculté de médecine de Grenoble).

Dans les études étiologiques de type cas-témoins on ne peut pas mesurer de RR. En effet, les patients sont sélectionnés selon leur statut malade ou non malade, cela n'a donc pas de sens de calculer le risque de maladie dans chaque groupe : il dépend du choix de l'investigateur ! Dans ce type d'étude, c'est l'Odds Ratio (OR) qui est estimé (voir chapitre cas-témoins). Sous certaines conditions, il constitue une bonne approximation du RR.

Les méthodes d'analyses statistiques font l'objet d'un chapitre spécifique.

4.7 LES RÉSULTATS (GÉNÉRALITÉS)

Dans l'interprétation des résultats se posent **deux questions essentielles** :

1. L'effet a-t-il une probabilité suffisamment basse (traditionnellement < 5 %) d'être dû uniquement au hasard ? Autrement dit, l'effet est-il « statistiquement significatif » ?

Cette question est en lien avec le petit p (p-value) et/ou l'intervalle de confiance à 95%.

Exemple : on observe dans notre essai précédent que dans le groupe des femmes traitées par bisphosphonates, le risque de fracture est réduit par rapport aux femmes sous placebo avec un risque relatif de 0,33 et un intervalle de confiance à 95 % de [0,25 – 0,38] avec $p < 0,01$.

$p < 0,01$ signifie que l'on a 99 % de chances que l'effet observé ne soit pas dû au hasard. Autrement dit, si l'on pouvait répéter l'étude à l'infini, le hasard ne ferait apparaître ce résultat que dans moins d'une étude sur 100.

L'IC 95 % signifie que l'on a 95% de chances que la valeur du risque relatif réel, c'est-à-dire présent dans la population, se trouve entre 0,25 - 0,38. Cela signifie que l'on aurait moins de 5% de chances d'observer un risque relatif en dehors de cet intervalle si l'expérience était répétée 100 fois.

Ces deux notions sont donc très proches, l'intervalle de confiance apportant des informations plus complètes que le « p ».

Mais attention, ni l'IC 95% ni le « p » ne renseigne sur l'importance de l'effet.

2. L'effet est-il important ? On parle de l'ampleur / de la taille de l'effet pour l'efficacité d'un traitement ou de la force de l'association entre facteur de risque et maladie.

Dans l'essai sur les bisphosphonates, l'effet est statistiquement significatif (donc non dû au hasard) car la taille de l'échantillon étudié est très importante, ce qui permet de mettre en évidence même une petite différence.

En revanche la différence de risque est-elle « cliniquement importante » ???

Cela dépend de la fréquence de la maladie :

Dans l'exemple précédent :

- le traitement permet une réduction absolue de $\approx 2\%$ (66 % de 3 %), soit 2 fractures évitées pour 100 patients traités soit $100 / 2 = 50$ patientes à traiter pour éviter une fracture.

Si cet essai était mené dans une population dont le risque de base est de 1%, par exemple chez des femmes plus jeunes, sous réserve d'une **efficacité identique du traitement**, on observerait :

- une réduction de 66 % de 1% soit une réduction absolue du risque de 0,66 %,
- donc dans la population traitée le risque absolu de nouvelle fracture passerait de 1% à 0,33%, il faudrait donc traiter 100 patientes pour éviter 0,66 fractures soit $100 / 0,66 = 152$ patientes à traiter pour éviter 1 fracture.

Une question possible en LCA est : **L'effet est-il cliniquement significatif ?**

Cette question repose premièrement sur l'appréciation de l'**ampleur d'effet** comme nous venons de l'évoquer mais également sur une **appréciation clinique**. Elle est donc très subjective. Cette appréciation se fait sur la pertinence du critère de jugement.

Par exemple, un traitement qui permet de faire gagner en moyenne 3 m de périmètre de marche à des patients souffrant d'artérite des membres inférieurs, est-ce un résultat cliniquement significatif ? Une chimiothérapie de troisième ligne qui permet de gagner en moyenne 1 mois de survie supplémentaire, est-ce un résultat cliniquement significatif ? Pour juger cela il faut également prendre en compte la lourdeur des traitements (tolérances et effets secondaires) ainsi que leur coût.

Cette notion repose en partie sur un jugement médical qui nécessite une réflexion prenant en compte plusieurs éléments cliniques et épidémiologiques (voire économiques).

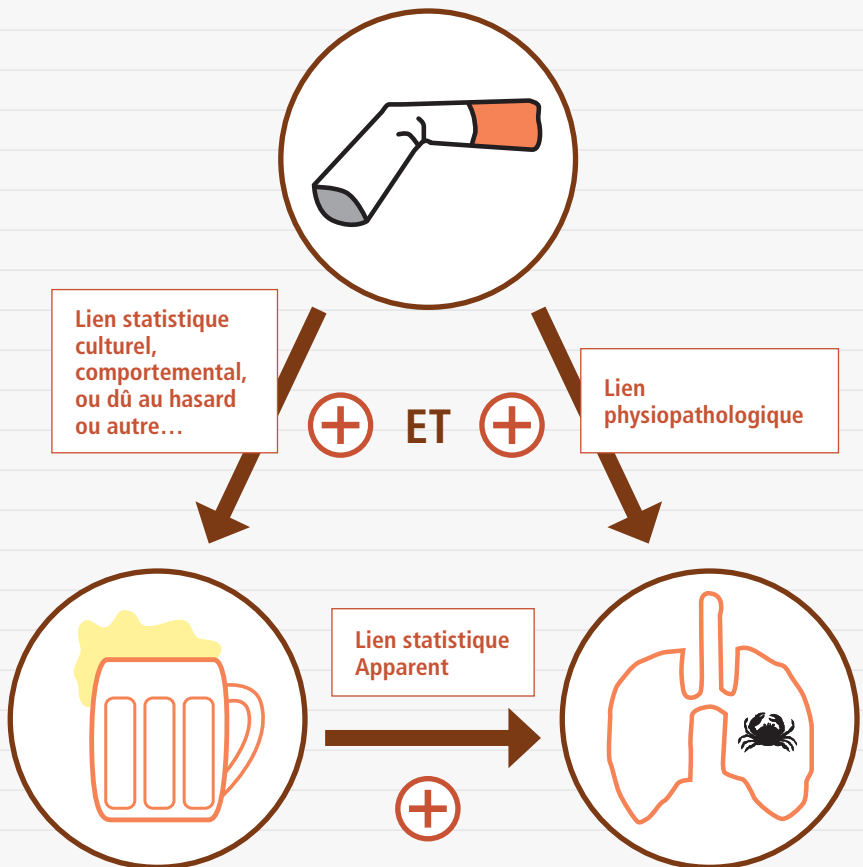
4.8 LES FACTEURS DE CONFUSION, LES BIAIS ET LES INTERACTIONS (GÉNÉRALITÉS)

LES FACTEURS DE CONFUSION

Un « facteur de confusion » est un **facteur qui peut introduire une confusion (erreur) dans la mesure de l'effet du facteur de risque sur le critère de jugement étudié.**

Si l'on s'intéresse par exemple à la relation entre consommation d'alcool et risque de cancer du poumon on va observer qu'il existe un lien statistique, les patients atteints de cancer du poumon ont tendance à avoir une consommation d'alcool supérieure à ceux qui n'ont pas de cancer du poumon. En réalité, ce lien n'existe pas, la consommation d'alcool n'a pas de lien direct avec le risque de cancer du poumon. La relation statistique apparente est un artéfact.

Exemple : Confusion liée à l'absence de prise en compte du tabagisme lorsqu'on étudie le lien entre alcool et cancer du poumon.



En réalité si les personnes qui consomment plus d'alcool ont un risque plus élevé de cancer du poumon ce n'est pas parce que la prise d'alcool augmente le risque de cancer du poumon mais parce que la prise d'alcool est souvent associée à une consommation tabagique qui, elle, augmente le risque de cancer du poumon.

On voit donc qu'un **facteur de confusion est à la fois lié au critère de jugement** (donc potentiellement tous les facteurs de risque réels de la maladie étudiée peuvent être des facteurs de confusion) et **au facteur de risque étudié** (ce lien peut être dû au hasard ou à d'autres raisons comme par exemple le style de vie ou le comportement...).

Un facteur (âge, sexe, tabagisme, consommation d'alcool, vie sociale, dépression...) n'est pas un « facteur de confusion » par nature, mais seulement vis-à-vis de l'association qu'on cherche à mettre en évidence : c'est une histoire à 3 !, un facteur apporte de la confusion dans l'estimation de la relation entre deux autres facteurs : le facteur de risque potentiel et la maladie impliqués dans l'hypothèse de recherche testée dans l'étude.

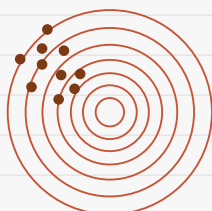
On peut annuler l'effet de confusion de plusieurs façons, notamment en « ajustant » les analyses statistiques pour les facteurs de confusion. Dans l'exemple ci-dessus, si l'on prend en compte le tabac dans l'analyse statistique alors le lien apparent entre alcool et cancer disparaît.

Donc tout facteur de risque de la maladie étudiée peut a priori se comporter comme un facteur de confusion s'il est aussi lié au facteur de risque étudié (par hasard ou parce qu'il existe un lien de style de vie par exemple les consommateurs de tabac sont en moyenne de plus grands consommateurs d'alcool, ou les femmes ont en moyenne moins de comportements à risque, etc.).

LES BIAIS (GÉNÉRALITÉS)

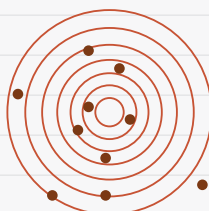
Les biais sont des erreurs qui vont systématiquement dans le même sens, qu'il faut distinguer de l'erreur aléatoire qui est dispersée en fonction du hasard.

Exemple de trois tireurs qui visent **le centre d'une cible**.



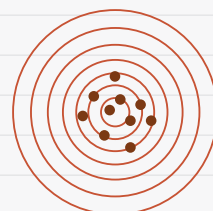
Tireur 1

Tirs **biaisés** par rapport au centre de la cible



Tireur 2

Tirs très dispersés de façon assez homogène autour du centre : forte **erreur aléatoire** mais pas forcément de biais



Tireur 3

Tirs très groupés, **faible erreur aléatoire**, pas de biais

Vous trouverez dans la littérature de nombreux noms de biais spécifiques.

Ce qu'il faut retenir à ce stade est qu'ils peuvent être classés dans deux grandes familles :

- **Biais de mesure** : biais qui concerne la mesure des facteurs de risque ou du critère de jugement),
- **Biais de sélection** : biais qui concerne la sélection des sujets étudiés, tant au début de l'étude qu'à la fin du suivi). Le biais de sélection peut correspondre à un défaut de représentativité, mais aussi à la non comparabilité des groupes étudiés liée à la sélection. Il faut noter que certains considèrent que le défaut de représentativité n'est pas un biais de sélection, mais un problème à part.

Un troisième type de biais est parfois évoqué : le **biais de confusion** qui est simplement la non prise en compte d'un facteur de confusion qui entraîne un biais au niveau de l'estimation du lien entre facteur de risque et maladie (il fait donc parti de la famille des biais de mesure).

(NIVEAU 2)

LES INTERACTIONS

À côté du phénomène de « confusion » que l'on peut observer, dans certaines conditions certains facteurs peuvent créer des phénomènes « **d'interaction** », ils sont alors appelés facteurs « **modificateurs d'effet** ».

Reprenons un exemple proche du précédent des relations entre tabac, alcool et cancer de l'oropharynx. Cette situation est différente car l'alcool augmenterait réellement le risque de ce cancer mais la valeur de cette augmentation est influencée par un troisième facteur. La consommation d'alcool augmente le risque de cancer de l'oropharynx mais l'importance de cet effet varie avec la consommation de tabac. **Le tabac est donc un modificateur de l'effet de l'alcool vis-à-vis du risque de cancer de l'oropharynx.** Autrement dit il y a une interaction entre le tabac et l'alcool vis-à-vis du risque de cancer de l'oropharynx.

Par exemple, si le tabac multiplie le risque de cancer de l'oropharynx lié au tabac par 1,3 (risque relatif de 1,3 (IC 95 % 1,1-1,4)) chez les patients qui ne consomment pas d'alcool et par 4,2 (risque relatif de 4,2 (IC 95 % 3,9-4,7)) chez les patients qui consomment de l'alcool (ces chiffres sont fictifs !) on voit bien

que l'alcool modifie l'effet du tabac sur le cancer de l'oropharynx. Le plus souvent, cette modification va dans le sens d'une augmentation de l'effet (il s'agit d'une synergie entre les deux facteurs de risque qui se potentialisent). Parfois cette modification va dans le sens d'une réduction de l'effet (il s'agit alors d'un effet antagoniste).

4.9 LA DISCUSSION - LA CONCLUSION

La discussion reprend en général les résultats de l'étude et les met en perspective avec les résultats d'autres études publiées. Elle discute également des limites et des forces de l'étude.

ATTENTION à être vigilant sur la discussion. Les auteurs ont souvent tendance à défendre leur point de vue même si les résultats ne sont pas en faveur de leur hypothèse de départ. En revanche la discussion peut apporter des éléments essentiels pour vous sous la forme d'informations issues d'autres études mais qui doivent **IMPERATIVEMENT** correspondre à des références bibliographiques d'études publiées.

La conclusion doit porter sur les résultats de l'étude. Elle doit répondre aux objectifs annoncés et en priorité à l'objectif principal.

REMARQUE : les auteurs ont souvent tendance à conclure sur des éléments qui ne sont pas directement issus de l'article.

Par exemple, si une étude de cohorte, donc observationnelle, a démontré qu'il y avait une association entre un faible taux de vitamine D sérique et l'augmentation du risque de maladie cardiovasculaire la conclusion EST : « un faible taux sérique de vitamine D est associé à une augmentation du risque de maladie cardiovasculaire » Mais la conclusion **NE PEUT PAS ÊTRE** : il faut donner de la vitamine D aux personnes à haut risque cardiovasculaire. En effet pour répondre à cette question il faudrait faire, chez des patients à haut risque cardiovasculaire, un essai randomisé pour comparer un groupe sous placebo à un groupe sous traitement par vitamine D.

LES MESSAGES CLEFS

Vous devez COMMENCER votre lecture en cherchant les éléments qui vous permettent d'identifier l'objectif principal (ou la question de recherche) et ainsi de classer l'article dans une des grandes catégories de question de recherche.

- La lecture critique sert à apprécier en premier lieu la « validité interne » de l'étude : quelles garanties sont apportées par les auteurs que les résultats sont réels et ne sont pas uniquement liés à des biais.
- Cette validité passe par la vérification d'un certain nombre de points, on utilise une check list pour ne pas en oublier.

5

LCA DES ESSAIS CLINIQUES, ÉVALUATION D'UNE INTERVENTION THÉRAPEUTIQUE

La majorité des essais sont destinés à répondre à la question : le traitement évalué est-il plus efficace que le placebo ou le traitement de référence ? **Ce sont des essais de supériorité.**

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

La prévalence de la consommation de tabac chez les patients schizophrènes est beaucoup plus élevée (70-80 %) que dans de la population générale (20 %). Le niveau de dépendance au tabac des patients schizophrènes est également plus élevé, ils ont plus de difficultés à arrêter de fumer même avec les substituts nicotiques ou le Bupropion qui ne semble pas efficace chez ces patients. Un nouveau traitement, le Varenicline semble beaucoup plus efficace que le Bupropion. En population générale, son efficacité s'est avérée 3 fois plus supérieure à celle du Bupropion dans les essais cliniques. Néanmoins, ces essais ont montré la possibilité d'effets secondaires neuropsychologiques chez certaines personnes traitées par Varenicline.

L'objectif principal de cette étude était d'évaluer l'efficacité du Varenicline versus placebo sur l'arrêt du tabac entre l'inclusion et 24 semaines chez les patients schizophrènes souhaitant arrêter de fumer. Les objectifs secondaires étaient d'évaluer la tolérance du Varenicline et les effets secondaires psychiatriques.

Un essai contrôlé randomisé en double insu versus placebo a été mené pour répondre à l'hypothèse : l'efficacité du Varenicline est supérieure à celle du placebo pour aider les patients schizophrènes à arrêter de fumer.

5.1 LA QUESTION

La question de recherche principale est donc : l'efficacité du Varenicline est-elle supérieure à celle du placebo pour aider les patients schizophrènes à arrêter de fumer ? La réponse à cette question permettra d'évaluer le bénéfice du traitement.

La question secondaire sur les effets indésirables est également très importante et devra être prise en considération dans un deuxième temps avant de pouvoir apprécier le rapport bénéfice-risque.

L'objectif principal correspondant comporte les éléments suivants :

P : Dans quelle population ?	Patients schizophrènes souhaitant arrêter de fumer
I : Quel est le traitement étudié (ou l'intervention) ?	Varenicline
C : A quelle référence est-il comparé ? (Autre traitement, placebo)	Placebo
O : Quel est le bénéfice attendu ?, Sur quel critère, quelle amplitude, à quel délai ?	Arrêt consommation de tabac à 24 semaines

Les études interventionnelles sont souvent des essais thérapeutiques visant à évaluer l'efficacité d'un nouveau traitement : médicament, intervention chirurgicale. Mais il peut s'agir d'autres types d'intervention, par exemple des interventions de prévention (dépistage, vaccination, éducation thérapeutique du patient), de formation des professionnels de santé ou encore d'introduction de nouvelles techniques ou organisations des soins.

Exemple d'intervention non médicamenteuse étudiée :

Leslie et al, BCM PH 2012 : un essai contrôlé randomisé a été mené pour évaluer l'efficacité d'un programme structuré de suivi diététique par rapport à un programme standard sur la prise de poids chez les personnes venant d'arrêter de fumer. Le groupe intervention recevait un programme complet de conseil et prise en charge diététique sur 24 semaines, le groupe contrôle recevait uniquement un programme de soutien sur 7 semaines.

5.2 LE TYPE D'ÉTUDE

L'essai contrôlé randomisé est le seul type d'étude adapté **lorsqu'on veut mesurer l'efficacité d'une intervention**, car il **permet de limiter les biais** et apporte ainsi le **niveau de preuve scientifique le plus élevé**.

Il s'agit en général d'un essai de phase 3 dont les résultats peuvent conduire à l'autorisation de mise sur le marché (AMM).

Autorisation de mise sur le marché

	Phase préclinique	Phase I	Phase II	Phase III	Phase IV
Population	Animal	Petit nombre de volontaires sains	Petit nombre de malades	Grand nombre de malades	Population générale
Population	Toxicologie Pharmaco-cinétique	Tolérance (dose maximale tolérée) Pharmaco-cinétique Calcul des doses pour la phase II	Efficacité pharmacologique Calcul des doses pour la phase III	Efficacité clinique	Pharmacovigilance Nouvelle indication

Rappel des différentes phases d'un essai portant sur le médicament.

(tiré de l'abrégé Masson Santé Publique)

Le terme contrôlé signifie qu'il y a un groupe contrôle (recevant par exemple un placebo).

Le plus souvent, l'essai contrôlé randomisé est un essai à 2 « bras » (2 groupes) parallèles :

- Le groupe intervention est le groupe recevant le traitement (ou l'intervention) évalué,
- Le groupe contrôle est le groupe qui va servir à comparer l'efficacité du traitement évalué : ce groupe reçoit soit un placebo, soit le traitement de référence (le meilleur traitement reconnu ou recommandé).

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

▀ Ici le groupe contrôle reçoit un placebo, on peut se poser la question de tester le nouveau traitement contre des substituts nicotiniques ou contre le Bupropion. Néanmoins, ceux-ci n'ont pas fait la preuve de leur efficacité dans la population de patients schizophrènes : on ne peut donc pas considérer qu'ils représentent un traitement de référence.

Pourquoi un groupe contrôle est nécessaire ?

Pourquoi ne pas simplement analyser un groupe de patients avant puis après la prise du médicament pour analyser l'efficacité de celui-ci ?

La présence d'un groupe contrôle **permet de s'assurer que l'évolution observée est liée à l'intervention testée** et non à d'autres facteurs tels que **1. l'évolution naturelle de la maladie** ou **2. l'effet placebo** ou **3. le phénomène de**

régression à la moyenne ou **4. l'effet de traitements pris de façon concomitante.**

1. L'évolution spontanée de la maladie peut ainsi être confondue avec l'effet du traitement, c'est donc un facteur de confusion potentiel qu'il faut éliminer grâce au groupe contrôle chez lequel on peut s'attendre à la même proportion de patients guéris spontanément et ainsi c'est la différence entre les deux groupes qui montrera l'efficacité réelle du traitement.

Par exemple : L'administration d'un nouveau traitement antiviral à un patient souffrant d'un rhume banal est suivie de la guérison complète du patient, est-ce une preuve de l'efficacité du traitement ?? **NON.** La guérison après l'administration du traitement ne peut pas être attribuée avec certitude au traitement car elle serait peut-être survenue spontanément.

2. L'effet placebo a été prouvé maintes fois scientifiquement et peut être expliqué par des effets d'auto et d'hétéro suggestion. Par exemple, des patients ne sachant pas s'ils reçoivent un placebo ou un anxiolytique avant une épreuve stressante sont moins anxieux s'ils pensent prendre un véritable anxiolytique même s'ils ont reçu sans le savoir un placebo. À l'inverse il existe l'effet nocebo : si ces patients ont connaissance de certains effets secondaires des anxiolytiques, ils peuvent ressentir ces effets secondaires alors même qu'ils prennent un placebo. **L'effet placebo peut être très important dans certaines pathologies. Par exemple dans les groupes placebo de 45 essais randomisés portant sur le colon irritable, une amélioration globale des symptômes a été observée chez 40 % des patients.**

3. La régression à la moyenne (« regression toward the mean ») est un phénomène purement statistique qui survient lorsque l'on s'intéresse à un groupe de sujets sélectionnés parce qu'ils présentent des valeurs d'un critère d'inclusion, supérieures ou inférieures à un seuil. Par exemple, un groupe de sujets hypertendus sélectionnés pour un essai si leur pression artérielle diastolique (PAD) est $>$ à 90 mmHg. **MAIS, la mesure** de la PAD, comme celle de bien d'autres phénomènes d'ailleurs, est soumise à **1. une variabilité intra-individuelle** et **2. des erreurs de mesure.**

Variabilité intra-individuelle : Chez un même sujet, la pression artérielle est variable d'un moment à l'autre sous l'effet de nombreux paramètres physiologiques, cependant ces valeurs oscillent autour d'une valeur moyenne caractéristique du sujet : sa vraie valeur de PAD. Comme on a sélectionné seulement les individus avec une valeur haute de PAD, dans de nombreux cas, cette valeur était plus haute que d'habitude et sera plus basse la prochaine fois.

Les erreurs de mesure (s'il s'agit bien d'une erreur aléatoire et non d'un biais) sont également réparties au-dessus ou en dessous de la vraie valeur de façon aléatoire. L'investigateur a pu surestimer la vraie valeur de la PAD du fait d'une erreur de mesure et lors de la prochaine mesure ces patients auront une mesure probablement plus proche de leur vraie valeur, donc inférieure au seuil. Ces valeurs vont tirer la moyenne du groupe vers le bas.

Ainsi, ce **phénomène de régression à la moyenne**, lié uniquement à des sujets sélectionnés à tort dans le groupe du fait de la variabilité des mesures, va entraîner une diminution de la moyenne du groupe sans aucune modification de la vraie valeur des sujets.

4. L'effet de traitement pris de façon concomitante.

Imaginons par exemple un essai sur un nouveau traitement antalgique. Si ce nouveau traitement est réellement plus efficace que le traitement du groupe contrôle, les patients du groupe contrôle ont tendance à augmenter leurs prises d'autres antalgiques et ainsi la différence entre les groupes sera moins importante. Les traitements concomitants qui peuvent interférer avec le traitement évalué doivent donc être recueillis et enregistrés.

LA RANDOMISATION

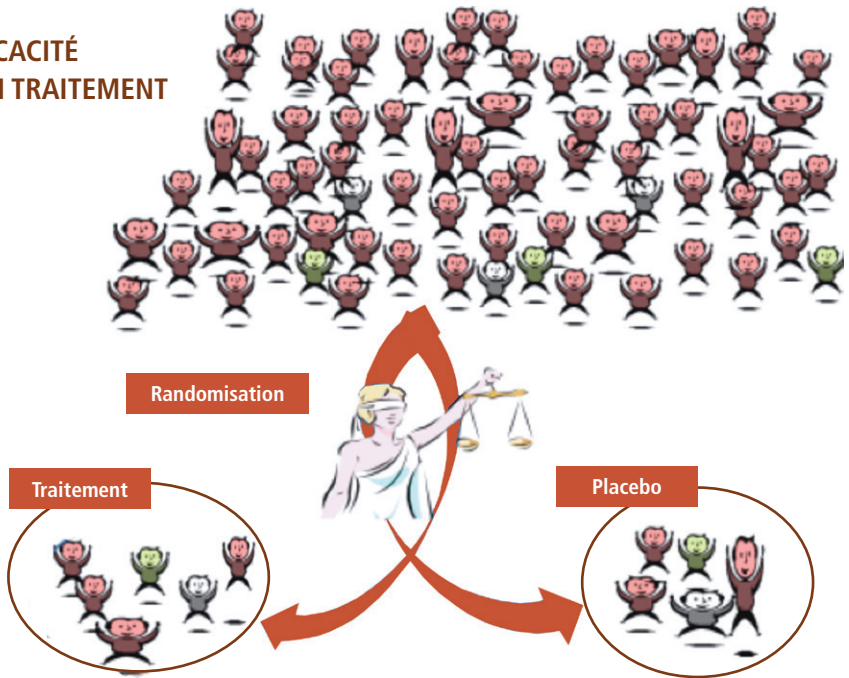
(= ALLOCATION ALÉATOIRE DES PATIENTS DANS CHAQUE GROUPE)

Il s'agit de la méthode permettant de **tirer au sort le groupe** dans lequel va se trouver chaque patient inclus dans l'étude. Cela signifie que les groupes ne sont pas constitués selon des critères cliniques ou de gravité mais uniquement par **le hasard (allocation aléatoire)**.

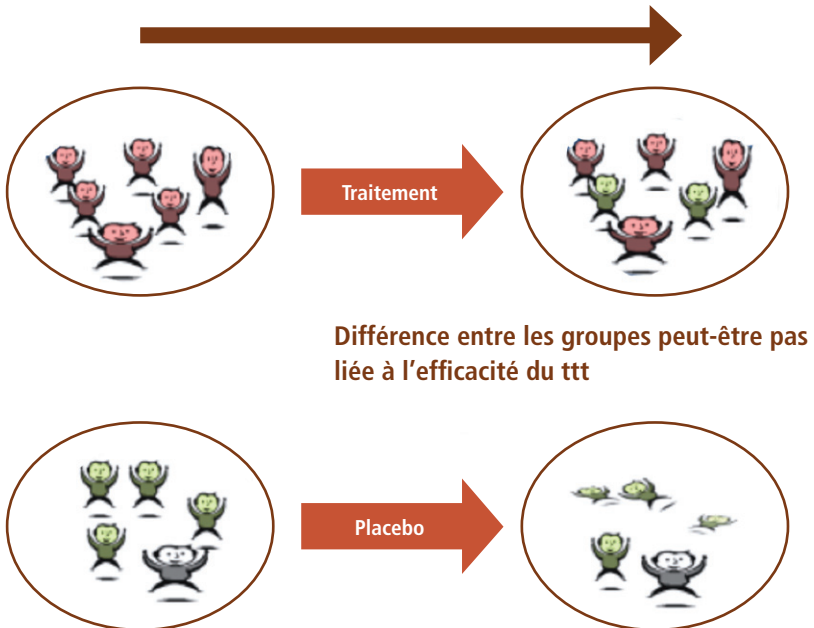
Une **randomisation bien faite** assure la **comparabilité initiale** des groupes, pour tous les **facteurs pronostiques connus et inconnus**, et **garantit l'absence de biais de sélection différentiel**.

Si on observe une différence entre les groupes à la fin de l'essai, c'est-à-dire si les patients sous traitement vont mieux que les patients sous placebo, cela ne peut pas être attribué au fait que les patients du groupe traitement étaient dès le départ en meilleure santé que les autres.

EFFICACITÉ D'UN TRAITEMENT



ABSENCE DE RANDOMISATION



Comment s'assurer que la randomisation a été bien menée ?

La qualité de la randomisation est un élément clef pour évaluer la validité interne d'un essai randomisé. Elle repose sur 3 points :

- **La génération de la séquence de randomisation,**
- **L'assignation secrète (la clause d'ignorance),**
- **La vérification de la comparabilité initiale des groupes.**

Si l'un de ces 3 éléments pose problème, il y a un risque de biais de sélection.

1) La génération de la séquence de randomisation

La méthode de randomisation doit permettre une allocation parfaitement **aléatoire** :

- table de nombres au hasard,
- séquence informatique sur ordinateur.

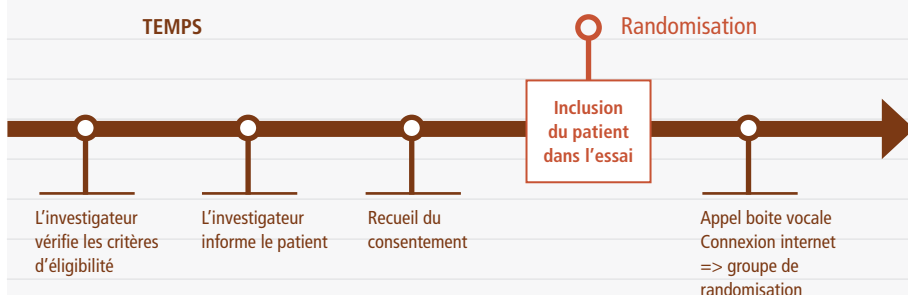
2) assignation secrète (synonymes : **allocation concealment, clause d'ignorance** (terme du glossaire du CNCI), non divulgation de l'allocation, masquage de l'allocation, **imprévisibilité**)

Il ne doit pas être possible pour l'investigateur qui inclut le patient de savoir à l'avance dans quel groupe ce dernier va être inclus. Le fait de pouvoir prévoir le groupe de randomisation du prochain patient risque d'influencer la décision d'inclure le patient dans l'essai.

Par exemple, si l'investigateur est persuadé de l'efficacité du nouveau traitement, il mettra plus volontiers les sujets jeunes et de bon pronostic dans le groupe traitement pour leur éviter une perte de chance. A l'inverse s'il pense que le traitement n'est pas efficace voire risqué, il peut décider de ne pas proposer au patient de rentrer dans l'essai s'il sait que le prochain patient de l'étude sera randomisé dans le groupe traitement. Au final, on ne pourra pas distinguer un effet réel du traitement d'un effet apparent lié à une différence de pronostic des groupes au départ.

La meilleure méthode pour assurer l'imprévisibilité (la clause d'ignorance) est la randomisation centralisée. Pour chaque patient éligible, le médecin investigateur contacte le centre de randomisation et reçoit en retour le groupe dans lequel le patient est alloué (par téléphone, fax ou via un site internet). L'utilisation d'enveloppes **N'EST PAS OPTIMALE** car elle ne garantit pas l'absence de « triche » (l'investigateur peut ouvrir plusieurs enveloppes jusqu'à avoir le groupe qu'il veut pour un patient donné...).

Chronologie de l'inclusion d'un patient pour respecter la clause d'ignorance :



Quelques exemples de ce que n'est pas la randomisation : choix du groupe selon la date de naissance, un patient sur 2 dans chaque groupe, les patients consultant les jours pairs dans le groupe intervention et les patients consultant les jours impairs dans le groupe contrôle... Le médecin investigateur connaît le groupe dans lequel va être inclus son patient (ce n'est plus imprévisible) et il va être tenté de mettre ses patients dans un groupe ou dans l'autre.

3) La vérification de la comparabilité initiale des groupes (voir aussi paragraphe 5.3 Comparabilité initiale)

Les caractéristiques initiales des patients par groupe de randomisation sont en général présentées dans le premier tableau de l'article. Toutes les caractéristiques importantes au regard de l'essai doivent être présentées (âge, sexe, sévérité de la maladie, antécédents, comorbidités, autres traitements pris par les patients...). Il permet de vérifier que les caractéristiques initiales des patients sont relativement comparables entre les 2 groupes. Cette évaluation se fait « à l'œil » (Eye Ball test) ce qui signifie sans faire de tests statistiques. Pour évaluer si les groupes sont comparables, il faut prendre en compte l'effectif de l'essai. En cas de faible effectif, on peut s'attendre à quelques déséquilibres entre les groupes liés aux fluctuations d'échantillonnage. En revanche, si l'effectif est élevé, les déséquilibres entre les groupes doivent être minimes. De plus, les déséquilibres liés aux fluctuations d'échantillonnage doivent aller dans les 2 sens (favorisant tantôt le bras expérimental tantôt le bras contrôle). Des déséquilibres favorisant toujours le même groupe doivent alerter et faire évoquer un risque de biais de sélection.

Remarques :

- L'absence de différence majeure entre les groupes randomisés est un indice de qualité mais n'est pas synonyme de randomisation bien faite, car les groupes peuvent différer sur des caractéristiques non renseignées dans l'essai.

- Une randomisation de bonne qualité peut ne pas aboutir à des groupes comparables, par le simple fait du hasard (fluctuations d'échantillonnage), surtout en cas de petit effectif.
- Quelle qu'en soit la raison (randomisation mal faite ou fluctuations d'échantillonnage), si les groupes diffèrent, on envisagera un ajustement statistique pour tenir compte des différences entre les groupes.

NIVEAU 2

Il existe plusieurs types de randomisation :

- **La Randomisation simple** est basée sur une simple séquence de nombres. Seul inconvénient : les 2 groupes peuvent être déséquilibrés surtout en cas de faible effectif. Par exemple, si on tire au sort le groupe de traitement pour 10 personnes, on peut obtenir 7 personnes dans le groupe intervention et 3 personnes dans le groupe contrôle.
- **La Randomisation par blocs** consiste à s'assurer qu'à tout moment de l'essai, le même nombre de patients est alloué dans chaque groupe. Par exemple, dans un essai comportant deux bras, une randomisation par blocs de taille 4 signifie que tous les 4 patients, 2 seront randomisés dans le groupe expérimental (A) et 2 dans le groupe contrôle (B). On aura donc au maximum une différence d'effectif de 2 entre les groupes à la fin de la randomisation. Un bloc de 6 signifie que tous les 6 patients on aura 3 patients dans le groupe intervention et 3 dans le groupe contrôle, avec un écart maximal de 3 entre les 2 groupes, etc.

Exemple de liste de randomisation par blocs de 4 :

1	A	9	B
2	A	10	B
3	B	11	A
4	B	12	A
5	A	13	B
6	B	14	A
7	A	15	B
8	B	16	A

Le risque de la randomisation par blocs est qu'un investigateur de l'étude qui coordonne l'essai puisse déduire quel va être le groupe dans lequel le patient suivant sera randomisé, si il connaît la taille des blocs (violation de la clause d'ignorance). Il est cependant possible de limiter ce risque en faisant varier la taille des blocs au cours de l'essai.

On peut également volontairement choisir de former des groupes d'effectif différent en modifiant le ratio d'allocation (**tttexp / ttref ou placebo**). **Le plus souvent, le ratio est 1:1**, c'est-à-dire **autant de patients randomisés dans le groupe expérimental et dans le groupe contrôle**, ce qui respecte le **principe d'équipoise** (ou **clause d'ambivalence**). Ce principe est lié au fait qu'on réalise l'essai car on ne sait pas quel est le traitement le plus efficace, sinon, ce ne serait pas éthique. **Dans certains essais, le ratio est 2:1**, c'est-à-dire qu'on va randomiser 2 fois plus de sujets traités par le nouveau médicament que de sujets traités par placebo. On parle alors de randomisation déséquilibrée. L'argument est d'augmenter le nombre de sujets traités par le traitement expérimental afin d'avoir davantage de données concernant la tolérance de ce traitement. Une autre raison moins fréquemment avouée dans les essais contre placebo est que cela va améliorer le recrutement dans l'essai car les patients ont plus de chances de recevoir le traitement expérimental que le placebo.

- La **Randomisation stratifiée** est utilisée pour limiter le risque de déséquilibre des facteurs pronostiques importants entre les groupes. Le principe est de s'assurer qu'un nombre égal de patients ayant certaines caractéristiques de mauvais pronostic (par exemple une maladie à un stade plus sévère que les autres, ou patients très âgés...) soit randomisé dans chaque groupe. La randomisation stratifiée consiste à faire **une liste de randomisation par strate**. Par exemple, si la randomisation est stratifiée sur le sexe et le stade de la maladie (précoce et avancé), il y a aura 4 listes de randomisation : 1 pour les femmes en stade précoce, 1 pour les femmes en stade avancé, 1 pour les hommes en stade précoce, 1 pour les hommes en stade avancé. Ce type de randomisation est intéressant pour prendre en compte des facteurs de confusion ou modificateurs de l'effet (influençant l'efficacité de l'intervention étudiée) car elle permet une répartition strictement identique de ces facteurs dans les deux groupes. Par exemple, si on stratifie sur l'âge 18-50 et >50, on aura une randomisation qui permettra d'avoir autant de patients

intervention et contrôle chez les 18-50 et chez les >50. Cette randomisation permet l'analyse en sous-groupe.

REMARQUE : il est fréquent dans les essais multicentriques de stratifier sur le centre.

- **La Randomisation par minimisation** est parfois utilisée, notamment quand les effectifs sont faibles et qu'il existe des facteurs pronostiques importants. Cette technique est fréquemment utilisée dans les essais en cancérologie. Cette méthode utilise un algorithme permettant de limiter les déséquilibres sur des facteurs pronostiques importants. Lors de l'inclusion d'un patient, l'investigateur rentre les facteurs pronostiques dans l'algorithme qui détermine alors le groupe du patient en fonction des caractéristiques renseignées afin de limiter les déséquilibres entre les groupes.

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

Les sujets ont été randomisés (2:1) entre le groupe Varenicline et placebo, et étaient stratifiés selon le traitement antipsychotique suivi (antipsychotiques typiques et antipsychotiques atypiques).

LE DOUBLE INSU (OU DOUBLE AVEUGLE, DOUBLE BLIND)

S'il connaît le traitement qu'il a reçu, le patient risque de modifier son comportement. Par exemple s'il est randomisé dans le groupe placebo, il pourra être déçu et quitter l'essai. À l'inverse s'il est randomisé dans le groupe nouveau traitement, il pourra redouter de présenter des effets secondaires. De la même manière, le médecin qui suit le patient va modifier son comportement (même de manière inconsciente) s'il sait quel traitement le patient a reçu. Par exemple, il pourra prescrire d'autres traitements s'il sait que le patient n'a pas reçu le traitement expérimental afin que celui-ci ne soit pas lésé. Toutes ces différences de comportement peuvent avoir un impact sur le critère de jugement et ainsi biaiser l'estimation de l'effet de l'intervention. En l'absence d'aveugle, il y a un risque de biais de suivi lié à des différences systématiques dans le suivi des patients.

