1 TABLE DES MATIÈRES

2 ASPECTS MÉTHODOLOGIQUES [1]

2.1 TYPOLOGIE DES ÉTUDES EN ÉPIDÉMIOLOGIE ET RECHERCHE CLINIQUE

2.1.1 Selon leur objectif

Une recherche débute avec une hypothèse que l'on souhaite tester. Cette hypothèse donne lieu à une formulation précise de l'objectif de la recherche. L'objectif conditionne la méthodologie à employer.

Objectifs	Schémas d'étude possible
Décrire la prévalence (1) ou l'incidence (2) d'un évènement de santé	1 : Etude descriptive transversale 2 : Etude descriptive longitudinale - cohorte
Rechercher les facteurs de risque d'une maladie ou les facteurs prédictifs (pronostiques) de la survenue d'une complication de maladie	Etude de cohorte à visée analytique Etude exposés/non-exposés Etude cas/témoins (Etude transversale) (Etude écologique)
Evaluer les propriétés d'un test diagnostique : Sensibilité/spécificité, validité (1) **Fiabilité (2)**	1 : Etude comparant les résultats du test à un gold standard 2 : Etude comparant les résultats du test répété dans plusieurs situations
Evaluer une thérapeutique : essai thérapeutique ou clinique (1) **Evaluer une intervention : action de prévention ou stratégie de dépistage (2)**	1, 2 : Essai contrôlé randomisé - RCT 2 : Etude quasi-expérimentale - ici/ailleurs, avant/après Etude de cohorte à visée analytique Etude exposés/non-exposés Etude cas/témoin (Etude transversale) (Etude écologique)

On trouvera en page 29 un tableau des schémas d'étude à privilégier selon les objectifs, validé par la Haute Autorité de Santé.

[1] Philippe Carrère, MD

2.1.2 Selon leur méthodologie

Une étude peut être :

- Interventionnelle ou expérimentale, si l'on mesure l'effet d'une exposition contrôlée par l'expérimentateur, ou intervention, sur un évènement ou maladie.

- Observationnelle, si l'exposition n'est pas sous contrôle de l'observateur.

Parmi les études observationnelles :

- Si les sujets se prêtant à la recherche sont ou ont été suivis au cours du temps avant survenue de l'évènement étudié, on parlera d'étude de cohorte.

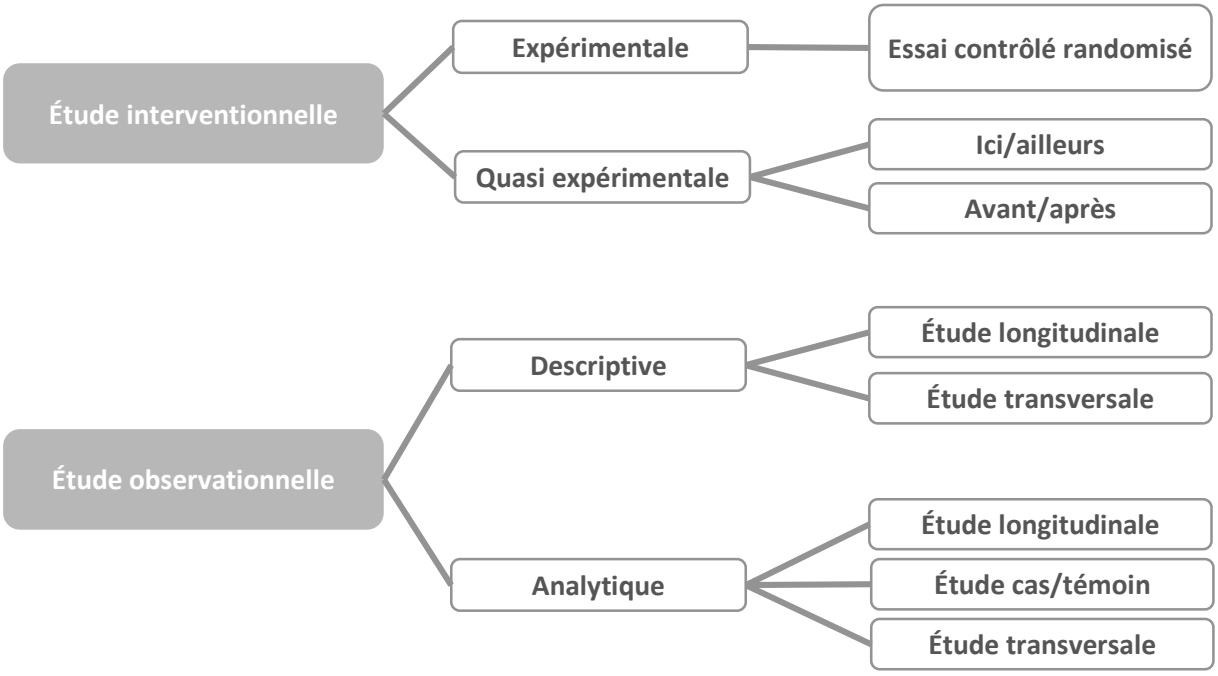
- Si les sujets n'ont pas été suivis au cours du temps et sont recrutés sur critère de survenue de l'évènement, on parlera d'étude cas/témoin.

- Dans tous les autres cas, on parlera d'étude transversale.

Les études observationnelles peuvent être :

- A visée analytique, si l'on teste une hypothèse d'association entre exposition et évènement.
  → *Etude de cohorte, cas/témoin, ou transversale*

- A visée descriptive, si l'on estime une incidence ou une prévalence.
  → *Etude de cohorte, ou transversale.*



2.1.2.1 Etude expérimentale

En recherche biomédicale, le but d'une étude expérimentale est d'évaluer l'efficacité d'une intervention sur un phénomène de santé.
→ *Par exemple un nouveau traitement, un dispositif médical, une technique chirurgicale, un programme d'éducation à la santé, une méthode diagnostique…*

Le gold-standard de l'étude expérimentale est l'essai contrôlé randomisé ou Randomized Controlled Trial (RCT).
→ *Par exemple, un RCT est systématiquement mis en œuvre en phase 3 des essais cliniques préalables à l'autorisation de mise sur le marché d'un nouveau médicament.*

2.1.2.1.1 Principes méthodologiques de l'essai contrôlé randomisé

Un RCT obéit à plusieurs principes méthodologiques fondamentaux.

- Position du problème correctement identifiée (revue de la littérature), objectif de l'essai précisément formulé, choix méthodologiques adaptés (type d'essai, intervention évaluée, conditions d'administration de l'intervention, population source, critère de jugement).

- Définition des critères d'éligibilité des sujets devant recevoir l'intervention, compte tenu de l'indication de l'intervention : critères d'inclusion et de non-inclusion ou exclusion. Ces critères doivent assurer l'homogénéité des groupes de sujets recevant ou non l'intervention évaluée. Plus la population éligible est homogène, moins la variabilité de la réponse est importante, moins le nombre de sujets nécessaires sera élevé (cf. chapitre 2.2.6.3).

- Calcul du nombre de sujets nécessaires (cf. chapitre 2.2.6.3) : le nombre de sujets inclus conditionne la puissance de l'analyse statistique, c'est-à-dire sa capacité à mettre en évidence une différence d'efficacité entre intervention évaluée et non-intervention ou intervention de référence. Ce calcul doit être effectué avant la mise en œuvre de la recherche.

- Comparaison : d'un groupe de sujets recevant l'intervention évaluée et d'un groupe témoin ne la recevant pas.
  → *Le groupe témoin peur recevoir un placebo (éthique ?) ou une intervention de référence.*

  Cette comparaison permet de faire la part entre l'évolution due à l'intervention évaluée, et l'évolution naturelle ou due à l'intervention de référence.

  Chaque sujet éligible et inclus doit être en mesure de recevoir ou non l'intervention évaluée (pas de contre-indication). C'est ce que l'on appelle la clause d'ambivalence.

- Equirépartition : les groupes intervention évaluée et témoin doivent être équilibrés en effectif. Cela permet de maximiser la puissance de l'analyse statistique.

- Randomisation : allocation aléatoire (par tirage au sort) de l'intervention aux patients se prêtant à la recherche. La probabilité de recevoir ou non l'intervention évaluée doit être identique pour chacun de ces sujets, indépendamment de leurs caractéristiques de base.

La randomisation assure la comparabilité initiale des groupes : ils ne doivent différer qu'à l'égard de l'intervention administrée.

La randomisation à lieu après l'inclusion des sujets éligibles à l'intervention. Elle peut être :

- Simple (pile ou face), mais l'équirépartition des sujets inclus et randomisés dans chaque groupe n'est pas garantie (espérance ≠ réalisation).
- Par bloc, pour obtenir des groupes équilibrés en effectif.
- Stratifiée sur un facteur, si ce facteur peut nuire à l'obtention de l'équirépartition dans les délais escomptés.
  → *Par exemple dans le cas des études multicentriques quand certains centres ont un faible niveau de recrutement.*
- Adaptative, une des autres méthodes pour obtenir l'équirépartition…

- Critère de jugement : il permet l'évaluation de l'efficacité de l'intervention. Il doit être précisément défini et sa mesure doit être standardisée (cf. chapitre 2.4.5).

- Aveugle ou insu : il assure l'égalité d'appréciation du critère de jugement au cours de l'essai, donc le maintien de la comparabilité des deux groupes. Tous les biais liés à la connaissance de la correspondance patient-intervention peuvent alors être évités.

  Cette clause d'ignorance n'est pas toujours possible. Un essai peut être :

  - Ouvert (pas d'insu).
    → *Par exemple, si l'intervention évaluée est chirurgicale vs médicale (l'éthique d'une « chirurgie blanche » est toujours discutable).*
  - En simple aveugle (le patient ne connait pas le traitement reçu).
  - En double aveugle (ni le patient ni le soignant ne connaissent le traitement reçu).
  - En aveugle complet (aucun des acteurs de la recherche ne connait la correspondance patient-traitement).

- Suivi : il doit être rigoureux et identique dans tous les groupes étudiés. Tous les moyens doivent être mis en œuvre pour limiter le nombre de perdus de vue et les données manquantes (cf. chapitre 2.4.6).

Ces principes limitent les risques de biais, ce qui permet d'imputer les résultats observés à l'intervention évaluée.

A noter qu'un diagramme de flux (flow chart) est souvent présenté dans les rapports de RCT : il décrit le nombre de patients à chaque stade de l'étude (sélection, inclusion, randomisation, allocation des traitements, suivi, analyse).

2.1.2.1.2 Essais de supériorité, d'équivalence ou de non-infériorité

L'objectif d'un RCT peut être…

- De comparer l'efficacité d'une intervention à un placebo ou une non-intervention. Cela peut poser des problèmes :
  - Ethiques, en particulier si l'on sait qu'une intervention efficace existe déjà.

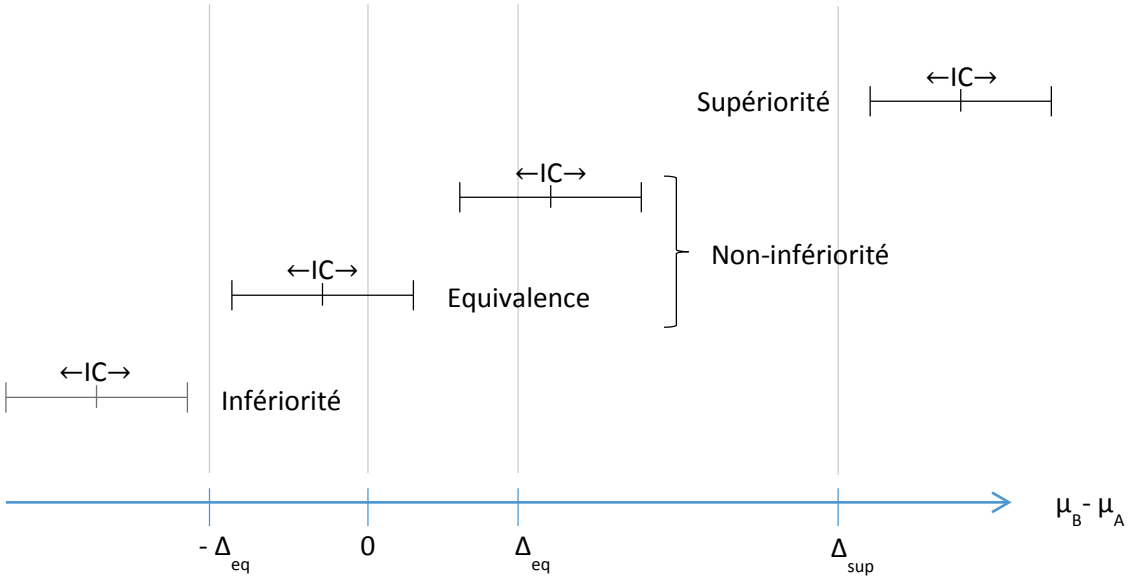
- D'attrition plus importante dans le groupe témoin, en particulier si l'inefficacité du placebo est perçue par les sujets se prêtant à la recherche.

- De comparer une nouvelle intervention A à une intervention de référence B, à condition que l'intervention de référence B ait déjà fait la preuve de son efficacité contre placebo dans le cadre d'une recherche bien conduite. Il peut alors s'agir :
  - D'un essai de supériorité, si l'hypothèse est que l'intervention A est plus efficace que l'intervention B. La supériorité sera démontrée si la différence d'efficacité entre les traitements A et B est supérieure à une certaine valeur jugée cliniquement pertinente (différence minimale cliniquement importante).
  - D'un essai d'équivalence, si l'hypothèse est que A et B sont d'efficacité équivalente.
  - D'un essai de non infériorité, si l'hypothèse est que B n'est pas moins efficace que A.

    Pour les essais d'équivalence ou de non infériorité, on tolèrera une marge d'équivalence.

    → *Les essais d'équivalence ou de non infériorité concernent les interventions dont on n'attend pas qu'elles soient d'efficacité supérieure, mais qui présentent d'autres avantages (tolérance, coût, facilité de mise en œuvre par exemple).*

Prenons pour exemple un essai thérapeutique dont l'objectif serait de comparer l'efficacité d'un nouveau médicament antihypertenseur (A) à celle d'un traitement plus ancien (B), le critère de jugement étant la Pression Artérielle Systolique (PAS). Le suivi des groupes A et B après administration des traitements va permettre de calculer la différence entre les valeurs moyennes des PAS observées dans chacun des groupes (μ_B-μ_A) ainsi que l'intervalle de confiance de cette différence (IC).



- On conclura à la supériorité de A sur B avec une différence minimale cliniquement importante Δ_{sup} si :

$$\Delta_{sup} \notin IC \text{ et si } \mu_B > \mu_A$$

On conclura à la non-supériorité de A sur B si :

$$\Delta_{sup} \in IC$$

- On conclura à l'infériorité de A sur B en tolérant une marge d'équivalence Δ_{eq} si :

$$IC \subset]-\infty;-\Delta_{eq}]$$

On conclura à la non-infériorité de A sur B si :

$$IC \subset]-\Delta_{eq};+\infty[$$

Sinon on ne conclue rien.

