

La vérité expérimentale

Randomisation, biais et mesures entropiques

Hugues Genvrin

Octobre 2026

Résumé

Les conclusions des expériences contrôlées — en médecine, en biologie, en économie — orientent des décisions publiques. On examine ici ce qu’elles établissent et à quelles conditions. On situe d’abord la question dans l’épistémologie : méthode expérimentale, réfutationnisme, tests statistiques, pragmatisme. On formalise ensuite une expérience par un ensemble de sujets, une attribution des traitements, des résultats et un *patron canonique* de covariables. On démontre ce que garantit la randomisation, et ce qu’elle ne garantit pas : elle assure la validité interne, pas la représentativité. On définit le biais de confusion et on en donne la décomposition exacte. On propose trois mesures entropiques : la dépendance entre critères de résultat, le déséquilibre entre groupes, mesuré par une information mutuelle, et le coût d’extension d’un résultat à une autre population, mesuré par une divergence. On traite enfin les difficultés propres à la chirurgie et la lecture des méta-analyses négatives.

Table des matières

1	Repères épistémologiques	2
1.1	De la méthode expérimentale au réfutationnisme	2
1.2	Le pragmatisme	2
1.3	Deux théories des tests	2
1.4	L’asymétrie entre hypothèse établie et hypothèse nouvelle	3
2	Le cadre formel d’une expérience	3
2.1	Sujets, traitements, résultats, covariables	3
2.2	Résultats potentiels et effet causal	3
2.3	Le test comme décision	4
3	La randomisation	4
3.1	Ce qu’elle garantit	4
3.2	Ce qu’elle ne garantit pas	5
3.3	Stratification et blocs	5
4	Le biais de confusion	5
4.1	Définition et décomposition	5
4.2	L’huile d’olive	6
4.3	Le paradoxe de Simpson	6
5	Mesures entropiques	6
5.1	La dépendance entre critères de résultat	6
5.2	Le déséquilibre entre groupes	7
5.3	Le coût d’extension d’un résultat	8
5.4	Hétérogénéité de l’effet	8

6	Le cas de la chirurgie	8
6.1	Les difficultés	8
6.2	L'opérateur central et l'effet de grappe	9
6.3	Deux réponses méthodologiques	9
7	Niveaux de preuve et méta-analyses	9
7.1	Niveaux de preuve	9
7.2	Lire une méta-analyse négative	9
8	Synthèse	10

1 Repères épistémologiques

1.1 De la méthode expérimentale au réfutationnisme

Claude Bernard, dans l'*Introduction à l'étude de la médecine expérimentale* (1865), décrit la méthode *hypothético-déductive* : on formule une hypothèse, on en déduit des conséquences observables, on institue l'expérience qui les met à l'épreuve, et l'on compare avec un témoin.

Deux conceptions de la confirmation se sont ensuite opposées.

- Le *vérificationnisme* du Cercle de Vienne tient qu'un énoncé empirique a un sens dans la mesure où il est vérifiable par l'expérience.
- Le *réfutationnisme* de Popper (*Logik der Forschung*, 1934) le rejette : aucun nombre d'observations favorables ne vérifie une loi universelle, mais une seule observation contraire peut la réfuter. Une hypothèse n'est jamais vérifiée ; elle est *corroborée* tant qu'elle résiste aux tentatives de réfutation.

Popper n'est donc pas vérificationniste : il est l'adversaire principal du vérificationnisme.

1.2 Le pragmatisme

Pour Peirce (« The Fixation of Belief », 1877), une croyance est une *habitude d'action* : croire, c'est être disposé à agir d'une certaine manière. Dewey (*Logic : The Theory of Inquiry*, 1938) remplace la notion de vérité par celle d'*assertabilité garantie* (*warranted assertibility*) : est assertable ce qu'une enquête bien conduite autorise à affirmer. Ces deux conceptions conviennent bien à la vérité expérimentale, qui débouche sur des décisions — rembourser un médicament, financer un programme — plutôt que sur des certitudes.

1.3 Deux théories des tests

Les tests statistiques reposent sur deux écoles distinctes, souvent confondues dans la pratique.