Le double insu ou double aveugle garantit que ni le patient, ni le médecin ne connaît le groupe de randomisation du patient tout au long du suivi. Il permet de maintenir la comparabilité des groupes randomisés au cours du suivi.

Il existe d'autres niveaux d'insu :

- **Simple insu** : l'investigateur sait dans quel groupe le patient a été randomisé. Le patient ne sait pas s'il prend le traitement étudié ou la référence,
- **Triple insu** : l'investigateur et le patient ne connaissent pas le traitement pris et le statisticien effectuant les analyses ne sait pas non plus quels patients sont dans le groupe intervention ou contrôle.

Remarques : ces définitions peuvent être ambiguës. Par exemple, le terme double aveugle implique que 2 protagonistes de l'essai sont en aveugle mais on ne sait pas forcément lesquels (habituellement, il s'agit du patient et du médecin, mais... on ne sait jamais). C'est pourquoi il est préférable de préciser qui est en aveugle du traitement reçu dans l'essai : le patient ? le médecin qui suit le patient ? la personne qui évalue le critère de jugement qui peut être soit le patient (par exemple pour la douleur) soit le médecin (par exemple infarctus du myocarde), soit un évaluateur extérieur, le statisticien ?...)

Seul le double insu, comportant l'insu du patient et l'insu de l'investigateur qui évalue le critère de jugement, garantit l'absence de biais de mesure et de biais de suivi.

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

Par exemple, si le médecin sait que le patient prend le Varenicline et non le placebo, il peut inconsciemment induire chez le patient des réponses plus favorables lors de l'estimation de sa consommation de tabac et des difficultés de sevrage (le patient a envie de faire plaisir au médecin, le médecin a envie de croire que le traitement marche mieux...).

Pour qu'un essai soit en double aveugle, il faut que le **placebo (ou le traitement actif de référence) ait les mêmes caractéristiques (apparence, goût, forme) que le traitement évalué**, ce qui n'est pas toujours possible pour un traitement actif de référence.

Cas particulier : si le traitement de référence est d'apparence différente ou a un mode d'administration différent (par exemple comprimés alors que le nouveau traitement est injectable), il faut utiliser un double placebo pour que l'essai soit en double aveugle. Tous les patients vont recevoir deux traitements : soit le traitement évalué et le placebo du traitement de référence (groupe nouveau traitement), soit le placebo du traitement évalué et le traitement de référence (groupe contrôle).

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

On souhaite comparer 2 traitements pour arrêter de fumer : le Varénicline et les patchs, on ne peut avoir d'insu en comparant des comprimés à des patchs. Avec la technique du double placebo, le groupe 1 reçoit des comprimés de Varénicline et patchs de placebo, le groupe 2 reçoit des comprimés de placebo et patchs actifs. Ainsi l'insu pourra être garanti car les 2 groupes recevront des traitements indiscernables.

Chaque fois qu'on aurait pu faire une étude en double insu, notamment pour les traitements médicamenteux, avoir réalisé l'étude en simple insu ou en ouvert est inacceptable.

MAIS le double insu n'est pas toujours possible du fait même de la nature de l'intervention pour des raisons pratiques ou éthiques.

Exemples de situations pour lesquelles il est difficile voire impossible de réaliser un double-insu :

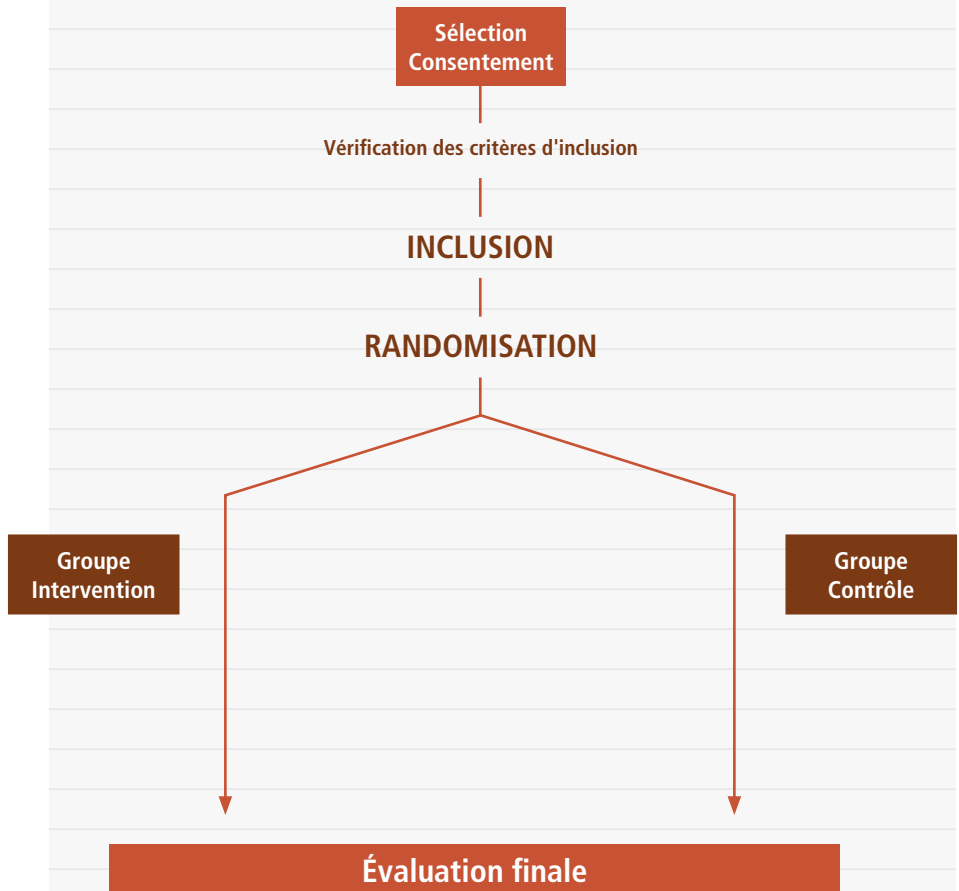
- le traitement est directement visible comme en chirurgie, kinésithérapie ou en radiothérapie où il est impossible de faire un simulacre, car le traitement factice risquerait d'avoir un effet néfaste (fausse chirurgie, faux massages...);
- les situations où il s'agit de comparer des stratégies de prise en charge (traitement à domicile versus traitement hospitalier);
- un des traitements comparés s'accompagne d'effets indésirables ou d'une toxicité évocatrice qui laisse deviner la nature du traitement dans presque tous les cas (chute de cheveux dans des chimiothérapies, augmentation de la diurèse pour les diurétiques...).

Dans ce cas, on parle **d'essai en ouvert**. Le **risque de biais de mesure et de suivi** est important : il faut vérifier que **le suivi des patients et l'évaluation des critères de jugement** sont effectués **de manière identique dans les deux groupes** (voir paragraphe 5.4 sur les biais de mesure).

La seule façon de limiter le risque de biais de mesure et de suivi est de faire en sorte que **l'évaluateur qui va mesurer le critère de jugement soit en insu**, il ne doit alors pas être le médecin investigateur qui a inclus et traité le patient.

LES DIFFÉRENTS SCHÉMAS D'ESSAIS RANDOMISÉS :

- **Essai en parallèle** : les 2 groupes sont définis initialement et reçoivent soit le traitement testé soit le traitement de référence. L'analyse porte sur la comparaison entre les 2 groupes qui ont chacun reçu un traitement différent.

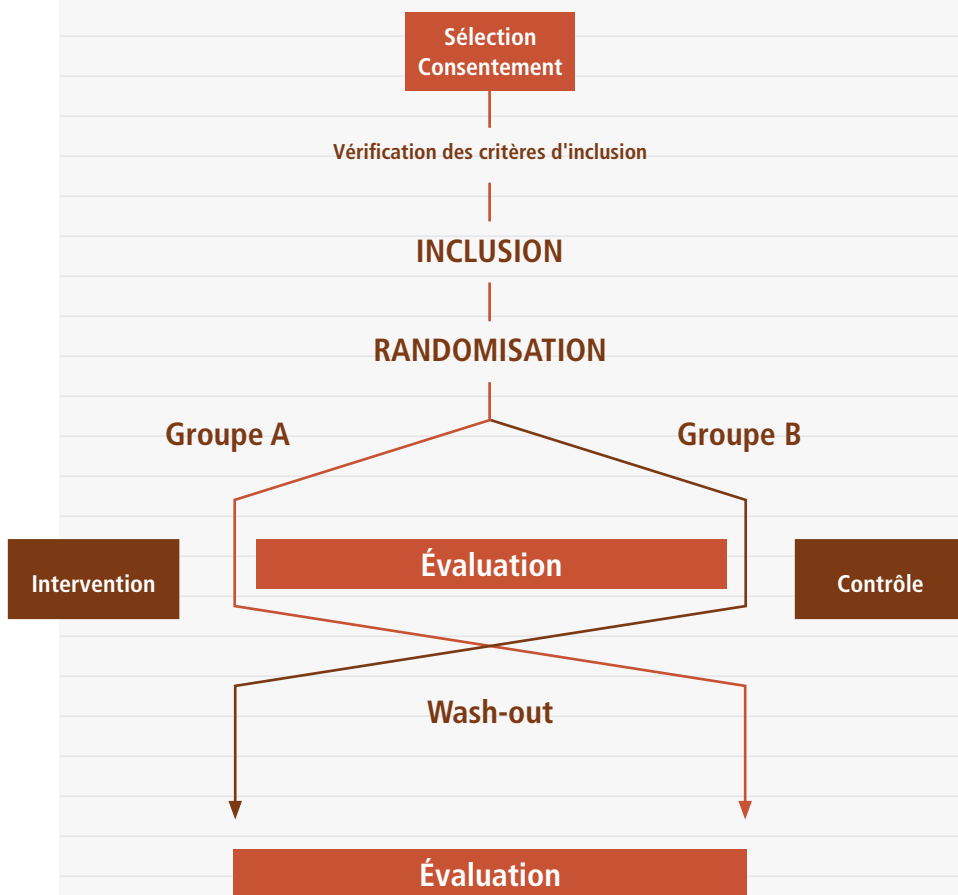


- **Essai en cross over ou essai croisé** : la randomisation forme deux groupes qui vont recevoir le traitement testé ou le traitement de référence dans une première partie de l'étude, puis chaque groupe reçoit l'autre traitement dans une 2^e partie de l'étude. À la fin de l'étude les deux groupes auront reçu successivement les 2 traitements. La randomisation porte sur l'ordre dans lequel le groupe va recevoir les traitements. La 1^e partie de l'étude permet de comparer les 2 traitements, la seconde de comparer la séquence de trai-

tements : traitement testé puis référence et référence puis traitement testé. Entre les 2 parties de l'étude on effectue une période de « wash-out » c'est-à-dire un temps sans traitement permettant d'éliminer l'effet du traitement préalablement reçu.

Ce format a l'avantage d'augmenter la puissance de l'étude. Cependant, il impose de respecter un certain nombre de contraintes :

- le critère de jugement doit pouvoir être mesuré plusieurs fois de façon indépendante,
- l'effet du traitement doit être réversible,
- la maladie ne doit pas évoluer entre les 2 périodes,
- les perdus de vue en 1^e période ne peuvent pas être comptabilisés,
- Il ne doit pas y avoir d'effet d'apprentissage,
- Il faut respecter une période de « wash-out » pour éviter que les effets des traitements se recourent.



5.3 LA POPULATION / L'ÉCHANTILLON ÉTUDIÉ

RAPPEL

(voir paragraphe 4.3 généralités)

La population source est définie par les critères d'inclusion et de non inclusion (section matériels et méthodes).

La population ou échantillon étudié :

- est décrite dans son ensemble au début du texte de la section résultats et/ou dans le tableau 1. Elle donne des informations sur l'extrapolabilité des résultats (validité externe) ;
- est également décrite dans le tableau 1 pour chaque groupe, ce qui apporte des informations sur la comparabilité des groupes (validité interne).

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

Les patients inclus étaient des patients non hospitalisés âgés de 18 à 75 ans présentant un diagnostic de schizophrénie confirmé par les critères du DSM IV révisé, fumant au moins 15 cigarettes par jour sans période d'abstinence de plus de 3 mois dans l'année précédente, désireux d'arrêter de fumer (mesuré par un résultat au score de contemplation (« Contemplation Ladder », validé) supérieur ou égal à 7, signifiant l'intention d'arrêter dans les 30 jours), cliniquement stables (pas d'hospitalisation ou exacerbation dans les 6 derniers mois) présentant un score à l'échelle PANSS inférieur à 70. Les femmes en âge de procréer devaient avoir un test de grossesse négatif et utiliser une méthode contraceptive fiable durant la période de l'étude.

EXTRAPOLABILITÉ (VALIDITÉ EXTERNE)

La population incluse et randomisée est définie par des critères d'inclusion et de non-inclusion strictement identiques pour tous, ces critères doivent être précisément décrits dans la partie méthodes.

Ces critères doivent nous permettre de bien cerner la condition de santé étudiée. Il est important qu'ils permettent d'avoir une population homogène MAIS sans être trop restrictifs afin d'être en mesure d'appliquer les résultats à l'ensemble des personnes atteintes de la pathologie (si on a étudié uniquement des patients

atteints d'un stade particulier de bon pronostic, on ne pourra pas appliquer les résultats à l'ensemble des patients atteints de la maladie). La population ciblée par l'essai (population source) est pertinente si elle correspond à la population à laquelle l'intervention étudiée va être proposée en dehors de l'essai (population cible).

L'extrapolabilité des résultats peut être évaluée grâce à la proportion des patients qui participent réellement à l'étude parmi les patients présélectionnés ou « screenés ». Si une grande proportion de ces patients n'est finalement pas incluse et randomisée dans l'étude, il y a un phénomène de **sur-sélection** de l'échantillon avec une **faible extrapolabilité** des résultats de l'étude à l'ensemble des patients ciblés. Les effectifs screenés doivent être présentés sur le flow-chart.

COMPARABILITÉ DES GROUPES

Comparabilité initiale

Une **randomisation** bien menée permet d'obtenir des groupes comparables et donc d'éviter le biais de sélection différentiel (*voir paragraphe 5.2 randomisation*).

On retrouve la description de la population incluse et les caractéristiques de chaque groupe au début des résultats et dans le tableau 1.

Ce **tableau** permet de vérifier que les caractéristiques initiales des patients sont relativement comparables entre les 2 groupes. Les recommandations actuelles sont plutôt de ne pas utiliser de test statistique pour comparer les groupes à l'état initial et de réaliser cette évaluation « à l'œil » (Eye Ball test). Cependant, les groupes sont parfois comparés par des tests statistiques, pour rechercher l'absence de différence statistiquement significative. En cas de gros effectif, cela n'a aucun intérêt. En cas de faible effectif, on peut s'attendre à quelques déséquilibres entre les groupes liés aux fluctuations d'échantillonnage. Le problème est que la réalisation de ces tests va s'accompagner d'une inflation du risque alpha.

Les paramètres qui sont différents entre les deux groupes (soit en apparence « à l'œil » soit d'après les tests statistiques) doivent être pris en compte dans l'analyse notamment au moyen d'analyses multivariées qui permettent « d'ajuster » les analyses pour ces facteurs.

Comparabilité au cours de l'étude : Suivi et perdus de vue

La comparabilité des groupes doit se maintenir jusqu'à l'évaluation du critère de jugement principal.

Une fois les patients randomisés, il faut pouvoir les suivre et connaître le nombre de patients de chaque groupe qui est allé jusqu'au bout du suivi et qui est donc pris en compte dans l'analyse.

L'attrition est la perte d'effectif au cours de l'étude, certains patients sont perdus de vue c'est-à-dire qu'on ne dispose pas des informations sur le critère de jugement principal (par exemple, on ne sait pas s'ils ont ou non été victimes d'un AVC durant l'étude). De fait, ils ne peuvent pas être pris en compte dans l'analyse. Il faut être attentif :

- aux caractéristiques des perdus de vue : sont-ils différents des personnes analysées ?
- Au nombre de perdus de vue total et dans chaque groupe : y a-t-il une proportion plus importante de perdus de vue dans l'un des 2 groupes qui pourrait modifier les résultats ?

Les conséquences de l'attrition :

- forcément une perte de puissance statistique liée à la baisse de l'effectif,
- possiblement un biais de sélection à cause de cette attrition = biais d'attrition. Les sujets peuvent être perdus de vue en raison d'effets secondaires du traitement, et exclure ces sujets de l'analyse entraînerait une surestimation de l'efficacité de l'intervention, en particulier lorsque la proportion de sujets sortant de l'étude varie selon les groupes de traitement.

Afin de limiter le risque de biais d'attrition, au moment de l'analyse, on effectue une analyse en intention de traiter, qui s'oppose à l'analyse per-protocole.

Une **analyse en intention de traiter (ITT)** consiste à analyser tous les patients randomisés dans le groupe dans lequel ils ont été randomisés :

- qu'ils aient ou non reçu le traitement,
- qu'ils aient ou non suivi le protocole,
- qu'ils soient ou non évaluable pour le critère étudié (on choisit alors la situation la plus défavorable pour le traitement testé afin de ne pas surestimer les qualités du traitement testé).

Une **analyse per-protocole** consiste à n'analyser que les patients compliant au protocole c'est-à-dire ayant reçu le traitement ou le placebo pendant toute la durée de leur suivi. Les patients sont analysés en fonction du traitement qu'ils ont réellement pris et des données réellement disponibles. **L'analyse per-protocole**

a tendance à surestimer l'efficacité du traitement évalué. L'analyse en ITT est plus conservatrice c'est-à-dire qu'elle a tendance à diminuer les différences entre les 2 groupes. L'analyse per protocole donne des résultats plus favorables car on exclut tous les patients qui n'ont pas respecté le protocole (on exclut ceux qui ont arrêté le traitement car il ne leur paraissait pas efficace, ceux qui ont arrêté le traitement en raison d'effets secondaires, etc.).

Dans les essais contrôlés randomisés de supériorité, seule l'analyse en intention de traiter est recommandée.

L'analyse en intention de traiter contribue à maintenir la comparabilité des groupes randomisés au moment de l'analyse.

Afin d'être sûr que la différence de résultats observée entre les 2 groupes n'est due qu'au traitement reçu, il est nécessaire de maintenir la comparabilité des groupes (qui a été obtenue avec la randomisation) pendant toute la durée de l'essai et ce jusqu'à l'analyse. Le double aveugle et l'analyse en intention de traiter permettent de maintenir la comparabilité des groupes pendant le suivi et l'analyse, respectivement et de limiter les biais.

Remarques :

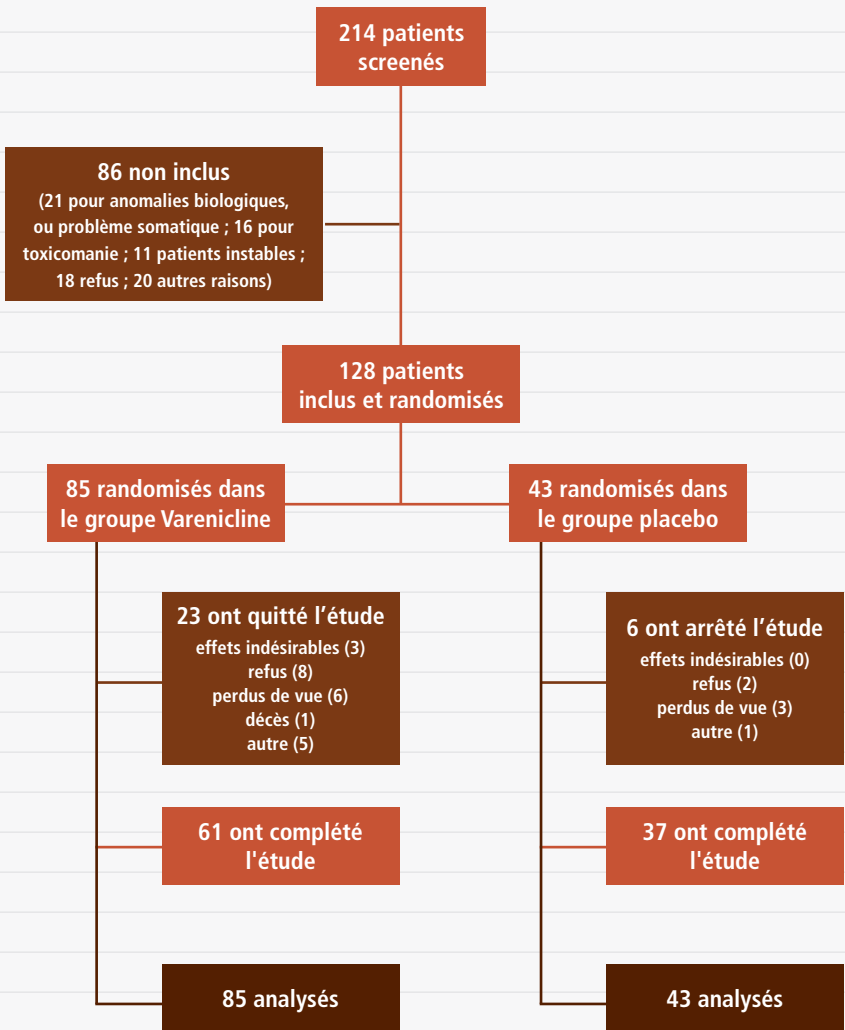
Mais Pourquoi considérer les patients qui n'ont pas pris le traitement comme l'ayant pris ?... Pour comprendre le concept d'intention de traiter, dites-vous qu'on ne compare pas des sujets traités et non traités, mais des sujets que l'on avait ou non l'intention de traiter. C'est une analyse plus proche de la « vraie vie » que l'analyse per protocole, car en pratique clinique on ne sait pas ce qui va arriver lorsque l'on prescrit un traitement (le patient peut ne pas prendre le traitement, prendre son traitement un jour sur deux, ne pas revenir en consultation, etc.).

Attention : les auteurs rapportent souvent une analyse en intention de traiter dans les méthodes mais quand on évalue le diagramme de flux ou les résultats, il est fréquent que des patients aient été exclus de l'analyse.

Il faudra donc vérifier que l'analyse est bien en intention de traiter par :

- le **diagramme de flux** (« flow-chart ») : permet de suivre la cohérence des effectifs à toutes les étapes de l'essai avec les raisons permettant de comprendre toute modification d'effectif, le nombre de patients analysés en bas du diagramme de flux doit correspondre au nombre de patients randomisés,
- les **résultats** : le dénominateur doit correspondre au nombre de patients randomisés.

Graphique des flux (flow-chart) : il représente l'effectif présent à chaque étape de l'étude



Attention : l'analyse en intention de traiter ne peut permettre d'éviter un biais d'attrition que si on a des informations concernant le critère de jugement à la fin de l'étude chez tous les patients, y compris ceux qui ont arrêté le traitement.

Or si les patients sont réellement perdus de vue on ne dispose pas de ces informations qui sont donc des données manquantes, de fait, ces patients ne font pas partie des sujets analysés puisqu'on n'a pas de données sur l'efficacité de leur traitement. On voit donc bien que **l'analyse en intention de traiter n'est donc pas suffisante pour éviter le biais d'attrition**, elle n'assure en aucun cas à

elle seule que tous les patients inclus sont analysés. Pour atteindre cet objectif il faut prendre en compte ces patients pour lesquels aucune valeur du critère de jugement n'est disponible en **remplaçant les données manquantes**.

C'est pourquoi l'analyse en intention de traiter doit être systématiquement associée à une stratégie de gestion des données manquantes (voir paragraphe 5.7 analyses statistiques).

Attention aux idées reçues : de nombreux étudiants font l'erreur de considérer que les perdus de vue ne sont pas un problème si on est en intention de traiter, C'EST FAUX ! il faut également gérer de façon adéquate des données manquantes.

5.4 LE CRITÈRE DE JUGEMENT

Il faut discuter si le choix du critère est **pertinent** pour répondre à la question, et si ce critère est **correctement mesuré**.

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

Le critère de jugement principal était l'arrêt du tabac à la semaine 24. On considérait qu'un patient avait arrêté si le patient déclarait avoir été abstinent sur les 7 derniers jours avec une mesure du monoxyde de carbone expiré inférieure ou égale à 10 ppm. Le nombre de cigarettes par jour était recueilli par interrogatoire sur les 7 jours précédant la visite (abstinence = aucune cigarette consommée dans les 7 derniers jours). La mesure du CO expiré se fait après une inspiration profonde gardée pendant 10 secondes suivie d'une expiration dans un spiromètre ? (VitalographInc, Lexena, Kansas). Le critère de jugement principal était recueilli à chaque visite. Un des critères de jugement secondaires était l'évolution des troubles psychiatriques évaluée par l'échelle PANNS (Positive and negative symptoms scale) mesurée à l'inclusion et à chaque visite. Les patients des 2 groupes étaient vus en consultation à l'inclusion, toutes les semaines jusqu'à la semaine 12. Puis après la visite de la semaine 12 (fin du traitement), ils étaient revus pour des visites de suivi aux semaines 13, 16, 20 et 24. Des appels téléphoniques de suivi étaient réalisés aux semaines 14, 18 et 22.

Les critères de qualité d'un critère de jugement

(voir paragraphe 4.4 pour les qualités d'un critère de jugement)

Le choix du critère de jugement principal doit se baser sur :

- la pertinence clinique,
- son objectivité, avec une méthode de mesure fiable, valide et reproductible.

Pertinence du critère de jugement

Les critères peuvent être **directement importants pour le patient** (décès, rémission d'un cancer) ou **intermédiaires** (mesure biologique ou pharmacologique par exemple qui n'est pas directement perçue par le patient mais qui est censée être liée à un critère perçu par le patient (par exemple, densité minérale osseuse et risque de fracture ostéoporotique, ou ulcération gastrique à l'endoscopie sans aucun symptôme clinique).

Les critères intermédiaires appelés encore critères de « substitution » ou « surrogate » en anglais. Il peut s'agir de mesures cliniques (ex. : pression artérielle), biologiques (ex. : clairance de la créatinine) ou radiologiques. Pourquoi les choisir ? Parce qu'ils permettent une durée de l'étude moins longue et un nombre de patients recrutés moins important (donc un coût moindre), cependant ils ne donnent pas d'information directe sur un effet thérapeutique, ils renseignent sur les mécanismes d'actions du traitement. Il faut toujours être prudent sur les conclusions en cas d'utilisation de critère de jugement intermédiaire car **l'effet sur le critère intermédiaire ne se traduit pas toujours directement et invariablement en effet clinique.**

Exemple : un essai visant à tester l'efficacité du Fluor dans le traitement de l'ostéoporose a montré que le fluor permettait d'augmenter la densité minérale osseuse (DMO). Par ailleurs, dans les études épidémiologiques en population générale, le risque de fracture ostéoporotique augmente quand la DMO diminue. Donc la DMO a été choisie comme critère intermédiaire de la solidité osseuse, le critère clinique étant la survenue de fracture. Les conclusions de l'essai étant que la DMO étant plus élevée dans le groupe traité par fluor, on en a déduit que le fluor permettait une diminution des fractures ostéoporotiques. Un second essai a été réalisé en utilisant la survenue de fracture en critère de jugement principal, les résultats ont été en faveur d'une **augmentation du risque de fracture périphérique** dans le groupe Fluor...

Il faut donc être vigilant lorsqu'un critère intermédiaire est utilisé, même si parfois les conséquences cliniques sont tellement lointaines qu'il est pratiquement

impossible de les utiliser comme critère de jugement principal. Par exemple dans l'hépatite virale chronique C, la « guérison virologique » (réponse virologique totale 24 semaines après traitement) est actuellement considérée comme un critère de jugement recevable même si in fine c'est l'évolution vers la cirrhose ou le cancer qui restent les critères les plus pertinents.

Définition du critère de jugement principal

S'il y a plusieurs critères de jugement (quasi-totalité des cas dans les essais), le critère de jugement principal doit être pré-spécifié dès le protocole (car c'est sur lui que repose le calcul d'effectif) et ne doit pas changer au cours de l'étude. La conclusion doit porter sur ce critère de jugement principal. Il est préférable d'avoir un seul critère de jugement principal pour maintenir un risque alpha à 5%.

Critère de jugement subjectif/objectif

Un critère de jugement est dit objectif si son évaluation ne peut pas être sujette à des interprétations différentes selon la personne qui l'évalue (par exemple, mortalité toutes causes).

Certains critères de jugement sont, à l'inverse, très subjectifs comme les critères rapportés par le patient tels que la douleur, la qualité de vie, qui sont utilisés pour évaluer des traitements symptomatiques. Dans ce cas, il est très important que le patient soit en aveugle du traitement reçu afin de limiter le risque de biais de classement et que le critère de jugement soit évalué à l'aide d'une échelle validée (par exemple, pour la qualité de vie, échelle SF36).

Entre les deux existent d'autres critères qui sont moins subjectifs que les précédents (par exemple, les événements cliniques comme l'infarctus du myocarde, les critères radiologiques comme une récurrence tumorale évaluée par scanner, la mortalité cause spécifique comme la mortalité cardiovasculaire). Ce sont des critères plutôt objectifs mais soumis dans une certaine mesure à l'interprétation d'un individu. Il faut que leur évaluation soit faite en aveugle du traitement reçu afin d'éviter les biais de mesure.

Il faut également essayer de limiter la variabilité entre les évaluateurs (améliorer la reproductibilité). Pour cela, la définition du critère de jugement doit être bien standardisée, et on peut prévoir une évaluation en double ou triple (par 2 ou 3 personnes de manière indépendante) voire centralisée par un comité indépendant (appelé comité d'adjudication).

Plus les critères de jugement reposent sur des données compliquées à obtenir, plus le risque de données manquantes est important. Par exemple, un critère qui repose sur un ou plusieurs examens radiologiques complexes sera associé à un plus grand nombre de données manquantes qu'un critère plus simple tel que le décès, l'infarctus ou l'hémorragie digestive (regarder sur le flow chart le nombre de patients ayant reçu l'examen...).

Risque de biais de mesure sur le critère de jugement (appelé aussi biais de classement)

Les modalités de suivi et de mesure dans les 2 groupes doivent être précisées et strictement identiques pour limiter le biais de mesure et de suivi (si on suit plus fréquemment un des 2 groupes, on risque de mesurer plus d'évènements dans ce groupe). Ce risque est éliminé par le double insu. Il est en revanche important en l'absence de double insu, et ceci d'autant plus que le critère de jugement est subjectif... dans ce cas il faut tout faire pour que l'évaluateur du critère de jugement soit différent de celui qui a inclus le patient et soit en insu du groupe de randomisation.

La mesure du critère de jugement doit se faire autant que possible en insu du traitement reçu pour limiter tout biais de mesure (voir paragraphe 5.6).

NIVEAU 2

Critères de jugement composites

Il s'agit de critères composés de plusieurs événements. Ces critères sont souvent utilisés dans les essais en cardiologie. Un critère composite peut être, par exemple, la survenue d'un infarctus du myocarde, d'un AVC ou d'un décès. On considère que le patient a présenté le critère de jugement s'il a eu au moins l'un de ces événements. Les critères composites permettent un gain de puissance en augmentant la probabilité de survenue de l'évènement (ou de réduire l'effectif nécessaire pour une même puissance). Ils permettent également de prendre en compte l'ensemble des événements importants notamment quand on veut évaluer la balance bénéfice-risque. Par exemple, pour évaluer un traitement anti-thrombotique, on peut utiliser un critère composite combinant des événements ischémiques (efficacité du traitement) et des événements hémorragiques (tolérance). Cependant, l'inconvénient est la difficulté de leur interprétation. Si le résultat pour le critère de juge-

ment composite (par exemple la survenue d'un infarctus du myocarde, d'un AVC ou d'un décès) est statistiquement significatif, il n'est pas possible de conclure que le traitement permet de diminuer les décès. La conclusion doit être : « le nouveau traitement permet de diminuer de manière significative la survenue d'un infarctus du myocarde, d'un AVC ou d'un décès ». Chaque événement clinique composant le critère composite doit être défini comme critère secondaire. Il faut vérifier que l'effet du traitement est le même pour tous les événements cliniques du critère composite.

5.5 L'INTERVENTION

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

La Varenicline était débutée par une période de titration de 0.5 mg / jour en comprimé par voie orale à prendre le soir de J1 à J3, puis 2 comprimés de 0.5 mg / jour (1 le matin et 1 le soir) les 4 jours suivants, puis de J8 à la visite de la semaine 12 les patients prenaient 2 comprimés de 1 mg par jour (1 le matin, 1 le soir). Entre les semaines 12 et 24 (évaluation finale), il n'y avait plus de traitement. À la visite de la semaine 12 il était demandé au patient de ramener les blisters pour évaluer l'observance (questionnaire d'observance et décompte des comprimés restants).

L'intervention doit être décrite très précisément afin de pouvoir reproduire cette intervention dans la pratique courante. Par exemple, pour un traitement, il faut préciser la molécule, la galénique, la voie d'administration, la posologie, la fréquence, la durée. Pour un dépistage par dosage biologique, il faut préciser les conditions de prélèvement, la technique de dosage, l'appareil utilisé, les seuils retenus.

Le **comparateur** (traitement/intervention de référence ou placebo) doit également être présenté précisément. Le choix de l'intervention dans le groupe contrôle doit être justifié.

Il est important de préciser toutes les conditions expérimentales et de discuter celles qui seront différentes en pratique courante afin d'évaluer si les résultats issus de l'expérience pourront en être modifiés. Par exemple, une intervention testée uniquement en secteur hospitalier qui sera amenée à être mise en œuvre en secteur ambulatoire peut avoir des résultats différents.

5.6 LES BIAIS ET FACTEURS DE CONFUSION

L'essai clinique contrôlé randomisé est le design d'étude qui permet de réduire au maximum les biais. Néanmoins certaines situations peuvent faire craindre la présence de biais même en adoptant ce design.

BIAIS DE SÉLECTION DIFFÉRENTIEL

Le biais de sélection différentiel survient lorsque les deux groupes de l'essai ne sont pas comparables et qu'une différence peut apparaître en dehors de tout effet lié à l'intervention ou au traitement.

Un biais de sélection différentiel peut se produire :

- **au départ par un défaut de randomisation**, une randomisation mal menée : **une bonne randomisation** garantit l'absence de biais de sélection différentiel ;
- **au cours de l'étude par un nombre trop élevé de perdus de vue** : c'est le biais d'attrition. L'arrêt précoce du traitement ou le fait de ne plus participer peut être lié à la survenue d'un évènement indésirable ou à l'aggravation de l'état de santé. Si ces sujets ne sont pas analysés, on peut sur ou sous-estimer l'action du traitement étudié (le plus souvent surestimer). Ce risque de biais est très important dans les essais. On le réduit grâce à **l'analyse en intention de traiter**. **Contrôle du biais d'attrition = analyse en intention de traiter + remplacement des données manquantes (+ éviter ++ les perdus de vue).**

Concernant la randomisation, il faut se poser les questions suivantes (voir paragraphe 5.2 randomisation) :

- **si la méthode de randomisation ne garantit pas l'imprévisibilité du traitement, le risque de biais est fort et remet en cause la validité interne du résultat ;**
- **si la méthode de randomisation a produit des groupes non comparables : il existe un déséquilibre entre les groupes au niveau des principaux facteurs pronostiques, il faut prévoir un ajustement dans l'analyse statistique.**

BIAIS DE MESURE = BIAIS D'ÉVALUATION OU D'INFORMATION

Ce biais survient lorsque la mesure du critère de jugement n'est pas réalisée de la même manière dans les deux groupes. On l'appelle aussi biais de classement des individus vis-à-vis du critère de jugement.

Le double insu bien réalisé fait disparaître le risque de biais d'évaluation (voir paragraphe 5.2 insu).

Ce biais doit être particulièrement recherché dans les essais ouverts ou lorsque le double aveugle n'est pas totalement garanti. Dans ce cas :

1. Le risque de biais de mesure est **d'autant plus grand que le critère de jugement est subjectif** et donc influençable (par exemple : douleur, ressenti, satisfaction), il est beaucoup plus limité avec des critères « durs » comme la survenue d'événements aigus (infarctus, AVC, décès).
2. Le risque de biais de mesure peut être limité en adoptant des mesures particulières telles que :
 - **l'évaluation du critère de jugement par un tiers indépendant** de l'étude et en insu du traitement que reçoit le malade. Cela reste imparfait s'il s'agit de critères cliniques « mous » car le patient connaissant le traitement peut en informer l'évaluateur. En revanche, l'évaluation en insu est très efficace s'il s'agit par exemple d'examens biologiques ou radiologiques qui peuvent être interprétés par un évaluateur indépendant sans contact avec le patient ;
 - l'existence de **procédures écrites et standardisées** pour la mesure du critère de jugement, l'utilisation d'outils validés (définition de l'infarctus, échelles standardisées, auto-questionnaires). Cela permet d'atténuer ce biais mais non de l'éliminer.

Donc, dans les essais en ouvert le biais de mesure doit être systématiquement évoqué et recherché d'autant plus que le critère de jugement est subjectif.

BIAIS DE SUIVI

Ce biais survient lorsque le suivi n'est pas réalisé de la même manière

dans les deux groupes. Le double insu fait disparaître le risque de biais de suivi.

Ce biais, comme le biais de mesure, peut survenir dans les essais ouverts ou lorsque l'insu n'est pas totalement garanti, dans ces cas le risque de biais est fort et peut remettre en cause la validité interne de l'étude. L'absence ou la mauvaise réalisation du double insu est susceptible d'entraîner différents biais : biais de suivi, biais de mesure. Si l'un des 2 groupes est mieux ou plus fréquemment suivi, on risque de recenser plus d'évènements dans ce groupe.

BIAIS LIÉS AUX FACTEURS DE CONFUSION

Comme pour les études étiologiques, il peut y avoir d'autres facteurs qui influencent l'effet de l'intervention étudiée. Le fait qu'il y ait un groupe contrôle permet de maîtriser l'effet de ces facteurs.

Lorsque les facteurs de confusion sont anticipés, ils peuvent être intégrés dans l'analyse multivariée, cependant, celle-ci ne pourra prendre en compte que les facteurs qui ont été anticipés et mesurés.

5.7 LES ANALYSES STATISTIQUES

Elles doivent porter en premier lieu sur le critère de jugement principal.