- On conclura à l'équivalence de A et B en tolérant une marge d'équivalence Δ_{eq} si

$$IC \subset]-\Delta_{eq};\Delta_{eq}]$$

On conclura à la non-équivalence de A et B si

$$IC \cap]-\Delta_{eq};\Delta_{eq}[= \emptyset$$

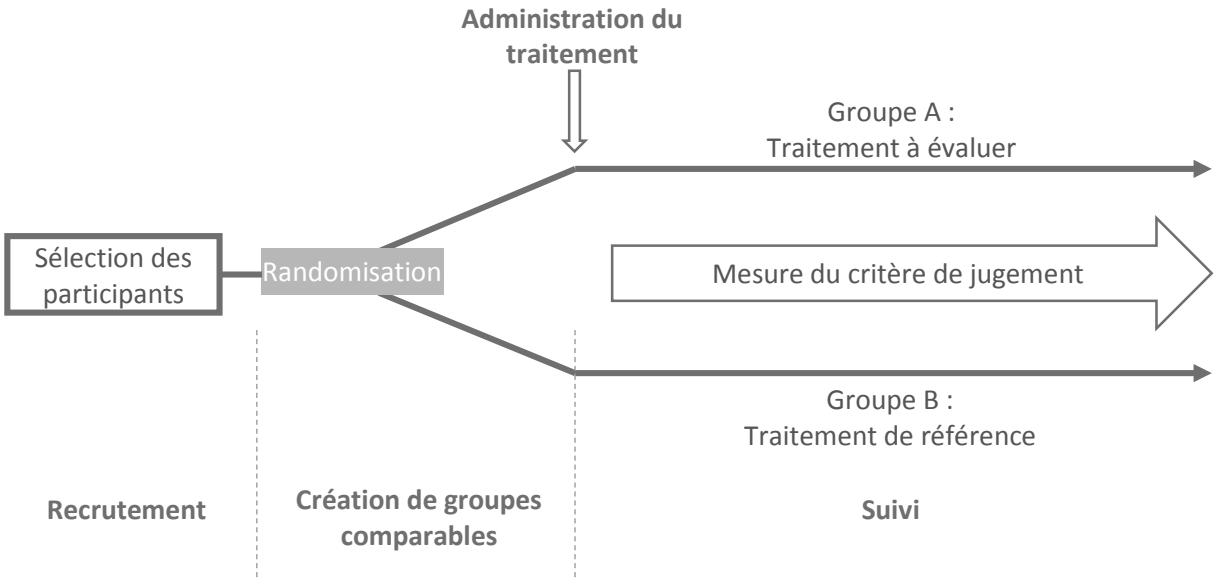
Sinon on ne conclue rien.

2.1.2.1.3 Population d'analyse

Pour les essais de supériorité, l'analyse doit être menée en Intention de Traiter (ITT) : tous les sujets inclus et randomisés seront pris en compte, même en cas de déviation au protocole.

Pour les essais de non-infériorité ou d'équivalence, l'analyse pourra être menée Per Protocole (PP) : seuls les patients ayant parfaitement suivi le protocole seront pris en compte. L'analyse PP favorise la détection d'une différence d'efficacité entre les interventions testées, elle pourra être complétée par une analyse en ITT.

2.1.2.1.4 Plans expérimentaux



La figure ci-dessus renvoie à un essai comparatif classique :

- deux groupes parallèles sont constitués,
- chacun d'entre eux reçoit la même intervention tout au long de l'étude,
- l'hypothèse de recherche est statistiquement testée à la fin d'un suivi dont la durée a été initialement fixée.

Cependant, il existe de nombreux autres plans expérimentaux. En voici quelques-uns :

- Dans un essai en cross-over, chaque sujet se prêtant à la recherche reçoit alternativement l'une et l'autre des interventions testées. Chaque sujet est donc son propre témoin.

  Deux groupes équilibrés sont constitués par randomisation, l'un recevant le traitement A puis le traitement B, l'autre recevant le traitement B puis le traitement A. Une période sans traitement (appelée washout) sépare la première et la seconde période de traitement pour limiter les interférences entre les deux traitements (appelé effet carry-over). L'analyse permet de distinguer l'effet traitement de l'effet-temps ou période, à condition que l'effet cary-over soit statistiquement nul.

  Ce type d'essai n'est adapté qu'aux pathologies chroniques, c'est-à-dire sans guérison, ni décès rapide, ni même amélioration durable après traitement. Il est particulièrement indiqué si la variabilité de la réponse interindividuelle est importante. Il permet de réduire le nombre de sujets nécessaires.

- Le plan factoriel complet permet d'évaluer l'efficacité de plusieurs traitements et de leur interaction. Il s'agit de tester toutes les combinaisons possibles de traitements. L'intérêt est que l'on obtient des estimations précises de l'effet de chaque médicament et une détermination directe de la meilleure association. L'inconvénient est le nombre d'essais nécessaires pour y parvenir.

- Des analyses intermédiaires peuvent être menées en cours d'essai afin de :
  - Détecter au plus tôt le bénéfice de l'intervention évaluée, ce qui peut conduire à un arrêt de l'essai pour efficacité.
  - Détecter au plus tôt un éventuel effet délétère de l'intervention évaluée, ce qui peut conduire à un arrêt de l'essai pour toxicité.
  - Vérifier le bon déroulement de l'essai (arrêt pour futilité).

  L'inconvénient est que la multiplication de tests statistiques induit une inflation du risque alpha. Les seuils de significativité de ces tests doivent donc être abaissés à chaque nouvelle comparaison.

  Les méthodes séquentielles permettent de répéter de nombreuses analyses intermédiaires et d'arrêter l'essai dès qu'il est possible de conclure.

2.1.2.1.5 Indices de l'effet de l'intervention et tests d'hypothèse

Si le critère de jugement correspond à une variable qualitative binaire ou dichotomique (survenue ou non de l'évènement), plusieurs indices de l'effet de l'intervention pourront être utilisés :

- (Odds Ratio, Risque relatif)
- La Différence de Risque (DR) ou différence absolue ou bénéfice absolu ou Absolute Risk Reduction (ARR) est la différence entre les risques de survenue de l'évènement dans le groupe intervention testée et dans le groupe témoin. L'évènement étudié peut être souhaité ou non souhaité, la DDR peut être positive ou négative.
- La Réduction Relative de Risque ou Relative Risk Reduction (RRR) est le rapport entre la DDR et le risque dans le groupe témoin.
- Le Nombre de sujets Nécessaires à Traiter ou Number Need to Treat (NNT) est le nombre de sujets qu'il est nécessaire de traiter pour éviter qu'un évènement survienne avant la fin du suivi. C'est l'inverse de la différence de risque.

Si le critère de jugement correspond à une variable quantitative continue ou qualitative ordonnée (mesure d'un paramètre physique ou biologique, score…), d'autres indices de l'effet de l'intervention seront utilisés :

- (Rapport des moyennes).
- Différence moyenne.
- Différence relative des moyennes (différence moyenne rapportée à moyenne dans le groupe témoin).
- Différence de moyenne standardisée ou effet standardisé (ou effect size). Par définition, l'effet standardisé est la différence des moyennes rapportée à l'écart type commun aux deux groupes. Cela correspond à une variable normale réduite, d'interprétation clinique délicate.

Le biais de confusion est par définition contrôlé si l'allocation des interventions a bien été réalisée de façon aléatoire (randomisation). Dans ce cas, il est inutile de procéder à une analyse multivariée. Les différents tests d'association possibles selon les variables rencontrées sont présentés au chapitre 3.

2.1.2.2 Cohorte

Une cohorte est une population de sujets qui répondent à une définition donnée et qui sont suivis dans le temps.

Une cohorte peut être constituée :

- à visée descriptive, si l'on se contente de mesurer la survenue d'un ou plusieurs évènements dans le temps.
  → *Etude d'incidence par exemple.*

- à visée analytique, si une ou plusieurs relations entre exposition(s) et survenue(s) d'évènement sont explorées.
  → *Etude de facteurs de risque ou pronostiques, étude exposés/non-exposés par exemple.*

Dans le temps, une cohorte se caractérise de deux manières :

- Selon la date de sélection des sujets pour l'étude, une cohorte peut être…

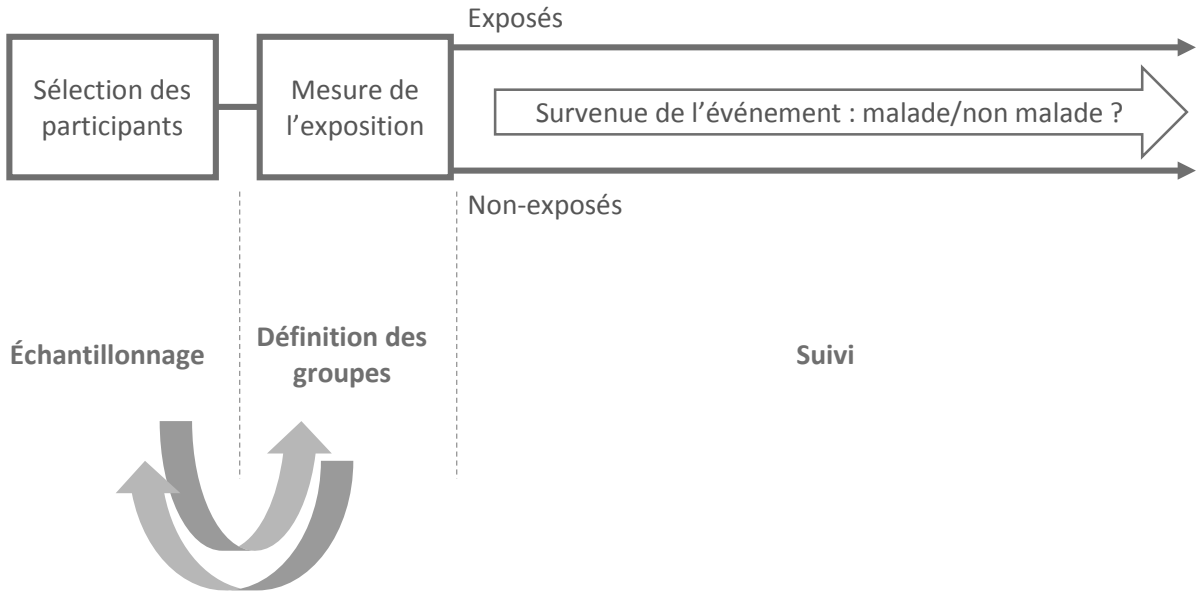
- historique, constituée de manière rétrospective, exposition et survenue de l'évènement étant alors antérieurs à la date de sélection des sujets.
- prospective, constituée de sujets exposés ou non, mais indemnes de l'évènement étudié lors de leur sélection pour l'étude.

- Selon la date de début de suivi de chaque sujet, une cohorte peut être…
  - fixe, les sujets la composant étant tous inclus à la date de constitution de la cohorte.
  - dynamique, les sujets la composant pouvant être inclus depuis la date de création de la cohorte jusqu'à la fin du suivi.

Le recrutement des sujets peut se faire :

- Sur critères d'appartenance à une population, quel que soit le statut à l'égard de l'exposition. → *Cohorte de naissance par exemple.*

- Sur critère d'exposition, c'est l'étude exposés/non-exposés (par définition à visée analytique).
  - Les sujets composant les groupes exposés ou non-exposés doivent être identiques en tout point, excepté à l'égard de l'exposition.
  - Ils doivent être indemnes de l'évènement étudié au début de leur suivi.
  - L'exposition peut être mesurée de manière rétrospective (et prospective), la survenue de l'évènement étudié doit être mesurée de manière prospective.



L'événement étudié doit être précisément défini, tout comme l'exposition (intensité, durée) si nécessaire. Leur mesure doit être standardisée, le cas échéant en aveugle de l'exposition.

Le suivi doit être rigoureux et identique dans tous les groupes étudiés. Tous les moyens doivent mis en œuvre pour limiter le nombre de perdus de vue.

A visée descriptive, les études de cohorte permettent d'estimer l'incidence d'un ou plusieurs évènements. Ces estimations peuvent être standardisées aux caractéristiques d'âge et de sexe d'une

population de référence, afin de permettre leur comparaison aux résultats obtenus à l'observation d'autres populations.

A visée analytique, les études de cohortes permettent de tester une ou plusieurs associations entre une ou plusieurs variables à expliquer (évènements, maladies) et une ou plusieurs variables explicatives (expositions, facteurs de risque ou pronostiques).

- La force d'une association peut être estimée grâce au rapport d'incidence de l'évènement entre exposés (éventuellement par sous-groupes d'intensité d'exposition) et non-exposés, c'est-à-dire un Risque Relatif (RR).
  → *Parmi les études observationnelles, seules les études de cohorte permettent l'estimation du RR.*

- Si l'on dispose de mesures de la survenue de l'évènement répétées dans le temps, et que l'incidence instantanée varie au cours du temps, l'analyse de survie permet de tester la significativité du RR (ou Hazard Ratio - HR).
  Dans un premier temps, l'analyse de la significativité de la relation entre une exposition et un évènement fait appel au Log Rank de courbes de survie (méthode actuarielle ou de Kaplan-Meier).
  A l'étape multivariée, le modèle de Cox permet de contrôler les facteurs de confusion intervenant dans la relation entre une exposition et un évènement, en proposant HR, intervalle de confiance du HR, et p-value ajustés sur ces facteurs de confusion. A condition bien sûr que les données ayant trait aux facteurs de confusion aient été recueillies au cours de l'étude.

- Si l'incidence instantanée est constante au cours du temps, le taux d'incidence peut se modéliser par régression de Poisson. La régression de Poisson produit également un RR et permet de contrôler les facteurs de confusion intervenant dans la relation entre une exposition et un évènement, en proposant RR, intervalle de confiance du RR, et p-value ajustés sur ces facteurs de confusion.

- Cas particulier : si l'on ne dispose pas de mesures répétées et si la durée de suivi est identique pour tous sujets inclus, en particulier si cette durée de suivi est courte, l'indicateur de risque utilisé sera plutôt l'Odds Ratio (OR). L'OR est une bonne approximation du RR si la prévalence de la maladie est rare dans la population. La régression logistique ou multiniveau sera alors utilisée pour contrôler les facteurs de confusion.

	Evénement	**Non-événement**
Intervention ou exposition	**a**	**b**
Non-intervention ou non-exposition	**c**	**d**

$$\mathbf{OR} = \frac{\mathbf{a} * \mathbf{d}}{\mathbf{b} * \mathbf{c}}$$

$$\mathbf{RR} = \frac{\mathbf{a} * (\mathbf{c} + \mathbf{d})}{\mathbf{c} * (\mathbf{a} + \mathbf{b})}$$

Les études de cohorte à visée analytique sont coûteuses en temps et en logistique, mais elles sont d'un haut niveau de preuve. Elles permettent une mesure très précise de l'exposition. Les études exposés/non-exposés sont particulièrement indiquées lorsque l'exposition est rare. Elles posent un problème de sélection des groupes.

