Poly de révision EPI

Annales d'épidémiologie : toutes les questions, toutes les réponses

Table des matières

Poly de révision EPI

Annales résolues, méthodes, lecture critique et réponses types

Document de travail orienté examen : **toutes les questions repérées dans les originaux EPI**, avec corrections didactiques, raisonnement méthodologique et réponses exploitables dans une copie.

Ce poly a un objectif simple : **permettre de répondre entièrement aux annales d'épidémiologie analytique** du dossier.

Le fil directeur est le même que dans le poly NSN :

1. repartir des **sujets originaux** ;

2. identifier toutes les **questions réellement posées** ;
3. revenir au **papier support** pour comprendre ce qu'il fallait aller chercher ;
4. produire une **réponse d'examen** complète, concise et défendable.

épidémiologie annales exhaustives registre causalité HTML + PDF

> **!** Point de méthode
>
> En EPI, la bonne discussion n'est pas toujours celle du "nombre de sujets nécessaires". Très souvent, la vraie question est ailleurs : **précision des estimations, nombre d'événements, validité causale, qualité des mesures, confusion résiduelle, biais écologique, limites du design**.

Comment utiliser ce poly

1. Lisez d'abord le **contexte de l'article** et le **type de design**.
2. Repérez ensuite **ce que l'examinateur veut vraiment tester**.
3. Si un calcul est possible, refaites-le au brouillon.
4. Si le calcul exact n'est pas possible, regardez comment construire une **réponse méthodologique solide** plutôt qu'un faux calcul.
5. Utilisez les **réponses types** comme gabarit de copie, pas comme texte à réciter mot pour mot.

1. Carte rapide des questions couvertes

Année	Article support	Type de questions dominantes	Nombre de questions EPI couvertes
2011	Kalager et al., NEJM 2010	quasi-expérience, personnes-années, puissance, interprétation	6
2012	Cooper et al., NEJM 2011	cohorte, exposition, modèle de Cox, surajustement, puissance	6
2013	Lasalvia et al., Lancet 2013	échantillonnage, modèle binomial négatif, précision, absence de contrôle	4
2014	Bao et al., NEJM 2013	confusion, causalité, puissance, exclusion des malades au départ	4
2016	Schnabel et al., Lancet 2015	cohorte, PAR, puissance hétérogène, données manquantes	4
2018	Autier et al., BMJ 2017	surdiagnostic, joinpoint, APC, critique des essais	5

Année	Article support	Type de questions dominantes	Nombre de questions EPI couvertes
2019	Shan et al., BMJ 2018	attrition, interaction addi-tive/multiplicative, sensibilité, taille d'effet	5
2020	Das-Munshi et al., Lancet Psychiatry 2019	registres, multi-level Poisson, biais écologique, interprétation	5
2021	Srinivasan et al., Lancet Psychiatry 2020	outcome, exposition continue, médiation, génétique, modification d'effet	5
2022	Niederkrotenthaler et al., BMJ 2021	SARIMA, ITS, causalité, chocs exogènes, présentation des résultats	5
2023	Brynge et al., Lancet Psychiatry 2022	registres, modèles emboîtés, contrôle négatif, siblings	4
2024	Hasin et al., Lancet Psychiatry 2023	répétitions transversales, staggered DiD, échelle additive, limites	5
2025	Salvatore et al., AJP 2024	DAG, contrôle négatif, modèle additif, importance clinique	4
2026	Rosenström et al., Lancet Psychiatry 2025	design naturaliste, TMLE, causalité, biais, interprétation, éthique	35

2. Réflexe n°1 : identifier le bon type de question

2.1. Question de précision ou de puissance

Quand l'énoncé parle de puissance :

- soit il veut vraiment un **calcul grossier** à partir d'un `IC95%`;
- soit il veut surtout que vous expliquiez pourquoi la bonne discussion porte en réalité sur la **précision**, le **nombre d'événements** ou la **validité causale**.