Les analyses suivent en général 2 ou 3 étapes : 1. analyses descriptives, 2. analyses unies ou bivariées testant seulement l'effet du traitement, puis, le cas échéant, 3. analyses multivariées mesurant l'efficacité du traitement indépendamment des facteurs de confusion potentiels. Il est à noter que dans les essais, si la randomisation est bien faite et l'effectif suffisant, les facteurs de confusion connus et non connus sont répartis de façon égale entre les deux groupes et on peut se passer des modèles d'analyse multivariée (ajustés pour les facteurs de confusion potentiels).

Les méthodes statistiques utilisées dépendent du type de la variable qui mesure le critère de jugement. Pour les critères binaires (ex. décès oui/non, infarctus oui/non), soit le critère est la survenue de l'évènement dans un délai défini (ex. récurrence d'un cancer à 5 ans), soit l'effet de la durée de suivi est pris en

compte et les comparaisons se font sous forme de courbes en fonction du temps
= courbes de survie. (voir chapitre 6.2 cohortes pronostiques)

LES DIFFÉRENTES ÉTAPES DE L'ANALYSE STATISTIQUE DES ESSAIS

Type de critère	Critère binaire Ex. : mortalité à 30 jours	Critère continu Ex. : douleur, qualité de vie	Critère de jugement censuré Ex. : mortalité, survenue d'un AVC
Analyse descriptive	Fréquences et pourcentages	Moyenne ET écart-type Médiane ET min-max ou Q1-Q3	Courbe de Kaplan-Meier (médiane de survie)
Analyse univariée ou bivariée Tests statistiques (1 seule variable explicative)	Comparaison des fréquences d'évènements : Test du Chi 2 (paramétrique) Test exact de Fisher (non paramétrique) Risque relatif brut Ou Régression logistique univariée (OR* brut)	Comparaison des moyennes : Test t de Student (paramétrique) Test de Wilcoxon (non paramétrique) Test de Mann Whitney (non paramétrique) ANOVA ou ANCOVA	Test du Log rank Modèle de Cox univarié (HR brut)
Analyse multivariée (plusieurs variables explicatives) ajustement sur n facteurs de confusion	RR ajusté ou Régression logistique multivariée (OR* ajusté pour n facteurs de confusion)	Régression linéaire	Comparaison des survies = délai avant la survenue de l'évènement Modèle de Cox (HR ajusté pour n facteurs de confusion)

*Bien qu'il s'agisse d'étude prospective, et qu'on puisse calculer le risque relatif, l'utilisation de modèles de régressions logistiques abouti au calcul d'OR et non de RR. L'explication de ce phénomène dépasse le niveau statistique de ce polycopié.

La perte de patients (attrition) au cours du suivi de l'étude **est susceptible d'induire un biais, le biais d'attrition**, surtout quand ces exclusions ne se font pas strictement au hasard mais avec une probabilité dépendant du traitement reçu et/ou de l'évolution du patient (il est néanmoins difficile de prouver qu'il y a un biais et difficile de prouver qu'il n'y en a pas, aussi il est nécessaire d'éviter d'être en situation d'avoir un biais donc éviter les perdus de vue chaque fois que possible...). Les sujets perdus de vue pour lesquels on ne dispose pas de la mesure du critère de jugement principal (abandon, visite ou examen non effectué) **représentent des données manquantes**.

L'analyse en intention de traiter doit être systématiquement associée à une stratégie de gestion des données manquantes.

En effet, imaginons un essai sur un traitement préventif de l'AVC, si 30 patients sont perdus de vue dans le bras du traitement évalué et 40 dans le bras placebo, le problème est qu'on ne sait pas si ces 70 patients ont fait un AVC ou non... Il existe des méthodes statistiques pour imputer les données de ces 70 perdus de vue.

NIVEAU 2

Les stratégies de remplacement des données manquantes recommandées sont :

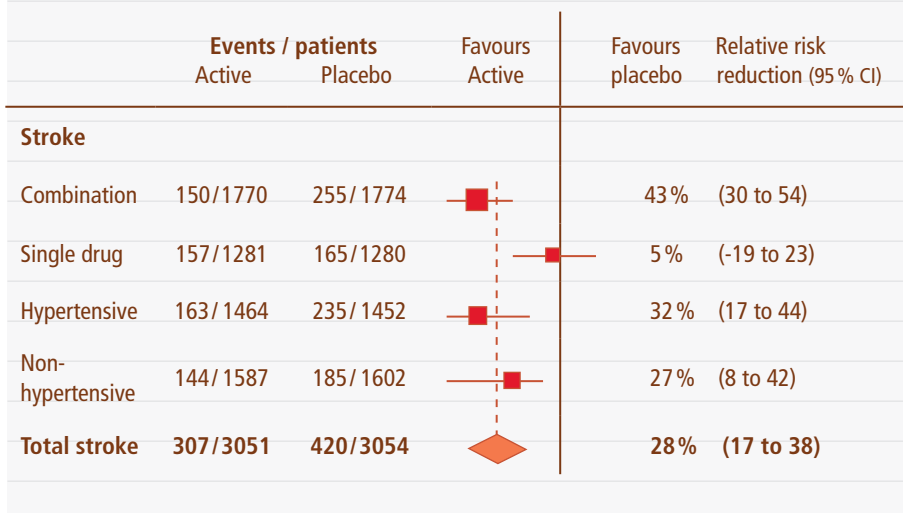
- **l'imputation multiple.** Il s'agit d'une analyse statistique permettant de remplacer les valeurs manquantes du critère de jugement en fonction des caractéristiques des patients perdus de vue,
- **la méthode du pire scénario (worse case scenario).** Cette méthode revient à considérer les données manquantes du bras expérimental comme des échecs (ou des non-réponses) et les données manquantes du bras placebo comme des succès (ou des réponses). Dans notre exemple, cela reviendrait à considérer que les 30 perdus de vue du groupe expérimental ont tous eu un AVC alors qu'aucun des 40 perdus de vue du groupe placebo n'aurait eu d'AVC. Cela est probablement faux mais si on arrive à montrer une différence avec cette méthode, c'est qu'elle existe vraiment. Cette méthode est toutefois peu utilisée car trop stricte.

Une autre méthode est fréquemment utilisée **mais n'est pas recommandée**. Il s'agit de la méthode LOCF (Last Observation Carried Forward) qui consiste à prendre la dernière valeur disponible pour le patient avant qu'il ne sorte de l'essai. Par exemple, si un patient est venu en consultation à 6 mois mais n'est pas revenu à 12 mois, on prendra son résultat à 6 mois...

INTERPRÉTATION DES ANALYSES EN SOUS-GROUPES

Souvent les investigateurs explorent si l'effet du traitement est différent en fonction de certains paramètres tels que l'âge, le sexe, les antécédents, la sévérité de la maladie, etc. Il s'agit alors d'analyser le critère de jugement principal au sein de sous-groupes de patients définis selon ces paramètres. Ces analyses permettent d'évaluer si l'effet du traitement est le même quelles que soient les caractéristiques de la population. Les analyses en sous-groupes doivent être prédéfinies, **et peuvent servir à générer des hypothèses mais en aucun cas à émettre des conclusions (surtout si elles n'ont pas été planifiées avant de faire l'étude)**, elles doivent être considérées comme des analyses exploratoires. **La conclusion de l'essai doit porter sur le résultat pour toute la population et non sur l'un des sous-groupes.**

Le fait de faire des analyses en sous-groupe entraîne également une **inflation du risque alpha** (plus on fait de tests, plus le risque alpha augmente), c'est-à-dire qu'il sera **donc fréquent d'avoir un résultat statistiquement significatif pour l'un des sous-groupes par le simple fait du hasard**. Les résultats des analyses en sous-groupe sont fréquemment présentés sous la forme de figures (appelées « Forest plot ») comme ci-dessous.



Dans cette figure on voit que l'efficacité du traitement pour éviter un AVC (Stroke) n'est pas significativement influencée par le traitement antihypertenseur. En revanche, le traitement combiné (combination) semble plus efficace qu'un traitement simple.

INTERPRÉTATION DES RÉSULTATS DE TOLÉRANCE

La tolérance d'un traitement est un élément essentiel mais fréquemment mal rapporté dans les essais. Tous les événements indésirables (décès, prolongation d'une hospitalisation) doivent être rapportés par groupe de traitement avec leur fréquence de survenue. **ATTENTION**, Les analyses de tolérance **manquent fréquemment de puissance** car les événements indésirables sont (heureusement !) rares et le calcul d'effectif a été fait sur le critère de jugement principal et non sur des critères de tolérance. Le fait qu'il n'y ait pas de différence statistiquement significative ne permet donc absolument pas de conclure à un bon profil de tolérance du traitement évalué. Il faudra apprécier si la différence des pourcentages entre les 2 groupes semble importante (appréciation assez subjective il est vrai...).

LE CALCUL DU NOMBRE DE SUJETS NÉCESSAIRES (NSN)

CALCUL DE L'EFFECTIF DE L'ESSAI

Ce calcul a dû être effectué *a priori*, avant le début de l'étude, et être précisé dans la partie méthodes / analyses statistiques. Il est basé sur une hypothèse qui doit être clairement exprimée et justifiée. La taille de l'échantillon détermine la puissance de l'étude, c'est-à-dire la capacité à mettre en évidence une différence qui existe entre les groupes : ici la **capacité de l'étude à mettre en évidence l'efficacité du traitement si elle existe réellement**.

La puissance de l'étude doit être supérieure à 80 % (si possible 90 %). Une puissance de 80 % signifie que si le traitement est efficace, on a 80 % de chances de le montrer, et donc 20% de risques de conclure que le traitement n'est pas efficace alors qu'il l'est (c'est le risque beta). Le risque alpha est la probabilité de conclure que le traitement est plus efficace que le placebo alors qu'il ne l'est pas.

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

Calcul de l'échantillon – D'après la littérature, la probabilité d'arrêt du tabac dans les six mois est en moyenne de 5 %. On fait l'hypothèse que dans le groupe placebo, la probabilité sera de 10 % par le simple fait de suivre les patients et que grâce à l'intervention cette probabilité va passer à 20 %. On choisit une puissance de 90 % et un risque alpha à 5 %.

Critère	% arrêt du tabac (abstinence et < 10 ppm de CO expiré) à semaine 24
Risque de base	5 % d'arrêt
Groupe Varenicline	20 % d'arrêt
Groupe placebo	10 % d'arrêt (on considère que l'effet placebo et le suivi vont améliorer le % d'arrêt dans le groupe placebo)
Différence attendue	On compare 20 % à 10 % soit une augmentation absolue de 10 % et relative de 100 %
Puissance (bêta)	90 % (bêta 10 %)
Alpha	5 %

À partir de ces éléments, une formule statistique adaptée permet de déterminer combien de patients doivent être inclus dans chaque groupe.

5.8 LES RÉSULTATS (AMPLEUR DE L'EFFET ET SIGNIFICATION STATISTIQUE)

LES RÉSULTATS SONT-ILS STATISTIQUEMENT SIGNIFICATIFS ? (EFFET DU HASARD)

Pour les bases, se reporter au chapitre 4.6 Analyses statistiques (généralités)

L'effet du traitement observé est-il réel ou peut-il être lié seulement au hasard ? La réponse à cette question est donnée par la valeur du « p » qui résulte du test statistique utilisé.

$P < 0.05$ signifie : l'effet observé a moins de 5% de chances d'être dû seulement au hasard. Cette information est similaire à celle apportée par l'intervalle de confiance à 95% autour du RR ou de l'OR, s'il ne comprend pas 1 alors il y a 95% de chances que l'effet observé soit réel et non dû au hasard. Autrement dit, si on répétait un grand nombre de fois ce type d'étude, on obtiendrait un résultat qui se trouve dans cet intervalle dans 95% des cas.

Si la différence entre les groupes n'est pas statistiquement significative (p n'est pas strictement inférieur à 0.05).

Il faut rechercher un manque de puissance de l'étude. Si le nombre de patients inclus et analysés est supérieur ou égal au nombre de sujets nécessaires (NSN) calculé dans la section matériel et méthodes, alors l'essai est *a priori* correctement calibré et ce n'est pas un manque de puissance qui explique l'absence de différence. On peut seulement conclure que le traitement n'a pas l'efficacité sur laquelle on avait basé la question et l'hypothèse de recherche.

De façon plus empirique si plusieurs milliers de patients ont été inclus, la question ne se pose pas. Si seulement quelques dizaines de patients ont été inclus dans chaque groupe, le manque de puissance est fort probable.

ATTENTION, dans les essais de supériorité une différence non significative ne permet pas de conclure à l'équivalence. **Il n'est pas acceptable de conclure que l'absence de différence significative = équivalence des traitements de l'étude.**

À l'inverse, en cas de comparaisons multiples, il existe une inflation du risque α (critères secondaires, sous-groupes, répétitions au cours du temps). Lorsque de nombreux tests statistiques sont utilisés, certains vont être significatifs par le simple fait du hasard, sans refléter une différence réelle.

Attention, plus le p est petit et plus la probabilité que le résultat soit lié au hasard est faible. En revanche cela n'a aucun rapport avec la taille de l'effet : ce n'est pas parce que le « p » est très petit que le traitement est très efficace. Ce qui évalue l'efficacité du traitement, c'est la valeur de la différence entre les groupes (c'est-à-dire, selon le type d'analyse, le RR ou l'HR ou l'OR ou la différence entre les groupes).

DISCUTER LA PERTINENCE CLINIQUE DES RÉSULTATS

La pertinence clinique dépend de deux dimensions : **pertinence du critère choisi** et **taille de l'effet**.

- **Pertinence clinique** : le critère de jugement choisi est-il important pour le patient ?

La douleur, l'asthénie, la qualité de vie par exemple, de même que toutes les maladies symptomatiques sont des dimensions directement appréciables par les patients donc elles sont pertinentes. La question se pose pour les critères de jugement intermédiaires de type critères biologiques, pression artérielle, lésions endoscopiques, images radiologiques etc. ;

- **Importance de l'effet** (= ampleur de l'effet, taille de l'effet, ou magnitude de l'effet).

La taille de l'effet est appréciée par la différence observée entre les groupes dans la mesure du critère de jugement.

Si le critère de jugement est une **variable binaire** (ex. : arrêt du tabac oui/non...) l'effet du traitement est quantifié par la différence de la proportion d'évènements entre les groupes (ex. : 30% d'arrêts du tabac en plus dans le groupe traité par rapport au groupe placebo) ou le risque relatif (ex. : risque relatif d'arrêt dans le groupe traité par rapport au groupe placebo = 1.6).

Si le critère de jugement est une **variable continue** (ex. : nombre de cigarettes fumées par jour, poids, pression artérielle systolique...), l'effet du traitement est quantifié par la **différence entre les moyennes des groupes** (moyenne de 4 cigarettes par jour dans le groupe traité versus 24 cigarettes par jour dans le groupe contrôle, PAS de 12 dans le groupe traité versus 13 dans le groupe placebo). **Si l'effet est de trop petite taille** (ex. : différence de 1 mm de mercure pour la TA entre les deux groupes, réduction de 2 fractures pour 1000 femmes traitées...), il a peu d'intérêt en pratique pour les patients.

Il faut s'assurer enfin que l'effet a été déterminé par rapport à un **comparateur adapté, placebo ou traitement de référence** validé.

Exemple : essai contrôlé avec critère de jugement principal binaire (hémorragie digestive : oui/non), nouveau traitement pour réduire le risque d'hémorragie digestive chez les patients sous AINS.

Risque de base (R) (placebo ou traitement de référence)	$20/100 = 20\%$
Risque avec un nouveau traitement (R_T)	$15/100 = 15\%$
Risque relatif ($RR = R_T/R$)	$0.15/0.20 = 0.75$
Réduction absolue de risque ($RAR = R - R_T$)	$0.20 - 0.15 = 5\%$
Réduction relative du risque ($RRR = 1 - RR$)	$(1 - 0.75) \times 100 = 25\%$
Nombre de sujets à traiter pour éviter un évènement : $NST = 1/RAR$	$1/0.05 = 20$

On peut exprimer un même résultat d'efficacité de différentes façons :

- Le nouveau traitement permet de réduire le risque d'hémorragie digestive de 25%.
- Le risque passe de 20 à 15% grâce au nouveau traitement par rapport au ttt de référence.
- Le risque d'hémorragie est multiplié par 0.75 avec le nouveau traitement.
- Le nouveau traitement permet une réduction absolue du nombre d'hémorragies de 5%.
- En moyenne il faut traiter 20 patients avec le nouveau traitement pour éviter 1 décès.

Attention, une erreur fréquente est de confondre la notion de risque (absolu) avec celle de risque relatif.

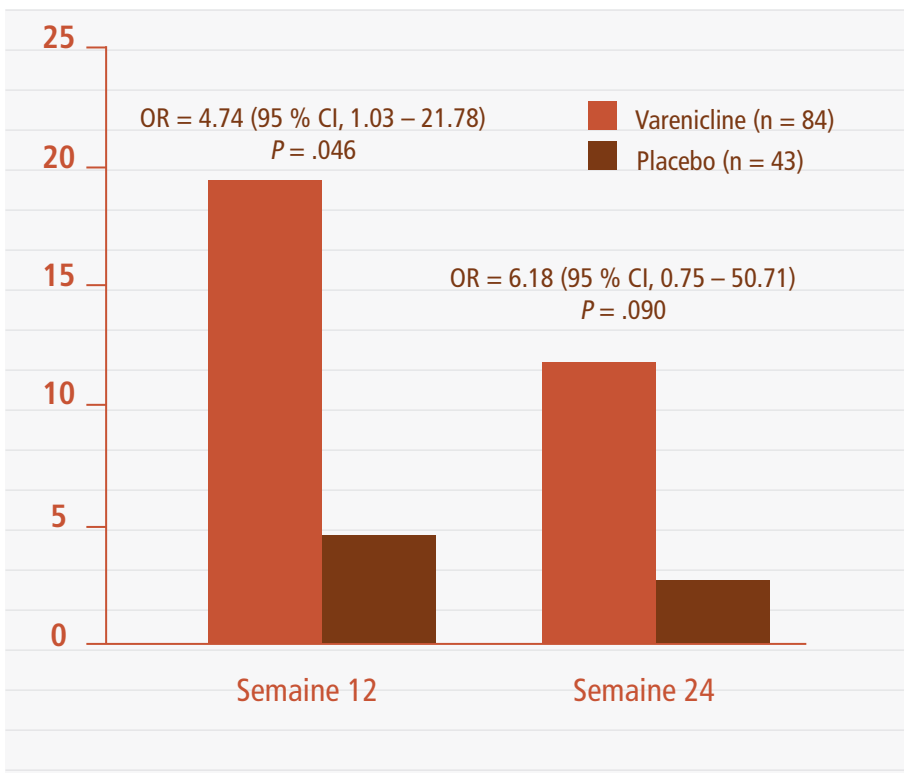
Le **risque** se réfère implicitement au risque absolu, c'est-à-dire à la probabilité de l'événement.

En revanche le **risque relatif est un ratio de deux risques**, et se réfère à une augmentation ou à la réduction du risque observée dans un groupe par rapport à celui du groupe de comparaison (groupe de référence).

Fil rouge :

Essai clinique
Varenicline
et arrêt
du tabagisme

À 24 semaines, 13% des patients du groupe traité ont arrêté de fumer, contre 3% des patients du groupe placebo. L'OR est de 6.18 [0.75-50.71], il n'est donc pas significatif (l'intervalle de confiance comprend le 1, ce qui signifie qu'on a plus de 5% de chances que le risque relatif observé soit l'effet du hasard, $p > 0.05$). Il faut noter que l'intervalle de confiance est extrêmement large, traduisant un manque de puissance dû à des effectifs trop faibles.



5.9 LA CONCLUSION

Vérifier la logique de la discussion et sa structure.

La discussion doit apprécier les résultats (ex. : si $RR = 0,5$ pour le décès, l'essai montre une diminution de 50 % de la mortalité en valeur relative).

Dans la discussion, il faut bien reconnaître ce qui relève des données de la littérature et ce qui est opinion personnelle de l'auteur.

Il faut s'assurer que les conclusions sont justifiées par les résultats : la conclusion d'un article doit porter uniquement sur les résultats observés.

Doivent être discutés :

- **la validité interne** : qualité de la randomisation, double insu ou sinon gestion des risques de biais de mesure, gestion des perdus de vue, qualité de la mesure du critère de jugement principal, qualité de l'analyse statistique, maîtrise des biais possibles, discussion des résultats : sont-ils significatifs ?, discussion des limites et des biais.

Y a-t-il une réponse à la problématique posée ? La réponse à la question

annoncée est un résultat significatif et sans biais. Il n'y a pas de réponse à la question si le résultat est non concluant ;

- **la validité externe** : nombre de patients screenés/inclus, lieux de recrutement (hôpitaux spécialisés ou universitaire ou généralistes), pays, patients etc., les résultats sont-ils extrapolables au contexte souhaité ?, quelle est la représentativité de l'échantillon étudiée par rapport à la population concernée ? ;
- **la pertinence clinique** : les résultats sont-ils importants pour le patient ? (Pertinence du critère de jugement et ampleur de l'effet) ;
- **la cohérence externe** : par rapport aux autres publications ;
- **et globalement les conséquences pour la pratique et la recherche.**

ATTENTION : en cas d'absence de significativité, l'absence de preuve n'est pas la preuve de l'absence d'effet. Mais les résultats ont toutefois une application concrète : il n'y a pas d'indication de traitement tant que son efficacité n'est pas clairement démontrée. Dans ce cas il faut vérifier si l'étude ne manque pas de puissance (dans le cas dans les petits essais non médicamenteux (ex. : physiothérapie ou kinésithérapie...) ou dans le cas des maladies rares avec des capacités de recrutement des patients limitées). Il n'y a pas non plus de réponse à la question s'il existe des biais.**

Vérifier le respect des règles d'éthique.

On doit être particulièrement attentif aux règles d'éthiques dans le cadre d'un essai clinique. En effet, il est indispensable de s'assurer que les bénéfices du traitement sont supérieurs aux risques attendus. En France aucun essai ne peut être mené sans l'aval d'un Comité de protection des personnes (CPP). Le consentement éclairé des patients est indispensable et réglementaire.

Les points à retenir :

- Comparabilité initiale = qualité de la randomisation
- Maintien de la comparabilité des groupes randomisés =
 - double aveugle
 - analyse en intention de traiter et remplacement des données manquantes
- Critère de jugement principal (important pour le patient, objectif ?)
- Différence cliniquement pertinente (taille d'effet = efficacité)
- Vérification de la cohérence : objectif principal-critère de jugement principal - résultats principaux - conclusion sur le critère de jugement principal

NIVEAU 2

5.10 LES ESSAIS D'ÉQUIVALENCE OU DE NON-INFÉRIORITÉ

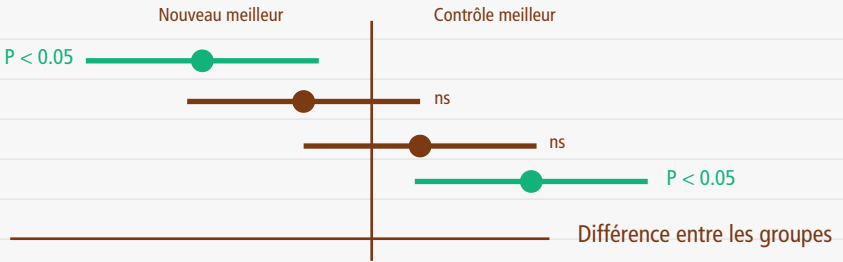
Il arrive que l'on soit dans une situation de médicaments qui ne sont pas plus efficaces mais présentent moins d'effets secondaires, sont moins coûteux, ou encore qui sont plus pratiques d'utilisation (ex. : comprimés au lieu d'injections). Dans ce cas il faut s'assurer que le nouveau traitement a une efficacité comparable à celui qui existait déjà (mais qui est moins bien toléré, plus cher ou moins pratique).

Comme nous l'avons vu précédemment, dans l'essai « classique » de supériorité, l'absence de différence statistiquement significative ne permet pas de conclure que les traitements sont comparables. **Il existe un type d'essai spécifiquement adapté pour évaluer si deux traitements sont équivalents ou si le nouveau traitement est non inférieur au traitement de référence.**

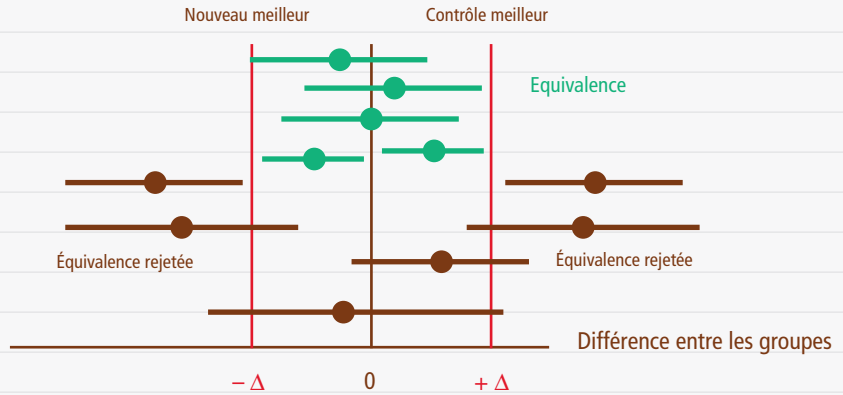
En pratique, **il est impossible de déterminer que deux traitements sont STRICTEMENT ÉQUIVALENTS** en termes d'efficacité. On montre donc qu'ils ne sont pas trop différents, c'est-à-dire **qu'on doit définir une borne à partir de laquelle on trouve acceptable de conclure à l'équivalence ou la non-infériorité**. Si la différence entre les 2 traitements (et son intervalle de confiance à 95 %) est comprise dans cette borne, on pourra conclure à l'équivalence ou à la non-infériorité, en acceptant que l'efficacité du traitement évalué puisse être légèrement inférieure à celle du traitement de référence, dans des proportions acceptables et fixées par les bornes choisies. Ce risque peut-être pris si le traitement évalué présente d'autres avantages (meilleure tolérance, plus grande facilité d'utilisation...).

Dans les figures suivantes, Δ représente les bornes d'équivalence ou la borne de non infériorité.

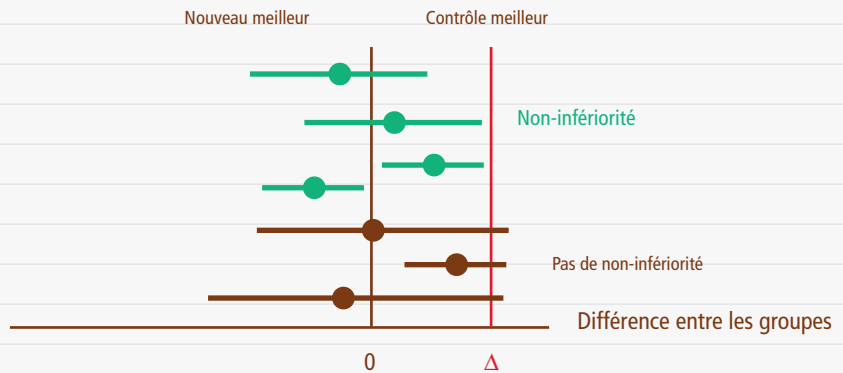
Essai de supériorité



Essai d'équivalence



Essai de non-infériorité



Dans l'essai de supériorité, pour conclure à la supériorité du traitement évalué (A), il faut que l'IC à 95 % de la différence soit strictement > 0 .

Dans l'essai d'équivalence, pour conclure à l'équivalence entre le traitement évalué (A) et le traitement de référence (B), il faut que l'IC à 95 % de la différence soit strictement inclus entre $-\Delta$ et $+\Delta$ (le traitement A ne doit être ni trop inférieur à B ni trop supérieur).

Dans l'essai de non-infériorité, pour conclure à la non infériorité de A par rapport à B, il faut que l'IC à 95 % de la différence soit strictement supérieure à $-\Delta$ (le traitement A peut être supérieur).

Le choix de la borne est très difficile et va conditionner le nombre de patients à recruter. Plus la borne est petite, plus il faut inclure de patients dans l'essai. D'un autre côté si la borne est trop large, ce serait erroné de considérer que les deux traitements ont une efficacité comparable. Dans un essai d'équivalence ou de non-infériorité, il faut mener non seulement une analyse en intention de traiter mais également une analyse per protocole et vérifier que les résultats sont cohérents entre ces 2 analyses. En effet, l'analyse en intention de traiter a tendance à diminuer les différences entre les 2 groupes, ce qui risque de faire conclure plus facilement à l'équivalence ou à la non-infériorité alors que l'analyse per protocole a tendance à augmenter la différence entre les groupes. Les essais d'équivalence doivent être d'une grande qualité et nécessitent souvent un très grand nombre de patients (supérieur aux essais de supériorité). En non-infériorité ou équivalence, on admet que le nouveau traitement puisse être moins efficace que le traitement de référence d'une quantité Δ (limite de non-infériorité) définie et faible et la perte d'efficacité consentie en non-infériorité ne peut se justifier qu'en échange d'un avantage dans un autre domaine.

Analyse en intention de traiter dans les essais d'équivalence ou de non-infériorité

L'Analyse en intention de traiter :

- teste la stratégie thérapeutique,
- biaisée si beaucoup de perdus de vue / valeurs manquantes,
- biaisée si mauvaise observance.

L'Analyse per protocole :

- teste l'effet d'une modalité de traitement,
- nécessairement biaisée.

Conclusion :

- faire deux analyses,
- privilégier l'analyse per protocole,
- interpréter selon le point de vue adopté.

Parfois les auteurs transforment au cours de l'étude l'essai de supériorité en essai de non-infériorité ou vice-versa. Dans quelles conditions est-ce acceptable ?

Interpréter un essai de non-infériorité comme un essai de supériorité

- ne pose pas de problème de multiplicité,
- doit être analysé en intention de traiter.

Interpréter un essai de supériorité comme un essai de non-infériorité

- il faut que la limite de non-infériorité ait été pré-définie AVANT de débiter l'essai,
- doit être analysé en per-protocole,
- le design doit être approprié (comparateur, doses, population, critères de jugement),
- l'essai doit avoir une sensibilité suffisante,
- les auteurs doivent apporter des preuves (directes ou indirectes) que le traitement contrôle a montré dans l'essai son efficacité habituelle.

Il est préférable de concevoir d'emblée comme essai de non-infériorité !

5.11 ÉVALUATION D'UNE INTERVENTION DE DÉPISTAGE

DÉFINITION

Une procédure de dépistage vise à identifier dans une population en bonne santé des sujets ayant une maladie inapparente ou à risque élevé de présenter une maladie, en vue d'examen complémentaires, d'un suivi plus rapproché ou de mesures de prévention.

On distingue le dépistage opportuniste ou individuel et le dépistage organisé. Le dépistage opportuniste ou individuel est proposé aux patients par leur médecin sous la forme d'un test susceptible de découvrir une maladie curable qui n'est pas l'objet de la consultation. Le dépistage organisé est mis en place sur décision de l'état. Les personnes ciblées reçoivent toutes systématiquement une invitation au dépistage (ex. : mammographie pour dépister le cancer du sein), il est pris en charge totalement par l'assurance maladie.

Avantages et inconvénients d'une procédure de dépistage

Avantages	Inconvénients
<ul style="list-style-type: none"> - Diminution de la mortalité ou de la morbidité - Traitements moins lourds - Réconfort des sujets négatifs 	<ul style="list-style-type: none"> - Risque de faux négatifs (faux réconfort) - Risque de faux positifs - Risques iatrogéniques des tests diagnostiques et des traitements après dépistage - Allongement de la période de maladie (marquage)

Un dépistage efficace permet une avance au diagnostic qui s'accompagne d'une guérison.

FORMULATION DE L'OBJECTIF

L'objectif doit reprendre les différents éléments du « PICO ». Par exemple, évaluer l'efficacité d'un dépistage organisé par frottis cervico-vaginal (I) par rapport à une absence de dépistage organisé (C) en termes de mortalité (O) chez les femmes de 25 à 65 ans (P).

TYPE D'ÉTUDE

L'essai contrôlé randomisé est la référence pour évaluer l'efficacité d'une procédure de dépistage car il permet de limiter les biais et apporte (en théorie) le niveau de preuve scientifique le plus élevé. Mais il est très difficile de réaliser un essai randomisé car l'impact sur la santé des personnes est beaucoup plus dilué et difficile à mettre en évidence dans ce type d'essai, du fait des personnes qui ne vont pas se faire dépister et de celles qui ne se font pas traiter après dépistage. Il faut donc en général des effectifs extrêmement importants suivis de nombreuses années.

L'évaluation de ce type d'étude se fait selon les notions précédemment vues dans les essais contrôlés randomisés. Il faudra accorder une attention particulière aux points suivants :

- **Qualité de la randomisation ;**
- **Critère de jugement principal :** pertinence clinique ? subjectivité ? mode d'évaluation ? ;
- **Analyse en intention de dépister :** c'est le même principe que l'analyse en intention de traiter mais pour une procédure de dépistage, on analyse les patients dans le groupe dans lequel ils ont été randomisés quel que soit leur suivi, leur compliance et l'intervention qu'ils ont effectivement reçue.

L'analyse en intention de dépister est le reflet de l'impact de la procédure de dépistage organisé dans la vraie vie car tous les individus qui reçoivent une invitation pour un dépistage ne vont pas systématiquement le faire (par exemple seulement 52% des femmes se rendent au dépistage organisé du cancer du sein par mammographie) et que certains individus randomisés dans le groupe absence de dépistage auront la procédure de dépistage dans le cadre d'un dépistage dit opportuniste via leur médecin ;

- **Interprétation des résultats** : Il faut évaluer si la différence est statistiquement significative et si celle-ci est cliniquement pertinente. Pour cela, il faut considérer la réduction absolue du risque et le nombre de sujets à dépister pour éviter un évènement. **Attention, il est normal que ce nombre soit plus élevé que dans un essai thérapeutique car les sujets ne sont pas malades.**

CRITÈRES JUSTIFIANT LA MISE EN PLACE D'UN DÉPISTAGE ORGANISÉ

La maladie et l'examen utilisé pour le dépistage doivent répondre à un certain nombre de critères :

- la maladie doit être un problème de santé publique par sa fréquence, sa sévérité ou son impact socio-économique ;
- l'histoire naturelle de la maladie doit être bien connue ;
- la maladie doit pouvoir être détectée pendant sa phase de latence asymptomatique ;
- il n'existe pas de facteurs de risque accessibles à la prévention primaire ;
- il doit exister un test de dépistage ayant une forte sensibilité et une forte spécificité ;
- le test de dépistage doit être simple à utiliser et avoir une bonne acceptabilité par les patients ;
- il existe un traitement efficace permettant de prendre en charge les patients dépistés ;
- les moyens appropriés de diagnostic et de traitement sont disponibles ;
- le coût du dépistage (y compris le diagnostic et le traitement) ne doit pas être disproportionné par rapport au coût global des soins médicaux ;
- le programme de dépistage doit avoir fait la preuve de son efficacité avec des essais contrôlés randomisés.



EXERCICE_ESSAI_CLINIQUE

CHECK-LIST DES MOTS CLÉS ESSAIS CLINIQUES

TYPE D'ÉTUDE

- Expérimentale / interventionnelle
- Essai contrôlé :
 - groupe intervention : traitement, intervention,
 - groupe contrôle : traitement de référence / placebo.
- Critères de qualité :
 - randomisation : aléatoire, imprévisible,
 - double insu.
- Type d'essai supériorité / équivalence / non infériorité, groupes parallèles, cross-over.
- Toujours prospectif !

OBJECTIF

Préciser : Population, Intervention, Contrôle / comparateur, l'Outcome ou critère de jugement, le délai de mesure = PICOT.

POPULATION ÉTUDIÉE

Critères d'inclusion et non-inclusion : vérifier l'absence de sur-sélection, la représentativité. Suivi des effectifs à chaque étape de l'étude par un flow chart, justification de tout changement d'effectif.

CRITÈRE DE JUGEMENT

Pertinent cliniquement, objectif, mesuré de manière identique dans les 2 groupes, en insu du groupe.

ANALYSES STATISTIQUES

Calcul du nombre de sujets nécessaires *a priori* (selon alpha, bêta/puissance et différence attendue entre les 2 groupes), analyse en intention de traiter. Comparaison directe de 2 groupes et non analyse avant après dans chaque groupe. RR, OR, nombre de sujets à traiter.

BIAIS À RECHERCHER

	BIAIS À RECHERCHER	MÉTHODE POUR RÉDUIRE LE RISQUE DE BIAIS
SELECTION	Groupes non comparables	Randomisation centralisée indépendante
ATTRITION	Groupes analysés non comparables, sorties d'étude non liées au hasard	Analyse en intention de traiter + remplacement données manquantes
CONFUSION	Présence d'un tiers facteur influençant la mesure de l'efficacité	Prise en compte des facteurs de confusion potentiels dans l'analyse
MESURE	Mesure des critères ou suivi différents entre les 2 groupes	Double insu

Exemple de grille de LCA d'un article visant à évaluer l'efficacité d'un traitement

LES RÉSULTATS SONT-ILS VALIDES ?	OUI	NON	NSP
Les groupes sont-ils a priori comparables à l'inclusion ? (Les groupes comparés ont-ils commencé l'étude avec un pronostic similaire ?)			
Qualité de la randomisation en début d'étude?			
Les sujets ont-ils été randomisés ?			
La randomisation a-t-elle été réalisée de manière appropriée ? (masquée, stratifiée, par blocs... ?)			
Les groupes sont-ils comparables à la fin de l'étude (lors de l'analyse) ?			
Qualité du contrôle pendant l'étude ? (Les groupes comparés ont-ils gardé un pronostic similaire dans le courant de l'étude ?)			
Les sujets connaissaient-ils le groupe auquel ils étaient assignés ?			

Les cliniciens investigateurs connaissaient-ils le groupe auquel étaient assignés leurs patients ?			
Les investigateurs qui ont évalué et mesuré le critère de jugement connaissaient-ils le groupe auquel étaient assignés les sujets ?			
Le suivi a-t-il été complet (proportion de perdus de vue raisonnable par rapport à l'incidence du critère de jugement et équilibrée dans les deux groupes,			
Les sujets ont-ils été analysés dans le groupe où ils ont été randomisés initialement (intention de traitement)			
QUELS SONT LES RÉSULTATS ?	VALEUR	IC 95	NSP
Quelle est la réduction absolue de risque obtenue par le traitement ?			
Quelle est la réduction relative du risque ?			
Combien faut-il traiter de patients pendant un temps équivalent à la durée de suivi ?			
QU'APPORTENT LES RÉSULTATS À MON MALADE ?	VALEUR	IC 95	NSP
Les résultats s'appliquent-ils à mon patient ?			
Sont-ils cliniquement importants ?			
Le bénéfice est-il obtenu pour un risque raisonnable ?			
Le bénéfice est-il obtenu pour un coût acceptable ?			