Au chapitre des cohortes, on peut évoquer le cas particulier des registres. Leur fonction est d'assurer « un recueil (prospectif) continu et exhaustif de données nominatives intéressant un ou plusieurs événements de santé dans une population géographiquement définie, à des fins de recherche et de santé publique ».
→ *Registre REIN ou projet Monica par exemple.*

2.1.2.3 Etude cas/témoin

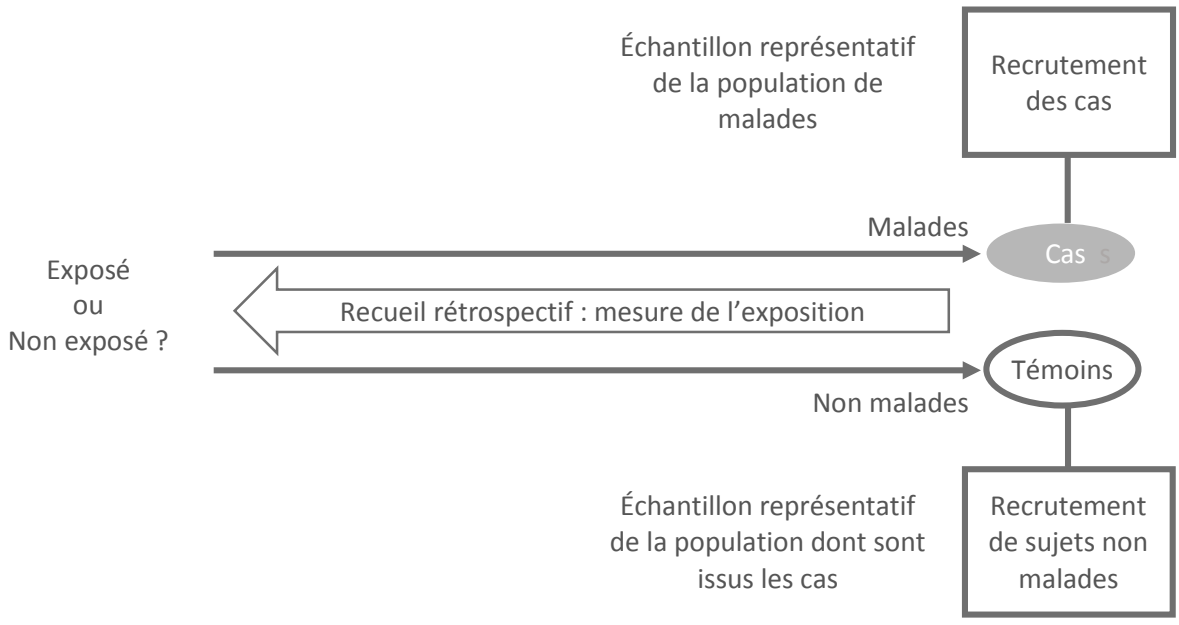
L'étude cas/témoin porte sur un échantillon de sujets sélectionnés en raison de leur statut vis-à-vis d'une maladie : cas (malades) ou témoins (indemnes). Cas et témoins peuvent être appariés selon des facteurs de confusion connus (cf. chapitre 2.3.2.4). L'effectif des témoins peut être supérieur au l'effectif des cas (doublé le plus souvent) afin d'augmenter la puissance.

La sélection des cas doit faire appel à une définition précise de la maladie et à l'utilisation de critères standardisés. Idéalement, il devrait s'agir d'un échantillonnage probabiliste sur une liste complète des patients affectés (type registre). Mais cette solution est rarement disponible, ce qui expose à un risque de biais de sélection.

Le groupe témoins devrait être constitué par échantillonnage probabiliste de la population dont sont issus les cas, afin d'estimer la fréquence d'exposition de référence. Mais cette solution est rarement mise en œuvre. Au minimum, les témoins choisis ne doivent pas être plus à risque que la population cible d'avoir subi l'exposition étudiée.
→ *Attention au choix de témoins en milieu hospitalier.*

La mesure de l'exposition est postérieure à la survenue de la maladie. Elle est rétrospective, ce qui expose au biais de mémorisation donc à un biais d'information/classement/mesure.

Maladie et exposition doivent être précisément définies. Le recueil des données doit être standardisé, la mesure de l'exposition réalisée si possible en aveugle de la maladie.



L'étude cas témoin ne permet pas d'estimer une incidence ou une prévalence.

Elle permet d'estimer une association entre une variable à expliquer (maladie) et une ou plusieurs variables explicatives (exposition). L'indicateur de risque est l'Odds Ratio (OR) et son intervalle de confiance, pas le Risque Relatif (RR). Les biais de confusion qui n'ont pas été traités par appariement peuvent être contrôlés par analyse statistique multivariée (régression logistique et multiniveau), à condition que les données ayant trait aux facteurs de confusion aient été recueillies au cours de l'étude.

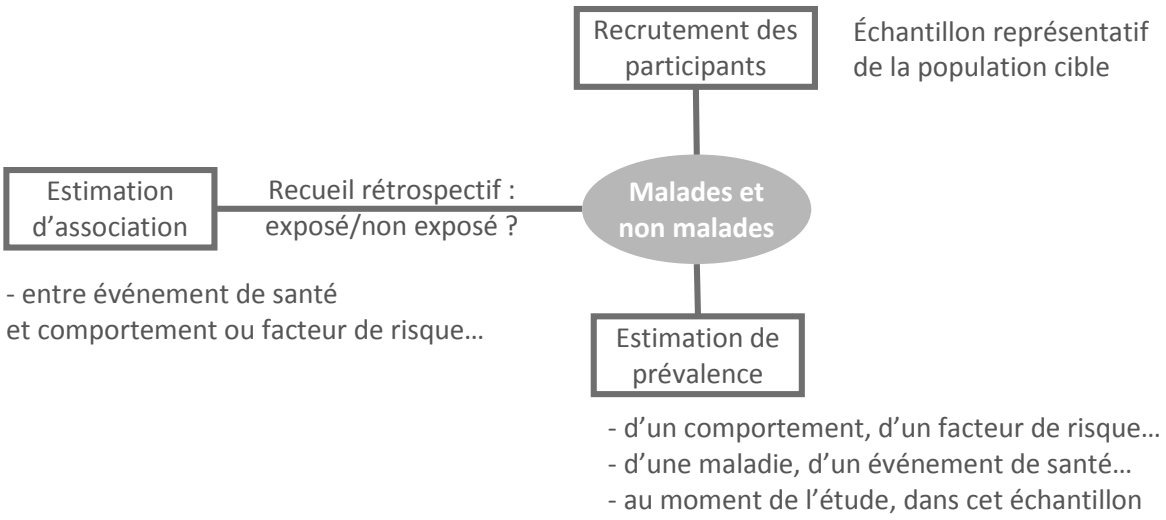
L'étude cas/témoin est d'un faible niveau de preuve, mais c'est le schéma indiqué quand on explore les facteurs associés à une affection rare ou qui survient longtemps après l'exposition. Ce schéma d'étude est peu coûteux en effectif, en logistique, et en durée.

2.1.2.4 Etude transversale

L'étude transversale porte sur un échantillon de sujets sélectionnés pour leur appartenance à une population, indépendamment de leurs éventuelles expositions ou maladies. Expositions et maladies sont ici mesurées simultanément.

A visée descriptive, une étude transversale permet d'estimer des prévalences ou des moyennes et leur variance, selon les données observées. Elle peut être utilisée en surveillance épidémiologique si elle est répétée au cours du temps. Mais l'évolution de la prévalence d'une affection peut résulter de variations de son incidence et/ou de sa durée d'évolution.

A visée analytique, une étude transversale peut permettre d'estimer une association entre une variable à expliquer et une ou plusieurs variables explicatives. L'indicateur de risque est l'Odds Ratio (OR) et son intervalle de confiance, pas le Risque Relatif. Les biais de confusion peuvent être contrôlés par analyse statistique multivariée (régression logistique et multiniveau), à condition que les données ayant trait aux facteurs de confusion aient été recueillies au cours de l'étude.



L'étude transversale est d'un très faible niveau de preuve, mais elle permet à moindre frais de préparer des études de plus grande valeur méthodologique.

2.1.2.5 Étude d'évaluation diagnostique

2.1.2.5.1 Définitions

L'objectif de ce type d'étude est d'évaluer la qualité d'un nouveau test diagnostique par rapport à un test de référence.

Définition : un test est un procédé de recueil d'information dont le résultat est utilisé dans une démarche de décision. Dans le domaine médical, l'objectif d'un test est de réduire l'incertitude clinique. Un test peut alors être mis en œuvre à des fins diagnostiques chez un sujet symptomatique, ou à des fins de dépistage chez un sujet apparemment indemne.

La qualité d'un test se définit par :

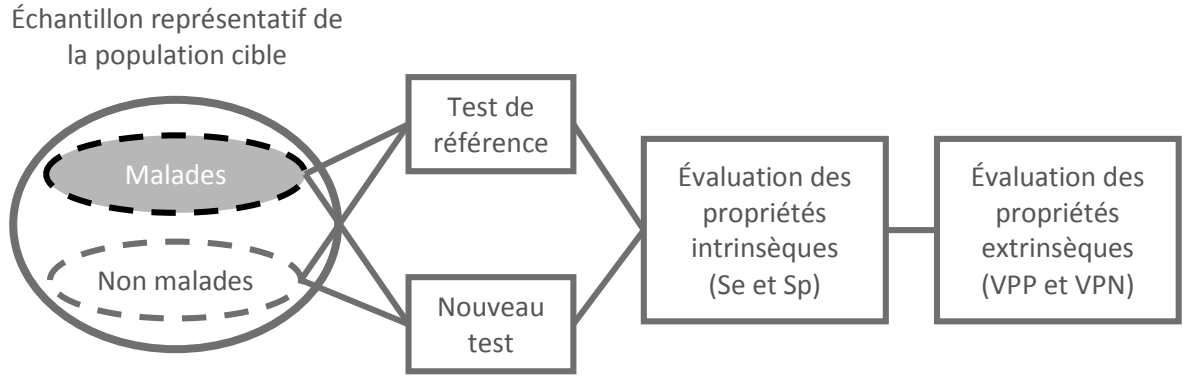
- Sa fiabilité, c'est-à-dire…
  - son exactitude (pas de biais systématique dans la mesure),
  - et sa précision (pas d'erreur aléatoire).

  → *Ce problème doit faire l'objet d'études spécifiques ayant trait à la fiabilité des outils de mesures. Nous ne le développerons pas ici.*

- Sa validité, c'est-à-dire sa capacité à classer des sujets soumis au test dans la catégorie à laquelle ils appartiennent (malades ou non-malades). On distingue…
  - Les propriétés intrinsèques du test, qui traduisent sa capacité informative. Elles sont indiquées par la Sensibilité (Se) et la Spécificité (Sp) du nouveau test, ainsi que par les rapports de vraisemblance.
  - Les propriétés extrinsèques du test, qui traduisent sa valeur décisionnelle. Elles sont indiquées par les Valeurs Prédictives Positives (VPP) et Négatives (VPN). Elles dépendent de la capacité informative du test et de la prévalence de la maladie dans la population testée, c'est-à-dire des conditions de l'expérience.

2.1.2.5.2 Plan expérimental

Le schéma classique pour l'évaluation de la validité d'un nouveau test est figuré ci-dessous.



Cette figure ne représente pas forcément l'ordre d'administration du test de référence et du test évalué.

Le problème principal de ce schéma d'étude est qu'il faut connaitre avec certitude le statut vis-à-vis de la maladie des sujets se prêtant à la recherche. C'est le test de référence qui est sensé identifier les sujets malades ou non-malades. Idéalement, ce test de référence devrait être présent et indiscutable
→ *par exemple coronarographie dans la maladie coronarienne*.

Ce n'est pas toujours le cas :

- Le diagnostic certain de la maladie peut être retardé,
  → *par exemple dans le cas de la maladie d'Alzheimer dont le diagnostic peut nécessiter autopsie cérébrale ;*
- l'apparition de la maladie jamais atteinte,
  → *par exemple dans le cas des dépistages de cancer*.
- Le test de référence peut manquer, le diagnostic relevant alors d'un consensus d'expert sur conjonction de divers critères,
  → *par exemple dans le cas de la dénutrition dont le diagnostic est fait sur critères cliniques, biologiques et nutritionnels.*

Les modalités d'échantillonnage dépendent de l'ordre d'administration des tests. Idéalement, chaque unité statistique devrait être simultanément diagnostiquée et testée. C'est difficile à réaliser en pratique. Deux autres solutions sont possibles :

- Le test de référence est tout d'abord appliqué à l'échantillon afin de déterminer malades et non-malades, puis le test à évaluer est administré dans chacun des groupes (rétrospectif).
- Le test à évaluer est d'abord administré à l'échantillon afin de déterminer test + et test -, puis le test de référence est appliqué dans chacun des groupes (prospectif).

Le test évalué ne doit pas avoir été utilisé dans le processus de sélection des malades et non-malades (pas de redondance).

L'interprétation du test évalué doit être effectuée en double lecture, la première à l'aveugle de la maladie pour contrôler au mieux le biais d'interprétation.

Notons que le test à évaluer peut être constitué de tests multiples, en parallèle ou en série.

2.1.2.5.3 Indices de validité

Propriétés intrinsèques, ou capacité informative :

- La sensibilité d'un test (Se) est la probabilité que ce test soit positif chez les vrais malades, c'est-à-dire l'aptitude du test à repérer les malades. Le taux de faux négatifs est égal à 1- Se.
  La spécificité (Sp) est la probabilité que ce test soit négatif chez les non-malades, c'est-à-dire l'aptitude du test à repérer les non-malades. Le taux de faux positifs est égal à 1- Sp.
  Sensibilité et spécificité varient en sens inverse.
  Plus le stade de la maladie est évolué, plus la sensibilité est élevée.

→ *Attention au biais induit par un test de référence trop restrictif (augmentation de la Se).*
→ *Attention au biais de sélection lié au choix des malades en milieu hospitalier : privilégier l'évaluation du test dans ses conditions courantes d'application.*

		Test de référence			
		Malades	**Non-malades**		
Test à évaluer	**Positif**	**A** Vrais positifs	**B** Faux positifs	**A + B** Total tests positifs	$\mathrm{VPP} = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}}$
	Négatif	**C** Faux négatifs	**D** Vrais négatifs	**C + D** Total tests négatifs	$\mathrm{VPN} = \frac{\mathrm{D}}{\mathrm{C}+\mathrm{D}}$
		A + C Total malades	**B + D** Total non-malades		
		$\mathrm{Se} = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{C}}$	$\mathrm{Sp} = \frac{\mathrm{D}}{\mathrm{B}+\mathrm{D}}$		

- Le rapport de vraisemblance d'un test positif (LR+) est le risque relatif de présenter un test positif chez les malades comparativement aux non-malades.
  Le rapport de vraisemblance d'un test négatif (LR-) est le risque relatif de présenter un test négatif chez les malades comparativement aux non-malades.
  Plus LR+ est élevé, plus le test permet de confirmer la maladie.
  Plus LR- est petit, plus le test permet d'exclure la maladie.
  Si LR+ = LR- = 1, on dit que la capacité discriminante du test est nulle.