2.2. Question de design

Beaucoup de questions EPI sont en réalité des questions de :

— schéma d'étude ;
— temporalité ;
— comparaison des groupes ;
— sélection des sujets ;
— facteurs de confusion ;
— validité interne et externe.

2.3. Question de causalité

Les indices classiques à examiner sont :

— temporalité ;
— ampleur et précision de l'effet ;
— cohérence externe ;
— dose-réponse ;
— plausibilité ;
— présence d'analyses négatives ou de comparaisons intrafamiliales ;
— persistance ou non après ajustement.

3. Les briques de base à connaître

3.1. Remonter à un SE à partir d'un IC95%

Sur une échelle linéaire :

$$SE \approx \frac{\text{borne sup} - \text{borne inf}}{3{,}92}$$

Pour un RR, 0R ou HR :

$$SE(\log \theta) \approx \frac{\log(\text{borne sup}) - \log(\text{borne inf})}{3{,}92}$$

3.2. Score de Wald et puissance observée grossière

Si l'on force un calcul de puissance a posteriori :

$$z \approx \frac{|\hat{\theta}|}{SE} \qquad \text{ou} \qquad z \approx \frac{|\log \hat{\theta}|}{SE(\log \theta)}$$

Puis :

$$\text{puissance} \approx \Phi(z - 1{,}96)$$

Cela reste un **calcul de coin de table**. En EPI, il faut presque toujours commenter en même temps la **largeur de l'IC95%** et la **pertinence clinique**.

3.3. Ce qui compte souvent plus que la taille totale de l'échantillon

— le nombre d'événements ;

— la fréquence de l'exposition ;
— la qualité des mesures ;
— la comparabilité des groupes ;
— le risque de confusion résiduelle ;
— la généralisabilité.

3.4. Rappels méthodologiques récurrents

`Personnes-années` somme des temps pendant lesquels chaque sujet contribue au risque.
`Biais écologique` erreur qui consiste à transposer au niveau individuel des associations observées sur des données agrégées.
`Interaction multiplicative` l'effet combiné diffère du produit des effets.
`Interaction additive` l'effet combiné diffère de la somme des effets absolus ; souvent plus pertinente en santé publique.
`Contrôle négatif` exposition ou outcome censé ne pas être causalement lié ; sert à détecter une confusion résiduelle.
`Sibling comparison` comparaison intra-fratrie pour neutraliser une partie des facteurs familiaux partagés.
`Repeated cross-sectional study` séries d'échantillons indépendants dans le temps ; bonne description des tendances, faible force pour la temporalité individuelle.
`Staggered difference-in-difference` différence-en-différences avec adoption de la politique à des dates différentes selon les unités.

4. Comment écrire une bonne réponse d'examen

Une bonne copie EPI suit presque toujours la même logique :

1. nommer correctement le **design** ;
2. dire ce que mesure réellement l'étude ;
3. identifier le **risque méthodologique principal** ;
4. commenter les résultats avec les **bons ordres de grandeur** ;
5. conclure sans surinterpréter.

Quand l'énoncé demande un calcul, une bonne copie doit dire :

1. quelle formule est utilisée ;
2. quelles données viennent de l'article ;
3. quelles données sont reconstruites ;
4. quel est le résultat numérique ;
5. ce qu'il faut vraiment en conclure.

5. Matrice de traçabilité des questions

Année	Fichier(s) original(aux) consulté(s)	Questions EPI repérées	Présence dans ce poly
2011	`Exam_2011_Epi.pdf` + `exam_2011_Epi_RC.d`	6	Oui, section 6.1