- *Fisher* (années 1920–1930) : on calcule, sous une hypothèse nulle H_0 , la probabilité p d'observer un résultat au moins aussi extrême. La valeur p est une mesure graduée de l'incompatibilité des données avec H_0 .
- *Neyman et Pearson* (1933) : on fixe à l'avance une règle de décision entre H_0 et une alternative H_1 , en contrôlant le risque de première espèce α (rejeter H_0 à tort) et le risque de seconde espèce β (ne pas la rejeter à tort). La quantité $1 - \beta$ est la *puissance*.

Une troisième approche, *bayésienne*, attribue une probabilité aux hypothèses elles-mêmes et la met à jour par les données.

1.4 L'asymétrie entre hypothèse établie et hypothèse nouvelle

Les tests traitent différemment l'hypothèse nulle et l'hypothèse nouvelle : c'est voulu. Le risque α , celui d'adopter à tort une nouveauté, est fixé bas ; le risque β est contrôlé par la taille de l'échantillon. L'approche bayésienne explique pourquoi cette prudence est rationnelle.

Proposition 1.1 (Valeur prédictive d'un résultat significatif). *Soit R le rapport, parmi les hypothèses testées dans un domaine, entre hypothèses vraies et hypothèses fausses. Un test de niveau α et de puissance $1 - \beta$ donne un résultat significatif qui correspond à une hypothèse vraie avec probabilité*

$$\text{VPP} = \frac{(1 - \beta)R}{(1 - \beta)R + \alpha}.$$

Démonstration. Sur $R + 1$ hypothèses, R sont vraies et donnent un résultat significatif avec probabilité $1 - \beta$; une est fausse et donne un résultat significatif avec probabilité α . On applique la formule de Bayes. $\square$

Avec $R = 0,1$ — une hypothèse vraie pour dix fausses —, une puissance de 0,8 et $\alpha = 0,05$, on trouve $\text{VPP} \approx 0,62$: près de quatre résultats significatifs sur dix sont des faux positifs. C'est l'argument d'Ioannidis (2005). Une hypothèse nouvelle, peu probable a priori, demande donc plus de preuves qu'une hypothèse déjà corroborée : le « biais initial » est la traduction rationnelle d'une probabilité a priori faible.

2 Le cadre formel d'une expérience

2.1 Sujets, traitements, résultats, covariables

Définition 2.1 (Système expérimental). Un *système expérimental* est la donnée de :

- un ensemble fini E de sujets, de cardinal n ;
- une *attribution* $G : E \rightarrow \{0, 1\}$, qui indique le groupe de chaque sujet (témoin ou traité) ;
- des *résultats* $Y_t : E \rightarrow F_t$, mesurés à une ou plusieurs dates t , à valeurs dans un ensemble F_t de dimension p ;
- un *patron canonique* de covariables mesurées $X = (X_1, \dots, X_m)$, où chaque $X_i : E \rightarrow E/\mathcal{R}_i$ est la surjection canonique d'une relation d'équivalence $\mathcal{R}_i$ (sexe, classe d'âge, stade de la maladie...).

Les *canons cachés* sont les autres caractéristiques U des sujets, non mesurées, qui peuvent influencer le résultat.

Le patron canonique définit l'application $x \mapsto (X_1(x), \dots, X_m(x))$ vers le produit $\prod_i E/\mathcal{R}_i$: deux sujets ayant les mêmes valeurs de covariables sont indiscernables pour l'expérience.

2.2 Résultats potentiels et effet causal

On adopte le cadre des résultats potentiels (Neyman 1923, Rubin 1974). Chaque sujet x possède deux résultats potentiels : $Y(1)$ s'il est traité, $Y(0)$ s'il ne l'est pas. On n'observe que l'un des deux, $Y = Y(G)$.

Définition 2.2. L'*effet causal moyen* du traitement sur la population E est

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \frac{1}{n} \sum_{x \in E} (Y_x(1) - Y_x(0)).$$

L'effet causal porte sur une comparaison *contrefactuelle*, sujet par sujet, qu'aucune observation ne réalise directement. Toute la méthodologie expérimentale vise à l'estimer à partir de comparaisons *entre groupes*.