Propriétés extrinsèques, ou valeur décisionnelle :

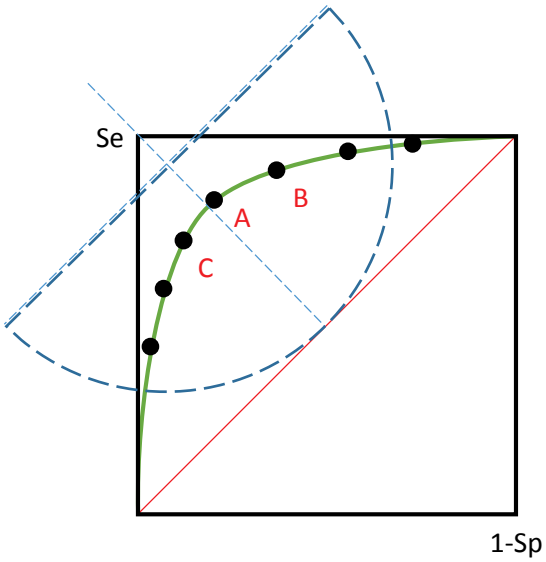
- La valeur prédictive positive est la probabilité d'être malade si le test est positif.
  La valeur prédictive négative est la probabilité d'être indemne si le test est négatif.
  La VPP augmente avec la spécificité, la VPN avec la sensibilité.
  La VPP augmente et la VPN diminue quand la prévalence de la maladie augmente, et inversement.

2.1.2.5.4 Courbe ROC

Le résultat d'un test peut être quantitatif ou qualitatif ordonné. Dans ce cas, son interprétation nécessite de décider d'un seuil de positivité ou seuil de décision, afin de dichotomiser ses résultats. La sensibilité et la spécificité du test dépendent alors de ce seuil de décision : plus le seuil de positivité est bas, plus la sensibilité est élevée et la spécificité basse.

L'outil statistique mis en œuvre pour décider de cette valeur seuil est le plus souvent graphique. La courbe ROC (receiver operating characteristic) donne le taux de vrais positifs (Se) en fonction du taux

de faux positifs (1 – Sp) si chaque valeur prise par le résultat d'un test appliqué à un échantillon était utilisée comme seuil de décision.

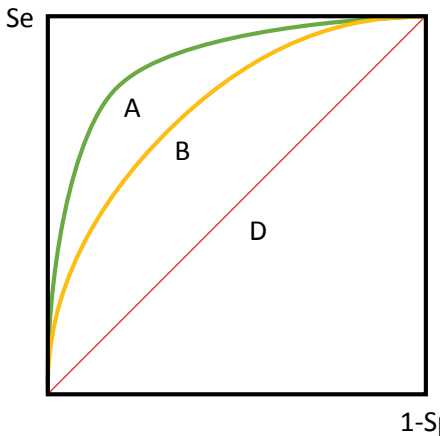


La valeur seuil peut être celle qui entraîne le minimum de mauvaises classifications. Dans le graphique ci-dessus, elle correspond au point d'inflexion de la courbe ROC le plus proche du coin supérieur gauche (A), c'est-à-dire le point où la somme des taux de faux positifs et de faux négatifs est la plus faible possible. Mais :

- Dans une approche de dépistage, et lorsque les faux négatifs sont inacceptables, la valeur seuil peut être choisie de manière à favoriser la Se, par exemple le point B dans la figure ci-dessus.
- Dans une approche diagnostique, et lorsque les faux positifs sont inacceptables, la valeur seuil peut être choisie de manière à favoriser la Sp, par exemple le point C dans la figure ci-dessus.

La courbe ROC permet aussi d'évaluer globalement la capacité discriminante d'un test de de la comparer à celle d'un autre test.

Si Se = 1 - Sp, la capacité discriminante est nulle. Cela correspond à la diagonale D dans la figure ci-dessous. Plus la courbe ROC se détache de la diagonale et s'enfonce dans l'angle supérieur gauche du graphique, plus grande est la capacité discriminante du test. Dans la figure ci-dessous, le test correspondant à la courbe A aura une plus grande capacité discriminante que le test correspondant à la courbe B.



La capacité discriminante globale du test peut donc être estimée grâce au calcul de l'aire sous la courbe ROC (ou Area Under Curve - AUC). La surface totale du carré est au maximum de 1, une AUC égale à 1 correspond donc à la meilleure capacité discriminante possible, et une AUC égale à 0,5 correspond à une capacité discriminante nulle. L'AUC traduit la probabilité que le résultat du test chez un sujet malade soit supérieur au résultat du test chez un sujet non-malade. Un intervalle de confiance de l'AUC correspondant à chaque test peut être calculé, et l'égalité des AUC peut être statistiquement testée.

2.1.2.6 Méta-analyse

2.1.2.6.1 Définitions

Glass, 1976 : Discipline ayant pour objectif le recensement critique de la littérature et l'intégration statistique des résultats des études antérieures portant sur une même question de recherche.

Fleiss & Gross, 1991 : Méthode augmentant la puissance statistique, permettant l'estimation de l'effet, expliquant les controverses des études individuelles.

Littell & coll. 2008 : Ensemble de procédures statistiques permettant de combiner (agréger) les résultats quantitatifs de multiples recherches afin de produire une synthèse des connaissances empiriques sur un sujet donné.

2.1.2.6.2 Buts et principes

Plusieurs études peuvent avoir été menées sur la même problématique afin de tester la même hypothèse de travail, c'est-à-dire avec le même objectif de recherche et des méthodologies similaires mais dans des échantillons différents. Les conclusions de ces études peuvent être discordantes…

- Certaines de ces études peuvent mettre en évidence une différence d'effet entre intervention testée et intervention de référence, ou entre exposition et non exposition. Il existe un risque que cette différence soit le fruit du hasard (risque alpha).
- Dans d'autres études, la différence d'effet peut manquer. Deux explications sont alors possibles :
    - absence réelle d'effet de l'intervention ou de l'exposition testée,
    - ou manque de puissance statistique par effectif insuffisant (risque beta).

Une méta-analyse permet d'agréger les résultats de ces différentes études en calculant un effet commun de l'intervention ou l'exposition testée, à partir des données de chaque étude.

2.1.2.6.3 Sélection des essais

La méta-analyse obéit à une méthodologie particulièrement rigoureuse, dont les normes ont été internationalement définies (Cochrane methods).

La première étape d'une méta-analyse est de formuler clairement et précisément son objectif.

La deuxième étape consiste à collecter les études dont les résultats vont être pris en compte dans cette méta-analyse. Cela nécessite une recherche systématique et exhaustive des études qui peuvent répondre à son objectif. Les résultats de cette recherche doivent être reproductibles. Les critères d'éligibilité des études à prendre en compte doivent être préalablement et précisément définis. Ces critères ont trait aux caractéristiques des études : population cible, intervention, critère de jugement. Ces caractéristiques doivent être en adéquation avec l'objectif de la méta-analyse.

Toutes les études répondant aux critères d'éligibilité doivent être identifiés, quels que soient leurs résultats. Cela nécessite de combiner plusieurs sources d'informations :

- Les bases bibliographiques informatisées (medline, embase…) permettent de retrouver les études ayant fait l'objet de publication. Mais certains articles peuvent être inadéquatement indexés.
- La plupart des essais cliniques non publiés peuvent être retrouvés dans le registre Cochrane des essais contrôlés. Mais aucune base similaire n'existe pour les études observationnelles non publiées.
- Comptes rendus de congrès, catalogues d'abstracts.
- Littérature grise (documents imprimés ou électroniques non contrôlés par des éditeurs commerciaux ou n'ayant pas transité par des comités d'évaluation).
- Contacts personnels auprès d'experts dans la thématique étudiée.
- …

La troisième étape consiste à évaluer la qualité méthodologique des études sélectionnées. Elle peut être approchée par une grille d'évaluation et un score global de qualité. Dans le cas des essais contrôlés, trois critères sont le plus souvent requis : randomisation, insu, et faible taux de perdus de vue.

- Les études de bonne qualité seront inclues dans la méta-analyse.
- Les études de qualité intermédiaire pourront être inclues et exclues de la méta-analyse afin de comparer les résultats finaux obtenus avec ou sans elles (analyse de sensibilité)
- Les études de faible qualité devront être exclues de la méta-analyse.

La recherche documentaire et l'évaluation de la qualité des études doivent être réalisées au moins par deux personnes, agissant en aveugle l'une de l'autre.

Le processus de sélection des études peut être présenté sous forme d'un diagramme de flux résumant les motifs d'exclusion.

Un premier problème est que les études dont les résultats ne sont pas concluants sont moins facilement publiées et moins accessibles que les études à résultats concluants. La recherche documentaire doit également porter sur les études non publiées. Tout rapport de méta analyse doit justifier l'exhaustivité de la recherche pour garantir un contrôle correct du biais de publication.

Un deuxième problème est que la méta-analyse ne peut améliorer la qualité des études qu'elle regroupe : ces études peuvent être biaisés, et donc conduire à une méta-analyse elle-même biaisée (garbage in, garbage out). Cependant…

- Les résultats de la méta-analyse seront moins biaisés que ceux des études biaisées qu'elle prend en compte (effet « tampon »).
- Des outils statistiques peuvent mettre évidence une hétérogénéité dans les résultats exploités, afin d'identifier les essais potentiellement biaisés et d'éventuellement les exclure de l'analyse.

2.1.2.6.4 Mesure de l'effet traitement

Littell & coll. 2008 : La taille de l'effet ou grandeur d'effet constitue la statistique de base de la méta-analyse. Elle consiste en une mesure standardisée indiquant l'ampleur et la direction d'une association entre deux variables.

La première de ces variables est en règle générale dichotomique : elle définit les groupes exposés/non-exposé ou intervention/témoin.

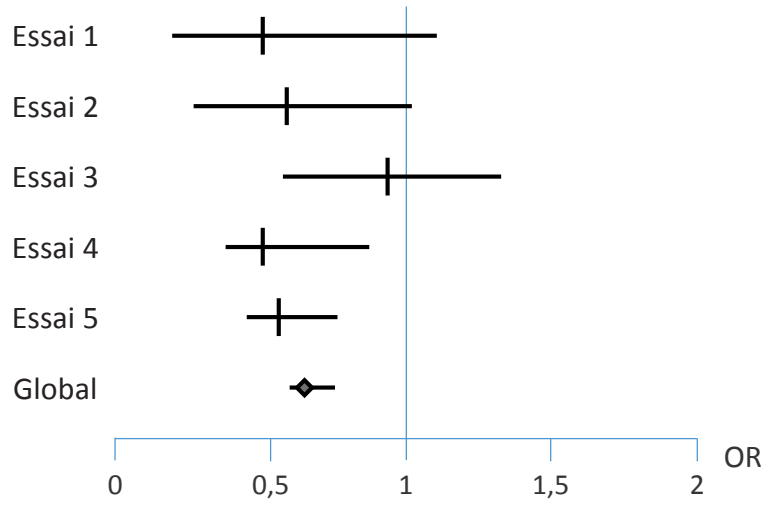
La seconde est une variable dépendante (critère de jugement) qui peut être qualitative (survenue ou non de l'évènement) ou quantitative (mesure d'un paramètre physique ou biologique par exemple).

- Dans le cas où la variable dépendante est qualitative binaire ou dichotomique, on s'intéresse à des proportions de sujets qui présenteront l'évènement dans les deux groupes. La Différence de Risque (DR), l'Odds Ratio (OR), le Risque Relatif (RR)… peut être utilisé comme indice de l'effet traitement.
- Dans le cas où la variable dépendante est quantitative, la différence de moyenne standardisée ou effet standardisé (ou effect size) sera utilisée. Par définition, l'effet standardisé est la différence des moyennes rapportée à l'écart type commun aux deux groupes. Cela correspond à une variable normale réduite, d'interprétation clinique délicate.

L'effet observé dans un essai est égal au vrai effet de l'intervention testée auquel s'ajoutent les effets des erreurs aléatoires et des éventuels biais propres à l'essai. Une méta-analyse de plusieurs essais doit permettre d'isoler le vrai effet du traitement (part commune à tous les essais), des erreurs aléatoires et systématiques (parts spécifiques à chaque essai). Théoriquement :

- L'erreur aléatoire est mécaniquement réduite par le nombre d'essais inclus dans la méta-analyse.
- Les erreurs systématiques peuvent être réduites par la rigueur méthodologique de sélection des essais, et par l'effet « tampon » propre à la méta-analyse.

Reste la part commune à tous les essais, nommée effet traitement commun. Voici comment on peut le représenter graphiquement (forest plot). L'indice de l'effet traitement est ici l'odds ratio (OR).



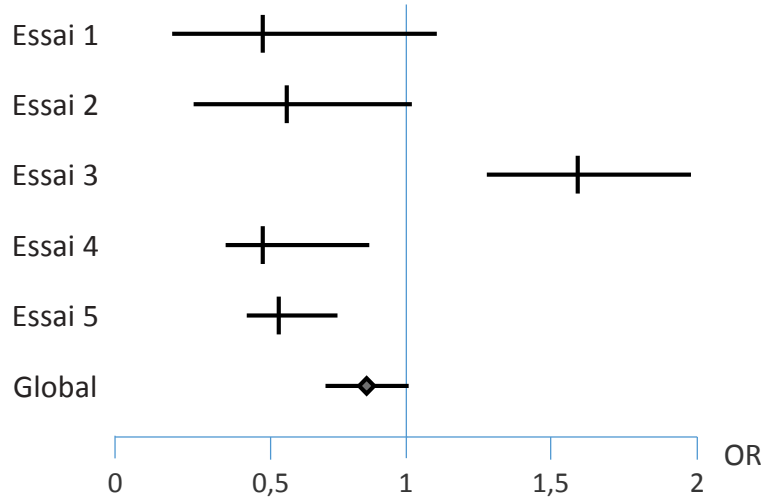
Dans l'exemple ci-dessus, l'intervalle de confiance de chaque effet observé peut inclure l'effet global, ou effet traitement commun. On dit qu'il y a « homogénéité », et c'est un principe fondamental de la méta-analyse : la variation inter-échantillon de l'effet doit rester suffisamment faible pour permettre le regroupement de ces échantillons dans une même méta-analyse.

S'il y a bien homogénéité, l'analyse de l'effet traitement commun pourra faire appel aux modèles à effets fixes. L'effet traitement commun est alors la moyenne des estimations de l'effet pour chaque essai, pondérée par l'inverse de leur variance. Un intervalle de confiance de l'estimation de l'effet traitement commun peut être calculé, et un test statistique d'association peut être appliqué.

Dans le cas d'un critère de jugement binaire, la méthode la plus utilisée est celle de Mantel-Haenszel. Elle permet d'obtenir une DR, un OR, ou un RR combiné et de tester sa significativité. Si la méta-analyse porte sur des données de survie, la méthode de Peto peut être utilisée. Le test d'association portera alors sur l'égalité des courbes de survie dans chaque essai inclus.

Dans le cas d'un critère de jugement quantitatif, la méthode la plus utilisée est celle de Hedges et Olkin. Elle permet d'obtenir un effet standardisé combiné et de tester sa significativité.