Année	Fichier(s) original(aux) consulté(s)	Questions EPI repérées	Présence dans ce poly
2012	`Exam_2012_Epi.pdf` + `exam_2012_Epi_RC.d`	6	Oui, section 6.2
2013	`Exam_2013_Epi.pdf` + `exam_2013_Epi_RC.d`	4	Oui, section 6.3
2014	`Exam_2014_Epi.pdf` + `exam_2014_Epi_RC.d`	4	Oui, section 6.4
2016	`exam_epi_20152016.` + `exam_2016_Epi_RC.d`	4	Oui, section 6.5
2018	`Exam_Epi_2018.pdf` + `exam_2018_Epi_RC.d`	5	Oui, section 6.6
2019	`M2MSR_Exam_Epi_201` + `exam_2019_Epi_RC_a`	5	Oui, section 6.7
2020	`Exam_Epi_2020_arti` + `exam_2020_Epi_RC_a`	5	Oui, section 6.8
2021	`Exam_Epi_2021.pdf` + `exam_2021_Epi_RC_a`	5	Oui, section 6.9
2022	`Sujet_exam_Epi_mar` + `exam_2022_Epi_RC_c`	5	Oui, section 6.10
2023	`Article_Exam_Epi_2` + `exam_2023_Epi_RC_c`	4	Oui, section 6.11
2024	`Article_Exam_Epi_2` + `exam_2024_Epi_RC_c`	5	Oui, section 6.12
2025	`Article_EPI2025.pd` + `exam_2025_Epi_RC.d`	4	Oui, section 6.13
2026	`Examen_EPI_M2MSR_2` `copie.pdf`	35	Oui, section 6.14

6. Annales résolues avec corrections détaillées

Dans chaque année, le bloc de révélation sert à pointer ce qu'il fallait aller chercher dans l'article ou dans le sujet. Les questions sont ensuite traitées **une par une**, sans fusionner plusieurs questions distinctes.

6.1. 2011 EPI

Article et contexte

Article support : **Kalager et al., NEJM 2010**, sur l'effet de la mammographie de dépistage sur la mortalité par cancer du sein en Norvège.

- quasi-expérience fondée sur le déploiement progressif du programme ;
- `40 075` femmes avec cancer du sein ;
- comparaison entre groupes contemporains et groupes historiques miroirs ;
- distinction essentielle entre **effet brut du programme** et **effet net attribuable spécifiquement au dépistage**.

Extrait de l'article à analyser

Question 1

> Quel est le principe de l'analyse principale ?

Réponse type

Le principe est une **logique de différence de différences** avant l'heure. Les auteurs ne comparent pas seulement un groupe dépisté à un groupe non dépisté ; ils comparent aussi chacun de ces groupes à son **groupe historique miroir**. Cela permet d'essayer de séparer l'effet propre du dépistage de l'amélioration générale des traitements et du diagnostic au cours du temps. La force de l'analyse vient de cette tentative d'isoler un effet net du dépistage, mais la faiblesse est que l'on reste dans un design quasi-expérimental, donc exposé aux biais temporels résiduels.

Question 2

> Les auteurs utilisent la notion de personne-année. Quelle définition en donneriez-vous et quel est l'intérêt de ce concept ?

Réponse type

Une personne-année correspond à **une personne suivie pendant un an au risque de l'événement**, ou à l'équivalent obtenu en additionnant des durées de suivi incomplètes. Par exemple, deux femmes suivies six mois chacune apportent une personne-année. L'intérêt est double : tenir compte de durées de suivi variables et permettre le calcul de **taux d'incidence** comparables entre groupes. Dans cet article, cet indicateur est particulièrement adapté parce que le suivi n'est pas parfaitement homogène dans le temps ni entre régions.

Question 3

> Dans la mesure du possible calculez une puissance a posteriori.

Résolution guidée

Si l'on raisonne sur l'effet **net** attribuable au dépistage seul, l'article donne environ `10%` avec `IC95% −4%` à `24%`.

$$SE \approx \frac{0{,}24 - (-0{,}04)}{3{,}92} = \frac{0{,}28}{3{,}92} \approx 0{,}071$$

$$z \approx \frac{0{,}10}{0{,}071} \approx 1{,}40$$

$$\text{puissance} \approx \Phi(1{,}40 - 1{,}96) = \Phi(-0{,}56) \approx 0{,}29$$

On obtient donc une puissance observée grossière d'environ **30%**.
Si l'on raisonne sur l'effet brut `RR = 0,72`, on trouve au contraire une puissance quasi maximale. La bonne copie doit expliquer cette différence.