LCA DES ÉTUDES DE COHORTES

Il existe deux types d'études de cohortes :

- Études de cohortes visant à mesurer un lien entre un facteur de risque supposé et la survenue de la maladie,
- Études de cohortes visant à analyser l'évolution d'une maladie et ses facteurs pronostiques.

Ces deux types d'études sont traités successivement dans ce chapitre.

6.1 LCA DES ÉTUDES DE COHORTES VISANT À MESURER UN LIEN ENTRE UN FACTEUR DE RISQUE SUPPOSÉ ET LA SURVENUE D'UNE MALADIE

Fil rouge :
mort subite du
nourrisson (MSN)
et position
de couchage

Dans les années 1980 on a commencé à coucher les nourrissons sur le ventre à partir d'une étude issue de deux séries de cas d'adultes ou d'enfants en insuffisance respiratoire qui semblaient mieux respirer sur le ventre. Parallèlement on a observé une augmentation progressive du nombre de MSN passant de 200 cas par an en 1979 à 1700 en 1991... Une étude de cohorte a été menée en 1991 en Australie dans laquelle 2000 nourrissons âgés de 3 à 24 mois ont été inclus pour étudier l'hypothèse selon laquelle la position de couchage ventrale pouvait être à l'origine d'une augmentation des MSN.

Il s'agit au final de répondre à la question : ce facteur augmente-t-il le risque de survenue de la maladie ?

On doit commencer par répondre à la question : Existe-t-il une association statistiquement et cliniquement significative entre le facteur de risque et la survenue de la maladie ?

- Première étape : **l'association est-elle au-delà du hasard ?**
(statistiquement significative ?)
- Deuxième étape : **l'association est-elle forte ?**
(taille de l'effet = cliniquement significative ?)

Troisième étape : **l'association observée correspond-elle à une réalité ?**
Ou peut-elle être due à des biais ?

Mise au point

- Deux types d'étude peuvent être appropriés pour répondre à cette question : **l'étude de cohorte** et **l'étude cas-témoins** (*traitée dans le chapitre suivant*).
- Ces deux designs d'étude ont des **forces**, des **limites** et des **sources de biais** potentiels très différentes, les deux types d'étude sont donc traités séparément.
- On appelle cohorte **un ensemble de personnes suivies dans le temps** et chez lesquelles le **recueil des données sur l'exposition au facteur de risque a précédé la survenue de la maladie étudiée**.
- Le terme « cohorte » peut être ambigu car peut correspondre à des types d'études différents en fonction de la question posée.
- **Question descriptive** : quelle est l'incidence du cancer du sein en France en 2010 ? on est alors dans le champ de l'épidémiologie descriptive traité dans le chapitre sur la LCA des études visant à mesurer un indicateur : prévalence, incidence, ou pronostic. Ce type d'étude n'est pas encore traité en détail dans ce polycopié (ce n'était pas notre priorité car il y a peu de chances d'avoir un tel article à l'ECN).
- **Question de nature étiologique** : la prise d'oestro-progestatifs à la ménopause augmente t'elle le risque de cancer du sein ? Ce type d'étude qui est traité dans le présent chapitre.
- **Question de nature pronostique** : la taille de la tumeur influence t'elle la survie des patients atteints de cancer du larynx ? Ce type d'étude est traité dans le présent chapitre.

LA QUESTION

Une étude de cohorte a pour but de répondre à la question « ce facteur de risque est-il associé à une augmentation significative du risque de maladie ? »

La façon la plus précise de décrire la question de recherche est le **PECO**
(voir chapitre : *Les étapes de la check-list*) qu'on remplace ici par le **PFCO**

- P** Population étudiée
- F** Facteurs de risque pressenti
- C** Comparaison (groupe non exposé = groupe de référence)
- O** Maladie ou état de santé étudié (= critère de jugement principal)

LE TYPE D'ÉTUDE

Ces études consistent en la formation d'une cohorte d'individus non malades, chez lesquels on recueille des données sur l'exposition au facteur de risque et que l'on va suivre en enregistrant la survenue de la maladie étudiée au cours du temps. À la fin du suivi, on analysera si la maladie est survenue plus fréquemment chez les patients exposés que chez les non-exposés.

Comment reconnaît-on une étude de cohorte ?

Deux critères sont systématiquement présents :

1. Les individus inclus dans la cohorte sont **indemnes de la maladie étudiée** (critère de jugement),
2. Ils sont **suivis dans le temps** (autrement dit on a pour chaque individu au moins deux temps de mesure, un temps initial pour la mesure de l'exposition et un temps au cours du suivi pour la mesure du critère de jugement).

LA POPULATION / L'ÉCHANTILLON ÉTUDIÉ

En fonction des modalités de recrutement (basé ou non sur l'exposition au facteur de risque) et en fonction de la temporalité de l'étude, on distingue trois grandes catégories d'études de cohorte analytique :

- la cohorte « classique » qui inclut les individus sans prendre en compte leur exposition au facteur étudié et les suit de façon prospective,
- la cohorte qui reprend ce design mais *a posteriori*, en se fondant sur des dossiers ou des bases de données médicales,
- la cohorte dont le groupe des personnes exposées est recruté du fait de leur exposition et est ensuite comparé à un autre groupe (comparateur externe) de personnes non exposées au facteur étudié.

La caractéristique N°1 des études de cohortes est que **les sujets, au moment de leur inclusion dans la cohorte, sont indemnes de la maladie étudiée**. Quel que soit le type de cohorte la première question sera donc de savoir comment les auteurs se sont assurés de l'absence de maladie. Les cohortes sont souvent de très grande taille et elles se font souvent en population générale, aussi on se contente généralement d'un questionnaire.

Exemple : on ne peut pas faire une mammographie aux 100 000 femmes suivies dans la cohorte E3N sur les facteurs de risque de cancer du sein (on se base sur les antécédents qu'elles déclarent dans le questionnaire et on récupère les mammographies si elles en ont eu).

▶ Ce problème ne se pose évidemment pas lorsque le critère de jugement est le décès toute cause.

Fil rouge :
mort subite du
nourrisson (MSN)
et position de
couchage

L'inclusion des individus dans la cohorte

a / Étude de cohorte prospective : personnes incluses prospectivement et indépendamment de leur exposition au facteur de risque

Les personnes sont incluses dans la cohorte **indépendamment de leur exposition** au facteur de risque.

L'exposition est mesurée par questionnaire après leur inclusion dans l'étude et indépendamment de la survenue de leur maladie (au départ ils sont tous indemnes de la maladie).

▶ Les nourrissons étaient inclus dans la cohorte sans connaissance *a priori* de leur position de couchage. C'est seulement une fois inclus dans la cohorte qu'on a fait remplir un questionnaire aux mères en leur demandant la position de couchage de leur enfant qu'on a saisi dans une base de données informatisée.

Fil rouge :
mort subite du
nourrisson (MSN)
et position de
couchage

Autres exemples :

- Étude EPIDOS, un courrier d'invitation a été envoyé à 15 000 femmes de plus de 75 ans inscrites sur les listes électorales de Lyon, 1500 ont accepté de par-

tipier. Les seuls critères d'inclusion étaient le sexe féminin, l'âge de 75 ans et plus, et l'absence de pathologie très grave ou très invalidante et bien entendu l'absence d'antécédent de fracture du col fémoral.

- Étude OFELY qui inclut 1000 femmes âgées de 20 à 90 ans adhérentes à la MGEN suivies depuis 1991.
- Cohorte Gazel débutée en 1989 parmi 20 000 volontaires d'Electricité et Gaz de France constituant un « Laboratoire épidémiologique ouvert », (<http://www.gazel.inserm.fr/>).
- Cohorte nutri-net santé : étude de 500 000 internautes sur les comportements alimentaires et les relations nutrition santé (<https://www.etude-nutrinet-sante.fr/fr/common/login.aspx>).

En général les personnes incluses dans l'étude ne sont pas représentatives de l'ensemble de la population française car les personnes qui acceptent de participer sont différentes de celles qui n'acceptent pas, ce qui constitue un biais de sélection quasi inévitable. Cependant, ceci ne menace pas la validité interne de l'étude tant que le biais n'est pas différentiel.

Exemple : les femmes ou les individus ayant un niveau d'études élevé ont tendance à participer plus volontiers à des études de recherche.

Ce qui est essentiel est que ce biais de sélection s'applique de la même façon pour tous les individus de la cohorte, sans différence entre les exposés et les non exposés au facteur de risque supposé. Il n'a donc pas d'impact sur l'estimation de l'association entre facteur étudié et maladie (on dit que ce biais n'est pas « différentiel »).

Dans ce type de cohorte le biais de sélection différentielle à l'inclusion du patient est très improbable puisqu'on ne sait pas si les personnes sont exposées avant qu'elles entrent dans la cohorte.

Dans ce type d'étude de cohorte, **le risque de biais de sélection est donc essentiellement lié aux perdus de vue (biais de sélection qui survient au cours du suivi)**. Il faut donc être particulièrement vigilant sur les procédures mises en place pour le suivi des personnes, le recueil des événements de santé surtout concernant le critère de jugement principal.

Il est souhaitable d'étudier les causes des arrêts de suivi et de comparer les caractéristiques démographiques et cliniques des patients perdus de vue à ceux ayant

eu un suivi complet (mais cela ne suffit pas si le nombre de perdus de vue est trop important).

Fil rouge :
mort subite du
nourrisson (MSN)
et position de
couchage

À l'inclusion des enfants dans la cohorte : *a priori* pas de risque que les nourrissons couchés sur le ventre soient systématiquement différents des nourrissons couchés sur le dos (sauf si par exemple la position de couchage avait un lien avec le niveau socio-économique ce qui *a priori* n'est pas le cas). À la fin du suivi il faut vérifier que les nourrissons qui sont perdus de vue ne sont pas plus nombreux dans un groupe que dans l'autre.

b / Études de cohortes exposés/non exposés = personnes incluses en fonction de leur exposition au facteur de risque

Dans les cas où l'exposition au **facteur de risque est rare**, on risque d'avoir trop peu de personnes exposées si on fait une étude en population générale. Dans ce type d'étude, **on sélectionne la population selon l'exposition au facteur étudié et on constitue un groupe de personnes exposées et un groupe non/exposées. Bien entendu, ces personnes sont toutes indemnes de la maladie au début de l'étude.**

Dans ce type de cohorte, il y a plus de **risque de biais de sélection différentiel** au départ (c'est-à-dire des biais qui s'appliquent différemment chez les exposés et les non-exposés).

Exemple : pour connaître l'impact de l'utilisation fréquente d'un sécateur sur la survenue d'un syndrome du canal carpien, on suit un groupe de viticulteurs en période de vendanges (exposés) et un groupe d'éleveurs (non-exposés) et on compare la survenue d'un syndrome du canal carpien dans ces deux groupes au cours du temps. Il est possible que les individus « exposés » (à la taille de la vigne) aient par ailleurs des caractéristiques différentes des individus non-exposés et que ces caractéristiques impliquent un risque systématiquement plus ou moins élevé de syndrome du canal carpien (en dehors de la taille de la vigne).

Il existe donc un risque de biais ayant un impact sur l'estimation de l'association entre facteur de risque et maladie, c'est un biais différentiel qui peut aller dans le sens d'une surestimation ou d'une sous-estimation de l'association.

REMARQUE : certains enseignants ne font pas la distinction et utilisent indifféremment les termes d'étude de cohorte et d'études exposées non exposées.

c / Étude de cohorte historique (ou « rétrospective »)

Dans les cas où :

- la maladie est rare et on a besoin d'un effectif important : grosse cohorte nécessaire pour avoir un nombre suffisant de cas,
- le délai entre l'exposition et la survenue de la maladie est long (suivi de cohorte trop long),
- il faut répondre rapidement à une question de recherche.

On peut mener une étude de cohorte historique : on reconstitue *a posteriori* une « cohorte » à partir de dossiers médicaux ou de données médicales initialement recueillies dans un autre but.

Le chercheur écrit son projet puis l'étude démarre mais il ne s'agit pas d'inclure des individus prospectivement mais le chercheur va essayer de vérifier son hypothèse dans des données existantes.

En pratique, **le chercheur reconstitue une cohorte d'individus qui étaient indemnes de la maladie au moment de la date d'inclusion dans la cohorte**, puis suit *a posteriori* au travers des visites successives figurant dans les dossiers médicaux ou les bases de données médicales la survenue ou non de la maladie étudiée pendant la durée du suivi qu'il a choisi.

Exemple : on décide en 2005 de débiter une étude de cohorte à partir d'individus ayant travaillé sur le site de Jussieu entre 1980 et 1990. À partir des dossiers de la médecine du travail de Jussieu, on reconstitue la cohorte et on mesure rétrospectivement en fonction de la profession et des contacts avec les matériaux amiantés, l'exposition à l'amiante. Ensuite, on suit dans le temps de l'inclusion (1980-1990) à 2005 pour recenser la survenue de la maladie.

Dans ce cas, il y a **peu de risque de biais de sélection au départ**. Les personnes sont incluses *a posteriori* dans la cohorte **indépendamment de leur exposition au facteur de risque et indépendamment de la survenue de leur maladie**

(ce dont il faut tout de même s'assurer car ils sont peut-être devenus malades au moment où l'on décide de mener l'étude).

Le principal problème est lié à la **qualité des données** : on ne maîtrise pas le recueil des données, les modalités de mesure, etc. On doit donc se contenter des données disponibles, il existe un risque important de **données manquantes** et un **manque de précision** par rapport à la question posée.

REMARQUE : les études exposés/non-exposés peuvent également être rétrospectives.

REMARQUE : il existe une ambiguïté sur le terme rétrospectif. La définition en épidémiologie classique est un recueil des données sur le facteur de risque qui est réalisé après la survenue de la maladie. Dans une cohorte historique, ce n'est pas le cas, le recueil a toujours été fait de façon prospective au moment où le patient était suivi dans une cohorte ou dans des dossiers de soins donc il a été fait de façon prospective. Ce qui est « rétrospectif » c'est la question de recherche qui n'était pas posée au moment du recueil de données sur l'exposition et ceci se traduit par une moins bonne qualité des données car elles n'ont pas été recueillies dans ce but initialement. Certains auteurs parlent également de cohorte rétro-prospective.

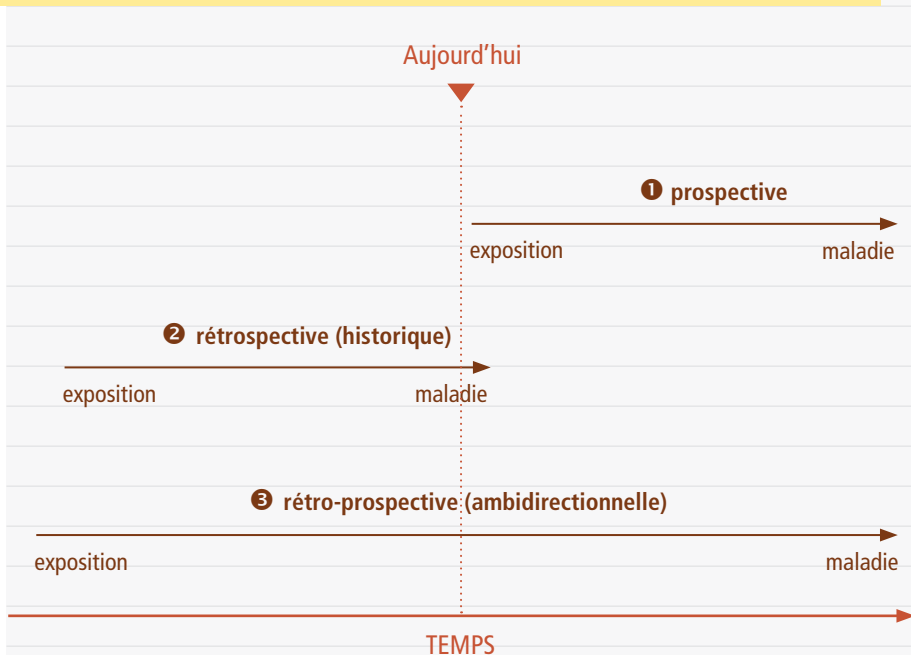


Figure du Pr J Labarère faculté de médecine de Grenoble.

On pourrait imaginer d'aller rechercher dans les dossiers médicaux d'une maternité la cohorte des enfants nés entre le 01/01/1989 et le 31/12/1990 et de récupérer les données sur leur position de couchage puis de retrouver soit en appelant les mères soit en croisant ces données avec un registre des décès sur les 2 ans suivants la naissance, les enfants victimes d'une MSN. Une cohorte historique ne pourra *a priori* pas être faite parce qu'il y a peu de chance qu'on ait recueilli des données fiables sur la position de couchage des enfants sur les deux ans de suivi. Ou alors on aura seulement la position de couchage des premiers jours à la maternité et les données seront donc de mauvaise qualité.

Le suivi : tous les individus sont suivis selon les mêmes modalités.

- Tous les individus inclus dans la cohorte initiale doivent être suivis selon les mêmes modalités.
- Ils doivent être suivis sur une période similaire.
- Idéalement l'état de santé de tous ces patients doit être connu à la fin du suivi.
- Attention aux perdus de vue : il faut rechercher si le fait d'être perdu de vue peut-être lié à la survenue de la maladie (ex. : si on recherche les facteurs de risque de maladie d'Alzheimer on peut craindre que les patients qui deviennent déments sont justement ceux dont on n'aura plus de nouvelles et qui seront perdus de vue). Il faut donc toujours étudier les causes des arrêts de suivi et comparer les caractéristiques démographiques et cliniques des patients perdus de vue à ceux ayant eu un suivi complet.

Exemple : les femmes de la cohorte EPIDOS sur les facteurs de risque de la fracture du col fémoral recevaient un questionnaire tous les 3 mois afin de savoir si elles avaient changé de domicile ou prévoyaient d'en changer, les coordonnées d'un proche et du médecin traitant avaient été recueillies à l'entrée dans la cohorte ce qui nous a permis de limiter à moins de 5% la proportion des femmes pour lesquelles nous n'avions plus de nouvelles (perdus de vue).

Lorsque le critère de jugement est un événement grave comme le décès, on arrive en général à retrouver les informations et il y a peu de risques de données manquantes dans les pays développés.

LE CRITÈRE DE JUGEMENT

(ISSUE CLINIQUE - ÉVÈNEMENT ÉTUDIÉ - ÉTAT DE SANTÉ ÉTUDIÉ)

- Mesure objective et non biaisée.
- Définition claire et précise des événements/état de santé étudiés avant le début de l'étude.
- Les événements peuvent être :
 - facilement **mesurables sans biais** (décès) : critère de jugement « dur »,
 - nécessiter un jugement **clinique/biologique/radiologique** (infarctus du myocarde, AVC),
 - nécessiter un jugement **subjectif** (qualité de vie, handicap) : critère « mou », plus influençable.
- Pour **minimiser les biais de mesure**, la mesure du critère de jugement doit se faire :
 - en aveugle (insu) du facteur de risque étudié (celui qui mesure le critère de jugement ne connaît pas l'exposition au risque des sujets),
 - avec des outils (ou instruments de mesure) standardisés et validés (questionnaires de qualité de vie SF-36, questionnaire sur la dépression, échelles de mesure de la douleur, de l'incapacité, de l'asthénie...).

Ceci est d'autant plus important que le critère de jugement est subjectif (ou « mou »), l'insu n'est pas nécessaire si le critère de jugement est le décès.

Tous les sujets doivent être soumis au **même suivi** (mêmes procédures diagnostiques, mêmes questionnaires), à intervalles réguliers et identiques pour tous **jusqu'à la fin du suivi**.

En l'absence d'insu : risques de biais de mesure.

Si la personne qui évalue la survenue de la maladie ou l'état de santé sait qu'un individu a été exposé à un facteur de risque.

- Il peut avoir tendance à le suivre de plus près (en particulier si suivi médical par exemple médecins du travail) et on pourra croire que les événements sont plus fréquents alors qu'ils sont seulement mieux recherchés.
- Sa mesure du critère de jugement peut être influencée (surtout si critère subjectif / mou, s'il s'agit de décès le risque est moins grand).

Le critère de jugement est le décès donc tous les décès sont recueillis de la même façon il n'y a pas de risque de biais, en revanche il peut y avoir un biais sur l'évaluation de la cause du décès. On peut imaginer que celui qui fait l'analyse des certificats de décès soit plus tenté d'attribuer le décès à une MSN s'il sait que les NRS étaient couchés sur le ventre. Le risque reste néanmoins très faible dans ce cas car à l'inverse des médicaments, il n'y a pas d'enjeux financiers liés à la position de couchage des enfants... néanmoins, pour lever tout doute il est alors bien de préciser que l'évaluateur était en insu de l'exposition des NRS.

La durée de suivi (fait partie du critère de jugement)

Les investigateurs doivent suivre les patients suffisamment longtemps pour que survienne l'événement. Dans certaines pathologies ce délai peut être long (exemple tabac et cancer du poumon).

a / Étude de cohorte prospective

Le chercheur écrit son projet puis l'étude démarre et les individus sont inclus prospectivement. Le critère de jugement est précisé avant le début de l'étude : ainsi, les méthodes de mesure sont parfaitement adaptées à la question posée. Des informations précises sur les méthodes de mesure, leur reproductibilité et la similitude du suivi pour tous les individus de la cohorte doivent figurer dans l'article, c'est ce qui permet d'évaluer le risque de biais de mesure (suivi, recueil des informations concernant notamment la maladie différentes selon l'exposition).

Dans ce type d'étude, comme dans toute étude dans laquelle les individus sont suivis au cours du temps, attention au biais de suivi et de recueil : Les modalités de suivi et la façon de rechercher et de mesurer le critère de jugement doivent être indépendants de l'exposition (c'est-à-dire identique), que les personnes soient exposées ou non au facteur étudié.

Exemple : il ne faut pas que le groupe des individus exposés ait des visites systématiques alors que le groupe des individus non exposés a des visites seulement en cas de symptômes d'appel.

b / Étude de cohortes exposés/ non exposés

Les individus sont inclus prospectivement ou il s'agit d'une cohorte historique, les deux situations sont possible. L'important dans ce type de design est que les

modalités de suivi et de recueil dans les 2 groupes exposés et non-exposés soient bien précisées dans l'article afin de s'assurer de l'absence de différence entre les 2 groupes. Attention au biais de suivi et de recueil : **possible +++ car le fait d'être exposé peut induire des méthodes de suivi différentes.**

Exemple : au sein du campus de la DOUA, sélection d'un groupe d'exposés à l'acétone (parmi les chimistes) et d'un groupe de non-exposés (parmi les secrétaires universitaires). Ces deux groupes sont totalement différents et il est possible que le médecin du travail ait instauré un suivi plus détaillé et plus rigoureux des chimistes.

c / Étude de cohorte historique

En général le **critère de jugement est mesuré à partir des données qui ont été précédemment recueillies** (dans un but qui était en général différent à l'époque) et dans ce cas il y a peu de risque de biais de mesure.

Donc attention au **manque de précision dans la mesure du critère de jugement et dans les modalités du suivi des individus**, attention au **possible biais de suivi** si les patients exposés avaient systématiquement un suivi différent des patients non-exposés.

Il est possible qu'au moment où le chercheur décide de faire l'étude, des individus aient déjà développé la maladie étudiée. Il est important que le lecteur n'ait pas de doute sur le fait que la sélection rétrospective des individus étudiés ait été faite par le chercheur sans savoir quels individus ont développé la maladie au cours du suivi (autrement dit sélection des individus indépendante du critère de jugement). Dans certains cas l'étude de cohorte est à la fois historique mais avec également un suivi prospectif et on se retrouve dans le même cas que la cohorte prospective.

LE FACTEUR DE RISQUE

Comment est mesurée l'exposition au facteur de risque supposé ?

a / Étude de cohorte prospective

- **L'exposition au facteur de risque est mesurée avant la survenue de la maladie** donc la mesure de l'exposition ne peut pas être influencée par le fait que l'individu a développé la maladie ou non (pas de risque de biais de mesure de l'exposition).

- La question de recherche est formulée avant de recueillir les données sur l'exposition, ce qui autorise un recueil précis et standardisé des facteurs de risque étudiés qui doit être détaillé dans l'article.

Dans ce type d'étude il n'y a pas de risque que la mesure de l'exposition au facteur de risque soit influencée par la survenue de la maladie étudiée.

Par exemple, si l'on étudie l'association entre consommation de café et survenue d'un diabète il n'y a pas de risque de surestimer la consommation de café chez les diabétiques puisque lorsqu'on recueille la consommation de café, aucun individu de la cohorte n'a encore développé un diabète.

b / Étude exposés / non exposés (« double cohorte »)

Dans ce type d'étude il n'y a pas de risque que la mesure de l'exposition soit influencée par la survenue de la maladie étudiée si l'étude est prospective (idem paragraphe précédent).

Si on utilise des cohortes exposés/non exposés historiques, se référer au paragraphe suivant.

c / Étude de cohorte historique

Dans ce type d'étude le risque que la mesure de l'exposition soit influencée par la connaissance du critère de jugement (développement de la maladie) est limité car l'exposition au facteur de risque a été faite avant la survenue possible de la maladie. Néanmoins même à partir de données rétrospectives il est toujours possible que leur interprétation soit biaisée si le chercheur sait si l'individu a finalement développé la maladie.

De plus, au moment où le chercheur décide de faire l'étude, la cohorte a déjà été constituée dans le passé et on doit se **contenter des données sur l'exposition qui ont été recueillies précédemment** (donc avant qu'on se pose précisément la question de recherche actuelle), souvent on ne dispose pas de toute les informations dont on aurait besoin et ceci peut générer des imprécisions).

Remarque valable quel que soit le type d'étude de cohorte analytique : Si dans toutes les études de cohorte, le recueil de l'exposition au facteur de risque précède le développement de la maladie il faut tout de même remarquer, que ce recueil porte souvent sur des expositions passées. Par exemple si l'on mesure le lien entre tabac et cancer du poumon, le plus souvent l'exposition au tabac n'est pas suivie prospectivement pendant 15 ans mais on reconstitue *a posteriori*

l'exposition en fonction de son intensité et de sa durée. Pour certains facteurs de risque en revanche on peut faire un recueil prospectif de l'exposition, c'est le cas de certaines grandes cohortes qui sont démarrées chez des sujets jeunes et qui sont suivies sur plusieurs dizaines d'années.

LES BIAIS ET FACTEURS DE CONFUSION

L'étude de cohorte est, parmi les études observationnelles, celle qui a le moins de risque de biais.

Risque de biais par ordre croissant : Cohorte prospective << cohorte historique << cohorte exposés/non exposés << cas-témoin nichées dans une cohorte << cas-témoins (cas incidents) << cas-témoins (cas prévalents).

Les principaux risques de biais dans les études de cohorte sont ceux liés aux perdus de vue car on redoute toujours que le fait d'être perdu de vue soit lié à l'exposition.

Type de biais susceptibles de se produire	Études concernées	Procédure pour limiter les risques de biais
BIAIS DE SÉLECTION		
À l'inclusion. Biais de sélection non différentiel :		
Les individus étudiés ne sont pas représentatifs de la population générale.	Toutes les études de cohorte.	Assez inévitable car seuls les volontaires participent. Pas très grave car s'applique à l'ensemble de la cohorte, ne menace pas la validité interne mais seulement l'extrapolabilité des résultats.
Biais de sélection différentiel :		
Les groupes sont différents en termes de facteurs de risque de la maladie étudiée ou de facteurs pronostiques de l'évolution de la maladie.	Seulement les cohortes exposés/non-exposés (un groupe d'individus sélectionnés sur la base de l'exposition vs un gp sélectionné sur la base de non-exposition).	Définition précise des critères d'inclusion et recrutement des deux groupes au sein de populations comparables en termes des autres facteurs de risque.

Type de biais susceptibles de se produire	Études concernées	Procédure pour limiter les risques de biais
BIAIS DE SELECTION		
Pendant le suivi :		
<p>Si nombreux perdus de vue, les groupes en fin du suivi (donc au moment de l'analyse) sont différents des groupes constitués au départ.</p> <p>Problème +++ si sont perdus de vue pour raison liée à la maladie</p>	<p>Problème n°1 de toutes les études de cohortes. (et plus globalement toute étude dans laquelle il y a un suivi des individus)</p>	<p>Méthode de suivi très rigoureuse</p> <p>(contacts fréquents et programmés régulièrement par multiples moyens, recherche des événements par plusieurs sources...)</p> <p>Etudier les causes des arrêts de suivi et comparer les caractéristiques cliniques et démographiques des patients perdus de vue à ceux ayant eu un suivi complet</p>
BIAIS DE MESURE		
Biais d'information		
<p>Recueil de l'exposition différent selon les groupes (mensonge sur l'exposition chez les exposés : sous-estimation si facteur de risque « politiquement incorrect » et surestimation si indemnisation).</p>	<p>Toutes les études de cohorte.</p>	<p>Pas très grave car en théorie non lié à la survenue de l'évènement de santé étudié</p> <p>Recueil fiable et standardisé, formation des enquêteurs.</p>
Pendant le suivi :		
<p>++ Survenue de la maladie au cours du suivi recherchée de façon différente selon les groupes (tous les cas connus chez les exposés vs seulement les cas symptomatiques chez les non-exposés).</p>	<p>Possible dans toutes les cohortes mais risque plus élevé pour les cohortes historiques et les cohortes exposés/non-exposés.</p>	<p>Suivi identique dans les deux groupes (même fréquence, même examens cliniques et para-cliniques)</p> <p>Recueil de la survenue de la maladie en insu du groupe d'exposition</p>

BIAIS DE CONFUSION		
Biais dans l'estimation du risque relatif dû à la non prise en compte de facteurs de confusion.	Toutes les études.	Recueil des autres facteurs de risques de la maladie (facteurs de confusion potentiels) et prise en compte dans l'analyse (analyse multivariée).

LES ANALYSES STATISTIQUES

Calcul du nombre de sujets nécessaires

La taille de l'échantillon détermine la puissance de l'étude, ici la capacité de l'étude à mettre en évidence l'impact du facteur de risque sur la maladie s'il existe réellement.

Une puissance de 80 % signifie que si le facteur étudié est bien un facteur de risque on a 80 chances sur 100 de le montrer avec cette étude, et donc 20 % de risques de conclure à tort que ce n'est pas un facteur de risque (c'est le risque beta). Le risque alpha est la probabilité de conclure que le facteur étudié est un facteur de risque alors qu'il ne l'est pas en réalité.

Ce calcul doit avoir été réalisé *a priori* à partir d'une hypothèse qui doit être clairement exprimée et justifiée et une puissance de l'étude au moins égale à 80 % (si possible 90 %).

Exemple, sur quels éléments calcule t'on la taille de l'échantillon nécessaire pour confirmer l'hypothèse suivante ? :



Le facteur de risque



Patients étudiés

La position de couchage ventrale chez des nourrissons âgés de 1 à 12 mois est associée à une augmentation d'au moins 20 % (Risque relatif de 1,2) du taux d'incidence des morts subites confirmées par autopsie par rapport aux autres couchages (latéral ou dorsal).



Comparaison au groupe de référence



L'effet attendu sur le critère de jugement

Ce risque relatif représente la différence entre les groupes (couchage ventral versus autres) minimum cliniquement intéressante et qui est compatible avec les données de la littérature.

On choisit une **puissance de 90%** (risque beta de 10%) et un **risque alpha de 5%**, et on pose l'hypothèse que l'incidence chez les enfants non exposés (couchage latéral ou dorsal) sera de 0,4/1000 nourrissons vivants/an et que le RR sera de 1,2 au moins à partir des données de la littérature.

REMARQUE : à partir de ces éléments le calcul de la taille de l'échantillon repose sur des formules qui en général ne sont pas présentées et qu'on ne vous demande pas de connaître.

Mesure d'association (entre Facteur de risque et maladie étudiée)

Le plus souvent dans les études de cohorte l'association entre facteur de risque et maladie est estimée par un risque relatif avec son intervalle de confiance à 95 %.

	Développent la maladie au cours du suivi	Ne développent pas la maladie au cours du suivi	
Exposés	A	B	Risque maladie si exposés = $A/A+B$
Non-exposés	C	D	Risque maladie si non-exposés = $C/C+D$
Risque Relatif = Risque chez les exposés / Risque chez les non-exposés.			

Dans certains cas, il existe plusieurs niveaux d'exposition. Pour l'analyse, on répartit les sujets en différentes classes en fonction de la quantité et/ou de la durée de leur exposition. Dans ce cas, on choisit une classe de référence, en général la catégorie la moins exposée, et le risque de toutes les autres classes d'exposition est exprimé par rapport au risque de base observé dans la classe de référence.

Remarque, par définition le risque relatif de cette classe de référence est de 1.

Exemple : risque relatif de cancer du poumon associé à plusieurs niveaux d'exposition au tabac.

Exposition cigarettes par jour	Personnes - années à risque	Cancers du poumon	Incidence /1000 p.a ¹	Risque relatif	ICC 95%	P tendance
> 25	25100	57	2,27	37,8	[33,5 – 41,6]	< 0.05
15 - 24	38900	54	1,39	23,2	[19,8 – 27,3]	
1 - 14	38600	22	0,57	9,5	[4,4 – 14,5]	
0	48800	3	0,06	1	catégorie choisie comme référence	

¹ p.a. = personnes-années.

Le Risque Relatif de la catégorie [15-24] = 23.2 => leur risque de cancer du poumon (1.39/1000 pa) est en moyenne multiplié par 23,2 **par rapport aux non-fumeurs** (0.06/1000 pa) soit $RR = 1,39/0,06 = 23,2$.

De même le RR de la catégorie >25 est le risque dans cette catégorie par rapport au risque de base = celui des non-fumeurs.

Dans cet exemple, les risques relatifs sont bien croissants d'une catégorie d'exposition à la suivante, et leurs intervalles de confiance ne se chevauchent pas, donc on peut en déduire qu'il y a bien un « **effet-dose** » **statistiquement significatif**. Néanmoins, la situation habituelle est souvent plus délicate car les intervalles de confiance se chevauchent d'une catégorie à une autre. Dans ces cas, seul le résultat du test de tendance permet de conclure s'il existe un effet-dose statistiquement significatif.

Lorsque l'on s'intéresse à l'évolution du risque en fonction de différents niveaux d'exposition (« l'effet-dose »), le test statistique approprié est le **chi-carré de tendance** qui permet de calculer le « **p tendance** » (p for trend en anglais).

Incidences et RR (avec leurs IC à 95 %) de SCA observés en fonction de l'IMC chez 28 991 femmes et 25 792 hommes

	IMC, kg/m ²			
	< 25	25 - 29,9	≥ 30	Par palier de 1 kg/m ²
FEMMES				
Incidence, cas (n)*	87 (101)	139 (106)	199 (62)	
RR brut (IC à 95 %)	1 (réfèrent)	1,48 (1,13 - 1,95)	2,08 (1,52 - 2,86)	1,05 (1,02 - 1,07)
RR après ajustements multivariés (IC à 95 %)**	1 (réfèrent)	1,54 (1,17 - 2,03)	2,06 (1,49 - 2,86)	1,05 (1,03 - 1,07)
Avec prise en compte des facteurs cliniques (rapportés par l'intéressé)***	1 (réfèrent)	1,33 (1,00 - 1,76)	1,63 (1,16 - 2,29)	1,03 (1,01 - 1,06)
Avec prise en compte des facteurs cliniques****	1 (réfèrent)	1,33 (1,00 - 1,76)	1,56 (1,11 - 2,20)	1,03 (1,01 - 1,06)
HOMMES				
Incidence, cas (n)*	340 (233)	456 (441)	644 (184)	
RR brut (IC à 95 %)	1 (réfèrent)	1,32 (1,12 - 1,54)	1,88 (1,55 - 2,28)	1,06 (1,05 - 1,08)
RR après ajustements multivariés (IC à 95 %)**	1 (réfèrent)	1,40 (1,19 - 1,64)	1,93 (1,58 - 2,35)	1,07 (1,05 - 1,08)
Avec prise en compte des facteurs cliniques (rapportés par l'intéressé)***	1 (réfèrent)	1,29 (1,09 - 1,52)	1,64 (1,33 - 2,02)	1,05 (1,03 - 1,07)
Avec prise en compte des facteurs cliniques****	1 (réfèrent)	1,19 (1,01 - 1,40)	1,43 (1,15 - 1,76)	1,04 (1,02 - 1,06)

* Incidence pour 100 000 années-personnes.

** Les modèles avec ajustements multivariés incluait l'activité physique (< 1, 1 à 3,5 et ≥ 3,5 h/semaine), le statut en matière de tabagisme (sujet n'ayant jamais fumé, ancien fumeur, arrêt récent, consommation présente comprise entre 1 et 14 g de tabac par jour, entre 15 et 24 g ou supérieure à 24 g et ancienneté du tabagisme), la durée de scolarité (inférieure à 8 ans, comprise entre 8 et 10 ans ou supérieure à 10 ans), le score de régime méditerranéen (3 degrés) et la consommation d'alcool (3 degrés). Les analyses ayant porté sur les femmes ont également été ajustées en fonction de leur statut en terme de ménopause et d'observance d'un éventuel traitement hormonal substitutif (préménopause, péri-ménopause, ménopause avérée sans prescription de traitement hormonal substitutif et ménopause avérée avec prescription d'un traitement hormonal substitutif).

*** Il s'agit du modèle multivarié décrit ci-dessus dans lequel des ajustements supplémentaires ont été pratiqués en fonction des indications fournies par les participants quant à l'éventuelle présence chez ces derniers d'une hypertension artérielle, d'une hypercholestérolémie et/ou d'un diabète médicalement documenté (oui/non).

**** Il s'agit du modèle multivarié décrit ci-dessus dans lequel des ajustements supplémentaires ont été pratiqués en fonction des chiffres tensionnels systoliques et diastoliques (variables continues), de la cholestérolémie (variable continue) et des indications fournies par les participants quant à l'éventuelle présence chez ces derniers d'un diabète médicalement documenté (oui/non).

Modèles de régression

Les risques relatifs (RR) sont le plus souvent estimés à partir de modèles de régression. Un des modèles fréquemment utilisé est la régression logistique dans laquelle la variable à expliquer (Y) représente un événement à deux modalités de réponse (variable binaire) oui/non. C'est le cas d'un grand nombre d'événements de santé tels que la fracture, l'infarctus, l'embolie pulmonaire, la chute...

Remarque : il est à remarquer qu'en fait, les modèles de régression logistique donnent un résultat sous forme d'un Odds Ratio (OR) et non d'un Risque Relatif (RR). Mais à ce stade cela ne modifie pas votre façon de raisonner. Nous verrons à un autre endroit du polycopié les relations entre OR et RR.