Prenons maintenant un nouvel exemple.



L'intervalle de confiance de l'effet observé dans l'essai 3 n'inclut pas la valeur de l'effet global ou traitement commun. On dit qu'il y a « hétérogénéité ». Cette trop grande variabilité peut s'expliquer par les facteurs suivants :

- Une population d'échantillonnage différente,
- Des caractéristiques d'intervention différentes,
- Une qualité méthodologique différente,
- Voire un biais dans l'essai induisant l'hétérogénéité.

Au-delà de l'aspect graphique, cette hétérogénéité peut être testée de manière statistique (test Q de Cochran par exemple). Si le test est positif ($p < 0,05$), l'hétérogénéité est démontrée. Il faut alors :

- Identifier l'essai induisant cette hétérogénéité, grâce à la méthode graphique ci-dessus.

- Rechercher les facteurs de variabilité ou hétérogénéité parmi les caractéristiques de l'échantillon, de l'intervention, ou de la méthode. Cela revient à chercher des interactions, c'est-à-dire des variations de la taille de l'effet en fonction de certaines de ces caractéristiques.
- Effectuer une analyse en sous-groupes, c'est-à-dire stratifier l'analyse sur le ou les facteurs de variabilité identifiés. Mais la multiplication des tests statistiques peut poser un problème d'inflation du risque alpha.

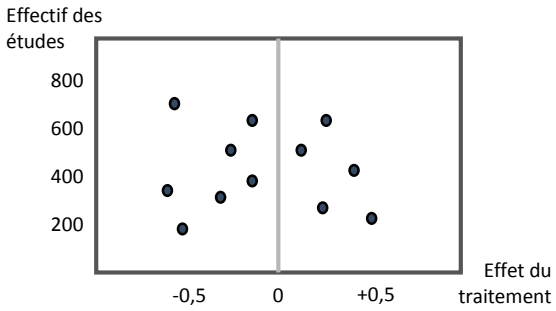
Si aucun facteur d'hétérogénéité ne peut être identifié ou si l'analyse en sous-groupe n'est pas possible, l'analyse de l'effet traitement commun nécessitera :

- Soit d'exclure de la méta-analyse l'essai induisant l'hétérogénéité,
- Soit de faire appel…
    - aux modèles à effets aléatoires, qui prennent en compte la variabilité inter-échantillon ;
    - ou aux modèles à effets mixtes, qui prennent en compte la variabilité inter-échantillon et les caractéristiques (échantillon, intervention, méthode) des études inclues.

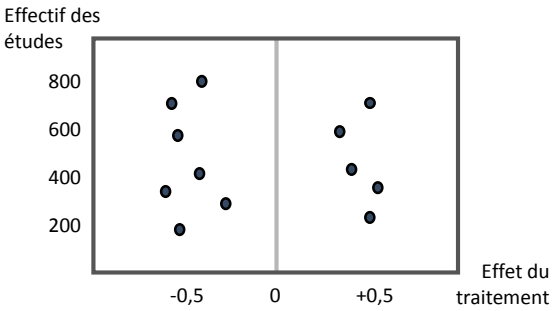
Notons que quelle que soit la méthode d'analyse de l'effet traitement commun, ce sont les études mettant en jeu des effectifs importants et/ou avec peu de variabilité intra-échantillon, et donc une précision d'estimation élevée, qui auront le plus de poids dans la méta-analyse.

2.1.2.6.5 Recherche du biais de publication

Un éventuel biais de publication peut être mis en évidence grâce à une méthode graphique (funnel plot) représentant la répartition des études selon l'effet observé.



Pas de biais de publication évident



Biais de publication très probable

Si peu d'études apparaissent dans la zone centrale du nuage de point (zone d'inefficacité du traitement), on peut suspecter un biais de publication.

2.1.2.6.6 Limites et apports de la méta-analyse

Une méta-analyse ne peut agréger les résultats d'études hétérogènes, c'est-à-dire :

- portant sur des patients dont les caractéristiques de base à l'égard de la maladie étudiée diffèrent,
- évaluant des interventions ou des expositions dont les caractéristiques diffèrent d'une étude à l'autre,

→ *Ces deux limites peuvent être levées si l'objectif de la méta-analyse est de discerner l'influence des caractéristiques des patients ou des interventions sur la réussite des interventions.*

- utilisant des critères de jugement différents d'une étude à l'autre,
- de qualité méthodologique diverse.

Une méta-analyse ne devrait pas agréger les résultats obtenus à partir d'un même échantillon (problème de non-indépendance des observations et des données), sauf à utiliser des outils statistiques spécifiques.

Si elle est bien conduite, la méta-analyse permet :

- De synthétiser de multiples informations.
- De clarifier des situations contradictoires.
- D'estimer plus précisément la taille d'un ou plusieurs effets associés à une intervention ou exposition.
- D'augmenter la puissance statistique.
- De gagner en représentativité.

⇨ De gagner en validité.

⇨ De décider quelle intervention doit être privilégiée pour faire face à une problématique.

2.2 ECHANTILLONNAGE

2.2.1 Définitions

Echantillonnage : procédé qui consiste à n'observer qu'une partie de la population étudiée (échantillon) et à tirer de cette observation des informations sur la population entière.

Echantillon : individus sélectionnés de manière à ce que leurs caractéristiques soient semblables à celles du groupe dont ils sont issus.

Population : groupe plus large à partir duquel les individus sont sélectionnés pour participer à l'étude.

Population cible : population que l'on veut observer.

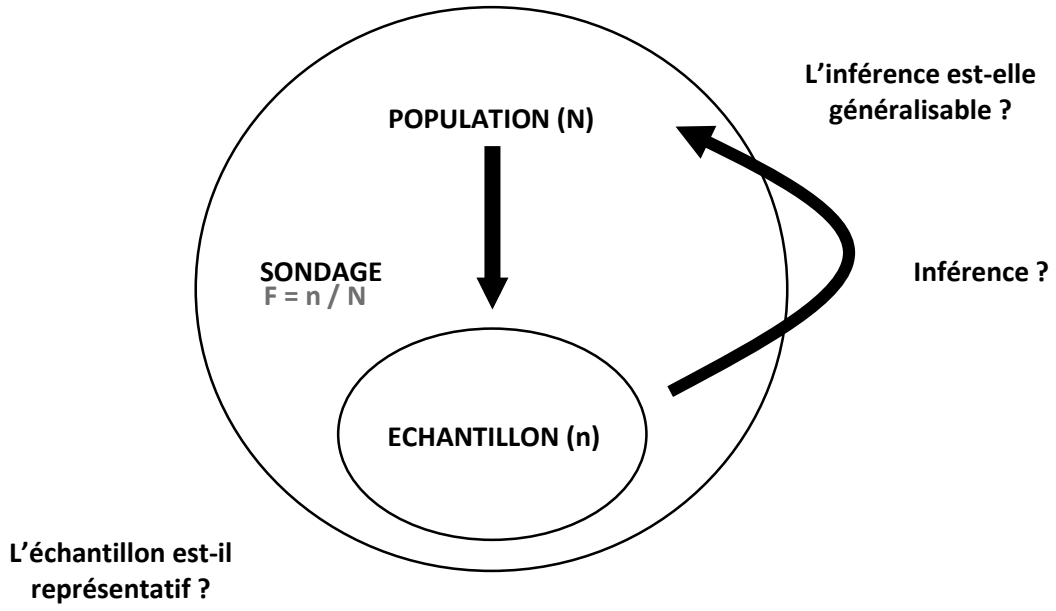
Population source : population que l'on peut observer. L'échantillon diffère de la population cible par l'application de critères d'inclusion-exclusion.

Fraction de sondage : c'est l'effectif de l'échantillon rapporté à l'effectif de la population source.
→ *Cette fraction est égale à 1 dans les rares cas d'échantillonnage exhaustif. L'unité statistique peut être un individu ou un groupe d'individu (étude écologique).*

Inférence statistique : estimation d'un paramètre de la population source (prévalence, moyenne et variance…) à partir des données de l'échantillon.

Généralisation : extrapolation des résultats observés dans un échantillon à une population cible voire externe.

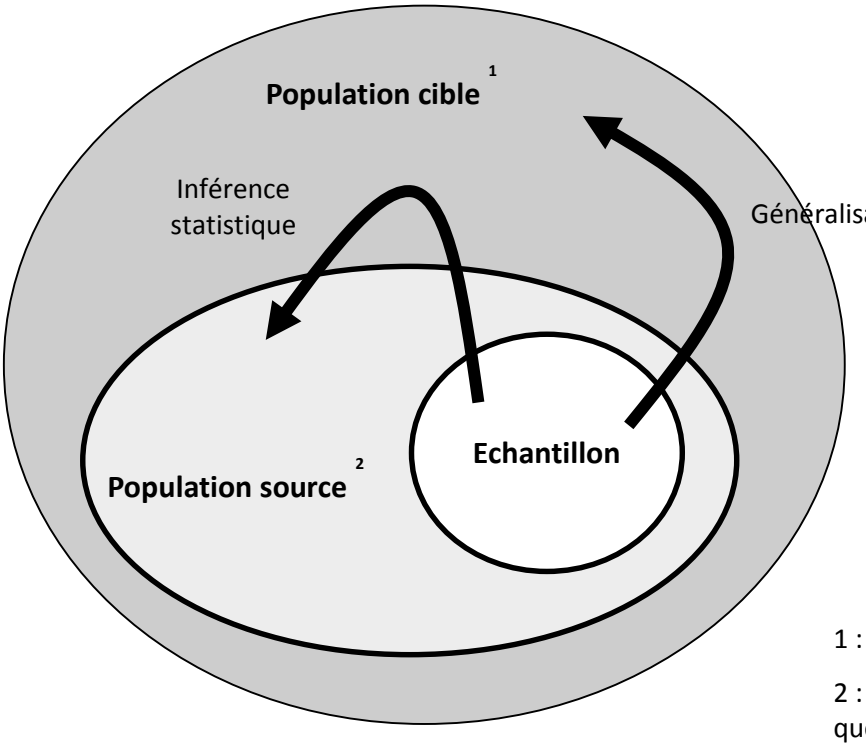
2.2.2 Représentativité



L'inférence ne peut être généralisée si la représentativité de l'échantillon n'est pas garantie. Seul un échantillonnage probabiliste peut garantir la représentativité. L'absence de représentativité peut constituer un biais de sélection (cf. chapitre 2.5.1).

2.2.3 Validité

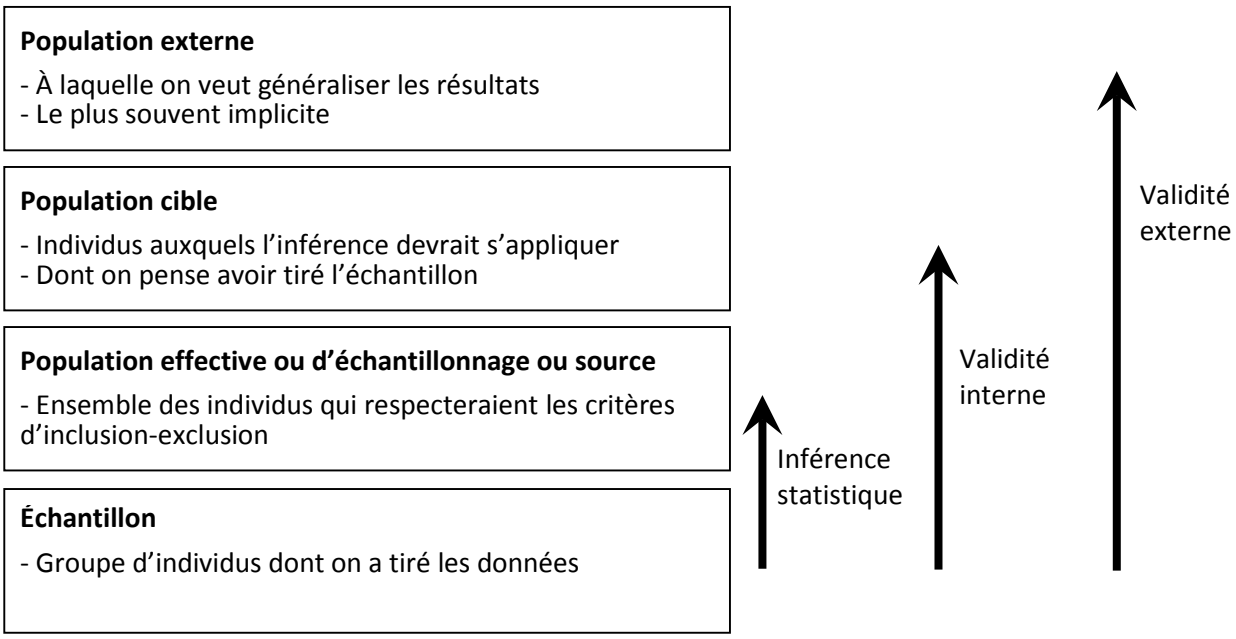
Un bon contrôle du biais de sélection, c'est-à-dire une bonne représentativité de l'échantillon, est indispensable pour assurer la validité d'une étude, c'est-à-dire la possibilité de généraliser ses résultats.



1 : Que l'on veut observer

2 : Population d'échantillonnage, que l'on peut observer

Si les biais de sélection mais aussi d'information et de confusion (cf. chapitre 2.5) sont bien contrôlés, les résultats peuvent être généralisés à la population cible (validité interne) voire à des populations externes (validité externe).



2.2.4 Erreur vs biais d'échantillonnage

Une méthode de sondage est bonne si elle permet d'obtenir des estimations en moyenne égales aux valeurs existant réellement dans la population cible.

Il faut distinguer erreur et biais d'échantillonnage :

- Erreur = variation ou fluctuation aléatoire => nuit à la précision de l'estimation.
  Pour limiter son impact, il faut augmenter la taille d'échantillon.
- Biais = variation non aléatoire => nuit à la représentativité de l'échantillon, donc à la validité.
  Dépend notamment de la méthode d'échantillonnage.

2.2.5 Méthode d'échantillonnage

L'échantillonnage doit assurer une correcte représentativité.

On distingue :

- Échantillonnage probabiliste…
  - Sélection aléatoire : chaque unité a une chance égale d'être sélectionnée.
  - Contrôle correct des biais d'échantillonnage.
  - Échantillonnage aléatoire simple, stratifié, à plusieurs degrés, en grappe, (systématique).
- Échantillonnage non probabiliste…
  - Sélection non aléatoire.
  - Pas de contrôle correct des biais d'échantillonnage.
  - Empirique, à choix raisonné, quota.

2.2.6 Taille d'échantillon : nombre de sujets nécessaires, précision et puissance

Le nombre de sujets à inclure dépend :

- De la faisabilité (coût, temps, éthique…).
- Du calcul du nombre de sujets nécessaires (précision et puissance).

Le nombre de sujets nécessaires doit être calculé préalablement à l'étude.

2.2.6.1 Définitions

Les fluctuations d'échantillonnage peuvent modifier les valeurs estimées à l'observation d'un échantillon. Elles peuvent faire apparaître ou disparaitre une différence qui n'existe pas ou qui existe entre deux groupes au sein de l'échantillon ou entre deux échantillons.