Réponse type

La réponse dépend de l'effet choisi. Si l'on prend l'effet brut du groupe dépisté (`RR = 0,72`, `IC95% 0,63 à 0,81`), la puissance observée est très élevée. Mais l'effet vraiment intéressant est l'effet **net** attribuable au dépistage seul, estimé à environ `10%` avec `IC95% −4% à 24%`. Dans ce cas, on reconstruit un `SE` d'environ `0,071`, puis un score de Wald autour de `1,40`, soit une puissance observée grossière d'environ **30%**. La conclusion attendue est donc que l'étude est grande, mais qu'elle reste peu puissante pour isoler proprement le bénéfice spécifique du dépistage.

Question 4

Quelle(s) information(s) utile(s) apportent les femmes âgées de 70 ans et plus ?

Réponse type

Les femmes de `70` ans et plus sont utiles pour au moins deux raisons. D'abord, elles servent de **point d'appui temporel** pour juger ce qui relève des progrès généraux de prise en charge plutôt que du dépistage lui-même. Ensuite, elles permettent de penser à un bénéfice potentiellement **différé** du dépistage, puisque des cancers dépistés plus tôt peuvent influencer la mortalité à un âge ultérieur. En revanche, leur interprétation n'est pas aussi simple qu'un vrai groupe témoin, car leur histoire d'exposition au dépistage peut être hétérogène.

Question 5

Que pourrait-on rajouter d'utile à la figure 3 ?

Réponse type

La figure gagnerait à montrer explicitement les **intervalles de confiance**, les **effectifs** ou au moins les **nombres d'événements** derrière chaque courbe, ainsi qu'un rappel plus visible de la distinction entre comparaison brute et effet net. Des nombres à risque ou des taux absolus auraient aussi aidé à juger la précision réelle des différences observées. Autrement dit, il manque surtout de l'information sur la **stabilité statistique** et la **lecture causale** des courbes.

Question 6

Si vous deviez résumer les résultats par un seul nombre, quel serait-il ? Quelle est sa valeur inférentielle ? Qu'en pensez-vous ?

Réponse type

Si l'on veut résumer **l'effet spécifique du dépistage**, le meilleur chiffre unique est la **réduction relative nette d'environ 10%**, avec `IC95% −4% à 24%` et `p = 0,13`. On pourrait aussi citer l'effet brut `RR = 0,72`, mais ce serait moins fidèle à la question causale réelle car cet effet mélange dépistage et progrès globaux de prise en charge. La bonne conclusion est donc : le bénéfice propre du dépistage est suggéré mais reste **imprécis** et non significatif dans cette approche.

6.2. 2012 EPI

Article et contexte

Article support : **Cooper et al., NEJM 2011**, sur les médicaments du TDAH et les événements cardiovasculaires graves chez l'enfant et l'adulte jeune.

- cohorte rétrospective sur bases médico-administratives ;
- `1 200 438` sujets ;
- `2 579 104` personnes-années ;
- seulement 81 événements cardiovasculaires graves confirmés ;
- comparaison des périodes d'usage actuel, ancien usage et non-usage.

Extrait de l'article à analyser

Question 1

> Il est indiqué que "Exclusion criteria included a hospital discharge during the preceding 365 days with a primary diagnosis of acute myocardial infarction or stroke." Pouvez-vous justifier ce critère d'exclusion ?

Réponse type

Cette exclusion vise surtout à éviter d'inclure des sujets déjà très fragiles sur le plan cardiovasculaire, chez qui la prescription de stimulants serait atypique et chez qui le risque de nouvel événement est élevé d'emblée. Elle sert donc à améliorer la **comparabilité** et à limiter la **confusion par indication** ou par contre-indication. Elle permet aussi de mieux se placer dans une logique d'événements incidents plutôt que de récidives immédiates.