2.3 Le test comme décision

Un test ne produit pas une vérité, mais une *décision* assortie de risques d'erreur.

	H_0 vraie (pas d'effet)	H_1 vraie (effet)
Décision : ne pas rejeter H_0	correcte, probabilité $1 - \alpha$	erreur de seconde espèce, β
Décision : rejeter H_0	erreur de première espèce, α	correcte, puissance $1 - \beta$

On ne peut donc pas écrire une équivalence entre la conclusion du test et la relation réelle entre les groupes. On a seulement des probabilités conditionnelles : $\Pr(\text{rejet} \mid H_0) \leq \alpha$ et $\Pr(\text{rejet} \mid H_1) = 1 - \beta$. Il faut distinguer la *décision statistique*, qui est le produit du test, et l'*énoncé sur le monde* — « le traitement réduit la mortalité » — qu'elle autorise à affirmer, avec un degré de garantie mesuré par α , β et, en termes bayésiens, par la valeur prédictive de la proposition 1.1.

3 La randomisation

3.1 Ce qu'elle garantit

Définition 3.1. L'attribution est *randomisée* si elle est tirée au sort, indépendamment de toutes les caractéristiques des sujets. Dans la randomisation *complète*, on choisit au hasard, uniformément, n_1 sujets traités parmi les n .

Théorème 3.2 (Interchangeabilité). *Si l'attribution est randomisée, alors*

$$G \perp\!\!\!\perp (X, U, Y(0), Y(1)).$$

Les deux groupes ont, en loi, la même distribution de covariables mesurées, de canons cachés et de résultats potentiels.

Démonstration. C'est la définition même : le tirage au sort est indépendant de tout ce qui caractérise les sujets avant l'attribution, mesuré ou non. $\square$

Théorème 3.3 (Absence de biais, Neyman 1923). *Sous randomisation complète, la différence des moyennes observées*

$$\hat{\tau} = \bar{Y}_{\text{traités}} - \bar{Y}_{\text{témoins}}$$

est un estimateur sans biais de l'effet causal moyen : $\mathbb{E}[\hat{\tau}] = \tau$.

Démonstration. Chaque sujet x est traité avec probabilité n_1/n . Donc

$$\mathbb{E}\left[\frac{1}{n_1} \sum_{x:G_x=1} Y_x(1)\right] = \frac{1}{n_1} \sum_{x \in E} \frac{n_1}{n} Y_x(1) = \frac{1}{n} \sum_{x \in E} Y_x(1).$$

De même pour les témoins avec $Y(0)$; on soustrait. $\square$

La randomisation neutralise ainsi *tous* les facteurs de confusion, mesurés ou cachés : c'est la seule méthode qui y parvienne sans les connaître.

3.2 Ce qu'elle ne garantit pas

L'équilibre exact. L'interchangeabilité vaut en loi, pas sur l'échantillon. Un déséquilibre fortuit est fréquent en petit effectif. Pour une covariable binaire de fréquence $\frac{1}{2}$, la probabilité que les deux groupes diffèrent de plus de dix points de pourcentage est :

Sujets par groupe	20	50	100	500	1000
Probabilité d'un écart > 10 points	0,43	0,27	0,14	0,001	$< 10^{-3}$

Ce déséquilibre n'introduit pas de biais — le théorème 3.3 vaut en moyenne sur les tirages — mais il augmente la variance de l'estimation.

La représentativité. Tirer au sort l'*attribution* ne rend pas l'échantillon représentatif de la population générale. Seul un tirage au sort de l'*échantillon* dans la population le ferait, ce que les essais cliniques ne pratiquent presque jamais : ils incluent des patients éligibles et consentants, souvent plus jeunes et moins malades que les patients traités en pratique courante. Il faut donc distinguer :

- la *validité interne* : l'effet est correctement estimé *dans* la population de l'essai ; la randomisation de l'attribution la garantit ;
- la *validité externe* : l'effet vaut aussi dans une autre population ; la randomisation de l'attribution ne la garantit pas (section 5.3).