On distingue :

- Le risque Alpha (α) ou erreur de première espèce :
  - Le risque alpha auquel on consent va conditionner le calcul des bornes de l'intervalle de confiance d'une estimation.
    → *Par exemple, choisir un risque alpha de 5% revient à proposer un intervalle de confiance à 95% (1-α) de l'estimation, c'est-à-dire un intervalle qui dans 95% des cas va contenir la valeur que l'on cherche à estimer.*
  - Dans toutes les études en épidémiologie et en recherche clinique, le nombre de sujets nécessaires va notamment être conditionné par le risque alpha auquel on consent.
  - Dans les études interventionnelles ou observationnelles à visée analytique, le risque alpha est le risque de conclure à une différence qui n'existe pas (faux positif).

- Le risque Beta (β) ou erreur de seconde espèce :
  - Il n'est considéré que dans le cas des études interventionnelles ou observationnelles à visée analytique.
  - C'est le risque de ne pas conclure à une différence alors qu'elle existe (faux négatif).
  - Le risque beta auquel on consent conditionne la puissance (1- β) des tests statistiques éventuellement utilisés pour juger d'une différence entre les valeurs estimées à l'observation de deux groupes, c'est-à-dire la capacité de ces tests à mettre en évidence une différence qui existe.
  - Dans les études interventionnelles ou observationnelles à visée analytique, le nombre de sujets nécessaires va notamment être conditionné par le risque beta auquel on consent.

2.2.6.2 Exemple d'une étude descriptive transversale

La variable étudiée est ici qualitative.

Pour information, la formule exacte permettant le calcul du nombre de sujets nécessaires *n* est :

$$n = \frac{N[\pi(1-\pi)]\left(\varphi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2}{(N-1)\varepsilon^2 + [\pi(1-\pi)]\left(\varphi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2}$$

La formule approchée est :

$$n = \frac{[\pi(1-\pi)]\left(\varphi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2}{\varepsilon^2}$$

Soit, pour un risque α à 5% :

$$n = {[\pi(1-\pi)] * 3{,}84}/{\varepsilon^2}$$

Où

π : prévalence attendue de l'événement étudié dans la population cible
α : risque de première espèce, $\varphi^{-1}(1-\frac{\alpha}{2}) \approx 1{,}96$ si α = 5%
ε : précision de l'estimation ou erreur consentie
N : taille de la population cible

On voit que dans le cas de l'étude descriptive transversale d'une variable qualitative, *n* dépend principalement :

- De la prévalence (π) attendue de l'événement étudié dans la population cible. Elle est fixée à 50% si aucune information a priori n'est disponible.
- Du risque de première espèce (α) choisi. Il est en règle générale fixé à 5%. Il peut être diminué si l'on souhaite une meilleure précision de l'estimation.
- De la précision de l'estimation (ε) souhaitée.

2.2.6.2.1 Relation entre risque alpha et nombre de sujets nécessaires

Imaginons que l'on cherche à estimer la prévalence d'une maladie chronique dans une population de 400 000 individus. On s'attend à ce que la prévalence soit autour de 50%, et l'on consent à une précision de l'estimation de 3%.

Obtenir un intervalle de confiance de l'estimation qui a 95% de chance de contenir la valeur réelle que l'on cherche à estimer nécessitera d'échantillonner 1064 sujets.

Restons dans le cas d'une étude de prévalence (sondage aléatoire simple).

Quel est « l'effet » du risque alpha auquel on consent sur le nombre de sujets nécessaires ?

α	1%	3%	5%
π	50%	50%	50%
N	400 000	400 000	400 000
ε	3%	3%	3%
n	**1 835**	**1 304**	**1 064**

Plus le risque alpha auquel on consent est petit, plus le nombre de sujets nécessaires grandit.

2.2.6.2.2 Relation entre précision souhaitée de l'estimation et nombre de sujets nécessaires

Imaginons que l'on cherche à estimer la prévalence d'une maladie chronique dans une population de 400 000 individus. Le risque α est fixé à 5%, et l'on s'attend à ce que la prévalence soit autour de 50%

Atteindre une précision de 3% nécessitera d'échantillonner 1 064 sujets. Si vous observez une prévalence à 50% dans cet échantillon, il y a 95% de chances que la prévalence réelle dans la population cible soit entre 47 et 53% (IC 95% : 47-53).

Restons dans le cas d'une étude de prévalence (sondage aléatoire simple).

Quel est « l'effet » de la précision souhaitée sur le nombre de sujets nécessaires ?

ε	1%	3%	5%
π	50%	50%	50%
N	400 000	400 000	400 000
α	5%	5%	5%
n	**9 378**	**1 064**	**384**

Plus on souhaite être précis, plus le nombre de sujets nécessaires grandit.

2.2.6.2.3 Relation entre taille de la population cible et nombre de sujets nécessaires

Observons maintenant, à taille d'échantillon fixée, « l'effet » de la taille de la population cible sur la précision de l'estimation.

N	400	4 000	40 000	400 000	40 000 000
π	50%	50%	50%	50%	50%
α	5%	5%	5%	5%	5%
ε	3%	3%	3%	3%	3%
n	**291**	**842**	**1 039**	**1 064**	**1 067**

Au-delà d'un certain seuil, la taille de la population cible ne modifie guère le nombre de sujets nécessaires pour parvenir à la précision souhaitée de l'estimation. C'est une traduction de la loi des grands nombres.

2.2.6.3 Exemple d'un essai de supériorité

La variable étudiée est ici quantitative. On cherche à démontrer qu'un traitement B est plus efficace qu'un traitement A.

Pour information, sous condition de normalité des distributions et d'égalité des variances, dans le cas d'un test bilatéral, la formule permettant le calcul du nombre de sujets nécessaires *n* par bras est :

$$n = \frac{2\sigma^2(\varphi^{-1}\left(1-\frac{\alpha}{2}\right)+\varphi^{-1}(1-\beta))^2}{(\Delta_{sup})^2}$$

Où

Δ_{sup} : différence minimale d'efficacité escomptée entre les deux traitements (minimal clinically relevant difference ou différence minimale cliniquement importante).
σ : écart type commun aux deux bras.
α : risque de première espèce. Il est en règle générale fixé à 5%. Il peut être diminué si l'on souhaite lutter contre l'inflation du risque alpha lié à la répétition d'un test statistique sur un même échantillon, ou plus généralement si l'on souhaite limiter le risque de faux positif.
β : risque de seconde espèce. Il est souvent fixé à 20%, mais peut être diminué afin d'augmenter la puissance des tests statistiques et de limiter le risque de faux négatif.

Ici, peu importe la taille de la population cible, ou la moyenne des valeurs prises par la variable que l'on observe (critère de jugement). Seuls comptent la différence minimale cliniquement importante, la dispersion des valeurs prises par le critère de jugement dans les deux bras, et les risques alpha ou beta auxquels on consent.

2.2.6.3.1 Relation entre risque beta et nombre de sujets nécessaires

Imaginons que l'on cherche à démontrer la supériorité d'un traitement antihypertenseur B sur un traitement A. Les informations dont on dispose a priori font supposer que l'écart type des pressions artérielles est semblable dans les groupes A et B, autour de 20. On s'attend à une différence d'efficacité entre les deux traitements de l'ordre de 10 mmHg de diminution de la pression artérielle. On limite le risque alpha à 5% (test bilatéral).

Voici le nombre de sujets qu'il faudra inclure selon le risque beta auquel on consent.

β	5%	10%	20%
Δ_{sup}	10	10	10
σ	20	20	20
α	5%	5%	5%
n total	**208**	**170**	**126**

Limiter le risque de passer à côté d'une différence qui existe réellement nécessite d'augmenter drastiquement le nombre de sujets à inclure.

2.2.6.3.2 Relation entre risque alpha et nombre de sujets nécessaires

Les données du problème sont les mêmes, mais on admet un risque beta de 20%.

Voici le nombre de sujets qu'il faudra inclure selon le risque alpha auquel on consent.

α	1%	3%	5%
Δ_{sup}	10	10	10
σ	20	20	20
β	20%	20%	20%
n total	**188**	**146**	**126**

2.2.6.3.3 Relation entre différence minimale cliniquement importante et nombre de sujets nécessaires

Le risque alpha est ici fixé à 5% et le risque beta à 20%. Voici le nombre de sujets qu'il faudra inclure selon la différence minimale d'efficacité escomptée entre les deux traitements.

Δ_{sup}	5	10	15
β	20%	20%	20%
σ	20	20	20
α	5%	5%	5%
n total	**504**	**126**	**56**

Plus la différence minimale d'efficacité que l'on cherche à démontrer entre deux traitements est petite, plus le nombre de sujets à inclure doit être grand. Un excès d'optimisme, qui ferait escompter une différence d'efficacité déraisonnablement grande, conduirait à inclure un nombre trop faible de patients et à nuire à la pertinence de l'expérience, donc à son éthique.

2.2.6.3.4 Relation entre écart type commun aux deux bras et nombre de sujets nécessaires

Le risque alpha est ici fixé à 5%, le risque beta à 20%, et la différence minimale d'efficacité escomptée entre les deux traitements à 10 mmHg. Voici le nombre de sujets qu'il faudra inclure selon la dispersion des valeurs de pression artérielle escomptée dans les deux groupes.

σ	**10**	**20**	**30**
β	20%	20%	20%
Δ_{sup}	10	10	10
α	5%	5%	5%
n total	**32**	**126**	**284**

Plus la variabilité de la réponse au traitement est grande, plus le nombre de sujets nécessaires est élevé. D'où l'importance :

- de définir des critères d'inclusion et d'exclusion permettant la sélection d'un échantillon homogène.
- De mettre en œuvre des outils de mesure du critère de jugement standardisés, fiables et valides.

2.3 CAUSALITÉ, MODÉLISATION

2.3.1 Causalité

Les outils statistiques utilisés en épidémiologie permettent de tester des associations entre divers variables. Mais que deux variables soient corrélées statistiquement ne signifie pas forcément que l'une est la cause de l'autre (association ≠ causalité).

Les outils statistiques permettent de réfuter une hypothèse, mais jamais de la prouver. La causalité ne peut être que suspectée, à condition qu'un certain nombre d'arguments soient réunis.

Ces arguments ont été résumés par Sir Bradford Hill en 1965 en huit à neuf points selon leurs présentations, certains d'entre eux caractérisant la nature de l'association, d'autres son contexte. Pris individuellement, aucun n'est suffisant pour démontrer une causalité

1. Temporalité de l'association :
   La cause, ou exposition, doit précéder l'effet, c'est-à-dire l'apparition de la maladie.
   → *Critère fort et indispensable.*

2. Force de l'association :
   Si une forte proportion des sujets malades ont été exposés à un facteur, mais que très peu des sujets non malades l'ont été, le risque relatif (ou l'odds ratio) est élevé. Plus le risque relatif (ou l'odds ratio) est élevé, plus la causalité de l'association peut être suspectée.
   → *Critère fort.*
   Des biais peuvent modifier la force d'une association, en particulier le biais de confusion.

3. Relation dose-effet, ou gradient biologique :
   Les sujets ayant subi l'exposition la plus forte ou la plus longue (dose) sont le plus fréquemment ou le plus gravement affectés par la maladie (effet), tandis que les moins exposés sont les moins affectés (relation monotone).
   → *Critère fort.*
   Fait défaut en cas de relation non linéaire.

4. Reproductibilité de l'association, fiabilité des résultats :
   L'association considérée est confirmée dans plusieurs études, dans des populations ou des contextes différents.
   → *Critère fort.*

5. Présence de données expérimentales :
   La fin ou la correction de l'exposition coïncide avec la diminution de survenue de la maladie.
   → *Critère fort, mais non nécessaire.*
   Certains processus pathologiques sont irréversibles, ce qui rend impossible la mise en évidence de ce critère à l'échelle individuelle.

6. Plausibilité et cohérence de l'association :
   Une association de nature causale repose sur des mécanismes biologiques (ou comportementaux).
   → *Critère fort, mais non nécessaire.*

L'absence d'explication biologique ou comportementale peut refléter des connaissances scientifiques pour l'heure insuffisantes.
Les données issues des expérimentations in vitro ou chez l'animal ne sont pas toujours transposables à l'homme.

7. Spécificité de l'association :
   Si une cause ne conduit qu'à un seul effet, cela laisse penser qu'il existe un mécanisme propre à la maladie étudiée, ce qui est en faveur d'une relation causale.
   → *Critère désuet (a trait à l'infectiologie, pas aux maladies chroniques multifactorielles) et critiquable.*

8. Analogie :
   Comparaison à d'autres relations causales et leurs mécanismes.
   → *Critère subjectif et critiquable.*

2.3.2 Modélisation

Les outils statistiques sont puissants, mais leur utilisation peut manquer de pertinence.

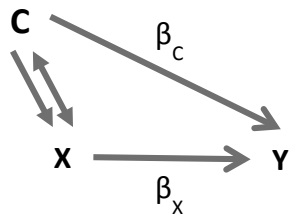
Nous avons évoqué la notion de confusion et le biais qui y a trait dans notre présentation des critères de Bradford Hill. La notion de confusion, mais aussi de médiation et d'interaction, est importante lorsque l'on cherche à comprendre la nature d'une relation statistique. Elle est cruciale quand on procède à une analyse multivariée des facteurs associés à une maladie (régression linéaire, logistique, modèle de Cox).

2.3.2.1 Notion de confusion

2.3.2.1.1 Définition

Un facteur C de confusion dans la relation causale existant entre une variable à expliquer Y et une variable explicative X est une variable explicative de Y associée à la variable X sans être conséquence de cette variable X.

En voici la représentation graphique (ou modèle théorique) :



Par exemple, dans la relation causale existant entre obésité (X) et accident vasculaire cérébral (Y), l'âge ou le niveau d'éducation peuvent jouer le rôle de facteurs de confusion (C).

2.3.2.1.2 Application statistique

Voici maintenant le modèle statistique correspondant au modèle théorique ci-dessus :

$$E(Y) = \beta_0 + \beta_X X + \beta_C C$$

Où :

E(Y) est l'espérance de Y
β_0 est une constante
β_X est l'effet causal total de X sur Y

C'est à partir de ce β_x que l'on va obtenir l'odds ratio ou le risque relatif de présenter la maladie Y lorsque l'on est exposé à l'agent X.

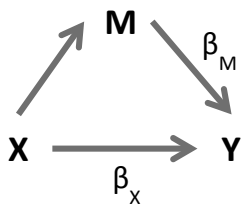
La variable C est un facteur de confusion. Si l'on ne prenait pas compte dans le modèle statistique, l'estimation du paramètre β_x serait fortement biaisée, et notre calcul d'odds ratio ou de risque relatif serait incorrect. On dit que l'on « ajuste » sur le facteur de confusion C.

2.3.2.2 Notion de médiation

2.3.2.2.1 Définition

Un facteur M de médiation dans la relation causale existant entre une variable à expliquer Y et une variable explicative X est une variable explicative de Y également conséquence de la variable X.