Question 2

> Expliquez ce que vous avez compris de la définition d'un sujet exposé et de celle d'un sujet non exposé.

Réponse type

L'exposition n'est pas définie une fois pour toutes à l'inclusion ; c'est une **exposition dépendante du temps**. Un sujet contribue du temps "exposé" pendant ses périodes d'utilisation actuelle du médicament, et du temps "non exposé" ou "ancien utilisateur" en dehors de ces périodes. Cette stratégie est bien meilleure qu'une classification binaire "a déjà été traité / n'a jamais été traité", car elle respecte la temporalité entre exposition et événement. La limite est qu'elle repose sur les données de prescription ou de délivrance, pas sur l'adhésion réelle au traitement.

Question 3

> Décrire le modèle statistique utilisé pour analyser le critère principal (et en particulier ce qui permet d'améliorer la comparabilité des exposés et des non exposés).

Réponse type

Le critère principal est analysé par un **modèle de Cox**, avec estimateur robuste de la variance. L'amélioration de la comparabilité ne repose pas seulement sur l'ajustement multivarié classique, mais aussi sur un **score de propension spécifique au site**, intégré dans le modèle, ainsi que sur des covariables

médicales et psychiatriques, certaines dépendantes du temps. L'idée est de rapprocher autant que possible exposés et non exposés sur leurs caractéristiques pronostiques, même si l'on reste dans une étude observationnelle non randomisée.

Question 4

> Dans l'analyse il y a un ajustement sur la variable "psychiatric condition". Est-ce que cet ajustement peut conduire à un surajustement ?

Réponse type

L'ajustement sur les antécédents psychiatriques est plutôt logique ici, car ils peuvent être liés à la probabilité d'être traité, au recours aux soins, à certains comportements à risque et à la qualité du codage. On est donc surtout dans une logique de **confusion par indication**. Le surajustement deviendrait un problème si la variable ajustée était un **médiateur** produit par le traitement ou une variable collidante. Ce n'est pas l'interprétation la plus naturelle ici ; l'ajustement paraît donc plutôt justifié, tout en restant discutable si le codage mélange histoire ancienne et état courant.

Question 5

> Une section "alternative analyses" est proposée. Citez les avantages et inconvénients de faire de telles analyses.

Réponse type

Les analyses alternatives ou de sensibilité sont utiles parce qu'elles testent la **robustesse** des résultats à plusieurs choix analytiques : définition de l'exposition, exclusions supplémentaires, sous-groupes, etc. Si elles convergent, elles renforcent la crédibilité du résultat principal. Leur limite est qu'elles multiplient les analyses, exposent au **cherry-picking** si elles ne sont pas bien cadrées, et peuvent brouiller le message si elles sont contradictoires. Il faut donc les lire comme un test de stabilité, pas comme une nouvelle pêche aux résultats significatifs.

Question 6

> Il est indiqué dans le texte que "The low number of events limited the statistical power of the study". Pouvez-vous retrouver au moins approximativement cette puissance statistique ?

Résolution guidée

L'article donne :

$$HR = 0{,}75 \qquad IC95\% = [0{,}31 ; 1{,}85]$$

$$SE(\log HR) \approx \frac{\log(1{,}85) - \log(0{,}31)}{3{,}92} \approx 0{,}456$$

$$z \approx \frac{|\log(0{,}75)|}{0{,}456} \approx 0{,}63$$

$$\text{puissance} \approx \Phi(0{,}63 - 1{,}96) = \Phi(-1{,}33) \approx 0{,}09$$

On retrouve donc une puissance observée très faible, de l'ordre de **10%**.

Réponse type

Malgré la taille énorme de la cohorte, l'information utile est faible parce que les événements cardiovasculaires graves sont très rares. À partir de HR = 0,75 et IC95% 0,31 à 1,85, on retrouve une erreur standard très grande et un score de Wald faible ($z \approx 0{,}63$). La puissance observée grossière est alors d'environ **10%**. La conclusion attendue est qu'une grande base de données n'est pas synonyme de grande puissance quand le nombre d'événements réellement informatifs est minuscule.