3.3 Stratification et blocs

- La *randomisation stratifiée* tire au sort séparément dans chaque classe d'une ou plusieurs covariables du patron canonique : l'équilibre est alors assuré pour ces covariables, et seulement pour elles.
- La *randomisation par blocs* tire au sort à l'intérieur de blocs successifs de taille fixe, par exemple quatre sujets dont deux traités : les effectifs restent équilibrés tout au long des inclusions.

Les deux techniques se combinent : on randomise par blocs à l'intérieur de chaque strate. Elles réduisent la variance, pas le biais, que la randomisation simple supprime déjà.

4 Le biais de confusion

4.1 Définition et décomposition

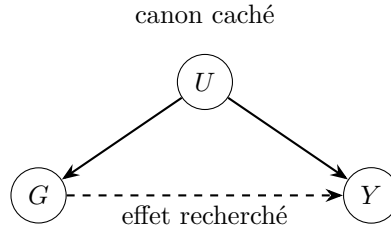
Sans randomisation, l'attribution peut dépendre des caractéristiques des sujets.

Proposition 4.1 (Décomposition). *Pour toute attribution,*

$$\underbrace{\mathbb{E}[Y \mid G = 1] - \mathbb{E}[Y \mid G = 0]}_{\text{différence observée}} = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid G = 1]}_{\text{effet chez les traités}} + \underbrace{\mathbb{E}[Y(0) \mid G = 1] - \mathbb{E}[Y(0) \mid G = 0]}_{\text{biais de sélection}}.$$

Démonstration. On écrit $\mathbb{E}[Y \mid G = 1] = \mathbb{E}[Y(1) \mid G = 1]$, on ajoute et retranche $\mathbb{E}[Y(0) \mid G = 1]$, et l'on utilise $\mathbb{E}[Y \mid G = 0] = \mathbb{E}[Y(0) \mid G = 0]$. $\square$

Le biais est nul si les deux groupes auraient eu le même résultat sans traitement, ce que garantit la randomisation (théorème 3.2). Il est non nul dès qu'une caractéristique U , non prise en compte dans le patron canonique, influence à la fois l'attribution et le résultat : c'est un *facteur de confusion*. Le « biais morphique » du texte d'origine — l'existence d'une relation entre le résultat et un canon caché — est exactement ce biais.



4.2 L’huile d’olive

Les études d’observation sur l’alimentation méditerranéenne, depuis l’étude des Sept Pays de Keys, ont associé la consommation d’huile d’olive à une longévité plus grande. Une association observationnelle ne permet pas de conclure à un effet causal : les populations concernées différaient aussi par leur activité physique, leur mode de vie, leur alimentation dans son ensemble. Chacun de ces facteurs est un canon caché possible.

L’essai *PREDIMED* a précisément cherché à lever cette ambiguïté par la randomisation : 7 447 personnes à haut risque cardiovasculaire ont été réparties entre un régime méditerranéen enrichi en huile d’olive vierge extra, un régime méditerranéen enrichi en fruits à coque et un régime témoin. Publié en 2013 dans le *New England Journal of Medicine*, il a été rétracté puis republié en 2018, car la randomisation n’avait pas été respectée partout : des membres d’un même foyer avaient été affectés au même groupe, et un centre avait randomisé des cliniques entières plutôt que des individus. Cet épisode illustre le théorème 3.2 : la garantie de la randomisation ne vaut que si le tirage au sort est effectivement indépendant des caractéristiques des sujets.

4.3 Le paradoxe de Simpson

Une étude de 1986 sur les calculs rénaux (Charig et al.) compare deux techniques. Le traitement A réussit mieux que B pour les petits calculs (93 % contre 87 %) et pour les gros (73 % contre 69 %), mais moins bien sur l’ensemble (78 % contre 83 %). La taille du calcul est un facteur de confusion : A a surtout été utilisé pour les gros calculs, plus difficiles. On en mesurera le déséquilibre à la section 5.2.