La régression logistique permet de mesurer l'association entre le facteur de risque supposé (variable « explicative » : X_{FDR}) et l'événement de santé (variable « à expliquer » : Y) tout en tenant compte des autres facteurs de risque qui peuvent se comporter comme des facteurs de confusion (autres variables « explicatives » dans le modèle par exemple : X_{age} , X_{sexe} , X_{tabac} , X_{alcool} , X_{BMI} ...

Lorsque seulement une variable explicative (facteur de risque supposé) est entrée dans la régression, on parle de régression univariée et le risque relatif que l'on obtient est un risque relatif « brut » ou « non ajusté » ($Y = \beta_1 X_{FDR}$) (appelé « crude » RR en anglais).

Lorsque plusieurs autres variables explicatives, correspondant aux différents facteurs de confusion possibles, sont entrées dans l'analyse, on parle alors d'analyse multivariée et le risque relatif obtenu est un risque relatif « ajusté » (« adjusted » en anglais, RRadj ou OR adj) pour toutes les variables explicatives ajoutées dans la régression. L'analyse multivariée permet d'estimer l'association entre le facteur de risque supposé et la maladie en éliminant l'effet des possibles facteurs de confusion (dans la survenue de l'évènement Y quel est le poids β_1 du FDR = $\beta_1 X_{FDR}$) une fois annulé l'effet des autres facteurs ($Y = \beta_1 X_{FDR} + \beta_2 X_{age} + \beta_3 X_{sexe} + \beta_4 X_{tabac}$).

Une autre méthode d'analyse est également souvent utilisée dans les études de cohortes, il s'agit des analyses de « survie ». Dans ces analyses on prend en compte le temps de suivi de tous les patients jusqu'à la survenue de l'évènement pour ceux qui développent la maladie étudiée, jusqu'à la date du décès pour ceux qui décèdent avant la fin du suivi, jusqu'à la date des dernières nouvelles pour

ceux qui sont perdus de vue avant la fin programmée du suivi et pour les autres, jusqu'à la fin du suivi programmé.

L'analyse repose alors sur des personnes-temps (personnes-années, personnes-mois...). Un modèle classiquement utilisé est le modèle de Cox. À partir de ce type d'analyse sont souvent présentés les risques relatifs ou Hazard Ratios (HR).

LES RÉSULTATS (ESTIMATION DE LA FORCE DE L'ASSOCIATION ET SIGNIFICATION STATISTIQUE)

Comme vu au chapitre 4.7 Les résultats (généralités), deux notions sont importantes pour interpréter les résultats :

- l'association est-elle **statistiquement significative** (intervalle de confiance à 95 % du RR et « p ») et...
- est-elle **forte** (valeur du RR ou du HR).

On attend de vous que vous soyez capables de donner une explication en français des différents résultats notamment concernant le RR et son ICC 95 % (voir tableau suivant sur les différents indicateurs pour mesurer et quantifier l'effet d'un facteur de risque). La conclusion doit porter sur le résultat le moins biaisé c'est-à-dire celui qui est issue des analyses multivariées (surtout si les résultats sont différents de ceux des analyses univariées).

Bien entendu la suspicion de biais majeur ou de non prise en compte de facteurs de confusion importants peut remettre en cause la validité des résultats, ces aspects sont vus au paragraphe suivant.

Dans les études « étiologiques, un troisième aspect doit être connu. Une fois établie l'association statistiquement significative ainsi que la force de cette association c'est en définitive une troisième question qui nous intéresse : **Comment prouve-t-on un lien de CAUSALITE entre un facteur de risque et une maladie ?**

Une des raisons de rechercher un facteur de risque est l'espoir que sa suppression permette de réduire l'incidence de la maladie. Or, s'il n'y a pas de lien de causalité, on ne peut espérer une réduction de la maladie par la réduction de l'exposition au facteur de risque.

Si une étude a montré une association statistique significative entre le risque de maladie et l'exposition au facteur considéré,

ET si cette étude ne comporte pas de biais importants,

ALORS ceci est un élément de présomption que ce facteur est réellement un facteur de risque.

MAIS, la preuve de la nature causale de la relation entre facteur de risque et survenue de la maladie n'est pas démontrée, seul un essai randomisé pourrait apporter la preuve... ce qui n'est pas envisageable (non éthique d'exposer volontairement des individus à un facteur de risque potentiel).

Pour étayer l'hypothèse d'une relation de causalité entre un facteur de risque et une maladie, différents arguments sont pris en compte : critères décrits par Hill.

Critères de Hill internes à l'étude :

- Forte **intensité** de l'association (Risque relatif/Hazard ratio/Odds ratio élevé),
- Existence d'une **relation de type « dose-effet »** entre l'exposition et la maladie,
- **Spécificité** (+/-) de relation exposition <-> maladie (par exemple amiante et mésothéliome),
- **Chronologie** (facteur de risque précède la survenue de la maladie).

Critères de Hill externes à l'étude (bibliographie) :

- **Concordance** entre les résultats des précédentes études (**Fil rouge : 17 études cas-témoins concluaient à un risque plus élevé chez les NRS couchés sur le ventre**),
- **Plausibilité** biologique (existence de mécanismes d'actions biologiques et physiopathologiques connus avant de mener l'étude),
- Concordance avec les **expérimentations** menées in vitro ou chez l'animal,
- **Diminution de l'incidence de la maladie lorsque l'exposition est supprimée** ou réduite (**Fil rouge : diminution du nombre de morts subites du nourrisson après la réalisation de campagnes visant à faire dormir les nourrissons sur le dos**).

Différents indicateurs pour mesurer et quantifier l'effet d'un facteur de risque

Quelle est la force de l'association ? De combien le risque est-il augmenté lorsqu'une personne est exposée au facteur de risque par rapport aux personnes non exposées ?

Exemple : Étude de cohorte avec critère de jugement principal binaire : survenue d'un AVC : oui/non et facteur de risque étudié : prise de pilules contraceptives.

Risque de base (R) (personnes non exposées)	$10/100 = 10\% (0.10)$
Risque si exposée à la pilule (RE)	$15/100 = 15\% (0.15)$
Risque relatif ($RR = RE/R$)	$0.15 / 0.10 = 1.5$
Augmentation absolue de risque ($RE - R$)	$0.15 - 0.10 = 5\% (0.05)$
Augmentation relative du risque $(RR - 1) \times 100$	$(1.5 - 1) \times 100 = 50\% (0.5)$
Nombre de sujets exposés pour observer un AVC supplémentaire : $N = 1/(RE - R)$	$1 / 0.05 = 20$

Traduction du résultat « en français » :

Le risque de base dans une population de femmes non exposées à la pilule est de 10 % (soit un AVC sur 10 patients).

Chez les femmes qui prennent la pilule, ce risque est de 15 % (soit un AVC sur 15 patientes).

Autrement dit le RR d'AVC lié à ce facteur de risque est de 1,5 : les femmes qui prennent la pilule ont en moyenne un risque multiplié par 1,5 par rapport à celles qui ne prennent pas la pilule.

Autrement dit l'augmentation absolue du risque d'AVC lié à la prise de la pilule est de 5 % ce qui signifie que pour 100 personnes exposées on aura 5 cas d'AVC supplémentaires.

On peut dire aussi que l'augmentation relative du risque est de 50 % (15 % représentent bien une augmentation de 50 % par rapport à 10 %), on voit que cette façon d'exprimer les résultats peut être trompeuse.

Enfin une autre façon d'exprimer le résultat est que l'exposition de 20 personnes à ce facteur de risque est suffisante pour créer un cas d'AVC supplémentaire.

LA CONCLUSION

Comme dans les autres types d'études la conclusion doit tenir compte des **résultats observés**, du **design** de l'étude et de ses **limites** (biais possibles, limites de puissance statistique...) qui vont influencer le **niveau de preuve** de la publication, viennent ensuite les notions de validité externe (résultats extrapolables dans notre contexte français ?) et de cohérence externe en fonction des **données déjà publiées**, enfin la notion de causalité doit également être argumentée.

Les messages clefs

- Les individus qui entrent dans une étude de cohorte sont indemnes de la maladie étudiée.
- Les études de cohortes peuvent être prospectives ou historiques.
- Les études de cohortes ont un plus fort niveau de preuve que les études cas-témoins car elles comportent moins de risque de biais.
- Le principal point à vérifier est la qualité du suivi et sa similitude entre les groupes ainsi que la proportion des perdus de vue.
- Comme dans les autres études observationnelles il est impératif que les facteurs de confusion aient été pris en compte dans l'analyse.

	Cohorte prospective	Cohorte historique	Cohorte groupe exposé / groupe non-exposé (issus de cohortes différentes)
Biais de sélection à l'inclusion	Peu de risque	Peu de risque	Risque + si exposés ≠ non exposés
Biais de sélection au cours du suivi (perdus de vue)	Risques +	Risques +	Risques +
Biais de mesure de l'exposition au FDR	Risque = 0, Exposition mesurée avant survenue évènement	Risque très faible Exposition recueillie avant évènement,	Risque = 0, Exposition mesurée avant survenue évènement
Biais de mesure du critère de jugement / biais de détection	Risque ± (selon CJ) ↘ si évaluation en insu de l'exposition et si CJ « dur »	Risque faible CJ recueilli avant question de recherche	Risque ++ car exposés souvent mieux suivis ↘ ± si évaluation en insu de l'exposition et si CJ « dur »
Qualité de mesure de l'exposition au FDR	La meilleure possible	Limité aux données existantes ± adapté à la question	La meilleure possible
Qualité de mesure du critère de jugement	La meilleure possible	Limité aux données existantes ± adapté à la question	La meilleure possible

6.2 LCA DES COHORTES VISANT À ANALYSER L'ÉVOLUTION D'UNE MALADIE ET SES FACTEURS PRONOSTIQUES

NIVEAU 2

Le pronostic se réfère au devenir (à l'évolution) d'une maladie.

LA QUESTION

Dans un groupe de patients chez lesquels on vient de diagnostiquer une maladie, afin d'adapter le traitement, j'ai besoin de savoir :

- la façon dont la maladie est susceptible d'évoluer en termes de survie et de complications,
- et/ou quels patients ont un « bon pronostic » ou un « mauvais pronostic ».

Certaines caractéristiques des patients et/ou de leur maladie ont un lien avec l'évolution bonne ou moins bonne de la maladie, elles peuvent être utilisées pour prédire de manière plus précise le devenir des patients. Ces caractéristiques sont appelées des facteurs pronostiques.

Exemple : un patient ayant une tumeur de stade T2 lors du diagnostic a un moins bon pronostic qu'un patient porteur d'une tumeur de stade T1 car son risque de récurrence locale et métastatique est plus élevé.

Évolution de la maladie en fonction de la présence d'un facteur pronostique

Exemple de question :

Un **volume tumoral supérieur à 2 cm** augmente-t-il de façon significative le risque de **récurrence métastatique dans les 5 ans** dans une population de **femmes âgées de 50 à 70 ans avec un premier cancer du sein non métastatique** ?



REMARQUE 1 : parfois, la question est seulement d'observer une cohorte de patients atteints de cette maladie pour analyser la durée et la qualité de la survie spontanée (« l'histoire naturelle de la maladie »). Dans certaines études on aura donc seulement une description de la survie.

Néanmoins le plus souvent cette survie est ensuite mesurée et comparée en fonction de la présence ou l'absence de certains facteurs supposés pronostique et on retrouve la problématique de l'association entre facteurs pronostiques et évolution d'une maladie.

REMARQUE 2 : on cherche constamment à prédire le risque de récurrence le plus précisément possible et on est souvent amené à combiner plusieurs facteurs en un « score de risque », plusieurs méthodes statistiques peuvent être utilisées mais sont au-delà de l'objectif de ce polycopié

Histoire naturelle de la maladie : événements susceptibles de se produire au cours de l'évolution de cette maladie et leur fréquence

Exemple de question :



Chez des patientes atteintes d'un premier cancer du sein invasif non métastatique, quelle est l'incidence des récurrences métastatiques à 5 ans ?



LE TYPE D'ÉTUDE

Ce sont DES ÉTUDES DE TYPE COHORTE : une « cohorte » d'individus que l'on suit prospectivement pour observer la SURVIE ou la SURVENUE D'ÉVÉNEMENTS de santé (récidives, complications...).

Par rapport aux études de cohorte à visée « étiologique » elles présentent les MÊMES FORCES et les MÊMES LIMITES.

Les principales DIFFÉRENCES :

- La population étudiée est une population de malades et non de sujets sains.
- Ces études sont utiles pour définir des stratégies thérapeutiques plus ou moins agressives en fonction du pronostic. Par Exemple, c'est ce type d'étude qui permet de décider une chimiothérapie et/ou radiothérapie en plus de la mastectomie dans le cancer du sein en fonction du stade initiale du cancer qui est un facteur pronostique.

REMARQUE : certains auteurs recherchent des facteurs pronostiques à partir d'études cas-témoins... ce n'est vraiment pas un design d'étude approprié... mais cela existe.

LA POPULATION ÉTUDIÉE

Les patients étudiés sont-ils à un stade identique de leur maladie (et si possible précoce) ?

Les maladies n'ont pas la même évolution chez tous les patients : Un cancer de la prostate peut évoluer vers un décès sur 20 ans chez certains patients ou sur 1 an chez d'autres.

Les pathologies chroniques s'aggravent au cours du temps (diabète, pathologies cardio-vasculaires...), un patient diabétique a un moins bon pronostic vital lorsque son diabète évolue depuis 20ans que lorsqu'il vient de débiter.

Pour que l'étude porte bien sur tout l'éventail possible des formes d'une maladie : Il est préférable que les patients soient inclus dans l'étude à un **stade identique** de la maladie et le **plus précocement possible** par rapport aux premiers symptômes.

Cela est d'autant plus vrai qu'il existe une grande variabilité de formes de la maladie, et vice versa, cela dépend donc de la maladie étudiée.

Dans la section Méthodes on doit trouver une description précise du stade d'évolution de la maladie au moment de l'inclusion des patients dans l'étude (exemple : **premier infarctus, premier cancer du sein invasif non métastatique, etc.**)

Les auteurs doivent décrire la durée d'évolution de la maladie lors de l'inclusion dans l'étude, puisque celle-ci est souvent associée au pronostic dans un sens ou dans l'autre.

Exemple 1 : on inclut dans une étude tous les patients consultant en urologie en 2011 pour un cancer de la prostate, quel que soit son stade et sa date de début (qu'il date de 10 ans ou d'1 mois).

Les patients avec un cancer plus agressif sont déjà morts (tous ceux dont la survie a été de 12 mois sont décédés sauf les cancers apparus en 2011 et fin 2010). On a donc dans cette cohorte une sur-représentation des patients porteurs des formes les moins agressives. Conséquences : si certains facteurs pronostiques sont liés uniquement aux formes les plus agressives, on ne pourra pas les identifier.

Exemple 2 : on peut avoir la situation inverse, si on étudie des patients porteurs d'une pathologie chronique (une polyarthrite rhumatoïde, un diabète, une bronchopneumopathie obstructive...) le pronostic est plus défavorable chez les patients dont la maladie évolue depuis plus longtemps.

Les patients étudiés sont-ils représentatifs de tous les patients présentant cette maladie ?

CRITÈRES A PRIORI (conditions du recrutement dans la cohorte).

Les patients devraient aussi être représentatifs que possible de l'ensemble des cas dans la population.

Par exemple les patients d'un **centre hospitalier spécialisé réputé** vont avoir des formes particulièrement graves et/ou complexes de la maladie, et donc un pronostic systématiquement plus mauvais que des malades suivis en consultations.

Des patients suivis par des **spécialistes de ville** auront peut-être un pronostic différent de celui des patients suivis en **médecine générale** (ceci peut également varier en fonction de la zone d'habitation rurale ou urbaine).

Dans la section Méthodes on doit trouver une **description précise du contexte** dans lequel les patients sont recrutés ; cabinet de ville, hôpital général, CHU ou

centre anticancéreux, centres de référence (centres hyperspécialisés)

En THÉORIE la seule manière de bien sélectionner la cohorte exhaustive consisterait à inclure systématiquement tous les cas éligibles à partir de toutes les filières de soins d'une région géographique. Il s'agit de repérer non seulement les cas existants à partir de leur dossier clinique mais aussi les nouveaux cas par la mise en place d'une surveillance.

En pratique, le plus souvent ce n'est pas le cas car ce serait très compliqué à réaliser.

Le minimum qu'on pourrait attendre est que TOUS LES PATIENTS CONSÉCUTIFS soient inclus dans l'étude.

Exemple 1 : « Les patients opérés pour un cancer du sein dans le service de chirurgie générale et oncologique de l'Hôpital Edouard Herriot entre Janvier 2004 et Décembre 2007 ont été inclus prospectivement dans l'étude. Nous avons exclu les patients qui avaient des métastases au moment du diagnostic et les patients qui avaient reçu une chimiothérapie néoadjuvante. »

Exemple 2 : « Deux cents patients avec un ostéosarcome de haut grade, non métastatique ont été inclus de façon prospective dans sept établissements de soins tertiaires entre 2000 et 2010, tous les patients traités consécutivement n'ont pas pu être inclus et l'échantillon étudié représente donc un sous-groupe des patients éligibles de chaque établissement. Tous les patients étaient nouvellement diagnostiqués et leur diagnostic était confirmé par biopsie.≈»

Exemple 3 : « L'étude a inclus 802 patients consécutifs avec un adénocarcinome colorectal stade II (n = 441) ou stade III (n = 361) ayant subi une résection chirurgicale curative entre Janvier 2000 et Décembre 2006 à l'Hôpital Hamilton (Toronto, Canada). Les patients âgés de plus de 85 ans et ceux porteurs de multiples carcinomes synchrones du colon, de maladie intestinale inflammatoire idiopathique, ou d'une polypose adénomateuse familiale ont été exclus. Les patients qui ont reçu une radiothérapie préopératoire et ceux avec une tumeur maligne détectée dans les 5 années précédentes ont aussi été exclus. »

A POSTERIORI (les patients qui constituent finalement la cohorte étudiée)

Les caractéristiques de l'échantillon étudié (âge, sexe, degré de sévérité de la maladie, présence ou non de facteurs pronostiques...) permettent au lecteur de savoir si les patients étudiés ressemblent à ses patients (validité externe).

Le suivi : tous les patients sont suivis selon les mêmes modalités

- Tous les patients inclus dans la cohorte initiale doivent être **suivis selon les mêmes modalités**.
- Ils doivent être **suivis sur une période et une durée similaire**.
- Idéalement **l'état de santé de tous** ces patients doit être **connu à la fin du suivi**.
- **Attention aux perdus de vue** car cela pourrait être lié au pronostic de la maladie : **Le pronostic des perdus de vue sera alors différent** de celui du reste de la cohorte. Les patients ne quittent pas en général une étude sans raison ils la quittent parce qu'ils sont guéris (et ne voient plus l'intérêt d'un suivi à l'hôpital), ou au contraire ont justement l'événement étudié, ou sont décédés).
- Comme dans toute étude avec un suivi individuel des patients, l'évolution des effectifs (entre inclusion et analyse) doit être clairement présentée, au mieux dans un diagramme de flux.
- **L'influence du nombre de perdus de vue** sur la validité de l'étude **dépend** de la **fréquence** de survenue de **l'événement** étudié et des raisons pour lesquelles ils sont perdus de vue (mais rarement connues...). Par exemple, une proportion de **5%** de perdus de vue aura **peu de conséquence** si **l'événement** survient chez **60% des patients**. Elle **remettra en cause la validité de l'étude** si **l'événement est rare** et survient chez seulement **10%** des patients.
- Il est également souhaitable d'étudier les causes des arrêts de suivi et de comparer les caractéristiques démographiques et cliniques des patients perdus de vue à ceux ayant eu un suivi complet (mais ce n'est pas suffisant si le nombre de perdus de vue est trop important).

Exemples d'articles

Exemple 3 (suite) : Parmi les 802 patients, 41 (20 stade II et 21 stade III) sont décédés en postopératoire et 13 (8 stade II et 5 stade III) ont été perdus de vue. 20 patients supplémentaires (14 stade II et 6 stade III) ont été exclus parce que les échantillons tumoraux pour l'analyse immunohistochimique n'étaient pas disponibles. Enfin, 10 (1,4%) des 728 patients examinés étaient exclus de l'étude parce

que la qualité de l'histologie a été jugée insatisfaisante.

« Tous les patients étaient suivis régulièrement pendant au moins 24 mois à partir du diagnostic ou jusqu'à une récurrence métastatique. Aucun patient n'a été perdu de vue ».

Les patients ont été suivis prospectivement chaque année. Des attachés de recherche clinique formés ont revu les dossiers et enregistré les événements d'intérêt, ainsi que les changements de traitement à l'aide d'un formulaire de recueil standardisé. Des entretiens téléphoniques, avec le patient ou un parent lorsque le patient était décédé, ont été réalisés pour compléter les données sur l'évolution de la maladie et le statut vital. 8 patients sur 772 ont été perdus de vue (1 %). En Décembre 2005, 357 patients (46,7 %) étaient vivants et indemnes de récurrence ou de progression (durée médiane de suivi, 62,6 mois, extrêmes 1 mois - 98 mois). Tous les décès (n = 200) ont été enregistrés. La survie a été calculée comme la période comprise entre le diagnostic et le décès ou le dernier contrôle.

« La majorité des patients était suivie dans le service d'oncologie de l'hôpital St Luc selon un protocole standardisé. Pour les autres patients, l'information concernant l'évolution clinique (et donc le critère de jugement) a été obtenue à partir de l'examen des dossiers et / ou d'entretiens téléphoniques directs avec les médecins traitants des patients. Le suivi moyen des patients survivants était de 93,9 mois (médiane 90,5 mois, extrêmes de 63 à 144 mois) ».

LE CRITÈRE DE JUGEMENT

(ISSUE CLINIQUE - ÉVÈNEMENT ÉTUDIÉ - ÉTAT DE SANTÉ ÉTUDIÉ)

Il doit idéalement présenter les caractéristiques suivantes :

- Mesure objective et non biaisée.
- Définition claire et précise des événements étudiés avant le début de l'étude.
- Les événements peuvent :
 - être facilement **mesurables sans biais** (décès) critère de jugement « dur »,
 - nécessiter un jugement **clinique/biologique/radiologique** (récidive d'infarctus du myocarde, récurrence d'AVC, survenue d'une épilepsie après un AVC, évolution métastatique d'un cancer, augmentation de taille des métastases, progression d'une sclérose en plaque...),
 - nécessiter un jugement **subjectif** (qualité de vie, handicap), critère « mou », plus influençable.
- Pour minimiser les biais de mesure,
 - la mesure du critère de jugement doit se faire **en aveugle (insu) des facteurs pronostiques étudiés** (celui qui mesure le critère de jugement

ne sait pas si le facteur pronostique était présent),

- avec des **outils (ou instruments de mesure) standardisés et validés** (critères explicites pour le diagnostic d'AVC, critères explicites pour mesurer les métastases pour établir la progression d'une sclérose en plaque...)

Ceci est d'autant plus important que le critère de jugement est plus subjectif, l'insu n'est pas nécessaire si le critère de jugement est objectif comme le décès.

- tous les patients doivent être soumis au même suivi (même procédures diagnostiques, mêmes questionnaires, à intervalles réguliers et identiques pour tous jusqu'à la fin du suivi.
- En l'absence d'insu, il y a un fort risque de biais.

Si le médecin sait qu'un patient a des facteurs de mauvais pronostic,

- il peut avoir tendance à le suivre de plus près donc et risque de conclure à tort à une aggravation plus rapide alors qu'il s'agit d'un recueil plus exhaustif des signes d'évolution péjorative (plus de scanners ou d'IRM, plus de dosages des marqueurs de cancer...);
- sa mesure du critère de jugement peut être influencée (=biais de mesure) d'autant plus que la part de subjectivité est grande, si c'est le décès le risque est moins grand;
- La durée de suivi doit être assez longue.

Les investigateurs doivent suivre les patients suffisamment longtemps pour que survienne l'événement. Dans certaines pathologies ce délai peut être long.

Exemple : la récurrence d'un cancer du sein de stade précoce peut survenir plusieurs années après le diagnostic.

Exemple : « la récurrence était définie comme l'apparition sur une période de 5 ans d'une nouvelle tumeur invasive à distance de la première tumeur après traitement de celle-ci, la progression était définie comme l'apparition d'une tumeur envahissant le muscle après traitement ».

LES FACTEURS PRONOSTIQUES

Les facteurs pronostiques peuvent être de plusieurs types :

- démographiques (comme l'âge, le sexe...),
- liés à la maladie (comme le stade tumoral lors du diagnostic, l'élévation des marqueurs de l'inflammation...),

- liés à d'autres conditions accompagnant la maladie (comme une maladie associée...).

Ils peuvent prédire :

- Un événement positif : survie, guérison, rémission (facteurs de bon pronostic),
- ou négatif : décès, rechute, complication (facteurs de mauvais pronostic).

Les facteurs pronostiques ne causent pas nécessairement les événements mais y sont associés de manière suffisamment forte pour prédire leur apparition.

Les critères pronostiques doivent être assez explicites, objectifs et précis pour que le lecteur puisse **les utiliser dans sa propre pratique**. Ces critères explicites sont appliqués de façon systématique et standardisée pour tous les patients.

L'ANALYSE STATISTIQUE

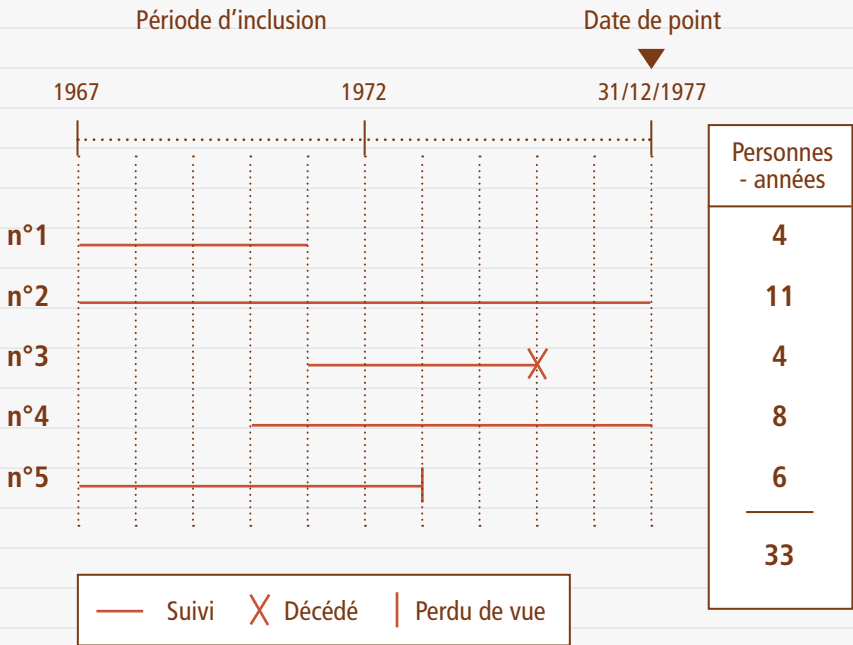
La taille d'échantillon

Voir 4.6 La taille de l'échantillon minimum nécessaire dans le chapitre Les grilles de LCA ou « Check-list ».

Le calcul du nombre de sujets nécessaires doit avoir été effectué *a priori* en justifiant des hypothèses et en demandant une puissance de l'étude supérieure à 80 % (si possible 90 %)

Au sein d'une cohorte, tous les individus n'ont pas la même durée de suivi (+++), le temps de suivi de l'ensemble des participants est décomposé en personnes-temps (personnes-années, personnes-mois, hommes-jours). Le principe est le même que celui des paquets-années dans la consommation de tabac, 1 personne suivie 2 ans représente 2 personnes-années, de même que 2 personnes suivies 1 an chacune, ou 4 personnes suivies 6 mois chacune.

Exemple de la somme des personnes-années qui détermine la taille de la cohorte



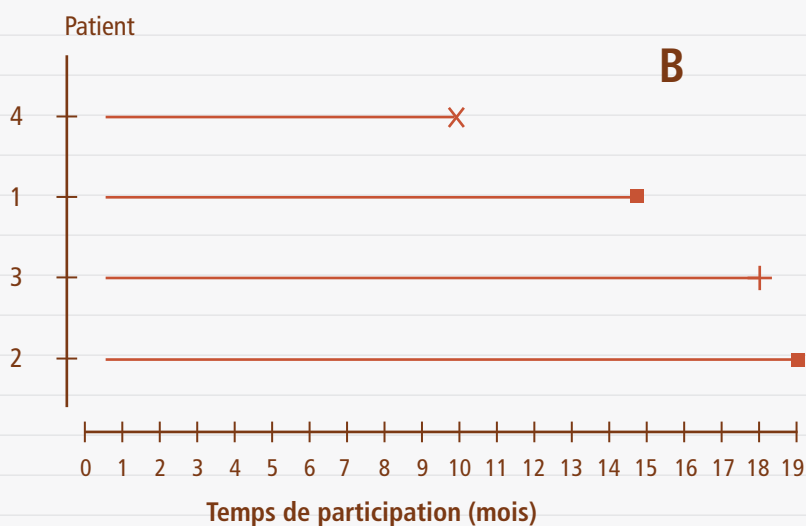
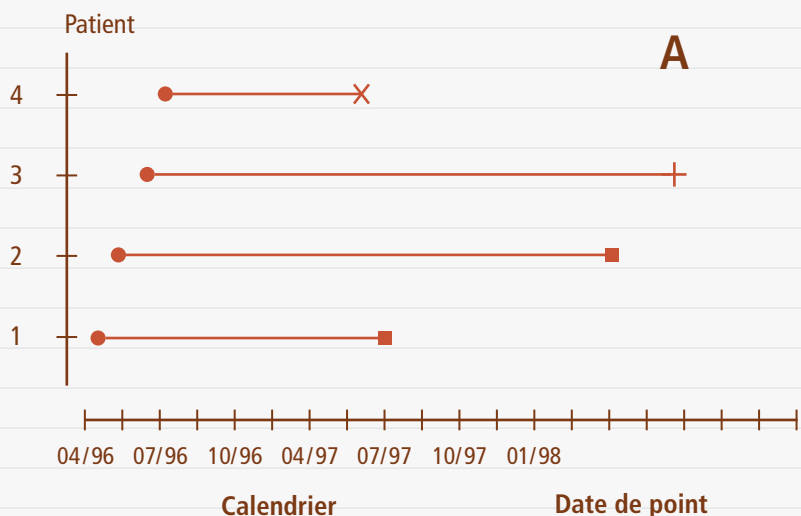
Généralités sur les analyses statistiques des études pronostiques

Les analyses statistiques des études pronostiques sont basées sur des méthodes « d'analyse de survie ». L'événement étudié n'est pas forcément le décès, mais peut être la survenue d'une maladie, la récurrence de symptômes après traitement... Lorsque l'événement d'intérêt est le décès, on peut soit s'intéresser au décès « toute cause », alors chaque décès compte comme un événement, soit ne s'intéresser qu'aux décès pour une cause spécifique (ex. : décès par accident coronaire), dans ce cas les décès d'autres causes (ex. : décès par cancer) ne comptent pas comme un événement, mais comme une « censure » (cf. ci-dessous).

La **date d'origine** (time origin) indique le point de départ du suivi, la date à laquelle les patients entrent dans l'étude. Cette date n'est pas identique pour tous les sujets (figure 1A). Il n'est pas possible d'inclure des dizaines ou des centaines de patients dans une étude le même jour !! (souvent les inclusions se font donc sur plusieurs mois voire plusieurs années). Toutes les dates d'origine vont être virtuellement ramenées à la même origine et vont définir le temps 0 pour calculer le temps de suivi de chaque sujet (figure 1B). Sur la figure 1A, la date d'origine du patient 1 est le 01/04/96 et celle du patient 2 le 01/06/96.

Figure 1. Temps de suivi et état de 4 patients de la cohorte

A. Date d'origine, date d'événement ou de dernières nouvelles, état aux dernières nouvelles. La date de point est le 01/01/1998. **B.** Temps de participation (en mois) de chacun des 4 patients.



● Date d'origine ■ Patient décédé + Patient vivant (censure) X Patient perdu de vue (censure)

La date de point est la date choisie pour faire le bilan de fin de suivi, au-delà de laquelle les informations recueillies ne sont plus considérées dans l'analyse. Aucune nouvelle donnée correspondant à une date ultérieure à la date de point ne sera prise en compte même s'il s'agit de la maladie étudiée ou du décès.

La date de dernières nouvelles est la dernière date à laquelle on a recueilli des informations sur le patient, notamment sur la survenue ou non de l'événement étudié.

Un sujet est dit **perdu de vue** lorsque sa surveillance est interrompue avant la date de point et que l'événement ne s'est pas produit.

Les données censurées (censored data) correspondent aux patients perdus de vue et à ceux qui sont vivants à la date de point (l'événement étudié n'est pas arrivé et le patient n'est pas décédé).

Le temps de participation (patient time) correspond à la durée de suivi pour chaque sujet. Trois situations peuvent se produire :

- l'événement étudié s'est produit au cours du suivi : le temps de participation est le délai entre la date d'origine et la survenue de l'événement ;
- l'événement ne s'est pas produit au cours du suivi ET le sujet est vivant à la date de point : son temps de participation est le délai entre la date d'origine et la date de point ;
- le sujet est perdu de vue : on n'a plus de nouvelles alors que lors de la dernière visite, il n'était ni décédé ni n'avait présenté l'événement étudié : dans ce cas, son temps de participation est défini par le délai entre la date d'origine et la date de dernières nouvelles.

Exemple : suivi d'une cohorte de patientes chez qui on vient de diagnostiquer un premier cancer du sein non invasif, non métastatique et sans atteinte ganglionnaire. La question est de comparer la survie sans progression métastatique (l'événement étudié est donc l'apparition de métastases) en fonction du caractère hormono-sensible de la tumeur.

La fonction de survie est la probabilité que l'événement étudié ne survienne pas avant une date t . S'il s'agit du décès, c'est donc la probabilité de survivre au moins jusqu'à la date t , si c'est la récurrence de symptômes après traitement, c'est la probabilité de survivre sans symptômes jusqu'à cette date (survie sans récurrence pour un cancer, ou survie du greffon pour un insuffisant rénal greffé, « disease free survival »). La fonction de survie est représentée graphiquement par une courbe de survie.

Deux méthodes d'analyse de survie sont utilisées : l'analyse actuarielle (qui n'est quasiment plus utilisée aujourd'hui) et la méthode de Kaplan-Meier, ce sont deux méthodes non paramétriques.

Les analyses descriptives de la survie

Analyse actuarielle (life-table estimate)

Pour chaque intervalle de temps, on représente l'estimation de la survie par un point. Les coordonnées du premier point sont 0 en abscisse qui représente le temps, et 1 (100 %) en ordonnée qui représente la proportion de patients vivants. Tous les points consécutifs sont reliés par un segment de droite.

La figure 2A présente l'analyse actuarielle des données de survie de patients en attente de transplantation cardiaque (d'après Kalbfleisch). L'illustration est réalisée sur les 34 patients non transplantés, pour des intervalles de 10 jours, sur les 100 premiers jours.

L'inconvénient majeur de cette méthode est qu'elle estime la survie à chaque borne supérieure des intervalles constitués a priori, et considère chaque censure, survenant dans un intervalle, de manière équivalente, dans cet exemple, un sujet suivi 21 jours apporte la même information qu'un sujet suivi 29 jours pour la survie à 30 jours. C'est la raison pour laquelle cette méthode est à réserver à de grands échantillons.

Méthode de Kaplan-Meier (product-limit estimate)

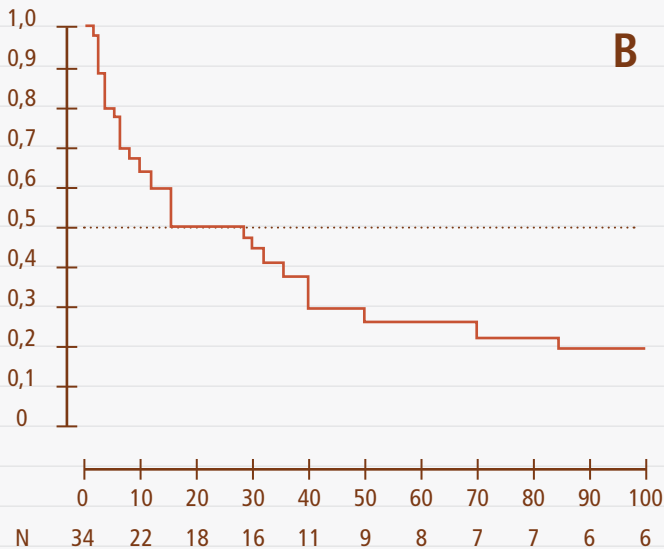
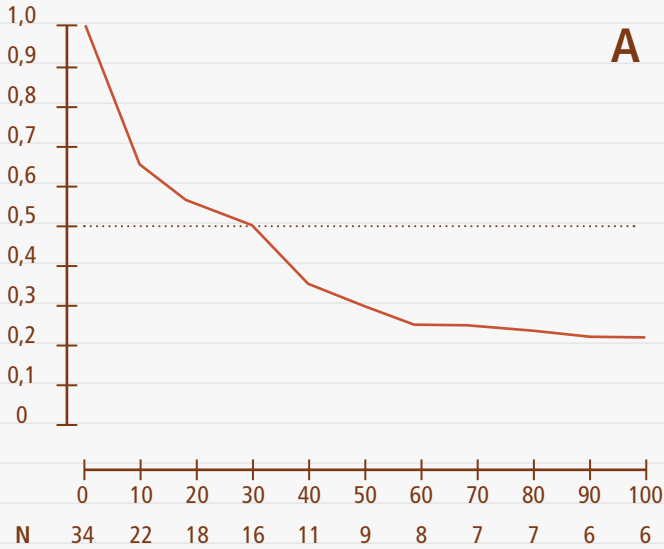
Contrairement à l'analyse actuarielle, les intervalles ne sont pas fixés a priori, mais sont définis par les instants auxquels les événements sont observés. Ces intervalles sont donc inégaux.

La courbe de survie se compose de paliers successifs, où les probabilités de survie sont constantes entre deux temps d'événements consécutifs. Le premier palier vaut 1 depuis l'origine jusqu'au délai de survenue du premier événement. Il s'abaisse ensuite à la première valeur calculée pour constituer un second palier jusqu'au délai de survenue de l'événement suivant, etc.

Il est possible de relier les paliers successifs par des segments verticaux, mais il n'est pas correct de les relier par des segments obliques. La courbe ainsi obtenue présente une allure en « marches d'escalier ». Les données censurées doivent être représentées par un trait vertical. La figure 2B présente l'analyse de Kaplan-Meier sur les données de suivi de patients en attente de transplantation cardiaque. On peut remarquer que la courbe est de forme analogue à celle obtenue avec la méthode actuarielle, sauf sur les 10 premiers jours, où survenaient 35 % des événements. Pour ces deux méthodes, il est recommandé d'indiquer, à intervalle de temps régulier, le nombre de sujets encore présents dans l'étude.