En voici la représentation graphique (ou modèle théorique) :



Par exemple, dans la relation causale existant entre obésité (X) et accident vasculaire cérébral (Y), l'hypertension peut jouer le rôle de facteur de médiation (M).

2.3.2.2.2 Application statistique

Voici maintenant le modèle statistique correspondant au modèle théorique ci-dessus, si l'on souhaite estimer l'effet causal direct de X sur Y :

$$E(Y) = \beta_0 + \beta_X X + \beta_M M$$

Où :

E(Y) est l'espérance de Y
β_0 est une constante
β_X est l'effet causal direct de X sur Y (différent de l'effet total)

En dehors des hypothèses de causalité, résumées par les modèles théoriques de confusion et de médiation présentés ci-dessus, rien ne permet de distinguer les deux modèles statistiques.

En fait, il est inutile d'ajuster sur les facteurs de médiation si c'est l'effet causal total de X sur Y que l'on étudie. Le modèle statistique devrait alors s'écrire :

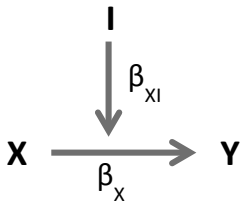
$$E(Y) = \beta_0 + \beta_X X$$

2.3.2.3 Notion d'interaction

2.3.2.3.1 Définition

Un facteur I d'interaction dans la relation causale existant entre une variable à expliquer Y et une variable explicative X est une variable modifiant l'effet de X sur Y.

En voici la représentation graphique (ou modèle théorique) :



Par exemple l'effet de la multiparité sur l'obésité est plus important chez les femmes de faible niveau socio-économique que chez les femmes de niveau socio-économique supérieur.

2.3.2.3.2 Application statistique

Voici maintenant les modèles statistiques correspondant au modèle théorique ci-dessus, si la variable I n'obéit qu'à deux modalités (codées 1 ou 2) :

$$E(Y|I=1) = \beta_{01} + \beta_{x1} X$$

$$E(Y|I=2) = \beta_{02} + \beta_{x2} X$$

L'effet de X sur Y varie selon I, c'est ce que l'on appelle une interaction entre X et I.

On peut également introduire un terme d'interaction dans notre modèle statistique, cela revient au même :

$$E(Y) = \beta_0 + \beta_X X + \beta_I I + \beta_{XI} X*I$$

2.3.2.4 En résumé…

Toute étude en épidémiologie ou en recherche clinique dont l'objectif est d'explorer un ou plusieurs facteurs de risque ou pronostiques (épidémiologie analytique) doit être précédée d'une réflexion sur les potentiels facteurs de confusion (qui nécessitent d'être pris en compte pour l'analyse) ou médiation (qui ne nécessitent pas forcément d'être pris en compte pour l'analyse).

Le contrôle du biais de confusion peut se faire de plusieurs manières :

- Avant la réalisation de l'étude :
  - Par restriction de la population d'étude (critères d'inclusion ou d'exclusion).
  - Dans les études cas/témoin, par appariement des groupes « témoins » et « cas » selon les principaux facteurs de confusion.
- Après la réalisation de l'étude :
  - Par analyse stratifiée sur les diverses modalités prises par les facteurs de confusion potentiels.
  - Par analyse multivariée permettant l'ajustement sur les facteurs de confusion potentiels (modélisation).

Dans le cas particulier des études interventionnelles, la randomisation permettant l'allocation aléatoire de l'intervention évaluée assure le contrôle des facteurs de confusion. La randomisation doit assurer la comparabilité des groupes. Il convient de présenter dans les résultats une comparaison des caractéristiques de base des sujets alloués aux deux bras. Si la comparabilité assurée par la randomisation n'est pas garantie, l'analyse statistique doit être ajustée sur les caractéristiques de base des sujets participant à l'essai.

2.4 RECUEIL DES DONNÉES

2.4.1 Spécificités des données

2.4.1.1 Situation dans le temps

Si les données concernant l'exposition et la maladie sont collectées à la même date chez un même sujet, on parle d'étude transversale. Il s'agit d'un instantané de l'état de santé des sujets participant à l'étude. Cela n'assure pas une détermination précise et fiable de la séquence des évènements.

Si la collecte des données concernant l'exposition et la maladie s'étalent au cours du temps chez un même individu, on parle d'étude longitudinale. Ce type d'étude permet la détermination de la séquence des évènements.

2.4.1.2 Mesures prospectives et rétrospectives

Si les données concernant la maladie sont collectées, grâce à un suivi des sujets, après les données concernant l'exposition à un facteur de risque ou pronostique, ou à une intervention, on parle de recueil prospectif des données.
→ *Cohortes prospectives ou RCT, par exemple.*

Si le recueil et l'analyse des données concernant l'exposition sont effectués après que l'on a eu connaissance de la survenue de la maladie, on parle de recueil rétrospectif des données. Ce mode de recueil a une moindre valeur méthodologique.
→ *Cohortes historiques ou études cas/témoin, par exemple*.

2.4.1.3 Données longitudinales et censurées

Dans le cas des études longitudinales, si les mesures sont répétées dans le temps, on a l'information nécessaire pour déterminer la date de survenue de l'évènement étudié. On parle alors de données censurées ; ces données permettent de procéder à une « analyse de survie » (cf. chapitre 3).

2.4.2 Modalités de recueil des données

Les données peuvent être collectées :

- (A partir de bases de données externes à la recherche).
- A partir de dossiers cliniques archivés ou informatisés.

→ Les données sont souvent incomplètes ou manquantes. Leur mesure n'est pas forcément standardisée, et leur codage peut manquer d'objectivité. Le problème et les hypothèses de recherche ne sont pas formulés avant enregistrement des données.

- A partir d'un registre.
  → Les données nécessaires à l'exploration de nouvelles hypothèses de recherche ne sont pas forcément disponibles.

- A partir de nouvelles observations, dans le cadre d'un protocole de recherche spécifique.
  → C'est la solution idéale : elle minimise les données manquantes et maximise la fiabilité si le protocole de recherche est bien construit et respecté.

Ces données peuvent être :

- (Administratives, externes).
- Obtenues par interrogatoire, déclaratives.
- Obtenues grâce un instrument de mesure.

2.4.3 Types de variables

Les variables observées peuvent être :

- Quantitatives, c'est-à-dire prenant des valeurs inclues dans l'ensemble des nombres réels et exploitables arithmétiquement.
- Qualitatives ordinales ou nominales, c'est-à-dire prenant des modalités non numériques ordonnées (niveau d'éducation…) ou non (sexe…). Ces variables doivent être recodées pour être exploitables statistiquement.

2.4.4 Outils de mesure

Les mesures doivent être standardisées, c'est-à-dire réalisée avec le même outil et selon la même procédure pour une même variable observée, quel que soit l'observateur et le sujet se prêtant à la recherche.

Deux types d'outils peuvent être utilisés :

- Questionnaire, qui peut être administré…
  - en auto-complétion,
  - ou par un enquêteur formé, en entretien en face-à-face ou à distance

  → Toute donnée déclarative expose à plusieurs biais.

- Instruments de mesure (de paramètres physiques, biologiques...).
  → Ce type de mesure est le plus souvent fiable (= exacte et précise), à condition qu'elle fasse appel à des appareils ayant démontré leurs qualités métrologiques et qu'elle soit standardisée.

Des scores ou échelles cliniques, psychométriques ou de comportement peuvent être calculés à partir de données recueillies par questionnaire et/ou instruments. On doit alors se poser deux questions :

- Le score ou l'échelle est-il valide, mesure-t-il bien ce qu'il est censé mesuré ?
- Le score ou l'échelle est-il fiable, donne-t-il des résultats comparables dans des conditions d'utilisation comparables (reproductibilité) ?

→ *De nombreux scores ou échelles de mesures ont été validés par des études spécifiques dans diverses populations. Si possible, il convient d'utiliser ces scores ou échelles validés.*

La performance diagnostique (propriétés intrinsèques : sensibilité et spécificité) des tests éventuellement utilisés doit être prise en compte. Une faible spécificité entraîne un biais de mesure, une faible sensibilité nuit à la précision et la puissance statistique.

Rappelons que la mesure devrait être effectuée en aveugle :

- du traitement alloué dans le cas des études interventionnelles,
- de l'exposition dans le cas des études exposées/non-exposés,
- de la maladie dans le cas des études cas/témoins.

2.4.5 Critères de jugement

Le problème des études interventionnelles est de faire la part entre l'effet de l'intervention évaluée et l'effet de l'évolution spontanée ou du contexte.

Cela nécessite la mise au point d'un protocole de recherche rigoureux. En particulier, un ou plusieurs critères de jugement doivent être choisis.

Il existe différents types de critères de jugement :

- Des critères cliniques (morbidité/mortalité), à privilégier. Ils doivent être les plus objectifs possibles. Ils peuvent être très forts (temps de survie, survie sans complication…) ou plus faibles (qualité de vie…).
- Des critères composites, combinaisons a priori de critères indépendants. Ils peuvent être subjectifs et difficiles à interpréter si l'effet du traitement diffère sur les diverses composantes de la combinaison choisie.
- Des critères de substitution, prédictifs de l'évolution clinique. Ils sont utilisés quand la durée de suivi l'impose. Ils sont plus faibles que les critères cliniques.
- Des critères intermédiaires, cliniques, physiques ou biologiques, souvent considérés de moindre qualité (niveau de pression artérielle, régression de la tumeur…).

Par ailleurs, on distingue :

- Le critère de jugement principal, correspondant à l'objectif principal de l'étude.
- Les critères secondaires, correspondant aux objectifs secondaires.

Les critères de jugement doivent être pertinents, c'est-à-dire correspondre à des diagnostics ayant une réelle importance pour la santé du malade ou sa prise en charge.

Les critères de jugement doivent être très précisément définis : type de donnée, temps et intervalle de mesure, modalités de recueil, outils de mesure.

2.4.6 Données manquantes

Les données manquantes ont trait aux observations que l'on avait l'intention de faire mais qui n'ont pas été faites (non-réponses). Ces non-réponses peuvent être totales (sujet décédé ou perdu de vue dès le début de l'étude, refus de participer, perte de cahiers d'observation…), ou partielles (sujet décédé ou perdu de vue au cours de l'étude, sortie du patient de l'essai par manque d'efficacité ou effets indésirables, impossibilité ou refus de répondre à certaines questions, incohérence des données, défaillance matérielle ou humaine…).

Elles sont inévitables, mais leur nombre doit être limité car elles sont source majeure de biais (biais d'attrition, de suivi…). Elles menacent la représentativité de l'échantillon, donc la validité de l'étude. Elles entrainent une perte d'information qui nuit à la précision des estimations et à la puissance statistique.

On distingue trois grands cas de figures :

- La donnée manquante l'est de manière totalement aléatoire, son absence ne dépend ni des variables observées ni de la valeur non observée (missing completely at random - MCAR).
  → *C'est le cas le moins grave.*
- La probabilité que la donnée soit manquante dépend des variables observées, mais non de la valeur non observée (missing at random - MAR).
  → *C'est un peu plus grave, mais cela peut être corrigé.*
- La probabilité que la donnée soit manquante dépend de la variable manquante (missing not at random - MNAR).
  → *C'est beaucoup plus grave, et c'est beaucoup plus compliqué à gérer…*

En pratique, il est en fait impossible de savoir si la variable manquante est MNAR puisque l'on ne peut pas l'observer.

Au minimum, l'analyse doit préciser si les données manquantes sont MAR, il suffit pour cela de comparer les caractéristiques des sujets présentant des données manquantes ou n'en présentant pas.

Différentes attitudes sont ensuite possibles pour faire face aux données manquantes :

- Analyse sur données complètes, c'est-à-dire sur les dossiers ne présentant aucune donnée manquante.
  → *Cela diminue la précision et la puissance statistique*
- Analyse sur données disponibles, c'est-à-dire sur les tous les dossiers ou les variables d'intérêt sont disponibles.
  → *L'échantillon diffère à chaque analyse, ce qui empêche leur comparaison.*
- Imputation des données manquantes, on attribue une valeur aux variables non observées :
  - Valeur aléatoire.
  - Valeur moyenne des données observées.

- Valeur arbitrairement fixée, par exemple dans le sens de survenue de la maladie chez les exposés et de non survenue chez les non-exposés (méthode du biais maximum).
- Valeur prédite à partir des données observées grâce à divers outils statistiques (régression, imputation multiple).

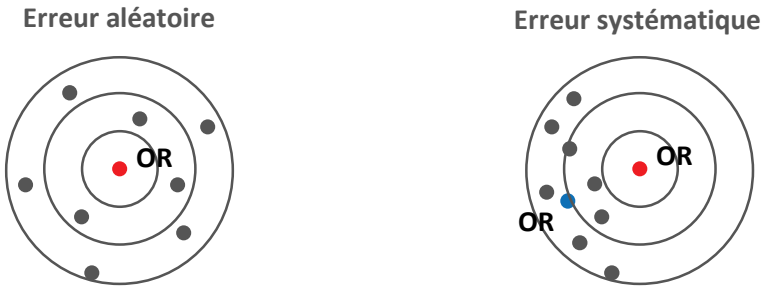
→ *Ces deux dernières solutions sont les moins mauvaises.*

Dans le cas des études interventionnelles, les données manquantes ayant trait au critère de jugement principal peuvent rendre inutilisables les dossiers affectés. Si l'analyse est effectuée en intention de traiter, la dernière valeur connue pourra être prise en compte, mais cela ne suffira pas forcément à contrôler le biais d'attrition. La solution la meilleure est d'avoir peu de données manquantes. Un seuil de 10% peut être proposé.

2.5 LES BIAIS

Les erreurs aléatoires, liées au hasard, par exemple aux fluctuations d'échantillonnage, induisent un manque de précision des estimations. On peut limiter le retentissement de ces erreurs en augmentant l'effectif des échantillons. Les erreurs aléatoires ne constituent pas de biais.

Les biais sont liés à des erreurs dites systématiques, car elles font différer les valeurs estimées des paramètres à l'observation d'un échantillon des valeurs réelles dans la population cible. Il est difficile voire impossible de les corriger a posteriori.



De nombreux biais sont évoqués dans la littérature. On peut les distinguer en trois grandes catégories, même si certains d'entre eux échappent à ce classement.

2.5.1 Biais de sélection

Dans la plupart des cas, les bases de données indexant l'intégralité d'une population cible dont on souhaiterait tirer un échantillon n'existent pas ou ne sont pas accessibles. Diverses méthodes de sondage permettent de faire face à ce problème. Si les méthodes choisies sont inadaptées, la représentativité de l'échantillon peut être contestée, donc l'inférence statistique et la validité de l'expérience. La généralisation des résultats en devient impossible. C'est le cas par exemple d'études où les participants sont recrutés sur la base du volontariat (*biais d'auto-sélection* ou de *volontarisme*).

On peut préciser plusieurs autres biais parmi les biais de sélection.