5 Mesures entropiques

On mesure les entropies en bits ; H désigne l’entropie de Shannon, I l’information mutuelle.

5.1 La dépendance entre critères de résultat

On découpe chacune des p dimensions du résultat à sa médiane : chaque sujet tombe dans l’une des 2^p cellules. Soit π_c la proportion de sujets dans la cellule c , et $H = -\sum_c \pi_c \log_2 \pi_c$.

Proposition 5.1. 1. $1 \leq H \leq p$ bits.

2. $H = p$ si et seulement si les p indicatrices « au-dessus de la médiane » sont indépendantes.
3. $H = 1$ si et seulement si elles sont toutes égales, ou opposées, deux à deux.
4. Le déficit $C = p - H$ est l’information multiple, ou corrélation totale, des p indicatrices : $C = \sum_i H(B_i) - H(B_1, \dots, B_p)$.

Démonstration. Chaque indicatrice B_i vaut 1 avec probabilité $\frac{1}{2}$, donc $H(B_i) = 1$. L’entropie jointe est au plus la somme des entropies, p , avec égalité si et seulement s’il y a indépendance ; elle est au moins l’entropie de chacune, 1, avec égalité si et seulement si chaque B_j est une fonction de B_1 . □

On peut alors définir le taux

$$\tau_H = 1 - \frac{H}{p} = \frac{C}{p} \in \left[0; 1 - \frac{1}{p}\right].$$

Remarque. Deux erreurs sont à éviter.

- Compter les cellules non vides au lieu d'utiliser les fréquences : le nombre de cellules vides dépend surtout de l'effectif. Avec $p = 10$ critères, il y a 1024 cellules, et 100 sujets en laissent au moins 924 vides, quelle que soit l'expérience. L'entropie de cellules non vides vaut d'ailleurs $\log_2(2^p - q)$ si q cellules sont vides, et non $p - q$.
- Interpréter τ_H comme un biais : il mesure la *redondance* entre critères de résultat. Un τ_H élevé signifie que les critères mesurent largement la même chose ; c'est une information utile pour choisir un critère principal, pas un défaut de l'expérience.

Comme toute entropie estimée sur un échantillon, H est biaisée vers le bas quand l'effectif est faible devant le nombre de cellules ; il faut au moins plusieurs sujets par cellule.

5.2 Le déséquilibre entre groupes

La bonne mesure entropique d'un biais de sélection compare les *groupes* sur les *covariables*.

Définition 5.2. L'*indice de déséquilibre* d'une expérience est l'information mutuelle $I(G; X)$ entre l'attribution et le patron canonique des covariables mesurées.

Proposition 5.3. 1. $I(G; X) \geq 0$, avec égalité si et seulement si la répartition des covariables est la même dans les deux groupes.

2. Si l'attribution est randomisée, $I(G; X) = 0$ pour la loi de population.

3. Sur un échantillon de n sujets, si G et X sont indépendants, la statistique $2n \hat{I}(G; X)$, avec $\hat{I}$ en nats, suit approximativement une loi du χ^2 à $(k - 1)$ degrés de liberté, où k est le nombre de valeurs du patron canonique (test G du rapport de vraisemblance). En particulier, $\mathbb{E}[\hat{I}] \approx (k - 1)/(2n)$ nats.

Démonstration. 1. C'est la propriété de l'information mutuelle : $I(G; X) = D(p_{GX} \| p_G p_X)$. 2. C'est le théorème 3.2. 3. C'est le théorème de Wilks appliqué au test d'indépendance dans un tableau $2 \times k$: la statistique du rapport de vraisemblance vaut exactement $2n\hat{I}$. $\square$

Exemple 5.4. Dans l'étude des calculs rénaux, avec G = traitement et X = taille du calcul,

$$\hat{I}(G; X) \approx 0,144 \text{ nat} \approx 0,207 \text{ bit}, \quad 2n\hat{I} \approx 201 \quad (n = 700, 1 \text{ degré de liberté}).$$

Sous randomisation, on attendrait $\hat{I} \approx 1/1400 \approx 0,0007$ nat : le déséquilibre observé est deux cents fois plus grand, et la valeur p du test est de l'ordre de 10^{-45} . Le traitement a été choisi *en fonction* de la taille du calcul : l'entropie le mesure directement. Rapporté à $H(G) = 1$ bit, le coefficient d'incertitude vaut $U(G | X) \approx 0,21$: la taille du calcul « explique » un cinquième de l'incertitude sur le traitement reçu.