Figure 2. Deux méthodes de représentation de la fonction de survie (données de survie de patients en attente de transplantation cardiaque)

A. méthode actuarielle, le « pas » de temps choisi est de 10 jours, **B.** méthode de Kaplan-Meier, données censurées représentées par des barres verticales, médiane de survie de 21 jours (N : nombre de sujets à risque).



Choix d'une valeur « résumée »

La fonction de survie peut être résumée soit en fixant un délai et en mesurant le taux de survie (ex. : survie à 1 an, survie sans récurrence à 5 ans, etc.), soit en fixant un taux de survie et en mesurant le délai correspondant (ex. : médiane de survie (median survival time) c'est la durée au bout de laquelle 50 % des patients sont encore en vie).

Médiane de survie

Il est utile de disposer d'indicateurs synthétiques ou résumés de cette courbe. La moyenne de survie n'est pas un bon indicateur, la médiane de survie, durée de suivi pour laquelle la probabilité de survie est de 50 % est souvent utilisée. À cause de la distribution par paliers de la fonction de survie, il est souvent impossible de connaître la durée correspondant à une survie exacte de 50 %. En pratique, la médiane est estimée par la plus petite durée pour laquelle la survie est inférieure à 50 %. Dans cet exemple la médiane de survie est de 21 jours (*figure 2B*).

Remarque : Il arrive que la fonction de survie soit toujours supérieure à 50 % (faible mortalité ou faible incidence de l'évènement étudié). Dans ce cas, la médiane ne peut être estimée. On estime alors les quantiles (Nb. : un quantile = 25 %) : pour le n-ième quantile on estime la durée pour laquelle la probabilité de survie est de 100-n. Par exemple, le 25^e quantile (ou 1^{er} quartile) correspond à la plus petite durée pour laquelle la survie est inférieure à 75 %. Il est de 6 jours dans l'exemple choisi.

Survie à date fixée

Un autre indicateur fréquemment utilisé pour résumer l'information d'une courbe de survie est l'estimation de la survie à un temps donné. Ainsi, la survie à 2 mois (60 jours) est de 26 % dans l'exemple présenté.

Comparaison entre deux ou plusieurs groupes

Très souvent, en plus d'évaluer la survie dans un groupe de malades, on cherche à la comparer à la survie d'un ou de plusieurs autres groupes de malades indépendants. Deux tests sont couramment utilisés pour comparer la survie entre deux ou plusieurs groupes.

Test du log-rank

Le principe est d'estimer le nombre attendu d'événements sur la période étudiée

si la survie était identique dans les deux groupes, puis le comparer au nombre d'évènements observés, pour chaque intervalle de temps (qu'il s'agisse de l'analyse actuarielle ou de Kaplan-Meier). Le test du log-rank est généralisable au cas de k groupes et permet de tester si globalement la survie est différente entre les groupes. En pratique, il est le test le plus souvent employé.

Test de Wilcoxon

Ce test repose sur le même principe que le test du log-rank, mais pondère la différence entre le nombre observé et le nombre attendu d'évènements. Ce test permet de donner moins de poids à cette différence lorsque le nombre de sujets encore à risque d'évènements est faible. Le test de Wilcoxon peut aussi se généraliser à k groupes.

La précision des taux estimés est calculée par l'intervalle de confiance à 95% autour de cette estimation. Plus il est étroit, plus l'estimation est utile (car proche de la vérité). La précision de l'estimation dépend du nombre d'observations sur lesquelles elle est basée (plus l'effectif est élevé, plus l'ICC est réduit).

Modèle de régression adapté aux données censurées : modèle de Cox

Ces modèles permettent de comparer plusieurs groupes et d'estimer si la différence est statistiquement significative. Ces modèles permettent également d'ajuster les analyses pour plusieurs facteurs de confusion dans des analyses multivariées. Les résultats de ce type de modèle sont exprimés par des risques relatifs instantanés ou Hazard Ratios (HR).

REMARQUE : ces analyses de survie peuvent être utilisées dans toute étude qui comporte un suivi individuel des individus si l'on s'intéresse à un évènement dichotomique (oui/non) qui ne se produit qu'une seule fois pendant la période de suivi. Elles peuvent donc être utilisées dans les études de cohortes et dans les essais cliniques randomisés, pour décrire la fréquence d'un évènement au cours du temps en tenant compte des décès et des perdus de vue et la comparer entre des groupes exposés ou non à un facteur de risque (cohorte) ou randomisés à un traitement ou un placebo (essais clinique).

RÉSULTATS

Comme vu aux chapitres précédents, deux notions sont importantes pour interpréter les résultats :

- l'association est-elle statistiquement significative (intervalle de confiance à 95 % du RR et « p ») et,
- est-elle forte (valeur du RR, de l'OR ou du HR).

Comme pour les autres types d'étude, la conclusion doit porter sur le résultat le moins biaisé c'est-à-dire celui qui est issue des analyses multivariées.

Dans les études pronostiques on ne s'intéresse pas forcément à une relation de causalité entre facteur étudié et critère de jugement. En effet, dans ce type d'étude l'idée sous-jacente est plus de cibler des groupes en fonction de leur pronostic ce qui permet d'envisager une surveillance ou un traitement plus ou moins lourds (par exemple le stade TNM pour les cancers).

Ici, c'est plutôt la force et la précision de la prédiction du devenir du patient qui sont importantes.

Une autre particularité est que les facteurs pronostiques sont parfois regroupés dans un score prédictif (obtenu par une méthode statistique plus ou moins complexe que vous n'aurez pas à comprendre en détail).

LES BIAIS ET FACTEURS DE CONFUSION

Facteurs de confusion

Ce sont les autres facteurs pronostiques connus ou potentiels.

Il faut les avoir 1) mesuré et 2) pris en compte dans l'analyse en ajustant les résultats pour ces facteurs pronostiques.

Biais de sélection

Ils peuvent concerner l'échantillon étudié, soit lors de sa constitution (inclusion des patients), soit lors du suivi (perdus de vue).

Ils sont évoqués dans le chapitre sur la population étudiée. *(voir 6.2)*

Biais de mesure

Ils peuvent concerner la mesure des facteurs pronostiques ou la mesure des événements de santé étudiés.

Dans le cas des études pronostiques on se méfiera particulièrement de modalités de suivi différentes en fonction du pronostic initial.

Les résultats (force de l'association et signification statistique).

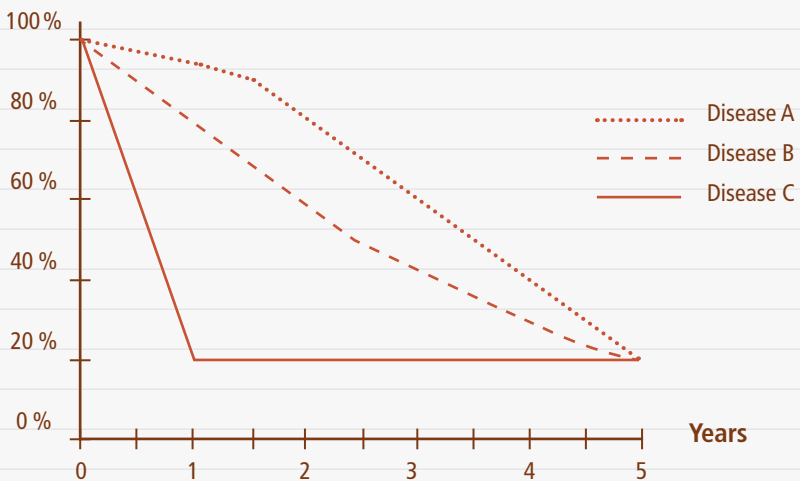
Quelle est la probabilité de l'évènement clinique étudié sur une période de temps bien définie ?

Exprimer le pronostic sous la forme d'un taux a des avantages. C'est simple, faci-

lement compris et mémorisé. MAIS, les taux moyens apportent peu d'informations et il peut y avoir des différences importantes de pronostic avec des taux similaires.

DONC, les courbes de survie sont utilisées pour estimer la survie d'une cohorte au fil du temps. La figure ci-dessous montre les courbes de survie pour trois maladies avec des taux de survie similaires à 5 ans. On remarque que les taux moyens occultent des différences pourtant importantes entre les patients au cours du temps.

Courbes de survie à 5 ans pour trois maladies différentes.



Quelle est la précision du pronostic estimé ?

Pour déterminer la précision des estimations, il faut regarder les intervalles de confiance à 95% (IC) autour de l'estimation. Plus il est étroit, plus l'estimation est utile (car proche de la vérité). La précision de l'estimation dépend du nombre d'observations sur lesquelles elle est basée (plus l'effectif est élevé, plus l'ICC est réduit).

Les périodes de suivi précoce incluent les résultats d'un plus grand nombre de patients que les périodes plus avancées dans le temps : les estimations sur le côté gauche de la courbe sont généralement plus précises. Les observations figurant sur le côté droit de la courbe (queue de la courbe) sont en général basées sur un plus petit nombre de patients en raison des perdus de vue, des décès et des

patients qui ont déjà présenté l'évènement étudié.

Par conséquent, les estimations de la survie à la fin de la période de suivi sont relativement imprécises et peuvent être affectées par ce qui arrive à quelques personnes seulement.

LA CONCLUSION

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

Exemple de check-list pour la LCA des études pronostiques :

1. L'étude est-elle valide?	OUI	NON	NSP
Y a-t-il un échantillon représentatif de patients à un stade similaire dans l'évolution de la maladie ?			
Y a-t-il eu un suivi suffisamment long et complet ?			
Des critères objectifs et non biaisés de mesure de l'issue clinique ont-ils été utilisés ?			
A-t-on ajusté les résultats pour l'effet d'autres facteurs pronostiques ?			
2. Quels sont les résultats ?	VALEUR	IC95 %	NSP
Quelle est la probabilité de l'évènement clinique étudié sur une période de temps bien définie ?			
Quelle est la précision de cette probabilité ?			
3. Quelle est l'applicabilité des résultats ?	OUI	NON	NSP
Les patients de l'étude sont-ils similaires à mes patients ?			
Les résultats entraînent-ils directement la sélection ou le rejet du traitement ?			
Les résultats sont-ils utiles pour rassurer ou conseiller mes patients ?			

MESSAGES CLEFS

- Respect de la chronologie : d'abord recueil du facteur pronostique puis pendant le suivi, recueil de l'évolution de la maladie.
- Cohorte de Malades (le plus souvent).
- Patients atteints d'une même maladie selon un critère diagnostique unique
- Patients au même stade de leur maladie (même durée d'évolution ++).
- Critère de jugement en insu des facteurs pronostiques.
- Attention aux perdus de vue (comme dans toutes les études où il y a un suivi).

Exemples : Les premières études pronostiques sur l'évolution des patients victimes d'infarctus ont été réalisées sur des cohortes de patients hospitalisés dans des unités de soins intensifs. Le pronostic qu'on a alors observé en suivant les patients était bien meilleur que le véritable pronostic après un infarctus. Notamment la survie était surestimée, POURQUOI ? :

Parce qu'on n'a pas tenu compte des décès préhospitaliers. Souvent les premières études sur une cohorte de malades sont faites dans des services cliniques et donc avec des patients qui arrivent dans ces services. Ensuite, d'autres études ont été réalisées en population telles que l'étude PRIMA réalisée en Rhône-Alpes dans les années 90 à partir de tous les patients de la région présentant un infarctus du myocarde (depuis le domicile jusqu'aux urgences en passant par le SAMU et les pompiers) est meilleur qu'il n'est en réalité car les patients décèdent pour moitié avant d'arriver aux urgences.

Au début du XX^e siècle, des études descriptives ont montré que la consommation de tabac était élevée chez les sujets atteints de carcinome bronchique. Cependant, on ne savait pas si la consommation de tabac était également élevée chez des sujets non atteints de carcinome bronchique.

La question était donc de savoir si l'exposition au tabac est un facteur de risque de cancer broncho-pulmonaire. Pour y répondre, des chercheurs décidèrent de prendre un groupe de sujets malades (des cas) et un groupe de sujets indemnes de la maladie (des témoins) et de rechercher si les individus de chaque groupe avaient été exposés dans le passé au tabac.

L'hypothèse des chercheurs était que si le tabac joue un rôle dans l'apparition de la maladie, alors la fréquence d'exposition au tabac parmi les cas sera plus élevée que parmi les témoins.

Les cas et les témoins étaient tous des hommes.

7.1 LA QUESTION

Une étude cas-témoins permet de répondre à la question de recherche suivante : **« Le facteur de risque est-il associé à une augmentation significative du risque de maladie ? »**.

L'étude cas-témoins apporte une réponse indirecte à cette question car en pratique elle répond à la question : « L'exposition passée au facteur de risque était-elle plus fréquente chez les personnes atteintes de la maladie (cas) que chez des personnes non atteintes de la maladie (témoins) ? ».

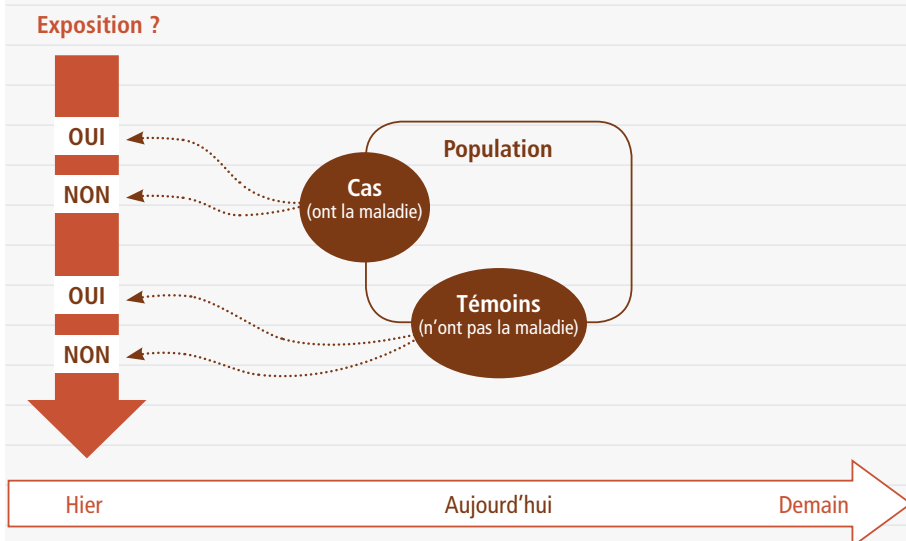
Une particularité sur ce type d'étude réside dans le fait que le ratio cas/témoins est fixé. (Nb. : cela signifie qu'en aucun cas on peut calculer la prévalence d'une maladie dans ce type d'étude, et cela explique pourquoi on ne peut pas estimer le risque relatif, mais l'odds ratio).

7.2 LE TYPE D'ÉTUDE

Une étude cas-témoins est une étude de type **observationnelle et étiologique** (= analytique).

Elle a la particularité que l'exposition au facteur de risque est toujours recherchée dans le passé, soit de façon **rétrospective** (figure 1).

Figure 1. Étude cas-témoins.



Fil rouge :
Cancer et tabac

L'étude comportait 200 sujets atteints de cancer broncho-pulmonaire (les cas), confirmés histologiquement et diagnostiqués au cours des 5 années précédant l'enquête. Ces patients avaient été pris en charge dans des hôpitaux et des cabinets privés situés dans les états de Californie, du Colorado, du Missouri, du New Jersey et de New-York.

L'étude comportait également 200 sujets indemnes de cancer du poumon (les témoins), et hospitalisés dans trois hôpitaux de la ville de Saint Louis de l'État du Missouri.

7.3 LA POPULATION ÉTUDIÉE

La sélection des cas et des témoins doit être indépendante des facteurs de risque étudiés. Le recrutement des cas et des témoins doit se faire sans connaître leurs éventuelles expositions à des facteurs de risque.

LES CAS

Représentativité des cas étudiés de l'ensemble des cas de la population générale

Sélection des cas : les cas peuvent être issus d'un enregistrement exhaustif de tous les cas sur une période donnée et un secteur géographique donné. C'est le cas des registres. Cela apporte une grande sécurité sur la bonne représentativité des cas étudiés par rapport à l'ensemble des cas dans la population générale.

Cas incidents et prévalents

- Cas incidents : si le délai entre la date du diagnostic de la maladie et la date d'inclusion dans l'étude est court, on parle plutôt de cas incidents. Dans ce cas, les stades précoces de la maladie peuvent être surreprésentés.
- Cas prévalents : si le délai entre la date du diagnostic de la maladie et la date d'inclusion dans l'étude est long, on parle plutôt de cas prévalents car les cas inclus seront ceux ayant survécu à la maladie. Dans ce cas, les formes lentes et/ou peu agressives peuvent être surreprésentées.

LES TÉMOINS

Il faut s'assurer que les témoins n'ont pas la maladie étudiée.

Le choix des témoins doit être indépendant de l'exposition. Idéalement, le groupe de témoins doit être représentatif de l'ensemble des sujets non malades de la population source d'où est issu le groupe de malades étudié. Les témoins peuvent aussi être issus de la population générale, ou être issus de populations particulières (patients hospitalisés pour une autre raison, personnes adhérentes à une mutuelle...). Chaque choix présente des avantages et des inconvénients, il n'y a pas de choix idéal. Le principal problème dans ce type d'étude est le choix des témoins.

L'APPARIEMENT

C'est une technique qui consiste à **contrôler *a priori* un (ou plusieurs) facteur de confusion** en formant des paires homogènes quant à ce facteur de confusion. Si on considère le l'âge et le sexe comme des facteurs de confusion dans une étude cas-témoins, on appariera à chaque cas qui est un homme de 50 ans un témoin de sexe masculin de 50 ans. À part pour la pédiatrie, on apparie rarement sur l'âge exact, on réalise plutôt une stratification : pour chaque cas de sexe masculin âgé de 50 à 55 ans on apparie un témoin de sexe masculin âgé de 50 à 55 ans. Pour accroître la puissance de l'étude on peut appairer chaque cas à plusieurs témoins (appariement individuel 1 cas pour 2 témoins, par exemple). **L'appariement nécessite l'utilisation de tests statistiques particuliers pour séries appariées.** (*voir biais, facteur de confusion, stratification*)

7.4 LE CRITÈRE DE JUGEMENT : LA MALADIE ÉTUDIÉE

Il faut s'assurer que les cas ont la maladie étudiée.

La maladie doit être établie sur des critères diagnostiques précis et identiques pour tous les cas (ex. : biopsie pour tous les cas).

Fil rouge : Cancer et tabac

Pour chaque patient, des enquêteurs entraînés et sans information sur le statut du patient ont conduit un entretien téléphonique à l'aide d'un questionnaire standardisé.

Des données sur les antécédents de pathologies pulmonaires, les métiers, l'exposition à la poussière ou aux fumées, la consommation alcoolique, le lieu de résidence, le niveau d'éducation, et les causes de décès des parents et des autres membres de la famille ont été collectées.

La consommation de tabac était évaluée au travers du nombre moyen de cigarettes fumées par jour, de la durée de la consommation de tabac, de l'âge au début de la consommation et du type de tabac fumé (brun, blond, roulées, pipe...). Pour les cas, leur consommation de tabac avant la date de diagnostic du cancer était demandée.

7.5 LE FACTEUR DE RISQUE

La mesure de l'exposition au facteur de risque porte sur une exposition antérieure et est donc sujette à des erreurs de mesure++.

Il faut distinguer plusieurs situations différentes :

- soit l'erreur est simplement liée au fait que les individus ne se souviennent pas avec exactitude de l'exposition, il s'agit alors d'une erreur aléatoire, entraînant un « bruit de fond » sur la mesure de l'exposition qui gêne la mise en évidence d'une association et va dans le sens d'un manque de puissance ;
- soit il s'agit réellement d'un biais qui s'applique de la même façon aux cas et aux témoins. Par exemple les patients ont tendance à sous-estimer systématiquement ou à l'inverse à surestimer systématiquement leur exposition passée. Si cela concerne les cas aussi bien que les témoins cela va encore une fois dans le sens d'un manque de puissance pour montrer une association.
- Le cas le plus problématique est celui, TRÈS FRÉQUENT, dans lequel les cas surestiment systématiquement leur exposition par rapport aux témoins. C'est très fréquent car lorsqu'une étude est décidée, c'est souvent parce qu'il existe déjà des présomptions sur un facteur de risque et les personnes en ont entendu parler. **Il s'agit alors d'un biais de mémoire différentiel.**

Par exemple, dans une étude cas-témoins sur vaccination contre l'hépatite B et risque de sclérose en plaque, 30% des cas qui se souvenaient d'avoir été vacciné contre l'hépatite B, ne l'avaient pas été (soit ils avaient été vaccinés plusieurs années plus tôt ce qui ne pouvait pas avoir déclenché la maladie, soit ils avaient été vaccinés pour une autre pathologie, soit ils n'avaient pas été vaccinés du tout). Cette surestimation n'était pas du tout observée dans le groupe témoin.

À côté de ce biais de mémoire, on peut aussi trouver des situations inverses, dans lesquelles les cas pourraient sous-déclarer certains comportements à risque qui seraient suspectés d'être des facteurs de risque pour la maladie qu'ils ont développée.

Ce biais entre dans la catégorie des biais de mesure qui sont aussi des biais de classement.

ATTENTION, ces informations peuvent être recueillies rétrospectivement dans des dossiers ou des fichiers existants : les auteurs doivent nous donner des informations sur la disponibilité et la qualité des informations (nombre et type d'informations manquantes).

La description des caractéristiques des témoins et des cas a montré que la distribution selon l'âge des témoins était différente de celles des cas (tableau 1). Or, la consommation moyenne de tabac par jour variait selon l'âge et le risque de survenue d'un cancer augmente avec l'âge.

TABLEAU 1. Effectifs de cas et de témoins en fonction du groupe d'âge

Groupes d'âge	Cas	Témoins
25 - 34 ans	1	30
35 - 44 ans	9	49
45 - 54 ans	46	43
55 - 64 ans	76	42
65 - 74 ans	55	28
75 ans et plus	13	8
TOTAL	200	200

Le risque de cancer broncho-pulmonaire a été calculé pour différents niveaux de consommation de tabac (tableau 2). Une consommation de < 9 grammes / jour était considérée comme le risque de base. Afin de prendre en compte l'effet de l'âge sur le niveau de consommation de tabac et sur le risque de survenue de cancer, on a appliqué aux deux groupes une structure d'âge identique.

TABLEAU 2. Risque de cancer broncho-pulmonaire en fonction de la consommation moyenne de tabac par jour

Consommation moyenne de tabac (en grammes par jour)	Cas	Témoins	OR ajusté pour l'âge
de 0 à 9	78	116	1,0
10 ou plus	122	84	2,2
10 - 19	58	50	1,7
20 - 29	33	26	1,9
30 et plus	31	8	5,9
TOTAL	200	200	

7.6 ANALYSES STATISTIQUES

La mesure d'association dans une étude cas-témoins est l'Odds-Ratio ou rapport de côtes.

Remarque : on peut le calculer quel que soit le type d'étude, contrairement au risque relatif.

Il s'agit d'un rapport des rapports ou d'un rapport des côtes.

	M+	M-	
E+	a	b	N1
E-	c	d	N0
	M1	M0	

$$OR = a/c/b/d = ad/bc$$

7.7 LES RÉSULTATS

L'association entre le facteur de risque et la maladie est mesurée par l'odds ratio (ou rapport de cote, RC).

L'ODDS RATIO (OR)

Dans une étude cas-témoins, le calcul de l'odds ratio permet d'estimer le risque relatif et ainsi de comparer les fréquences d'exposition dans les 2 groupes :

	Cas	Témoins
Exposé > 10 g / jour	122	84
Non exposé 0 à 9 g / jour	78	116

LA FORCE DE L'ASSOCIATION

Odd d'exposition chez les cas : cas exposés/cas non exposés = 122/78.

Odd d'exposition chez les témoins : témoins exposés/ témoins non exposés = 84/116.

Odds ratio (rapport des cotes), $OR = (122/78) / (84/116) = 2,7$.

Le risque de survenue de cancer est plus élevé chez les fumeurs que chez les non-fumeurs. L'association ici est modérée.

Ordre de grandeur des OR

- association entre consommation de café et cancer de la vessie : $OR = 1,5$.
- association entre exposition à l'amiante et cancer du poumon : $OR = 5$.
- association entre consommation de tabac et cancer du poumon : $OR = 10$.
- association entre exposition au monochlorure de vinyle et cancer du foie : $OR = 500$.

LA PRÉCISION DE L'OR

IC95 % [1,5 – 3,0]

Cette différence est statistiquement significative car l'intervalle de confiance de l'OR ne comprend pas la valeur 1 : il y a 95 % de chance que l'intervalle compris entre 1,5 et 3,0 contienne la vraie valeur de l'OR.

Le degré de significativité statistique

$p < 0,05$, Cela signifie qu'en acceptant l'hypothèse d'un lien entre tabac et cancer, le risque que l'on se trompe, et qu'il n'y ait en fait pas de lien est inférieur à 5 %.

Interprétation de l'odds ratio (OR)

L'OR est moins facile à interpréter que le risque relatif. Lorsque la prévalence de la maladie est faible dans la population cible, l'OR est proche du RR. On dit que c'est un bon estimateur du RR. C'est important car dans les modèles de régression on ne peut estimer que des OR.

Remarque : Si la prévalence de la maladie est faible, c sera proche de $n1$ et d de $n0$. Donc l'OR sera proche du RR et on pourra interpréter l'OR comme un RR.

Conséquences :

- si la prévalence de la maladie dans la population cible est de 8% et que l'on estime un OR à 2 sur un échantillon, alors on peut dire sans trop se tromper « le risque de maladie est 2 fois plus élevé chez les exposés que chez les non exposés » (phrase qui correspond normalement au risque relatif) ;
- si la prévalence de la maladie dans la population cible est de 35 % et que l'on estime un OR à 3,2 sur un échantillon, alors il est probable que la phrase « le risque de maladie est 3,2 fois plus élevé chez les exposés que chez les non exposés » soit fausse. Il faudra dire « il y a 3,2 fois plus de malades par rapport aux non-malades chez les exposés que de malades par rapport aux non-malades chez les non exposés ».

7.8 LES BIAIS ET FACTEURS DE CONFUSION

BIAIS DE SÉLECTION

Biais de sélection des témoins

(problème classique +++ des études cas-témoins)

Les témoins avaient-ils d'autres raisons (que le fait de ne pas avoir la maladie) pour être exposés ou ne pas être exposés ?

Exemple : on compare chez des patients atteints de cancer de l'œsophage (cas) et des sujets indemnes de cancer (témoins) la fréquence d'exposition à la caféine (facteur de risque).

	Cas	Témoïn
Exposé	200	100
Non exposé	100	100

L'OR = 2, le café est donc un facteur de risque... sauf si... le groupe témoins est mal choisi et qu'il s'agit de personnes qui pour une autre raison avaient moins de risque que la population générale d'être exposés à la caféine (leur consommation est alors sous-estimée) ; le groupe témoins peut aussi être moins à risque de cancer à cause d'autres facteurs (facteurs de confusion).

Les témoins sont des patients hospitalisés en service de gastroentérologie pour des pathologies bénignes. Ils ont souvent des œsophagites, des gastrites ou des

colopathies fonctionnelles et ont arrêté de boire du café depuis de nombreuses années.

Il y a un biais de sélection : on a sélectionné des témoins dont la probabilité d'avoir été exposé est systématiquement inférieure à celle de la population générale pour une autre raison (œsophagites ou colite). Ce biais a pour effet que l'on va donc conclure à tort à un lien entre cancer de l'œsophage et exposition au café. Un biais de sélection inverse pourrait se produire si on sélectionnait des témoins dont la probabilité d'avoir été exposée est supérieure à la population générale. On aurait alors une sous-estimation de l'effet du facteur de risque.

Biais de sélection des cas

En dehors du cancer, les registres sont rares. Le plus souvent les cas étudiés ne représentent pas l'exhaustivité des cas sur une période ou une zone donnée : il faut se poser la question du biais de sélection.

Le choix de cas prévalents entraîne un biais de sélection dans le cas de maladies létales comme le cancer et dans le cas où le facteur de risque étudié provoque la maladie mais aussi diminue la durée de survie.

Exemple : on a pu répertorier tous les patients résidants dans le Rhône qui ont développé un cancer du poumon de 1990 à 2000. Si on doit interroger ces patients pour avoir des informations sur leur consommation passée de tabac, on ne peut le faire qu'avec les patients toujours en vie. Le groupe des cas va donc présenter une surreprésentation des formes les moins graves de cancer et une sous-représentation des formes les plus graves.

Biais de mesure

Le biais de mesure qu'il faut redouter dans les études cas-témoins concerne la mesure de l'exposition au facteur de risque. Sa mesure repose sur la déclaration par les cas et les témoins de leurs consommations passées ce qui entraîne un biais de mémoire (ou biais de souvenir).

Exemple : étude cas-témoins de l'exposition au vaccin contre l'hépatite B chez les patients atteints de sclérose en plaques (cas) comparés à des témoins.

Les patients sont très concernés par la maladie, ils se posent beaucoup de questions et se demandent comment ils sont devenus malades. Ils ont tendance :

- à faire des efforts pour se souvenir de tous les vaccins qu'ils ont eus,
- à rapporter en excès des vaccins contre l'hépatite alors qu'il s'agissait d'autres vaccins.

Les témoins au contraire ne vont pas faire d'effort pour retrouver leurs vaccinations et risquent de sous-estimer la fréquence de leur vaccination contre l'hépatite B.

Cela induit une surestimation du lien entre vaccination contre l'hépatite B et survenue d'une sclérose en plaque.

Biais de confusion

Ce biais est lié à l'influence de tiers facteurs (l'âge, le statut socio-économique, l'exposition à d'autres facteurs de risque que celui étudié comme l'alcool par exemple...) sur l'association entre l'exposition et la maladie étudiée.

Exemple : dans l'étude sur l'association entre consommation de tabac et survenue du cancer du poumon, l'âge est un facteur de confusion car il est à la fois associé à la consommation de tabac et au risque de survenue de cancer :



Le biais de confusion peut être pris en compte à deux moments :

- à la conception de l'étude : **appariement** des cas et des témoins sur l'âge afin de les rendre comparables pour l'âge ;
- au moment de l'analyse : **ajustement** sur l'âge afin de mesurer la relation tabac-cancer à « âge égal ».

NIVEAU 2

L'ajustement sur un tiers facteur n'est pas justifié si l'effet du tabac sur le risque de cancer diffère selon la valeur du tiers facteur. C'est le cas par exemple pour le cancer de l'œsophage où il y a un effet combiné de l'alcool et du tabac : la relation entre tabac et cancer de l'œsophage n'est pas la même selon les classes de consommation d'alcool. On parle d'une **interaction**.

Exemple : OR de cancer de l'œsophage en fonction des consommations d'alcool et de tabac chez les hommes.

CONSOMMATION MOYENNE D'ALCOOL EN GRAMMES PAR JOUR	CONSOMMATION MOYENNE DE TABAC		
	0 – 9	10 – 19	20 ET +
0 – 40	1,0	3,4	5,1
41 – 80	7,3	8,4	12,3
81 et +	18,0	19,9	44,4

7.9 LA CONCLUSION

Elle doit tenir compte des résultats observés, des limites de l'étude et des données déjà publiées.

À l'instar des études de cohortes, et plus encore étant donné le plus faible niveau de preuve des études cas-témoins, pour étayer l'hypothèse d'une relation de causalité entre un facteur de risque et une maladie, on utilisera les critères décrits par Hill (voir chapitre études de cohortes).

Critères de Hill internes à l'étude :

- Forte **intensité** de l'association (Odds ratio élevé).
- Existence d'une **relation de type « dose-effet »** entre l'exposition et la maladie.
- **Spécificité** (+/-) de relation exposition <-> maladie.
- (par exemple amiante et mésothéliome).
- **Chronologie** (facteur de risque précède la survenue de la maladie).

Critères de Hill externes à l'étude (bibliographie) :

- **Concordance** entre les résultats des précédentes études (**Fil rouge : 17 études cas-témoins concluaient à un risque plus élevé chez les NRS couchés sur le ventre**).
- **Plausibilité** biologique (existence de mécanismes d'actions biologiques et physiopathologiques connus avant de mener l'étude).

- Concordance avec les **expérimentations** menées in vitro ou chez l'animal.
- **Diminution de l'incidence de la maladie lorsque l'exposition est supprimée** ou réduite.

7.10 ÉTUDE CAS-TÉMOINS NICHÉE DANS UNE COHORTE (« NESTED CASE-CONTROL STUDY »)

Il s'agit d'un type particulier d'étude cas-témoins dans lequel les cas et les témoins sont tous issus de la même cohorte.

Le design de l'étude princeps est une étude de cohorte puis au sein de cette cohorte, à la fin de l'étude on analyse le groupe des cas, c'est-à-dire des sujets de la cohorte qui ont développé la maladie au cours du suivi, et on tire au sort un groupe de témoins parmi les sujets indemnes en fin de suivi.

Les deux avantages majeures +++ par rapport à une étude cas-témoins classique sont que :

- a / le recueil sur l'exposition au facteur de risque précède la survenue de la maladie **annulant ainsi totalement le risque de biais de mémoire différentiel.**
- b / Les cas et les témoins sont issus d'une même cohorte et ainsi le **risque de biais de sélection différentiel est également écarté.**

Le niveau de preuve est donc plutôt identique à celui d'une cohorte que d'une étude cas-témoins.

Alors pourquoi réaliser une étude cas-témoins nichée plutôt qu'une étude de cohorte ??!

Pour des raisons de faisabilité essentiellement, pour mesurer un facteur de risque coûteux (on va mesurer à la fin de l'étude uniquement les sérums des sujets qui ont développé la maladie (« cas ») et d'un échantillon aléatoire des sujets indemnes de la maladie (« témoins »).

Par exemple, dans une étude de cohorte visant à mieux connaître les facteurs de risque d'ostéoporose nous avons inclus 7500 femmes suivies 4 ans, chacune avait eu une prise de sang et les sérums ont été congelés. Le budget pour faire 7500 dosages des bio marqueurs de l'ostéoporose arrivait sans doute aux alentours du million d'euros. Il était donc seulement faisable de doser les sérums des 350

femmes ayant eu une fracture au cours du suivi et 350 témoins tirées au sort n'en ayant pas eu.

Cela peut être le cas également si certaines données cliniques n'ont pas été saisies sur informatique et que leur saisie pour l'ensemble de la cohorte coûterait très cher.

Remarque : on peut considérer qu'une étude cas-témoins nichée dans une grande base de données telle que celle de l'assurance maladie par exemple, est une étude cas-témoins nichée dans une cohorte.

7.11 COMPARAISON ÉTUDE CAS-TÉMOINS ET ÉTUDE DE COHORTE

CHOIX ENTRE ÉTUDE CAS-TÉMOIN ET ÉTUDE DE COHORTE

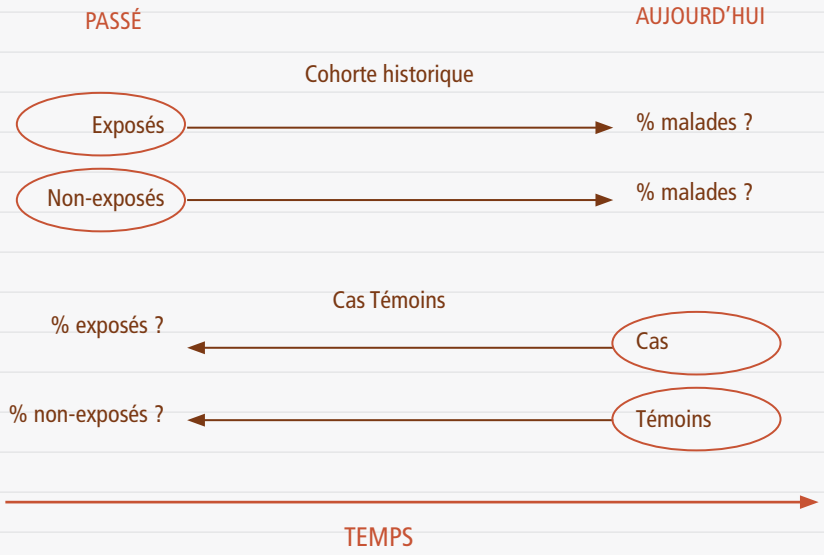
Le choix entre cohorte et cas-témoïn dépend de la question posée. Chaque type d'étude présente des avantages et des inconvénients.

Comparaison des avantages et des inconvénients des études de cohortes et des études cas-témoins.

	Avantages	Inconvénients
Cohortes*	<ul style="list-style-type: none"> – Possibilité d'étudier plusieurs maladies – Adaptées aux expositions rares – Permet de calculer l'incidence d'une maladie chez les exposés et les non exposés – Recueil prospectif des données 	<ul style="list-style-type: none"> – Long avec le risque d'un grand nombre de perdus de vue – Coûteux
Cas-témoins	<ul style="list-style-type: none"> – Économique – Rapide – Adaptée aux maladies rares – Adaptée aux maladies ayant un long temps de latence – Possibilité d'étudier plusieurs expositions 	<ul style="list-style-type: none"> – Risque de biais de sélection (comparabilité des cas et des témoins) – Biais de mesure du fait d'un recueil rétrospectif des données – OR biaisé si prévalence élevée de la maladie

* Un autre avantage de la cohorte est que l'on peut constituer une cohorte historique (par opposition à contemporaine) : ce type de cohorte est notamment adapté pour les maladies ayant un long temps de latence.

La figure ci-dessous montre la différence entre cohorte historique et étude cas-témoins.



EXERCICES_ÉTUDES_CAS_TÉMOINS



CHECK-LIST DES MOTS CLÉS

ÉTUDE CAS-TÉMOINS

TYPE D'ÉTUDE

- Épidémiologique.
- Observationnelle.
- Étiologique.
- Rétrospective.

POPULATION D'ÉTUDE

- Risque de biais de sélection des cas, des témoins si leur sélection dépend de leur niveau d'exposition au facteur de risque étudié.

CRITÈRE DE JUGEMENT : MESURE DE LA MALADIE

- Risque de biais de mesure (par erreur de classement en cas et en témoins).

MESURE DE L'EXPOSITION AU FACTEUR DE RISQUE

- Risque de biais de mesure (par erreur de classement en exposé-non exposé des cas et des témoins et par biais de mémoire),
- Limité par :
 - mesure du facteur standardisée et identique pour tous,
 - enquêteur en insu du statut cas-témoins,
 - utilisation de données objectives plutôt que déclaratives.

INTERPRÉTATION DES RÉSULTATS

Comparaison des groupes

- Risque de non comparabilité des cas et des témoins sur des facteurs de confusion.