- Dans le cas des études transversales :
  - Le *biais de non réponse*. Une fois sélectionnés pour participer à l'étude, un certain nombre de patient peuvent s'abstenir d'y participer. Les non-répondants peuvent avoir des caractéristiques d'exposition et de maladie différentes des répondants.
    → *Pour traiter ce biais, il faudrait comparer les caractéristiques des répondants et des non répondants, ce qui est souvent impossible.*
    Reste à obtenir un taux de réponse supérieur à 80%...
  - Le *biais de survie sélective*, si l'exposition diminue fortement la durée de survie.
    → *L'observation des seuls survivants peut alors conduire à une sous-estimation de la force de l'association entre exposition et maladie.*
    La seule méthode pour l'éviter consiste à choisir un schéma d'étude longitudinal avec une date de début de suivi préalable à la date d'exposition.

- Dans le cas des études cas/témoin :
  - Le *biais d'admission* concerne tout particulièrement les études portant sur des patients malades recrutés en milieu hospitalier.
    → *L'exposition est une cause d'hospitalisation en soi et ce biais peut conduire à une surestimation de la force d'association entre exposition et maladie.*
  - Le *biais de détection* (ou *de surveillance* ou *de diagnostic*) concerne les patients dont le suivi pour une maladie ou une autre augmente la chance de bénéficier d'un diagnostic de pathologie intercurrente.
    → *Ce biais peut conduire à une surestimation de la force d'association entre exposition et maladie.*
  - Le *biais de sélection des témoins*. L'échantillonnage des témoins devrait être représentatif de la population dont sont issus les cas.
    → *C'est l'un des multiples problèmes posés par les études cas/témoin…*

- Dans le cas des études de cohorte :
  - Le *biais lié aux travailleurs sains* (*healthy worker effect*) désigne la réduction de risque d'évènement de santé que l'on peut observer chez certains sujets exerçant une activité professionnelle comparativement à la population générale. En effet, par sélection ou auto-sélection, ces sujets peuvent présenter un meilleur état de santé que les autres.
  - Le *biais d'attrition* (perdus de vue ou données manquantes), les non participants ou non répondants étant potentiellement plus à risque de développer la maladie étudiée.
    → *Le taux de non-réponses ou de perdus de vue doit systématiquement être comparé entre les groupes.*

2.5.2 Biais d'information

Ce terme désigne une erreur systématique de mesure de l'exposition ou de la maladie (*biais de mesure*) qui peut conduire à classer à tort un sujet parmi les exposés/malades ou les non-exposés/non malades (*biais de classement*). Les erreurs d'information peuvent être :

- Différentielles, les erreurs de mesures différant alors entre les groupes constitués. Côté participant, cela peut relever d'un *biais de mémorisation* (malades et non malades ont une mémoire différente de leur exposition), voire *de prévarication* (omission volontaire). Côté enquêteur, cela peut relever d'un *biais de subjectivité* (s'il sait la nature de l'exposition ou de la maladie), *d'évaluation* (mesure du critère de jugement différente selon les groupes), ou *de suivi* (les groupes ne sont pas suivis de la même manière)
  → *Les biais de subjectivité, d'évaluation et de suivi peuvent conduire à une surestimation ou une sous-estimation de la force d'association entre exposition et maladie. Pour les limiter, le double aveugle et des procédures de recueil standardisées sont nécessaires.*
- Non différentielles, la probabilité d'erreur ne différant pas entre les malades et les non malades. Cela peut relever d'une erreur systématique de mesure due à un appareil défectueux.
  → *Les erreurs d'information non différentielles conduisent toujours à une sous-estimation de la force d'association. Pour les éviter, les instruments de mesure doivent être sélectionnés sur critère de fiabilité, leur procédure d'utilisation et de contrôle doit être respectée, les enquêteurs doivent être bien choisis et formés, le cadre d'observation doit être rigoureusement défini.*

2.5.3 Biais de confusion

C'est le seul biais qui peut être corrigé après le déroulement de l'étude, c'est-à-dire après le recueil des données, grâce à l'analyse multivariée. Nous l'avons déjà développé au paragraphe 1.3.2.

2.5.4 Et les autres…

On n'ajoutera ici que le biais de publication : les résultats les plus souvent publiés sont des résultats positifs, les échecs étant rarement mis en avant par les auteurs comme par les éditeurs. Ce biais peut avoir un impact particulièrement néfaste sur les méta-analyses.

2.6 NIVEAU DE PREUVE ET GRADATION

Nous reproduisons ici l'état des lieux sur le niveau de preuve et la gradation des recommandations réalisé par la HAS en 2013 à partir du guide d'analyse de la littérature publié par l'Anaes en 2000.

2.6.1 Niveau de preuve

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

La capacité d'une étude à répondre à la question posée est jugée sur la correspondance de l'étude au cadre du travail (question, population, critères de jugement) et sur les caractéristiques suivantes :

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

Classification générale du niveau de preuve d'une étude

Niveau de preuve	Description
Fort	- le protocole est adapté pour répondre au mieux à la question posée ; - la réalisation est effectuée sans biais majeur ; - l'analyse statistique est adaptée aux objectifs ; - la puissance est suffisante.
Intermédiaire	- le protocole est adapté pour répondre au mieux à la question posée ; - puissance nettement insuffisante (effectif insuffisant ou puissance a posteriori insuffisante) ; - et/ou des anomalies mineures.
Faible	Autres types d'études.

Selon le domaine exploré (diagnostic, pronostic, dépistage, traitement, etc.) un fort niveau de preuve peut être donné par des études dont le type de protocole sera différent.

Type de protocole préférentiellement proposé pour une question donnée

Question	Protocole
THÉRAPEUTIQUE Efficacité	Étude contrôlée randomisée
THÉRAPEUTIQUE Sécurité	Étude contrôlée randomisée ou suivi de cohorte
DIAGNOSTIC Reproductibilité/Variabilité	Transversal comparatif avec répétition de mesure
DIAGNOSTIC Sensibilité/Spécificité	Transversal comparatif avec étalon-or
DIAGNOSTIC Efficacité/Utilité	Étude contrôlée randomisée
DIAGNOSTIC Stratégie	Étude contrôlée randomisée ou arbre décisionnel
CAUSALITÉ Phénomène contrôlable fréquent	Étude contrôlée randomisée
CAUSALITÉ Phénomène non contrôlable fréquent	Suivi de cohorte (exposés/non-exposés)
CAUSALITÉ Phénomène rare	Étude cas/témoin
PRONOSTIC Maladie fréquente	Étude contrôlée randomisée ou suivi de cohorte
PRONOSTIC Maladie rare	Étude cas/témoin

2.6.2 Evidence scientifique

La gradation de l'évidence scientifique s'appuie sur :

- L'existence de données de la littérature pour répondre aux questions posées.
- Le niveau de preuve des études disponibles.
- La cohérence de leurs résultats.

Pour une question donnée, il est possible de classer les études en fonction de leur niveau de preuve.

La gradation proposée est la même que les recommandations soient d'ordre thérapeutique, diagnostique.

- Une recommandation de grade A est fondée sur une preuve scientifique établie par des études de fort niveau de preuve : PAR EXEMPLE, essais comparatifs randomisés de forte puissance et sans biais majeur, méta-analyse d'essais contrôlés randomisés, analyse de décision fondée sur des études bien menées.
- Une recommandation de grade B est fondée sur une présomption scientifique fournie par des études de niveau intermédiaire de preuve : PAR EXEMPLE, essais comparatifs randomisés de faible puissance, études comparatives non randomisées bien menées, études de cohortes.
- Une recommandation de grade C est fondée sur des études de moindre niveau de preuve : PAR EXEMPLE, études cas/témoin, séries de cas.

Grade des recommandations

Grade	Niveau de preuve scientifique
A **Preuve scientifique établie**	**Niveau 1** - essais 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** - essais comparatifs randomisés de faible puissance ; - études comparatives non randomisées bien menées ; - études de cohortes.
C **Faible niveau de preuve scientifique**	**Niveau 3** - études cas/témoins.
	Niveau 4 - études comparatives comportant des biais importants ; - études rétrospectives ; - séries de cas ; - études épidémiologiques descriptives (transversale, longitudinale).

2.7 ETHIQUE, DROIT ET RÉGLEMENTATION

Toute recherche biomédicale fait courir des risques aux patients qui s'y prêtent. Après les nombreux abus constatés au cours du 20e siècle, le premier acte de l'harmonisation des pratiques en matière de recherche portant sur les humains a été proclamé en 1964 par l'Association Médicale Mondiale (déclaration d'Helsinki). En 1990, les autorités de réglementation et les représentants de l'industrie pharmaceutique d'Europe, du Japon et des Etats-Unis ont constitué la Conférence Internationale sur l'Harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain (ICH). C'est cette conférence qui a rédigé les normes de bonne pratique clinique (Good Clinical Practice - GCP) qui régissent aujourd'hui la recherche biomédicale.

2.7.1 Bonne pratique clinique

«

Le respect de ces normes garantit au public que les droits, l'innocuité et le bien-être des sujets participant à l'essai sont protégés, conformément aux principes découlant de la Déclaration d'Helsinki, et que les données sur les essais cliniques sont fiables.

- Les essais cliniques doivent être réalisés conformément aux principes éthiques découlant de la Déclaration d'Helsinki, aux bonnes pratiques cliniques et aux exigences réglementaires applicables.
- Avant d'entreprendre un essai, il faut évaluer les inconvénients et les risques prévisibles en fonction des avantages prévus pour le sujet et la société. Un essai doit être entrepris et poursuivi uniquement si les avantages prévus l'emportent sur les risques.
- Les droits, l'innocuité et le bien-être des sujets ont préséance et doivent l'emporter sur les intérêts de la science et de la société.
- Des renseignements cliniques et non cliniques sur le produit de recherche doivent être fournis pour étayer l'essai clinique proposé.
- Les essais cliniques doivent être scientifiquement sûrs et décrits selon un protocole clair et détaillé.
- Un essai doit être réalisé conformément au protocole ayant reçu l'approbation/opinion favorable préalable du comité d'examen de l'établissement/comité d'éthique indépendant.
- Les soins médicaux dispensés aux sujets ainsi que les décisions prises en leur nom doivent toujours être supervisés par un médecin qualifié ou, le cas échéant, par un dentiste qualifié.
- Toute personne participant à la réalisation d'un essai doit posséder les connaissances, la formation et l'expérience requises pour exécuter les tâches qui lui sont confiées.
- Il faut obtenir le consentement libre et éclairé de tous les sujets avant que ces derniers puissent participer à un essai clinique.

- Toutes les données concernant l'essai clinique doivent être enregistrées, traitées et stockées de manière à ce qu'elles puissent être correctement présentées, interprétées et vérifiées.
- La confidentialité des dossiers pouvant servir à identifier les sujets doit être protégée, conformément aux règles relatives à la protection des renseignements personnels et à la confidentialité établies dans les exigences réglementaires applicables.
- Les produits de recherche doivent être fabriqués, manipulés et conservés conformément aux bonnes pratiques de fabrication applicables. Ils doivent être utilisés conformément au protocole approuvé.
- Des systèmes comportant des procédures visant à assurer la qualité de tous les aspects de l'essai doivent être mis en place.

»

2.7.2 Application en France

Plusieurs structures sont impliquées :

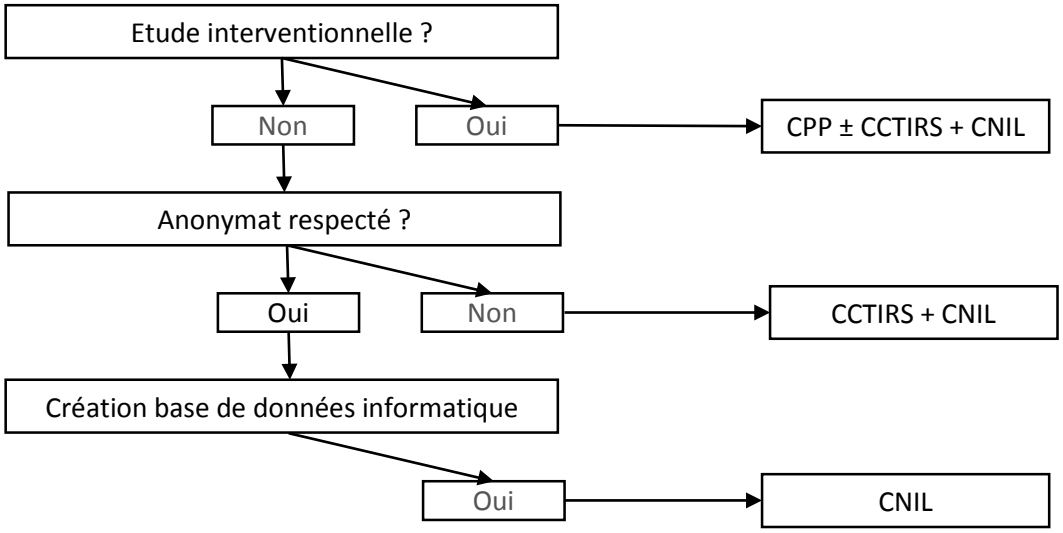
- Le Comité de Protection des Personnes (CPP).
- Le Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé (CCTIRS).
- La Commission Nationale de l'Informatique et des Libertés (CNIL).

D'autres structures peuvent jouer un rôle, hospitalière ou ayant trait l'utilisation d'échantillons biologiques par exemple. En outre, tout essai thérapeutique doit faire l'objet d'une demande d'autorisation à l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM).

Très schématiquement, deux questions sont préalables à la mise en œuvre d'un projet de recherche :

- L'étude est-elle interventionnelle ? Si oui, l'avis du CPP est indispensable. Cet avis portera sur la pertinence et l'éthique de la recherche (rapport bénéfice/risque), la méthodologie et les procédures employées, les modalités d'information et de consentement des participants.
- Les données à caractère personnel recueillies dans le cadre de la recherche permettraient-elle d'une quelconque manière d'identifier les participants ? Si oui, l'avis du CCTIRS est requis. Cet avis portera sur la méthodologie (pertinence de la question, qualité de la méthode), la confidentialité (nécessité du recours à des données à caractère personnel, procédure d'anonymisation et circulation de donnés), et la protection des personnes (respect des droits, information, consentement).

Dans tous les cas, une déclaration ou une demande d'autorisation devra être effectuée auprès de la CNIL. L'avis éventuellement donné portera sur la nécessité de déroger au secret professionnel pour les besoins de la recherche et sur les mesures techniques mises en œuvre pour garantir la confidentialité des données. Il est nécessaire avant de procéder au traitement de ces données.



2.7.3 Application aux Etats-Unis

Le système est décentralisé et fait intervenir des comités d'éthique mandatés par l'Etat fédéral. Désignés sous le nom de « institutional review board » (ou « independent ethics committee », « ethical review board », « research ethics board »), ces comités sont tenus par des institutions académiques, médicales ou commerciales indépendantes, régulées par un bureau du « department of health and human services », appelé « office for human research protections ».

L'avis d'un de ces comités est exigé pour toute recherche biomédicale ou comportementale impliquant des humains. Cet avis porte sur l'éthique de la recherche (rapport bénéfice/risque), sa méthodologie, les modalités d'information et de recueil du consentement des participants, et leur sécurité.

En outre, tout essai thérapeutique doit faire l'objet d'une demande d'autorisation à la « food and drug administration ».