Cet indice permet de construire le tableau croisé suggéré par le texte d'origine, mais avec un sens précis :

	$I(G; X)$ faible	$I(G; X)$ élevé
Effet favorable	résultat crédible	résultat suspect : confusion possible
Effet défavorable ou nul	résultat crédible	résultat ininterprétable

Un indice faible ne garantit rien pour les canons *cachés* : seule la randomisation le fait. Mais un indice élevé sur les covariables mesurées est un signal d'alerte objectif.

5.3 Le coût d'extension d'un résultat

On veut étendre une conclusion établie sur la population de l'essai, de loi de covariables P , à une population cible $\mathcal{C}$, de loi P^* . Si l'effet dépend des covariables, l'effet dans la population cible est

$$\tau^* = \sum_x \tau(x) P^*(x), \quad \tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

que l'on estime en repondérant les sujets de l'essai par $w(x) = P^*(x)/P(x)$. Cette extension n'est possible que si toute valeur de x présente dans la cible l'est aussi dans l'essai.

Proposition 5.5 (Coût de l'extension). *La repondération réduit l'effectif utile de n à l'effectif efficace de Kish*

$$n_{\text{eff}} = \frac{n}{\mathbb{E}_P[w^2]} = \frac{n}{1 + \chi^2(P^* \parallel P)} \leq \frac{n}{\exp(D(P^* \parallel P))},$$

où $\chi^2(P^* \parallel P) = \sum_x (P^*(x) - P(x))^2 / P(x)$, et D est l'entropie relative en nats.

Démonstration. $\mathbb{E}_P[w^2] = \sum_x P^*(x)^2 / P(x) = 1 + \chi^2(P^* \parallel P)$ par développement du carré. Par l'inégalité de Jensen appliquée à la fonction concave $\ln$ sous la loi P^* , $D(P^* \parallel P) = \mathbb{E}_{P^*}[\ln w] \leq \ln \mathbb{E}_{P^*}[w] = \ln \mathbb{E}_P[w]$. $\square$

Plus la population cible s'écarte de celle de l'essai, au sens de l'entropie relative, plus l'information disponible pour conclure sur la cible diminue : c'est la forme exacte du coût d'extension, que le texte d'origine rapprochait de l'entropie de mélange. Si la cible contient des patients absents de l'essai — $P(x) = 0 < P^*(x)$ —, la divergence est infinie et l'extension n'est pas justifiée par les données.

5.4 Hétérogénéité de l'effet

Lorsque l'effet $\tau(x)$ varie avec les covariables, l'effet moyen mélange des effets différents. Réunir des sous-populations hétérogènes en une seule estimation crée une incertitude qui ressemble à une entropie de mélange : l'effet moyen peut être nul alors que le traitement est bénéfique pour les uns et nuisible pour les autres. C'est la raison d'être des analyses en sous-groupes, à condition qu'elles soient prévues à l'avance et testées par un test d'interaction, et non par la comparaison de valeurs p entre sous-groupes.

6 Le cas de la chirurgie

6.1 Les difficultés

McLeod et al. (1996) ont recensé les difficultés propres aux essais randomisés en chirurgie. On peut les regrouper ainsi :

- *la compétence de l'opérateur* : l'apprentissage de la technique, les différences d'habileté entre chirurgiens et le rôle de l'équipe ; un médicament, au contraire, agit de la même façon quel que soit le prescripteur ;
- *la standardisation* : une intervention est plus difficile à décrire et à reproduire qu'une posologie ;
- *l'aveugle* : le double aveugle est en général impossible, le simple aveugle difficile ;
- *le calendrier* : le délai entre randomisation et intervention laisse l'état du patient évoluer, et certaines non-éligibilités ne se révèlent qu'en cours d'opération ;
- *le recrutement* : un patient peut obtenir une technique chirurgicale hors essai en changeant de chirurgien, alors qu'un médicament nouveau n'est souvent accessible que dans l'essai.