6.3. 2013 EPI

Article et contexte

Article support : **Lasalvia et al., Lancet 2013**, enquête transversale internationale sur la discrimination vécue par les personnes ayant un trouble dépressif majeur.

- 1082 participants ;
- 39 sites dans 35 pays ;
- instrument principal : DISC-12 ;
- outcome descriptif majeur : 855/1082, soit 79%, rapportent une discrimination vécue dans au moins un domaine.

Extrait de l'article à analyser

Question 1

> Décrivez la procédure d'échantillonnage des sujets inclus dans l'étude. Qu'en pensez-vous ?

Réponse type

L'échantillonnage est **pragmatique** et multicentrique, avec recrutement dans des services ou réseaux de soins dans 39 sites, chaque site devant inclure au moins un petit nombre minimal de participants. Ce n'est donc pas un échantillonnage aléatoire de toutes les personnes dépressives dans le monde. Ses forces sont la diversité internationale et la faisabilité ; ses limites sont un **biais de sélection** important, une représentativité discutable et une hétérogénéité potentiellement forte entre sites. La bonne conclusion est que l'étude décrit bien une population de patients recrutés dans des contextes de soins, mais pas "la dépression dans le monde" au sens strict.

Question 2

> Comment est construit le modèle multivarié expliquant la variable « experienced discrimination » ? Qu'en pensez-vous ?

Réponse type

Le papier utilise une **régression binomiale négative** parce que la variable à expliquer est un **score de comptage** de discrimination vécue, potentiellement surdispersé. Le modèle multivarié part d'une série de variables explicatives liées aux grandes questions de recherche : sexe, âge, épisodes dépressifs, anticipation de la discrimination, divulgation du diagnostic, emploi, traitement, etc. C'est un choix cohérent parce qu'un modèle linéaire serait peu adapté à un score entier asymétrique. En revanche, comme souvent dans les modèles multivariés de ce type, il faut garder en tête le risque de colinéarité, la difficulté d'interprétation causale et la dépendance à la qualité des variables mesurées.

Question 3

> Dans l'article on ne trouve pas de justification de la taille de l'échantillon. A posteriori, il n'est pas facile de calculer une puissance statistique. Essayez tout de même de proposer un tel calcul, même très rudimentaire.

Résolution guidée

Le calcul le plus simple consiste à raisonner en **précision** autour de la proportion principale :

$$p = \frac{855}{1082} \approx 0{,}79$$

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0{,}79 \times 0{,}21}{1082}} \approx 0{,}0124$$

$$IC95\% \approx 0{,}79 \pm 1{,}96 \times 0{,}0124$$

soit environ :

$$[0{,}766; 0{,}814]$$

La marge d'erreur est donc d'environ `± 2,4` points.

Réponse type

Le papier n'est pas construit comme un essai à hypothèse principale unique ; la bonne approche n'est donc pas un faux calcul de puissance sophistiqué mais une discussion de **précision**. Avec `855/1082`, soit `79%`, on obtient un `SE` d'environ `0,012` et un `IC95%` autour de `76,6%` à `81,4%`. Cela montre que l'étude décrit la proportion principale avec une précision correcte. Si l'on force malgré tout une discussion de puissance, on peut dire qu'elle est très élevée pour montrer que cette proportion est très différente de `50%`, mais ce n'est pas la vraie question scientifique du papier.

Question 4

> Dans cette étude il n'y a pas de groupe contrôle. Est-ce un problème selon vous ?

Réponse type

L'absence de groupe contrôle n'est pas forcément un défaut si l'objectif est **descriptif** : ici, le but principal est de documenter la fréquence et les formes de discrimination chez des patients dépressifs. En revanche, cette absence devient un problème dès que l'on veut faire une interprétation comparative ou causale. Sans groupe contrôle, on ne peut pas dire si la discrimination est plus élevée que dans une autre population ni attribuer certaines associations à la dépression elle-même. La bonne réponse consiste donc à dire : **pas gênant pour décrire, gênant pour comparer et expliquer causalement**.