Mesure de l'association facteur - maladie

- OR (odds ratio) ou RC (rapport de cotes) :
 - sa force,
 - son intervalle de confiance à 95 %,
 - statistiquement significatif ou non.

Prise en compte (contrôle) d'autres expositions ou d'autres facteurs importants

- Risque de biais de confusion :
 - contrôlé par l'ajustement sur les facteurs de confusion,
 - contrôlé par une analyse multivariée.

Discussion de l'influence des biais de sélection, de mesure et de confusion sur les résultats

Fil rouge :
origine
bactérienne des
conjonctivites

Les conjonctivites infectieuses aiguës sont une pathologie fréquente avec une incidence annuelle de 1,5 à 2 % en médecine générale. Les études ne retrouvent une origine bactérienne à la conjonctivite (prévalence de bactéries pathogènes) que dans 50 % des cas (IC95 % : 45 % à 54 %). En pratique, plus de 80 % des patients reçoivent des antibiotiques. Ainsi, de nombreux traitements antibiotiques locaux sont prescrits de façon inutile et inappropriée source de coûts inutiles et surtout de possibles développements de résistances des bactéries aux antibiotiques. Pour sélectionner les patients qui pourraient bénéficier le plus d'un traitement antibiotique local, les généralistes ont besoin d'un test diagnostique pour identifier les causes bactériennes. La plupart des généralistes font la distinction entre une cause bactérienne et une autre cause sur la base des symptômes. Les investigations diagnostiques supplémentaires, telles que la culture de prélèvements conjonctivaux, sont rarement faites, essentiellement à cause du délai nécessaire pour obtenir les résultats.

Est-il possible de différencier les origines bactériennes des origines virales sur la base de l'anamnèse et de l'examen clinique seuls ? Nous vous souhaitons évaluer la capacité d'un score clinique composé de plusieurs symptômes et antécédents pauvres le diagnostic étiologique dans la conjonctivite aiguë.

Rappels sur les tests diagnostiques

Ce sont des outils de mesure qui doivent permettre de classer les sujets comme «malade» (présentant la pathologie qui est recherchée) ou «non malade» (ne présentant pas la pathologie qui est recherchée).

Toute la nécessité d'évaluer ces tests diagnostique vient du fait qu'il s'agit de tests **imparfaits**, qui ne donnent jamais ou très rarement une certitude quant à la présence ou à l'absence de la maladie rechercher.

Il existe deux aspects complémentaires concernant la capacité d'un outil diagnostique à atteindre son objectif :

- la **reproductibilité** ou **fiabilité** : le test diagnostique donne-t-il les mêmes résultats quels que soient les conditions d'application,
- la **validité** ou **exactitude** (accuracy en anglais): le test diagnostique mesure-t-il ce qu'il est censé mesuré ?

Remarque : ce chapitre concerne les outils diagnostiques au sens large : toute méthode qui apporte de l'information pour faire évoluer la probabilité qu'un patient ait une maladie donnée et qui va apporter une aide à la décision thérapeutique. Ce sont donc non seulement les tests diagnostiques proprement dit (les dosages biologiques, l'imagerie, examen clinique,...) Mais également tous les outils de mesure comme les questionnaires, les grilles d'évaluation, etc..

Dans ce chapitre, nous traitons les études de **validité** (ou **exactitude**) **diagnostique** permettant de mesurer les **performances** d'un test diagnostique c'est-à-dire globalement sa capacité à classer des sujets en « malade » ou « non malade ». Ceci dépend donc à la fois de la **reproductibilité** (**fiabilité**) et de la **validité** (**exactitude**) du test.

La reproductibilité peut être mesurée par les indices de concordance : le pourcentage de concordance et le kappa, la validité est mesurée par les paramètres d'exactitude diagnostique qui sont la sensibilité, la spécificité, les ratios de vraisemblances positif et négatif et les valeurs prédictives positive et négative. Les ratios de vraisemblance permettent de quantifier l'information apportée par le test quand il est positif ou négatif à partir de la sensibilité et de la spécificité du test. C'est une façon d'exprimer les qualités intrinsèques du test.

Le principe pour évaluer tous ces paramètres et de comparer les résultats du test évalué (appelé parfois test index) à ceux d'un test de référence. Ce texte de référence (ou Gold standard) doit donner un diagnostic de certitude sur la présence ou l'absence de la maladie.

NIVEAU 2

Il ne faut pas parler d'efficacité d'un test diagnostique. Un test n'est pas « efficace » comme peut l'être un traitement, un test est fiable et exact. On peut parler en revanche de l'efficacité d'une stratégie diagnostique et thérapeutique. Par exemple la recherche par l'utilisation systématique d'un test diagnostique (questionnaires) qui est censé identifier les personnes dépressives permet-il d'améliorer la qualité de la vie des patients atteints de cancer du poumon métastatique. La réponse à cette question dépendra donc non seulement de l'exactitude et de la fiabilité du test mais également des traitements et de la prise en charge entreprise à

partir des résultats de ce test. Les études permettant de mesurer l'efficacité d'une stratégie diagnostique utilisent la méthodologie de l'essai clinique randomisé car on se trouve alors dans le domaine de l'évaluation de l'efficacité d'une intervention (stratégie diagnostique) sur un critère clinique (voir le chapitre LCA des essais cliniques).

8.1 LA QUESTION DE RECHERCHE

Une étude évaluant la performance d'un test diagnostique consiste à quantifier sa capacité à discriminer les malades et non malades. Pour pouvoir faire cela, il est nécessaire de connaître le statut malade ou non malade des sujets et donc de disposer d'un gold standard. L'étude comparera le test étudié et la méthode de référence ou gold standard.

Ici, le **PICO** ou le **PFCO** ne fonctionnent pas. Les composantes de la question de recherche sont la **Population étudiée**, le **Test évalué**, et le **test de référence (Gold Standard)** qui permet d'affirmer le statut malade ou non malade.

La table suivante (**tableau de contingence**) est très précieuse pour repérer la question de recherche. Un conseil : dès que vous avez identifié une étude d'évaluation d'un test diagnostique, construisez cette table avec les éléments dont vous disposez dans le texte ou dans les tableaux et si vous voulez être sûrs de ne jamais vous tromper, construisez-la toujours dans le même sens (test évalué en horizontale gold standard, en verticale).

Test évalué	Malade : GS +	Non malades : GS -	
Positif	Vrai positif	Faux positif	Valeur prédictive positive : VP/Positifs
Négatif	Faux négatif	Vrai négatif	Valeur prédictive négative : VN/Négatifs
	Sensibilité = VP/malades	Spécificité = VN/non malades	

8.2 LE TYPE D'ÉTUDE

Fil rouge :
origine
bactérienne des
conjonctivites

Neuf médecins généralistes travaillant dans 25 centres de soins dans la région d'Amsterdam et d'Alkmaar ont inclus des patients présentant un œil rouge et des sécrétions muco-purulentes ou des paupières collées.

Les études diagnostiques sont le plus souvent **transversales** car les sujets inclus sont malades ou non malades au moment de leur inclusion dans l'étude. Les informations obtenues par le test évalué d'une part et le Gold standard d'autre part sont recueillies en même temps ou dans un intervalle de temps assez court pour que le statut du patient n'ait pas eu le temps de changer (ces deux tests sont censés mesurer la même chose). Il n'y a pas de notion de suivi des patients, mais dans certains cas, le suivi est nécessaire pour confirmer ou infirmer la présence de la maladie au moment de l'inclusion.

Dans certains cas, notamment lorsque que le Gold standard est un examen invasif on ne peut pas le réaliser chez tous les sujets pour des raisons éthiques évidentes. Une option est de suivre les patients pendant une durée définie afin de voir s'ils développent la maladie, et d'obtenir ainsi un équivalent de Gold standard différé. Dans ces cas il y aura alors un suivi longitudinal et prospectif des patients.

Selon les modalités de recrutement des individus participant à l'étude, on peut distinguer 2 types de schéma d'étude **TRÈS différents**.

- **Le recrutement des individus se fait selon leur statut « malade » ou « non malade »** : le nouveau test est étudié au sein de deux groupes distincts, l'un constitué de « malades » et l'autre de « non malades ».

Attention ce type d'étude est fortement exposé au risque de biais de sélection car souvent le groupe des malades est assez sélectionné et non représentatif de toutes les formes de la maladie, de même le groupe des non malades peut-être très différent du groupe des personnes à qui on proposera ce test dans la vraie vie. **Dans ce type d'étude, les valeurs prédictives ne sont donc pas interprétables car la prévalence est imposée par la sélection et ne représente pas la réalité.** Cette méthode est en général utilisée au départ du processus de développement d'un nouveau test diagnostique mais

doit être ensuite complétée par des études dont la méthode figure dans le paragraphe suivant.

- **Le recrutement des individus se fait sans connaître leur statut « malade » ou « non malade »** et le nouveau test est étudié au sein d'une « cohorte » d'individus dont on ne connaît pas le statut vis-à-vis de la maladie. Les groupes comparés ne sont donc pas constitués *a priori*, mais *a posteriori* une fois passés les deux examens (le test évalué et le gold standard) de façon indépendante. Ce schéma d'étude permet de constituer un échantillon représentatif d'une population donnée, en particulier en termes de prévalence de la maladie. Le Gold standard est particulièrement important dans ce schéma d'étude car il permet de déterminer le statut malade ou non malade des individus, permettant d'estimer la sensibilité chez les malades et la spécificité chez les non malades. **C'est la meilleure méthode, celle qui apporte le moins de risque de biais de sélection.**

Fil rouge :
origine
bactérienne des
conjonctivites

Ici les médecins ont inclus les patients présentant des signes de conjonctivite avant que l'origine bactérienne ou virale ne soit déterminée par le Gold standard (culture bactérienne). Les patients vont recevoir les 2 tests et seront ensuite classés en « malade » (ici culture bactérienne positive) ou « non malade » (culture négative). Une autre méthode (néanmoins beaucoup moins performante) aurait été de constituer un groupe de patients dont le prélèvement bactériologique était positif (les « cas »), et de le comparer à un groupe de patients dont le prélèvement bactériologique était négatif (les « témoins »). On aurait alors estimé la sensibilité du nouveau test chez les malades et la spécificité chez les non malades.

Retenir :

- Existence d'un gold standard : étude transversale, Se, Sp, LR, courbe ROC.
- Gold standard invasif : étude transversale + prospective (gold standard réalisé chez les T+ +/- chez un échantillon de T-).
- Pas de Gold standard : étude prospective avec outcome clinique ou essai randomisé comparant des stratégies diagnostiques et thérapeutiques ou méthodes statistiques pour estimer les performances du test.
- Nb. Dans les études comparant un groupe de malades à un groupe de non-malades, c'est l'investigateur qui choisit le nombre des malades ainsi que le nombre des non malades à qui il va faire passer le test évalué. Il est donc évident que les valeurs prédictives n'ont aucun sens et qu'il est interdit de les calculer. Dans ce type d'étude seule la sensibilité qui est mesurée chez les malades (en verticale sur le tableau de contingence) et la spécificité mesurée chez les non malades peuvent être mesurées.

Remarque

Ces deux types d'études correspondent aux différentes phases de l'évaluation de tests diagnostiques. Les études comparant des malades à des non-malades sont adaptées à la phase précoce d'évaluation d'une nouvelle méthode diagnostique. Les études sur un échantillon représentatif dont le statut vis à vis de la maladie est inconnu sont adaptées à la phase tardive de l'évaluation, quand l'objectif est d'estimer les performances du test dans la population dans laquelle le test sera utilisé. Par exemple le développement d'un nouveau dosage biologique passe par plusieurs étapes successives.

La première étape est de mesurer les performances diagnostiques du nouveau test dans des sous-groupes connus de malades et de non malades. La deuxième étape est de la valider dans la population à laquelle s'adressera le test.

Par exemple, on veut développer un outil de dépistage de la dépression utilisable en médecine générale. Il s'agit d'un questionnaire comportant seulement de deux questions. Dans un premier temps on applique ce questionnaire à un groupe de patients dépressifs pour lesquelles le tableau est évident et en un groupe de personnes extrêmement en forme et dynamique. Il s'agit d'avoir une première estimation grossière de la capacité de ce test à distinguer les dépressifs des non dépressifs. En effet si cette première étape n'est pas validée il faut revoir le questionnaire. En revanche une fois cette première étape validée il est impératif d'évaluer dans une deuxième étape les capacités de ce test dans une cohorte d'individus dont on ne sait pas a priori s'ils sont dépressifs ou non.

8.3 LA POPULATION ÉTUDIÉE

Fil rouge :
origine
bactérienne des
conjonctivites

Les critères d'inclusion étaient un âge supérieur ou égal à 18 ans, des symptômes durant depuis moins de cinq jours, les critères de non inclusion étaient une baisse brutale de la vision, le port de lentille de contact, l'utilisation d'antibiotique par voie générale ou locale dans les deux semaines précédentes, kératite, traumatisme de l'œil, et des antécédents de chirurgie oculaire. Les patients étaient recrutés pendant les heures ouvrables seulement. Ces données ont été collectées lors de la consultation initiale.

La constitution de l'échantillon étudié doit absolument tenir compte des conditions d'utilisation du test en pratique. Il faut se poser la question suivante : à qui et quand le test sera-t-il proposé et réalisé en pratique ? les conditions d'évaluation du test sont-elles bien celles dans lesquelles l'examen sera réalisé en pratique ?

Le classement en « malade » et « non-malade » est déterminé par le résultat du Gold-standard.

LES « MALADES »

Il faut s'assurer qu'ils sont représentatifs des patients atteints de la maladie étudiée dans la population cible dans laquelle le test sera utilisé en pratique.

Si le test évalué doit permettre de diagnostiquer une maladie quel que soit son stade, il faut que tous les stades soient représentés dans l'échantillon étudié et non seulement certains stades ou formes de la maladie. La population de l'étude doit donc représenter un large spectre de la maladie. Par exemple pour le cancer du sein : tous les stades infra-clinique aussi bien que ceux cliniquement palpables.

LES « NON-MALADES »

Il faut s'assurer qu'ils sont représentatifs des sujets sains de la population cible dans laquelle le test sera utilisé en pratique.

Fil rouge :
origine
bactérienne des
conjonctivites

▀ Ici la population est recrutée en médecine générale ce qui constitue un élément en faveur de la représentativité de l'échantillon étudié, tant au niveau des personnes « malades » (porteuses d'une conjonctivite bactérienne) qu'au niveau des personnes non malades. En revanche certains patients à risque de conjonctivite sont exclus comme les enfants et les porteurs de lentilles de contact ce qui peut être un problème.

8.4 LE TEST ÉVALUÉ

Fil rouge :
origine
bactérienne des
conjonctivites

▀ À l'inclusion de chaque participant le généraliste remplissait un questionnaire standardisé et un examen clinique (index test). Le questionnaire contenait des questions sur les antécédents médicaux, la durée des symptômes (en jours), l'auto médication et l'auto traitement, le prurit, les sensations de brûlure ou de corps étranger dans l'œil, et le nombre d'yeux avec les paupières

colées le matin (0, 1 ou 2). L'examen clinique incluait l'évaluation du degré de rougeur de l'œil (périphérique, toute la conjonctive, ou toute la conjonctive et en périphérie de la cornée) la présence d'un œdème péri orbitaire, la nature des sécrétions (aqueuses, muqueuses ou purulentes), l'atteinte bilatérale (oui ou non). Le généraliste pratiquait alors un prélèvement par écouvillon au niveau conjonctival pour chaque œil pour mise en culture bactérienne (standard de référence). Pour chaque patient un œil était désigné comme l'œil étudié. Dans le cas de l'atteinte de deux yeux, l'œil présentant les symptômes les plus importants était l'œil étudié. Dans le cas où les deux yeux étaient affectés de façon similaire, l'œil affecté en premier était l'œil étudié.

- Il doit être **précisément décrit** afin que le lecteur soit en capacité de les reproduire à partir des données présentées dans l'article.
- Il doit être réalisé pour tous les individus **avec la même méthode**.
- On doit connaître sa **reproductibilité**, un même observateur/examineur qui mesure plusieurs fois un même patient (reproductibilité intra-observateur) et pour plusieurs observateurs/examineurs différents qui mesurent le même patient (reproductibilité inter-observateurs).

Fil rouge : ici la reproductibilité du score clinique n'est pas mentionnée.

- Acceptable pour les patients
- Utile pour les patients

Certaines règles doivent être respectées :

- il faut s'assurer que **tous les individus** ont eu le test étudié **et** la méthode de référence,

Fil rouge : tous les patients ont eu l'examen clinique et l'établissement du score clinique et la culture de prélèvement conjonctival.

- ils doivent être réalisés avec un **délai entre les 2 tests le plus court possible** dans le temps afin de s'assurer qu'il n'y ait pas d'évolution de la maladie entre les 2.
- **Attention point très important ++ :**

Le test évalué doit être réalisé **en insu du statut de l'individu et donc en insu des résultats de la méthode de référence/Gold standard** : ce point est capital. En l'absence d'insu, l'interprétation du test peut être influencée (**biais de mesure**) par la connaissance du statut « malade »

ou « non malade » et on risque une sur-estimation des performances du test. Remarque : difficile d'assurer l'insu lorsque le design est « de type cas témoins ».

Fil rouge : Le généraliste ne recevait pas le résultat des cultures

8.5 LE GOLD STANDARD (OU TEST DIAGNOSTIQUE DE RÉFÉRENCE)

Les résultats du test de référence sont censés être exacts à 100 % puisqu'ils définissent le statut malade *versus* non malade même si, en pratique, aucun test n'est fiable à 100 %...

Fil rouge :
origine
bactérienne des
conjonctivites

Dans notre exemple, il existe probablement quelques cultures bactériennes qui n'ont pas poussé et quelques cultures qui sont dues à une contamination.

- Le test de référence s'approche en fait plus ou moins d'un véritable Gold standard en fonction de sa capacité à classer sans erreur. Par exemple l'examen anatomopathologique est un des meilleurs gold standard mais parfois le gold standard est très imparfait par exemple pour le diagnostic de la sclérose en plaques ou de la polyarthrite débutante. Là aussi on doit parfois se servir de l'évolution ultérieure pour valider *a posteriori* le Gold standard.

Fil rouge : Procédure microbiologique.

Le généraliste pratiquait alors un prélèvement par écouvillon au niveau conjonctival pour chaque œil pour mise en culture bactérienne (standard de référence).

Les médecins généralistes pratiquaient le prélèvement en roulant une coton tige (Laboratoire Service Provider, Velzen-Noord, Pays-bas) sur la conjonctive au niveau du cul de sac inférieur. Ils mettaient ensuite le coton tige dans un milieu adéquat pour le transporter et l'envoyer au laboratoire de l'étude à Alkmaar. Directement après l'arrivée du prélèvement, nous avons inoculé les cotons tiges sur de l'agar enrichi avec 5 % de sang de mouton, de l'agar MacConkey et de l'agar chocolat. Tous les milieux étaient fabriqués au laboratoire avec des ingrédients standards (Becton Dickinson, Cockeysville, MD, USA). Après les inoculations standards, nous avons incubé l'agar avec des plateaux d'agar McConkey pendant 48 heures à 35 °C, nous avons incubé les plateaux

d'agar chocolat pendant la même durée et à la même température mais dans une atmosphère à 7% de CO₂. Nous avons analysé les cultures quotidiennement en suivant les recommandations de pratiques de la Société Américaine de Microbiologie. Nous avons identifié tous les germes pathogènes en utilisant les procédures biologiques de routine standardisées. Toutes les colonies suspectes d'être pathogènes ont été sélectionnées et investiguées par une coloration de Gram. En fonction des résultats de la coloration de Gram, nous avons fait des tests supplémentaires. En cas de cocci Gram plus, on a fait un test de la catalase suivi par un test de détection de la coagulase (staphylocoque) ou la recherche de la sensibilité à l'Opochine (pneumocoque). Dans le cas des bâtonnets et Coques Gram négatif, nous avons fait des tests au sucre.

- **Le test de référence doit aussi être réalisé en insu du statut de l'individu et des résultats du test à évaluer.** Ce point est surtout important si le gold standard est imparfait. S'il s'agit par exemple d'un examen d'imagerie, comme l'I.R.M., et que le test évalué est une échographie, la connaissance du résultat de l'échographie peut induire une modification à la fois dans la réalisation de l'I.R.M. (exploration plus poussée des zones qui apparaissent anormales en échographie) et dans son interprétation. Le risque est moins grand s'il s'agit d'un gold standard entaché d'un très faible risque d'erreur comme l'anatomopathologie, seul les cas les plus difficiles et litigieux seront susceptibles d'être influencés par la connaissance des résultats du test évalué.

Fil rouge : le microbiologiste qui analysait les cultures n'avait pas connaissance des résultats des tests cliniques.

Exemple : si on compare 2 techniques d'échographie pour le diagnostic de kystes ovariens, endovaginale et pelvienne, alors il ne faut pas que le même examinateur réalise les 2 examens, la réalisation et l'interprétation du second examen seraient influencée par le 1^{er}.

- Il doit être réalisé pour tous les individus avec la même méthode.
- **Le délai entre test étudié et GS le plus court possible** dans le temps afin de s'assurer qu'il n'y ait pas d'évolution de la maladie entre les 2.
- Reconnu : par exemple l'examen anatomopathologique (suite à une exérèse chirurgicale ou une biopsie).

REMARQUE 1 : la situation est particulièrement complexe lorsqu'il n'y a pas de bon Gold standard. C'est le cas en imagerie, chaque fois qu'on développe un examen qui est censé améliorer la capacité de classement par rapport aux examens existants. Par exemple une I.R.M. par rapport à un scanner, les cas de discordance IRM positive /scanner négatif, cela traduit-il une meilleure sensibilité de l'I.R.M. (moins de faux négatifs) ou à l'inverse un manque de spécificité par rapport au scanner (plus de faux positifs) ? il est difficile d'avoir un niveau de preuve élevé et on doit souvent attendre plusieurs études avant de progresser ou avoir recours à des méthodes statistiques complexes au-delà du niveau de ce cours.

REMARQUE 2 : comme évoqué précédemment dans un certain nombre de cas il n'est pas possible de disposer des résultats du Gold standard chez tous les patients. C'est souvent le cas pour le diagnostic des cancers. Il est alors difficile de déterminer si un test diagnostique est meilleur qu'un autre. Par exemple pour évaluer les performances de l'I.R.M. pour détecter un cancer du sein, on aura la confirmation histologique que dans les cas où au moins un des deux examens est positif (si l'imagerie est négative on ne peut ni techniquement ni éthiquement proposer une biopsie...). Ceci permettra d'estimer la valeur prédictive positive de l'I.R.M. (ligne supérieure du tableau de contingence correspondant aux tests positifs) mais pas d'estimer la sensibilité ni la spécificité (qui nécessitent de savoir si les tests négatifs sont des faux négatifs ou des vrais négatifs). Une possibilité est alors de suivre les femmes pendant un an afin d'identifier celles chez lesquelles un cancer du sein a été finalement diagnostiqué dans les mois suivants et qui étaient probablement des faux négatifs.

NIVEAU 2

En l'absence de test de référence reconnu, il peut s'agir d'un faisceau d'arguments cliniques et para-cliniques. Le diagnostic certain de sclérose en plaque par exemple reste parfois très difficile dans les formes débutantes.

Fil rouge :
origine
bactérienne des
conjonctivites

Pour le test évalué - score clinique - les items à recueillir sont bien décrits. Le Gold standard est reconnu et validé. Les techniques employées sont décrites très précisément, il est tout à fait possible de les reproduire avec les informations présentées.

8.6 ANALYSE STATISTIQUE

Des acquis sont à maîtriser, ci-dessous brièvement les principaux points mais nous vous encourageons vivement à reprendre ces notions plus en détail, elles vous ont été enseignées dans les cours d'épidémiologie et statistiques.

Rappels

- Dans le cas des tests diagnostiques, le but est de classer des individus en « malade » et « non malades ». On classe les résultats de manière dichotomique : test positif et test négatif.
- Sensibilité et spécificité : ce sont les caractéristiques intrinsèques du test. Elles doivent être présentées avec leur intervalle de confiance à 95%.

Elles permettent de comparer les performances des tests car elles sont « intrinsèques » et **théoriquement indépendantes du contexte** c'est-à-dire du **risque de base de la population** dans laquelle le test est appliqué. Cependant, ces caractéristiques intrinsèques du test dépendent souvent des caractéristiques des malades pour la sensibilité, et des caractéristiques des non malades pour la spécificité.

Test à évaluer	Malades	Non malades
Positif	VP	FP
Négatifs	FN	VN
	Sensibilité = parmi les malades % test +	Spécificité = parmi les non malades % test –

Sensibilité : chez les malades, % des tests positifs (probabilité que le test soit positif chez les malades).

Spécificité : chez les non malades, % des tests négatifs (probabilité que le test soit négatif chez les non malades).

Exemple : capacité de l'I.R.M. à faire le diagnostic de cancer du sein devant une « boule » dans le sein. (exemple issu du cours de l'ISPED, Bordeaux)

Sensibilité :

I.R.M. cancer du sein
(tumeur maligne)

Positive	71	
Négative	3	$Se = \frac{71}{74} = 96\%$
Total	74	

Parmi ces 74 patientes atteintes d'un cancer du sein, l'I.R.M. préopératoire réalisée avant de connaître les résultats de la chirurgie et de la biopsie, était positive dans 71 cas et négatifs dans trois cas. Il y avait donc 71 vrais positifs (l'I.R.M. avait donné la bonne réponse sur la présence et cancer) et trois faux négatifs (l'I.R.M. était négative alors qu'il s'agissait bien d'un cancer). On peut dire que l'I.R.M. a fourni le bon résultat chez 96 % des femmes atteintes d'un cancer du sein.

Spécificité :

I.R.M. Pas de cancer du sein
(tumeur bénigne)

Positive	28	
Négative	80	$Sp = \frac{80}{108} = 74\%$
Total	108	

Parmi les 108 femmes qui ne présentent finalement pas de cancer du sein, l'I.R.M. a donné un résultat vrai chez 74 %, mais elle était positive à tort chez 28 de ces 108 patientes..

Les valeurs prédictives positive et négative

Les valeurs prédictives d'un test dépendent non seulement de ses qualités intrinsèques mais aussi et surtout de la prévalence de la maladie dans la population étudiée. Elles estiment la probabilité d'être malade si le test est positif (VPP) ou de ne pas être malade si le test est négatif (VPN), Quand il s'agit d'un contexte de diagnostic, c'est-à-dire de la confirmation de la présence d'une maladie chez un patient qui consulte pour des symptômes de cette maladie, la prévalence est

en général relativement élevée. En revanche, dans le contexte d'un dépistage systématique de masse où on réalise le test chez des sujets qui n'ont pas de symptômes, la prévalence de la maladie est en général très faible. (C'est pourquoi il est essentiel de vérifier un certain nombre de critères supplémentaires avant de proposer un test diagnostique pour un dépistage de masse). Les valeurs prédictives sont donc dépendantes de la prévalence. La prévalence de la maladie dans la population testée peut être différente de la prévalence retrouvée dans la pratique du clinicien, ce qui peut en limiter l'interprétation.

Valeur prédictive positive (VPP): chez les individus ayant un test positif % de malades (probabilité qu'un individu ayant test positif soit malade).

Valeur prédictive négative (VPN): chez les individus ayant un test négatif % de non malades (probabilité qu'un individu ayant test négatif soit réellement indemne de la maladie).

VPP et VPN doivent être présentés avec leurs intervalles de confiance à 95 %.

Test évalué	Malades	Non malades	Valeurs prédictives
Positif	VP	FP	VPP = parmi les tests + % de malades

Test évalué	Malades	Non malades	Valeurs prédictives
Négatifs	FN	VN	VPN = parmi les tests - % de non malades

Dans l'exemple précédent :

Valeurs prédictives (dans un contexte de confirmation du diagnostic) :

Se = 96 %, Sp = 72 %, prévalence

(= probabilité d'avoir la maladie : $\frac{74}{182} = 41\%$, ce qui est élevé)

I.R.M.	Cancer du sein	Tumeur bénigne	Total	
Valeur Prédictive				
Positive	71	28	99	$VPP = \frac{71}{99} = 72\%$
Valeur Prédictive				
Négative	3	80	83	$VPN = \frac{80}{83} = 96\%$
<p>Valeur prédictive positive : L'I.R.M. est positive chez 99 patientes, mais parmi elles seulement 71 ont réellement une tumeur maligne du sein, autrement dit quand l'I.R.M. est positive, il y a une probabilité de 72 % que la tumeur du sein soit maligne.</p>				
<p>Valeur prédictive négative : L'I.R.M. est négative chez 83 patientes, et chez 80 d'entre elles, l'examen histologique retrouve une étiologie bénigne, en revanche l'I.R.M. est faussement négative chez 3 femmes qui ont en réalité une tumeur maligne.</p>				
<p>REMARQUE :</p> <p>impact de la prévalence de la maladie (contexte d'utilisation du test).</p> <p>Dans l'exemple précédent la prévalence de cancer est élevée à 41 % car ces femmes ne sont pas arrivées par hasard elles ont été sélectionnées à la suite d'un certain nombre de symptômes évocateurs ou de facteurs de risque génétique ou d'antécédents familiaux... Si on se place en revanche dans le contexte du dépistage systématique la plupart des femmes n'ont aucun symptôme clinique, ni d'antécédents familiaux particuliers et la prévalence est beaucoup moins élevée. Imaginons qu'elle soit par exemple de 0,25 %. En conservant exactement les mêmes sensibilité et spécificité de l'I.R.M. et en les appliquant à ces 50 tumeurs malignes et à ces 20 000 femmes sans tumeur maligne, cette prévalence beaucoup plus basse induit une modification radicale des valeurs prédictives. Autrement dit si l'on utilisait l'I.R.M. en dépistage de masse lorsque, celle-ci est positive il n'y aurait que $\approx 1\%$ de chances que la femme ait réellement un cancer du sein...</p>				

Valeurs prédictives (dans un contexte de confirmation du dépistage)

Pas de modifications de la Se = 96 % (et Sp = 72 %)

Mais Prévalence $\ll P = 50./20.000 = 0.25 \%$

I.R.M.	Cancer du sein	Pas de cancer du sein	Total
Positive	48	5200	5248
Négative	2	14800	14802
Total	50	20000	20050

$$VPP = \frac{48}{5248} = 0.9 \% \quad VPN = \frac{14800}{14802} = 99.9 \%$$

NIVEAU 2**Les rapports de vraisemblance**

La troisième catégorie de paramètres d'évaluation d'un test diagnostique correspond aux **rapports/ratios de vraisemblance (likelihood ratios)** qui traduisent la vraisemblance d'un résultat donné chez des sujets malades comparée à la vraisemblance de ce résultat chez des sujets non malades. Ces ratios sont basés sur les caractéristiques intrinsèques du test :

Le rapport de vraisemblance positif LR+ : correspond au rapport de la probabilité d'avoir un test positif chez les malades (sensibilité) sur la probabilité d'avoir un test positif quand on est non malade (1-spécificité). Il a une forte valeur discriminante pour affirmer le diagnostic.

$$LR+ = \frac{\text{probabilité d'avoir un test positif chez les malades}}{\text{probabilité d'avoir un test positif chez les non malades}} = \frac{\text{sensibilité}}{1 - \text{spécificité}}$$

Pour info :

LR+ > 10 \Leftrightarrow valeur importante ; LR+ = 5-10 \Leftrightarrow modérée ;LR+ = 2-5 \Leftrightarrow faible ; LR+ = 1-2 \Leftrightarrow très faible

Le rapport de vraisemblance négatif LR- : correspond au rapport de la probabilité d'avoir un test négatif chez les malades (1 - sensibilité) sur la probabilité d'avoir un test négatif chez les non malades (spécificité). Il a

une forte valeur discriminante pour éliminer le diagnostic.

$$LR- = \frac{\text{probabilité d'avoir un test négatif chez les malades}}{\text{probabilité d'avoir un test négatif chez les non malades}} = \frac{1 - \text{sensibilité}}{\text{spécificité}}$$

Pour info :

LR- < 0,1 \Leftrightarrow valeur importante ; LR- = 0,1-0,2 \Leftrightarrow modérée ;

LR- = 0,2-0,5 \Leftrightarrow faible ; LR- = 0,5-1 \Leftrightarrow très faible.

Les ratios de vraisemblances permettent d'exprimer l'information apportée par le test lorsque le test est positif et lorsque le test est négatif.

Nb. Un test diagnostique dont les LR sont égaux à 1 n'a pas d'intérêt diagnostique.

Ratios de vraisemblance (dans un contexte de confirmation du diagnostic)

IRM	Tumeur maligne	Tumeur bénigne	Total
Positive	71	28	99
Négative	3	80	83
Total	74	108	182
	$LR + = \frac{71/74}{28/108} = 3,7$ $LR - = \frac{3/74}{80/108} = 0,05$		

Interprétation, en règle générale :

- On considère que le test présente un intérêt **important en clinique**,
 - pour confirmer un diagnostic lorsque le test est positif :
quand $LR+ \geq 10$,
 - pour exclure un diagnostic lorsque le test est négatif :
quand $LR- \leq 0,1$.
- On considère le test comme **acceptable** du point de vue de son intérêt clinique,
 - pour confirmer un diagnostic lorsque le test est positif :
quand $5 \leq LR+ \leq 10$,
 - pour exclure un diagnostic lorsque le test est négatif :
quand $0,1 \leq LR- \leq 0,2$.

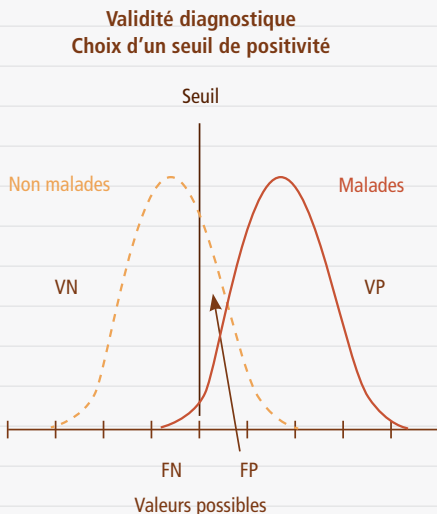
Intérêt des rapports de vraisemblance :

1. Ils sont indépendants de la prévalence de la maladie car ils dépendent de la sensibilité et de la spécificité du test qui sont indépendants de la prévalence.
2. Ils permettent de quantifier l'information apportée par le test.
3. Ils permettent de passer de la probabilité pré-test de la maladie à la probabilité post-test grâce au nomogramme de Fagan.
(Nb. : on utilise le LR+ si le test est positif et le LR- si le test est négatif)
4. On peut combiner ces rapports de vraisemblance quand on réalise plusieurs tests indépendants de façon consécutive car ils permettent le calcul de la probabilité post-test lorsqu'on connaît la probabilité pré-test (prévalence dans la population testée).

Remarques : la probabilité pré-test de la maladie est estimée par la prévalence de la maladie en population générale si le test est appliqué en population générale ou bien dans la population correspondant à celle à laquelle le test sera appliqué.

Exemple : Dépistage anténatal de la trisomie 21, technique utilisée pour interpréter le triple-test sérologique (HCG, AFP, E3 libre). La prévalence initiale est le risque moyen observé dans la population en fonction de l'âge maternel, à laquelle on applique les LR de l'HCG, AFP et E3 libre.

Prévalence initiale x LR clarté nucale x LR sérologie = prévalence ou risque final.

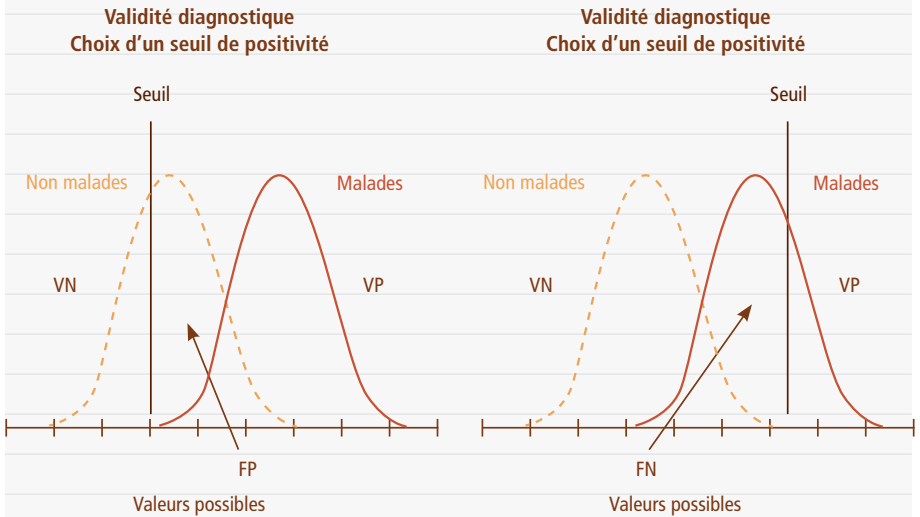


La sensibilité et la spécificité ne sont calculables que si le résultat du test est une variable dichotomique (positif ou négatif). Quand ce test donne un résultat quantitatif, comme un dosage biologique par exemple, il sera alors nécessaire de choisir un seuil de positivité.

En général, des distributions des valeurs du test chez les malades et les non malades se recouvrent plus ou moins, générant des faux positifs et des faux négatifs. Il s'agit alors de trouver le meilleur compromis entre

les 2. Le corollaire du déplacement du seuil est que lorsqu'on le déplace, la sensibilité et la spécificité vont varier en sens inverse. En quelque sorte, on n'a rien sans rien ! Le choix du seuil de positivité est délicat et nécessite de considérer les

conséquences des faux positifs et des faux négatifs. Si la maladie est très grave et qu'il existe des traitements très efficaces, un test faussement négatif aura des conséquences graves surtout s'il existe un traitement efficace aux stades précoces de la maladie. À l'inverse, si la maladie est peu grave, si les examens diagnostiques sont très invasifs ou si les traitements proposés sont très agressifs alors les faux positifs peuvent avoir également des conséquences lourdes...



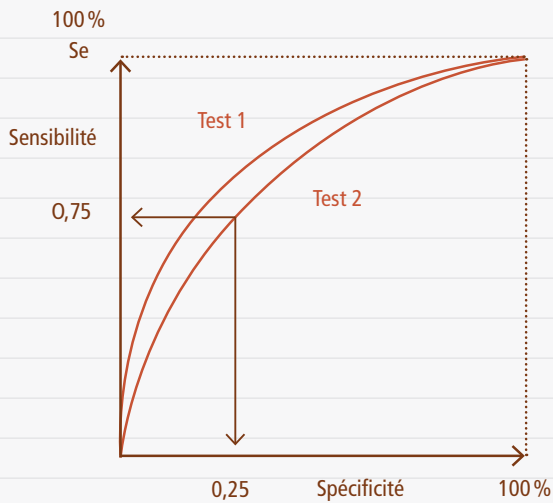
On pourrait déplacer le seuil de telle façon qu'il n'y ait plus de faux négatifs... Mais on voit bien sur la figure ci-dessus que cela aurait comme conséquence d'augmenter énormément les faux positifs. À l'inverse, on pourrait déplacer le seuil pour éliminer les faux positifs mais alors une grande partie des malades serait négatif et donc considéré à tort comme non malades....