6.2 L'opérateur central et l'effet de grappe

Lorsque chaque chirurgien opère plusieurs patients, les résultats de ses patients sont corrélés : ils partagent son habileté. Si chaque chirurgien opère m patients et que la corrélation intra-chirurgien est ρ , la variance de l'estimation est multipliée par l'*effet de plan*

$$\text{DEFF} = 1 + (m - 1)\rho.$$

Avec $m = 20$ patients par chirurgien et $\rho = 0,05$, l'effectif nécessaire est presque doublé ($\text{DEFF} = 1,95$). L'« opérateur central » du texte d'origine a donc une traduction statistique précise.

6.3 Deux réponses méthodologiques

- Les *essais fondés sur l'expertise* (Devereaux et al., 2005) randomisent les patients entre des chirurgiens experts de la technique A et des chirurgiens experts de la technique B, plutôt qu'entre deux techniques réalisées par le même chirurgien. Ils évitent le biais d'expertise différentielle et réduisent les changements de technique en cours d'intervention.
- Le cadre *IDEAL* (McCulloch et al., 2009) distingue les étapes d'une innovation chirurgicale — idée, développement, exploration, évaluation, suivi au long cours — et propose pour chacune le type d'étude adapté. On n'évalue pas de la même façon une technique à ses premiers essais et une technique candidate au rang de référence.

7 Niveaux de preuve et méta-analyses

7.1 Niveaux de preuve

La Haute Autorité de santé (2013) gradue les recommandations selon le niveau de preuve des études : grade A pour les essais randomisés de forte puissance sans biais majeur et les méta-analyses d'essais randomisés ; grade B pour les essais randomisés de faible puissance, les études comparatives non randomisées bien menées et les études de cohorte ; grade C pour les études cas-témoins et les séries de cas. L'accord d'experts intervient en l'absence de preuves suffisantes. Les études de cohorte et cas-témoins sont des études *d'observation*, non des expériences : l'investigateur n'y attribue pas le traitement.

7.2 Lire une méta-analyse négative

Une méta-analyse qui ne met pas en évidence d'effet ne démontre pas l'absence d'effet. Quatre explications sont possibles :

1. l'effet est nul ou négligeable ;
2. la puissance est insuffisante : l'intervalle de confiance de l'effet combiné reste compatible avec un effet important ;
3. l'effet est hétérogène entre études ;
4. des biais — de publication, de sélection, de mesure — faussent les études.

Seul l'examen de l'intervalle de confiance permet de conclure à l'absence d'un effet *cliniquement important* : si cet intervalle exclut tout effet de taille pertinente, l'absence d'effet est établie à ce degré.

Exemple 7.1 (Le décompte des votes). Cinq études de puissance 0,5 portent sur un effet réel. Le nombre d'études significatives suit une loi binomiale $\mathcal{B}(5; 0,5)$: avec probabilité $\frac{1}{2}$, deux études au plus sur cinq sont significatives. Conclure « la majorité des études est négative, donc pas d'effet » serait faux une fois sur deux. C'est pourquoi la méta-analyse combine les estimations au lieu de compter les résultats significatifs.

Définition 7.2 (Hétérogénéité). Pour k études d'estimations $\hat{\theta}_j$ et de poids $w_j = 1/\sigma_j^2$, la statistique de Cochran est $Q = \sum_j w_j(\hat{\theta}_j - \bar{\theta})^2$, avec $\bar{\theta} = \sum_j w_j \hat{\theta}_j / \sum_j w_j$. L'indice de Higgins et Thompson

$$I^2 = \max\left(0, \frac{Q - (k - 1)}{Q}\right)$$

estime la part de la variabilité entre études qui dépasse le hasard.