On utilise pour trouver la valeur seuil optimale la courbe ROC (receiver operating characteristics) :

Les résultats des couples Se-Sp en fonction des seuils sont représentés sur une courbe ROC dont les axes sont : $(1-Sp)$ en abscisse, et (Se) en ordonnée. L'aire sous la courbe ROC (Area under the Curve ou « AUC » en anglais) permet de déterminer l'intérêt du test. On recherche une aire sous la courbe supérieure à 0.50, une aire égale à 0.5 signifie que le test n'apporte rien. Plus la courbe se rapproche du coin supérieur gauche, meilleure est la performance du test (combinaison Sp et Se). Lorsque plusieurs tests sont comparés ; ils peuvent l'être par la comparaison des aires sous la courbe ROC.

Exemples :

- Pour une pathologie grave, que l'on sait soigner, mais peu fréquente : on choisit un seuil de positivité du test qui donne une grande sensibilité pour favoriser les vrais positifs (sensibilité élevée), quitte à augmenter le nombre de faux positifs (spécificité basse).
- Pour une maladie fréquente dont le traitement est lourd : on choisit un seuil de positivité du test qui donne une grande spécificité pour favoriser les vrais négatifs et éviter de traiter des malades dépistés à tort (faux positifs), quitte à augmenter le nombre de faux négatifs (sensibilité basse).



Dans cette figure, 2 tests sont comparés, chaque courbe correspond à 1 test. Le Test 1, dont la courbe ROC est constamment au-dessus de celle du Test 2, semble donc avoir une meilleure performance globale (à la fois Se et Sp). Il faudra ensuite appliquer des méthodes statistiques spécifiques pour savoir si cette différence entre les 2 tests est au-delà du hasard (statistiquement significative).

Le calcul du nombre de sujets nécessaires

Il doit avoir été effectué *a priori* en explicitant les hypothèses utilisées. Par exemple, la taille de l'échantillon a été calculée pour être en capacité d'estimer la sensibilité du test avec une précision de plus ou moins 5%. Pour calculer le nombre de malades nécessaire pour atteindre cet objectif on a besoin de faire une hypothèse sur la sensibilité attendue. Il est à noter que dans les études d'évaluation de tests diagnostiques il est nécessaire de déterminer le nombre de malades et le nombre de non malades.

Les méthodes dépendent du type d'étude groupe de malades et de non-malades ou échantillon représentatif d'une population sont le statut malade ou non-malade n'est pas connu *a priori*.

Reproductibilité du test : Évaluation de la concordance entre plusieurs observateurs

Pour mesurer la reproductibilité d'un test (dans quelle mesure il donne le même résultat quand il mesure la même chose) on utilise les indices de concordance. Pour une variable binaire (présent/absent), on peut mesurer le pourcentage de concordance observée et la concordance prévisible par pur hasard. Le coefficient Kappa est plus approprié que le pourcentage de concordance observé car il tient compte de la concordance attendue par le simple fait du hasard (à pile ou face). Ne pas prendre en compte l'effet du hasard peut conduire à de fausses conclusions en surestimant la concordance réelle.

Le coefficient kappa est défini ainsi :

$$\text{Kappa} = \frac{\text{concordance observée} - \text{concordance attendue}}{1 - \text{concordance attendue}}$$

Lorsque deux mesures sont concordantes à un degré qui n'est pas supérieur au hasard, la valeur de kappa est 0. Lorsque les deux mesures sont parfaitement concordantes, kappa est égal à 1.

À titre indicatif : si k 81-100 % \Leftrightarrow concordance excellente ; k entre 61 et 80 % \Leftrightarrow concordance bonne ; k entre 41 et 60 % \Leftrightarrow concordance moyenne ; k entre 21 et 40 % \Leftrightarrow concordance faible ; k entre 0 et 20 % \Leftrightarrow concordance mauvaise.

Exemple : les radios du rachis d'un groupe de patients sont analysées indépendamment par 2 radiologues qui recherchent la présence de tassements de vertèbres.

Pourcentage de concordance due au hasard =

$$((21 \times 39) + (196 \times 178)) / (217)^2 = 0,76$$

Pourcentage de concordance observée = $(14 + 171) / 217 = 0,85$

$$\text{Kappa} = (0,85 - 0,76) / (1 - 0,76) = 0,39$$

		Radiologue 2		
		Tassement vertébral présent	Tassement vertébral absent	
Radiologue 1	Tassement vertébral présent	14	7	21
	Tassement vertébral absent	25	171	196
		39	178	217

8.7 LES RÉSULTATS : PERFORMANCES DU TEST ÉTUDIÉ

Les résultats doivent présenter au minimum la sensibilité et la spécificité du test diagnostique évalué.

La sensibilité et la spécificité doivent être présentées avec leur IC à 95 %. Il faut pouvoir retrouver toutes les données pour pouvoir refaire le tableau de contingence 4 x 4 et recalculer ces indicateurs.

La sensibilité et la spécificité sont des paramètres essentiels pour comparer la performance des examens diagnostiques entre eux.

En revanche, leur interprétation clinique est peu intuitive. En effet la sensibilité par exemple représente, sachant que le sujet est malade la probabilité que son test soit positif ce qui ne correspond pas à la situation clinique habituelle. Ce sont donc deux paramètres qui intéressent surtout le soignant.

Les paramètres qui représentent la situation clinique sont les valeurs prédictives positives et négatives. La VPP par exemple donne, pour un patient dont le test est positif, la probabilité qu'il soit réellement malade. C'est donc bien ce paramètre qui semble utile pour la pratique, puisqu'il intéresse le patient. Malheureusement, les valeurs prédictives sont très influencées par la population dans laquelle est utilisé le test. Même si on ne vous demande pas d'être un spécialiste du théorème de Bayes, il faut impérativement en comprendre les conséquences. Malgré un même résultat du test diagnostique, de personnes peuvent avoir une probabilité très différente d'être malades en fonction de leur probabilité initiale (elle-même très difficile à connaître).

Par exemple un test VIH positif n'est pas associé à la même probabilité d'être réellement malade lorsqu'il est observé chez un jeune toxicomane ou chez une femme de 80 ans résidant en maison de retraite.

Vous devez être capable de donner l'interprétation des indicateurs de performance en « français ». Par exemple, en cas de VPP=70 %, la probabilité qu'un patient dont le test est positif soit réellement malade est de 70 %.

Fil rouge :
origine
bactérienne des
conjonctivites

Les couples Se-Sp correspondant à différents seuils du score clinique obtenus sur l'ensemble des sujets inclus dans l'étude sont présentés dans le tableau 1. La courbe ROC sera construite en reportant pour chaque seuil la valeur de la Se sur l'ordonnée et celle de 1-Sp sur l'abscisse.

Tableau 1. Sensibilité et spécificité du score clinique pour le diagnostic d'une conjonctivite d'origine bactérienne.

Score clinique	Sensibilité % (IC95%)	Spécificité % (IC95%)
30	100	54 (43-64)
40	96 (91-98)	73 (62-81)
50	90 (83-94)	88 (79-93)
60	76 (68-83)	90 (82-95)
70	69 (61-76)	95 (88-98)
80	63 (54-70)	98 (92-99)
90	61 (53-69)	100

8.8 LES BIAIS

BIAIS DE SÉLECTION

Le biais de sélection survient lorsque la population étudiée ne correspond pas à la population à laquelle le test sera appliqué.

Soit les malades ne sont pas représentatifs de la population des malades. Par exemple tous les stades de la maladie ne sont pas représentés, seul un stade particulier a été étudié. Exemple : on étudie une nouvelle technique de mammographie uniquement chez des femmes présentant une tumeur palpable. C'est souvent le cas si l'étude est menée dans un centre hyperspécialisé (centre anticancéreux, centre de référence, centre hospitalo-universitaire...).

Soit les non-malades ne sont pas représentatifs de la population saine.

De façon générale le risque de biais est beaucoup plus important lorsque les groupes sont constitués sur la base de leur statut « malade » ou « non malade ». Dans ce type d'étude le groupe des malades est général très sélectionné et n'est pas représentatif de toutes les formes de la maladie. Ce risque est d'autant plus fort si l'on étudie des malades sans tenir compte de l'ancienneté de leur maladie. En effet dans ce cas seront sélectionnés des patients dont la maladie n'est pas trop agressive car les patients atteints des formes les plus sévères sont peut-être déjà décédés. Dans ce type d'étude on ne se trouve pas dans la situation clinique réelle puisqu'il ne s'agit pas de diagnostiquer des maladies débutantes mais des maladies déjà présentes parfois évoluées, les stades précoces de la maladie n'ayant pas encore été confirmé, sont souvent peu représentés Les sujets non malades sont par ailleurs, souvent particulièrement sains. Ceci conduit souvent à surestimer les performances d'un nouveau test à la phase précoce de son évaluation.

L'avantage des études « de type cohorte » est qu'on ne sait pas à l'avance quels sont les sujets qui sont atteints de la maladie ainsi tous les patients sont à un stade initial de leur maladie ce qui correspond bien à la situation clinique réelle.

BIAIS DE MESURE = BIAIS D'ÉVALUATION

Le biais de mesure peut survenir si les tests ne sont pas réalisés en insu et de manière indépendante. La connaissance du résultat du test de référence permet de savoir si le patient est malade ou non et va introduire un biais majeur dans

l'interprétation du test à évaluer.

Dans une moindre mesure, la connaissance des résultats du test à évaluer peuvent influencer l'interprétation du test de référence. Le risque est surtout présent lorsque le test de référence n'est pas Gold standard très solide (on parle de Gold standard imparfait).

8.9 LA CONCLUSION

La discussion doit apprécier les résultats (Se, Sp) et les conséquences en pratique selon la prévalence de la maladie (VPP et VPN).

La discussion doit évaluer les applications cliniques :

- selon la fréquence de la maladie en pratique clinique,
- selon les autres stratégies diagnostiques existantes (définir la place du nouveau test),
- selon les conditions de faisabilité technique et économique (acquisition, installation et réalisation du test en pratique clinique courante),
- selon les conséquences des faux positifs et des faux négatifs.

Le test doit apporter une information utile pour la décision diagnostique et thérapeutique, et son utilisation doit entraîner une amélioration de l'état de santé. Cette amélioration de l'état de santé pourra être décrite dans des études complémentaires visant à démontrer l'efficacité d'une stratégie diagnostique par un essai contrôlé randomisé.

PRINCIPES GÉNÉRAUX DE LA LECTURE CRITIQUE D'ARTICLES ORIGINAUX

9.1 LES QUATRES POINTS CARDINAUX DE LA LCA

1. La validité interne :

- Est-ce que j'ai confiance dans les résultats ?
- Les résultats sont-ils fiables ?
- Le résultat observé est-il réel et ne résulte-t-il pas seulement d'un biais de l'étude ?

2. La pertinence clinique (ampleur ou taille de l'effet, précision de l'estimation) :

- Ce traitement apporte-t-il un bénéfice important/intéressant/cliniquement pertinent pour les patients?
- Ce facteur de risque est-il associé à une forte augmentation du risque de maladie ?

3. La cohérence externe :

- Ce résultat est-il concordant avec les autres connaissances sur le sujet ?
- Ce résultat est-il isolé ou existe-t-il tout un faisceau de preuves concordantes ?

4. La représentativité (ou validité externe) :

- Ce résultat est-il extrapolable ? Les patients de l'étude sont-ils représentatifs des (ressemblent-ils aux) patients que je rencontre couramment? (dans mon pays, mon établissement, mon cabinet, ma région ?

QUELLE EST LA VALIDITÉ INTERNE DE L'ÉTUDE ?

C'est une question essentielle à laquelle tous les étudiants doivent savoir répondre.

Si la validité interne est trop mauvaise, les autres questions perdent de leur intérêt. Les informations se trouvent dans les sections « matériel et méthode » et « résultats ».

QUELLE EST LA PERTINENCE CLINIQUE DE CETTE ÉTUDE ?

Les questions sur ce point sont difficiles pour vous donc entraînez-vous (ce seront les questions les plus intéressantes lorsque vous serez devenus praticien !)

Elles nécessitent votre jugement personnel basé sur des notions acquises dans d'autres matières sur :

- la pertinence des critères de jugement. Par exemple : Gagner en moyenne 2 mètres de périmètre de marche ou 3 jours de survie, Gagner 2 points sur une échelle de qualité de vie, => Quelle est la pertinence de ces résultats pour les patients ?
- L'importance de l'effet observé. Par exemple un RR à 1,15 [1,07-1,20] est-il pertinent ?

Les informations se trouvent dans les sections « matériel et méthode » et « résultats ».

QUELLE EST LA COHÉRENCE EXTERNE ?

Une étude ne suffit en général pas à porter une conclusion formelle, surtout si son niveau de preuve est bas.

Les informations sont à rechercher dans « introduction » et « discussion ».

Elles sont généralement sous la forme de données **quantifiées, objectives et référencées**.

QUE VAUT LA VALIDITÉ EXTERNE (OU EXTRAPOLABILITÉ, TRANSPOSABILITÉ) ?

Puis-je m'attendre à avoir les mêmes résultats si j'applique cela à la population que je prends en charge ?

Ceci fait appel à la **notion de risque de base** (incidence et/ou prévalence de la maladie dans la population étudiée).

Avec le nouvel ECNi plus intégré, on peut s'attendre à ce type de question qui fait appel à des connaissances médicales. De plus, on peut également trouver des **informations dans « introduction » et « discussion » pour ce qui concerne la population et dans la partie « résultats » pour ce qui concerne l'échantillon étudié.**

9.2 CONCLUSION

Peut-on répondre OUI aux 4 questions précédentes ?

Autrement formulé, existe-t-il des réserves sur la réalité et/ou la pertinence du résultat ?

Si oui, il est nécessaire d'attendre les résultats d'autres études.



ÉTUDE DIAGNOSTIQUE

EXEMPLE DE GRILLE DE LECTURE

Adaptation d'après "KT Clearinghouse" funded by the Canadian Institute of Health Research (CIHR)

1. LE TEST DE RÉFÉRENCE : GOLD-STANDARD (DÉFINITION DU STATUT MALADE OU NON MALADE)	OUI	NON	NA
Le test étudié est-il comparé à un examen de référence « gold standard » qui permet de classer malades et non malades ?			
Le test de référence est-il précisément décrit ?			
Le test de référence est-il validé (qualité du test et choix du seuil définissant malades et non malades) ?			
Le test de référence est-il pratiqué chez tous les patients inclus dans l'étude malades et non malades ?			
Si oui, est-il réalisé dans les mêmes conditions chez les malades et les non malades ?			
Le test de référence est-il réalisé en insu des résultats du test étudié ?			
2. LE TYPE DE L'ÉTUDE « PLAN EXPÉRIMENTAL »	OUI	NON	NA
Un seul échantillon de personnes dont on ne connaît pas le statut « malade » ou « non malade » ?			
Si oui, la fréquence de la maladie dans l'échantillon est en accord avec les données épidémiologiques connues ?			
Un échantillon de « malades » et un échantillon de « non malades » ?			
Les échantillons de malades et de non malades sont-ils représentatifs d'une population à laquelle le test sera appliqué ?			
L'échantillon de patients étudiés comporte un « panel » de patients suffisamment varié (en termes de formes de la maladie et de niveaux d'évolution et de sévérité) ?			

3. LES PERFORMANCES DU TEST ÉTUDIÉ (À ÉVALUER)	OUI	NON	NA
Les conditions de réalisation du test étudié sont précisément décrites ?			
Le test étudié est interprété indépendamment du test de référence			
Le test étudié présente un intérêt (plus rapide, plus simple, moins coûteux, moins invasif...) ?			
Sensibilité \pm IC 95 %			
Spécificité \pm IC 95 %			
Rapport vraisemblance (test positif)			
Probabilité pré test			
Probabilité post test (test positif)			
4. QUELLE EST L'APPLICABILITÉ DES RÉSULTATS ?	OUI	NON	NA
Le lieu de l'étude est décrit			
La méthode de réalisation du test est décrite et compatible avec votre pratique			
La reproductibilité du test est évaluée			
Les risques de conséquences négatives du test sont mesurés			
l'utilisation du test améliore l'état de santé des patients			

ÉTUDE DIAGNOSTIQUE

CHECK-LIST DES MOTS CLÉS

TYPE D'ÉTUDE

- Étude transversale le plus souvent.
- Recrutement de type cas / témoins : un groupe de malades et un groupe de non-malades.
- Recrutement de type cohorte : un groupe d'individus est étudié indépendamment du statut malade/non malade qui n'est pas connu initialement.

LES TESTS

- Description précise des 2 tests.
- Test de référence reconnu.
- Passation des 2 tests à l'ensemble des individus dans les mêmes conditions.
- Réalisation et interprétation des tests en insu l'un de l'autre – indépendamment.
- Étude de la reproductibilité intra et interobservateur.

LA POPULATION ÉTUDIÉE

- Représentative de la population chez qui le test sera réalisé en pratique clinique.
- Représentation de l'ensemble des stades de la maladie pour lesquels le test sera proposé.

LES ANALYSES ET LES RÉSULTATS

- Se, Sp, VPP, VPN, intervalle de confiance à 95 %.
- LR+ et LR-.
- Courbe ROC, aire sous la courbe.

LES BIAIS À DISCUTER

- Biais de sélection, biais de mesure ou d'évaluation.

STATISTIQUES

Proposition d'après le polycopié de LCA de Paris Descartes (2014-2015)

Ce chapitre a pour vocation de vous aider à comprendre quelques bases statistiques afin de pouvoir interpréter au mieux les résultats des différentes analyses auxquelles vous pourriez être confrontés lors de la lecture critique d'articles.

10.1 ORGANISATION GÉNÉRALE DE L'ANALYSE STATISTIQUE

L'analyse statistique suit souvent le même plan quel que soit le type d'article :

- **Une première phase descriptive.** Souvent trouvée dans le tableau 1, cette partie permet de décrire la population étudiée dans son ensemble et les 2 groupes comparés.
- **Une deuxième phase d'analyse univariée ou bivariée** qui s'intéresse à l'association entre une variable (un facteur de risque, un facteur pronostique, un traitement ou une autre intervention) et le critère de jugement. Cette phase permet de trier quelles variables ont a priori associées statistiquement au critère de jugement. Elle permet ainsi de sélectionner celles qui seront intégrées dans l'analyse multivariée (en général les variables associées au CJ avec un « p » inférieur à 0,2 ou même plus petit, entre 0,2 et 0,05).
- **Une troisième phase d'analyse multivariée** qui cherche à trouver quelles variables sont toujours associées au critère de jugement (= indépendamment associées au CJ), une fois qu'on a pris toutes les autres en compte (c.-à-d. une fois qu'on a « ajusté » l'analyse pour les autres variable.

10.2 DIFFÉRENTES CATÉGORIES DE VARIABLES

- **Variables quantitatives** : elles mesurent des « quantités ». Vous devez pouvoir dire : « le patient A a une valeur de cette variable supérieure au patient B ». Elles peuvent être **discrète** ou **continue** :
 - **Une variable discrète** a une valeur finie. Il est possible de les énumérer (« 1, 2, 3, ... »). On peut généralement l'énoncé sous la forme « Le nombre de... ». **Exemples** : nombre d'items dans une liste, nombre de personnes dans une salle.

- **Une variable continue** peut prendre, en théorie, une infinité des valeurs, formant un ensemble continu. **Par exemple, le temps de réussite d'une tâche sera compris entre 0 et 300 secondes, et pourra prendre les valeurs 12,235689 ou 12,235699999.**
- **Variables qualitatives ou catégorielles** : elles mesurent juste des « états », des catégories, des données sans notion de grandeur : Oui ou non, homme ou femme, Code postal, numéro de téléphone. Elle est binaire si elle ne peut prendre que 2 valeurs (ex. : oui/non, absent/présent, masculin/féminin, etc.).
- **Variables censurées** : elles expriment un délai avant la survenue d'un évènement (décès, rechute, rémission, etc). Ces variables sont surtout retrouvées dans les études de cohortes et les essais thérapeutiques. On les appelle censurées car à la fin de l'étude, pour un certain nombre de patients, on n'aura pas un temps jusqu'à l'évènement, mais un temps jusqu'à la fin de leur suivi ou la fin de l'étude. **Par ex. : dans une étude de survie on aura un temps jusqu'au décès pour les patients décédés, mais un temps jusqu'à la date de point (= date d'arrêt de l'étude) pour les patients non décédés, et un temps jusqu'à la date des dernières nouvelles pour les patients perdus de vue.**

10.3 LES DIFFÉRENTES ANALYSES STATISTIQUES POSSIBLES

LES ANALYSES DESCRIPTIVES

L'analyse descriptive sert à décrire les différents paramètres étudiés.

Variables qualitatives

Prévalence

Proportion de malades à un instant T :

$$\text{prévalence} = \frac{\text{nombre de malades}}{\text{population totale}}$$

La prévalence est une proportion, on l'exprime souvent sous forme de pourcentage. Elle est influencée par la durée de la maladie (plus la maladie est longue, plus il y aura de malades) et la vitesse d'apparition des nouveaux cas (plus la maladie se propage vite, plus il y a de malades). La prévalence est estimée sur un échantillon de population, elle est donc exprimée avec un intervalle de confiance à 95 % qui représente l'intervalle dans lequel on trouverait 95 % des estimations si on les répétait à l'infini.

Incidence

Nombre de nouveaux cas pendant une période donnée :

$$\text{Incidence} = \frac{\text{nombre de nouveaux cas sur la période}}{\text{nombre de personnes} \times \text{temps de suivi}}$$

L'incidence est une estimation de la vitesse moyenne d'apparition de nouveaux cas, on l'exprime sous forme de nombre de cas par personne-temps (personnes-années, personnes-mois, personnes-jours...). L'incidence est aussi estimée sur un échantillon de population, et s'exprime donc avec un intervalle de confiance à 95%.

Variables quantitatives

Les variables quantitatives sont décrites par un **paramètre de position** et un **paramètre de dispersion**.

Paramètres de position

- **Moyenne** : valeur moyenne d'une variable,
- **Médiane** (= 2^e interquartile = 50^e percentile) : valeur pour laquelle la moitié de l'effectif est au-dessus de cette valeur, et l'autre moitié en dessous.

La médiane est moins sensible que la moyenne aux valeurs extrêmes (**ex. : dans une population de 100 personnes, quelques personnes très jeunes feront que la moyenne peut être modifiée de plusieurs années, tandis que la médiane sera peu affectée par quelques valeurs extrêmes**).

La moyenne est un bon paramètre de position lorsque la répartition des valeurs suit une distribution normale. Sinon, il vaut mieux utiliser la médiane.

Paramètres de dispersion

- **La Variance et l'Écart-type** (racine carrée de la variance, standard deviation en anglais) sont associés à une moyenne et représente l'écart moyen des valeurs par rapport à la moyenne. Plus ils sont faibles, plus les valeurs sont regroupées autour de la moyenne (**Par exemple pour la répartition des notes d'une classe, plus l'écart type est faible, plus la classe est homogène**). À l'inverse, s'ils sont plus importants, les notes sont moins resserrées autour de la moyenne.
- **Intervalle interquartile** : Le 1^{er} quartile d'un paramètre représente la valeur qui sépare la population en 25 % en dessous et 75 % au-dessus, le 2^e quartile (= médiane) correspond à la valeur qui sépare la population en 50 % en

dessous et 50 % au-dessus, le 3^e quartile à la valeur qui sépare la population en 25 % en dessous et 75 % au-dessus. (Par exemple, une population avec un poids minimum de 40 kg, dans laquelle le 1^{er} quartile est 50 kg, le 2^e 60 kg, le 3^e 80 kg et le maximum 95 kg. L'intervalle entre le 1^{er} et le 3^e quartile = entre le 25^e et le 75^e percentile, exprime le fait que 25 % de la population présente une valeur inférieure à l'intervalle, et 25 % une valeur supérieure). L'intervalle interquartile est souvent présenté à côté de la médiane.

Variabes censurées

Les variables censurées sont décrites par des **courbes de survie**, souvent représentées par la méthode de **Kaplan-Meier**. On représente :

- **En abscisse** : la durée de « survie »,
- **En ordonnée** : la probabilité de ne pas avoir présenté l'évènement ou la proportion de patients n'ayant pas encore présenté l'évènement.

Remarque : les courbes de survie ne sont pas utilisées uniquement lorsque l'évènement est le décès, on peut aussi s'intéresser au délai avant une rechute ou avant la survenue d'un évènement.

Lorsque l'on s'intéresse à la survie à proprement parler, la courbe démarre à 1 (tous les patients sont en vie) et « descend ». Lorsque l'on s'intéresse à la survenue d'un évènement, la courbe démarre à 0 (personne n'a présenté l'évènement) et « monte ».

Grâce aux courbes de survie, on peut estimer la médiane de survie qui correspond à la durée pour laquelle 50 % des sujets n'ont pas présenté l'évènement.

ANALYSES UNIVARIÉES (OU BIVARIÉES)

Les analyses bivariées étudient l'association entre 2 variables : un facteur de risque/facteur pronostique/traitement et un critère de jugement. Elles sont néanmoins souvent appelées analyses univariées dans les articles par opposition aux analyses multivariées.

Pour réaliser ces analyses, on émet des hypothèses et on fixe des risques d'erreurs concernant ces hypothèses.

Hypothèses

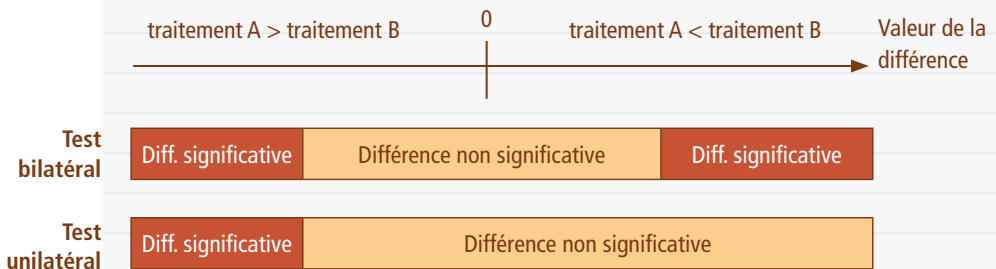
- **Hypothèse nulle (H0)** : absence de différence entre les deux groupes.
- **Hypothèse alternative (H1)** : différence entre les deux groupes.

Risques d'erreurs

- **Risque de 1^e espèce = risque alpha** = risque de conclure à tort à une différence = risque d'accepter H1 alors que H0 est vraie -> souvent fixé à 5%. Le risque alpha est dit bilatéral quand la différence peut aller dans les deux sens, unilatéral sinon.

Nb. : il faut faire attention à l'inflation du risque alpha lorsque les tests sont répétés, ou lorsqu'il existe des analyses intermédiaires. Ceci peut être corrigé grâce à la méthode Bonferroni qui consiste à « répartir » le risque alpha sur les différents tests.

Concrètement, fixer le risque alpha à 5% signifie qu'on accepte de conclure qu'il y a bien une différence entre les deux groupes avec un risque d'erreur de 5%.



- **Risque de 2^e espèce = risque bêta** = risque de conclure à tort à une absence de différence = risque d'accepter H0 alors que H1 est vraie -> $1 - \text{puissance}$
- **Puissance** = probabilité de montrer une différence qui existe réellement = probabilité d'accepter H1 alors que H1 est vraie -> fixé à au moins 80% (parfois 90%)

Concrètement, fixer la puissance à 90% signifie qu'on se donne 90% de chance de montrer qu'une différence existe bien. Plus le nombre de sujet est élevé, plus la puissance est grande.

		CONCLUSION	
		Différence (= on conclue que H1 est vraie)	Pas de différence (= on conclue que H0 est vraie)
RÉALITÉ	Différence (H1 est vraie)	1 – bêta = puissance	Risque bêta
	Pas de différence (H0 est vraie)	Risque alpha	

Pour les analyses univariées (bivariées), on fait souvent appel à des tests statistiques. Les résultats de ces tests sont exprimés par le « p », qui représente le degré de significativité. **Ce « p » correspond à la probabilité que le hasard puisse expliquer une différence au moins aussi grande que celle qui est observée dans les deux groupes.**

Lorsque « p » est inférieur à alpha, cela signifie que la probabilité que le hasard puisse expliquer la différence observée est inférieur au seuil de risque que l'on s'est fixé de conclure à tort à une différence. On peut donc conclure que la différence trouvée est statistiquement significative et qu'elle n'est pas due au hasard.

Si $p \geq \alpha$, cela signifie que la probabilité que le hasard puisse expliquer la différence observée est supérieure au seuil de risque que l'on s'est fixé de conclure à tort à une différence. Selon ce seuil fixé, on ne peut donc pas conclure que la différence trouvée n'est pas due au hasard. Cependant, on ne peut pas non plus conclure qu'il n'y a pas de différence.

Nb. : Une association statistiquement significative n'est pas forcément cliniquement significative. Il ne faut pas oublier de porter un regard critique sur les données et l'association mesurée, et les remettre dans le contexte de la pratique clinique.

Les tests statistiques

Les tests statistiques permettent de comparer 2 valeurs afin de déterminer s'il existe une différence « statistiquement significative » entre 2 ou plusieurs groupes.

Comparaison d'une variable qualitative entre 2 groupes :

- test du Chi 2 (test paramétrique -> sur des grandes populations, ou lorsque la répartition est normale),
- test exact de Fisher (non paramétrique).

Test paramétrique : hypothèses faites sur la distribution des valeurs.

Comparaison d'une variable quantitative entre 2 groupes :

- test t de Student (test paramétriques -> distribution normale des valeurs ou grand groupes),
- test de Wilcoxon (test non paramétrique).

Comparaison d'une variable quantitative entre plus que deux groupes :

- ANOVA (test paramétrique -> distribution normale des valeurs ou grand groupes),
- test de Friedmann (test paramétrique).

Comparaison d'une variable censurée entre 2 groupes :

- test du log rank

Association entre deux variables

Lorsque l'on veut évaluer l'association entre deux variables, on peut utiliser différents paramètres.

- Association entre deux variables qualitatives

Risque relatif : c'est le rapport des proportions de malades dans deux groupes. Il exprime le risque d'être malade dans un groupe par rapport au risque d'être malade dans l'autre groupe. On ne peut le calculer qu'à partir d'étude prospective donc cohortes et essais.

Odds Ratio : c'est le rapport des proportions d'exposés entre le groupe des cas et celui des témoins, il mesure de l'association entre un facteur de risque et une maladie et peut être utilisé quel que soit le type d'étude contrairement au risque relatif. Lorsque la prévalence de la maladie est faible (<10%) dans

la population cible, l'Odds Ratio est un bon estimateur du risque relatif. Plus la prévalence augmente, plus l'OR surestime le RR.

- **Hazard Ratio** : il s'utilise pour les variables censurées. Il correspond à l'équivalent du risque relatif pour des données censurées, c'est le rapport du risque instantané dans le groupe traité (h_1) divisé par le risque dans le groupe contrôle (h_0).

Ces paramètres sont établis sur un échantillon de personnes, ils représentent donc une estimation des valeurs « réelles ». Il faut donc les exprimer avec un intervalle de confiance, souvent fixé à 95%. Lorsque l'IC à 95% ne contient pas la valeur 1, on peut dire que le RR, l'OR ou le HR est statistiquement significativement différent de 1.

- Association entre une variable qualitative et une variable quantitative
Elle correspond à l'étude de l'association entre un facteur de risque quantitatif et une maladie. On utilise aussi les RR ou OR avec un intervalle de confiance à 95%, qui expriment l'augmentation du risque pour chaque unité de la variable quantitative. (Par exemple, si on considère le risque de diabète associé au poids, l'OR = 1,12 ICC [1,07-1,23] associé correspondra à l'augmentation du risque de diabète associé à l'élévation de 1kg du poids. Dans certains cas, les analyses sont faites non pas pour 1 unité mais pour 5 ou 10 unités ? Dans cet exemple les auteurs pourraient rapporter l'augmentation du risque de diabète correspondant à chaque élévation de 10kg de poids).
- Association entre deux variables quantitatives
On calcule un coefficient de corrélation compris entre -1 et 1. On teste ensuite si ce coefficient de corrélation est significativement différent de 0.
Concrètement :
Coefficient de -1 \Leftrightarrow données parfaitement corrélées, évoluant en sens inverse.
Coefficient de 0 \Leftrightarrow absence d'association, mais attention, cela ne veut pas dire que les deux variables sont indépendantes.
Coefficient de 1 \Leftrightarrow données parfaitement corrélées, évoluant dans le même sens.

High correlation : .5 to 1.0 or -0.5 to -1.0
 Medium correlation : .3 to .5 or -0.3 to -.5
 Low correlation : .1 to .3 or -0.1 to -0.3



ANALYSE MULTIVARIÉE

L'analyse multivariée permet de prendre en compte dans l'analyse ou un plusieurs tiers facteurs, c'est-à-dire autres que le facteur étudié (facteur de risque ou intervention). Cette analyse multivariée est justifiée par le fait que plusieurs facteurs peuvent influencer l'analyse univariée. On distingue différents types de facteurs :

- **Facteur intermédiaire** : facteur causé par l'exposition, et lui-même causant le critère de jugement. (ex. : traitements anxiolytiques qui augmentent les risque de chute, chute qui augmente à son tour la risque de fracture).
- **Facteur de confusion** : facteur associé au facteur d'exposition et au critère de jugement principal sans être un facteur intermédiaire.
 (ex. : si on étudie le lien entre teintures capillaires et cancer de la vessie (les teintures contiennent des facteurs cancérogènes qui sont suspectés d'augmenter le risque de cancer de la vessie chez les coiffeurs. Si la probabilité d'être fumeur (le tabac est un facteur de risque du cancer de la vessie) est associée à la probabilité de se teindre les cheveux (pour une raison X ou Y, comportement ou autre) alors il faut ajuster l'analyse pour la consommation de tabac afin de pouvoir dire : à consommation de tabac égale, la teinture capillaire est toujours associée à l'élévation du risque de cancer de la vessie ou l'inverse : à consommation de tabac égale, la teinture capillaire n'est plus significativement associée à l'élévation du risque de cancer de la vessie et dans ce cas, le tabac était bien un facteur de confusion, et la teinture n'est pas un facteur de risque de cancer de la vessie).
- **Facteur d'interaction** : facteur modifiant l'association entre le facteur d'exposition et le critère de jugement principal (appelé facteur modificateur d'effet).

(ex. : il a été montré que le lien entre teintures capillaires et cancer de la vessie était différent en fonction de certaines caractéristiques génétiques. Chez les personnes avec un phénotype N-acétyltransférase-2 (NAT2) qui sont acétylateurs lents la réalisation de teintures capillaires était associée à une multiplication du risque de cancer de vessie par 2.9-fold (OR=2.9 (95% CI = 1.2-7.5), alors que chez les acétylateurs rapides, l'OR était de 1.3 (95% CI = 0.6-2.8), donc non significativement différent de 1).

Pour faire la différence entre un facteur de confusion et un facteur d'interaction, on peut stratifier l'analyse selon ce facteur :

- si la relation entre le facteur d'exposition et le critère de jugement est indépendante du facteur, on la retrouvera dans toutes les strates ;
- si le facteur est un facteur de confusion, il explique en partie l'association entre le facteur d'exposition et le critère de jugement, l'analyse stratifiée fait disparaître ou atténue l'association dans chaque strate ;
- si le facteur est un facteur d'interaction, l'association sera différente dans chaque strate.

En cas d'interaction comme dans l'exemple ci-dessus, les OR sont différents chez les acétylateurs rapides et lents, cela n'a pas de sens d'en calculer une « moyenne ». Les résultats seront présentés séparément chez les acétylateurs rapides et les acétylateurs lents.

Unique facteur de confusion

On peut utiliser l'**ajustement de Mantel-Haenszel** qui permet d'estimer un OR ou un HR ajusté sur le facteur de confusion. On fait une sorte de moyenne des OR calculés pour chaque strate du possible facteur de confusion. On compare l'OR ou l'HR obtenu ainsi au paramètre brut :

- si ils sont proches, la relation facteur d'exposition-critère de jugement est indépendante du facteur, qui n'est donc pas un facteur de confusion ;
- si ils sont différents, la relation facteur d'exposition-critère de jugement est expliquée par le facteur qui est donc un facteur de confusion.

Il faut d'abord s'assurer que le facteur n'est ni un facteur intermédiaire, ni un facteur d'interaction.

REMARQUE : si le facteur est un facteur d'interaction, on fait une analyse en sous-groupe car il n'y a pas de sens à faire une analyse ajustée.

REMARQUE : la méthode d'ajustement de Mantel-Hanzel est de moins en moins utilisée dans les articles au profit des modèles de régression multivariés.

Modèles multivariés (plusieurs facteurs de confusion)

On utilise pour cela des **modèles de régression**. Ces modèles expriment le **critère de jugement (variable à expliquer)** en fonction d'**autres variables (variables explicatives)**.

Ce modèle permet d'identifier des facteurs indépendamment associés au critère de jugement, et « d'annuler » l'impact de facteurs de confusion.

On utilise différents modèles selon le critère de jugement.

	Critère de jugement binaire	Critère de jugement continu	Critère de jugement censuré
Analyse descriptive	Effectifs et pourcentages	Moyenne et écart-type Médiane et Q1-Q3	Courbes de Kaplan-Meier
Analyse univariée /bivariée (tests statistiques pour déterminer si la différence entre les groupes est statistiquement significative)	Test du Chi-2 (paramétrique) Test exact de Fisher (non paramétrique) Régression logistique univariée (OR bruts crude OR)	Test t de Student (paramétrique) Test de Wilcoxon (non paramétrique) Test de Mann Whitey (non paramétrique) Régression linéaire univariée (β coefficients)	Test du Log rank Modèle de Cox univarié (HR bruts)
Analyse multivariée (test statistique ajusté sur un ou plusieurs facteurs de confusion)	Régression logistique multivariée (OR ajustés *adjusted OR, adjOR)	Régression linéaire multivariée	Modèle de Cox (HR ajustés, adjHR)

* ajusté ne veut rien dire si on ne précise pas sur quelles variables de modèle est « ajusté »

Conception graphique :
Emmanuel Besson

Suivi de fabrication :
Irène Chatelus

Imprimé par :
Graphiscann / Vaulx-en-Velin

Octobre 2017