Un I^2 élevé signale que les études ne mesurent pas le même effet : des canons propres à chaque étude — population, dose, technique — modifient le résultat. C'est, au niveau des méta-analyses, l'équivalent de l'hétérogénéité de l'effet.

8 Synthèse

Notion	Forme méthodologique
Patron canonique	Covariables mesurées X
Canons cachés	Facteurs de confusion non mesurés U
Biais morphique	Biais de confusion (proposition 4.1)
Vérité logique du test	Décision statistique, risques α et β
Biais initial contre le nouveau	Probabilité a priori, valeur prédictive
Première randomisation	Échantillonnage aléatoire : validité externe
Seconde randomisation	Attribution aléatoire : validité interne
Randomisation segmentée	Randomisation stratifiée, par blocs
Indicateur entropique des résultats	Information multiple des critères (redondance)
Indicateur entropique de biais	Information mutuelle $I(G; X)$
Extension d'une vérité	Transportabilité, coût n_{eff}
Opérateur central	Effet de grappe, essais fondés sur l'expertise

Conclusion. La vérité expérimentale n'est pas une vérité au sens logique : c'est une assertion garantie, au sens de Dewey, dont la garantie se mesure. La randomisation en est le fondement, parce qu'elle seule neutralise les facteurs de confusion inconnus ; mais elle garantit la validité interne, non la portée du résultat hors de la population étudiée. L'entropie fournit des mesures précises des trois fragilités d'une conclusion expérimentale : le déséquilibre entre groupes, par l'information mutuelle entre attribution et covariables ; le coût d'extension à une autre population, par l'entropie relative entre populations ; et l'hétérogénéité de l'effet. Ces mesures ne remplacent pas le jugement, mais elles le rendent explicite et comparable d'une expérience à l'autre.

Références

- [1] C. Bernard, *Introduction à l'étude de la médecine expérimentale*, Baillière, 1865.
- [2] C. R. Charig et al., « Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy », *British Medical Journal* 292, 1986.
- [3] P. J. Devereaux et al., « Need for expertise based randomised controlled trials », *British Medical Journal* 330, 2005, p. 88.
- [4] J. Dewey, *Logic : The Theory of Inquiry*, Holt, 1938.

- [5] R. Estruch et al., « Primary prevention of cardiovascular disease with a Mediterranean diet », *New England Journal of Medicine* 368, 2013 ; rétracté et republié, *ibid.* 378, 2018.
- [6] R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, 1935.
- [7] Haute Autorité de santé, *Niveau de preuve et gradation des recommandations de bonne pratique. État des lieux*, 2013.
- [8] J. P. T. Higgins, S. G. Thompson, « Quantifying heterogeneity in a meta-analysis », *Statistics in Medicine* 21, 2002.
- [9] J. P. A. Ioannidis, « Why most published research findings are false », *PLoS Medicine* 2, 2005.
- [10] L. Kish, *Survey Sampling*, Wiley, 1965.
- [11] P. McCulloch et al., « No surgical innovation without evaluation : the IDEAL recommendations », *The Lancet* 374, 2009.
- [12] R. S. McLeod et al., « Randomized controlled trials in surgery : issues and problems », *Surgery* 119, 1996.
- [13] J. Splawa-Neyman, mémoire de 1923 sur l'application des probabilités aux expériences agricoles (en polonais), traduction anglaise partielle : « On the application of probability theory to agricultural experiments. Essay on principles. Section 9 », *Statistical Science* 5, 1990.
- [14] J. Neyman, E. S. Pearson, « On the problem of the most efficient tests of statistical hypotheses », *Philosophical Transactions of the Royal Society A* 231, 1933.
- [15] J. Pearl, *Causality*, Cambridge University Press, 2^e éd., 2009.
- [16] C. S. Peirce, « The fixation of belief », *Popular Science Monthly* 12, 1877.
- [17] K. R. Popper, *Logik der Forschung*, Springer, 1934 ; trad. fr. *La Logique de la découverte scientifique*, Payot, 1973.
- [18] D. B. Rubin, « Estimating causal effects of treatments in randomized and nonrandomized studies », *Journal of Educational Psychology* 66, 1974.