6.4. 2014 EPI

Article et contexte

Article support : **Bao et al., NEJM 2013**, étude de cohorte sur consommation de noix et mortalité.

— deux grandes cohortes prospectives ;

- 118 962 participants ;
- 3 038 853 personnes-années ;
- 27 429 décès ;
- résultat marquant pour les plus gros consommateurs : HR ≈ 0,80, IC95% 0,73 à 0,86.

Extrait de l'article à analyser

Question 1

> Pourquoi les auteurs ont-ils exclus les sujets ayant des antécédents de cancer, de maladie cardiaque ou d'accident vasculaire cérébral ?

Réponse type

Cette exclusion sert principalement à limiter la **causalité inverse** et la confusion liée à une mauvaise santé préexistante. Des sujets déjà atteints de maladies graves peuvent avoir modifié leur alimentation, y compris leur consommation de noix, avant l'inclusion. Ils ont aussi un risque de décès beaucoup plus élevé d'emblée. Les exclure améliore donc l'interprétation prospective de l'association entre consommation de noix et mortalité.

Question 2

> Qu'est-ce qu'un facteur de confusion ? Citez-en qui pourraient être pertinents ici. Quelle stratégie les auteurs ont utilisée ici pour les gérer ?

Réponse type

Un facteur de confusion est une variable associée à la fois à l'exposition et à l'issue, sans être un intermédiaire causal. Ici, les confondeurs plausibles sont nombreux : tabac, activité physique, BMI, qualité globale de l'alimentation, niveau socio-économique, alcool, comorbidités, recours aux soins. Les auteurs essaient de les gérer par exclusions initiales, ajustements multivariés détaillés, mise à jour répétée des variables dans le temps et analyses de sensibilité. Cela améliore la crédibilité des résultats, mais ne supprime pas le risque de **confusion résiduelle**.

Question 3

> Dans l'article on ne trouve pas de justification de la taille de l'échantillon. A posteriori, il n'est pas facile de calculer une puissance statistique. Essayez tout de même de proposer un tel calcul, même très rudimentaire.

Résolution guidée

À partir de HR = 0,80 et IC95% 0,73 à 0,86 :

$$SE(\log HR) \approx \frac{\log(0{,}86) - \log(0{,}73)}{3{,}92} \approx 0{,}042$$

$$z \approx \frac{|\log(0{,}80)|}{0{,}042} \approx 5{,}3$$

$$\text{puissance} \approx \Phi(5{,}3 - 1{,}96) \approx \Phi(3{,}34) \approx 1$$

On conclut à une puissance observée pratiquement égale à 100%.

Réponse type

Si l'on force un calcul de puissance a posteriori, on retrouve une puissance quasi maximale. Le message essentiel est donc que cette étude n'est **pas limitée par la puissance**. Avec plus de 27 000 décès, le vrai problème méthodologique n'est pas de savoir si l'on peut détecter une association, mais si cette association est **causale** ou encore influencée par un biais de sujet sain ou une confusion résiduelle.

Question 4

> Dans la discussion, les auteurs évoquent à plusieurs reprises la nature causale ou non du lien qu'ils ont mis en évidence. Qu'en pensez-vous ?

Réponse type

Les arguments favorables à une interprétation causale existent : temporalité prospective, relation dose-réponse, cohérence biologique modérée et résultats robustes après ajustement. Mais ils ne suffisent pas à eux seuls. Il persiste un risque évident de **healthy user bias** : les gros consommateurs de noix peuvent aussi avoir, globalement, une meilleure hygiène de vie et un meilleur état de santé. La bonne position d'examen est donc nuancée : l'association est cohérente, potentiellement causale, mais une étude observationnelle de ce type ne peut pas démontrer définitivement la causalité.

6.5. 2016 EPI

Article et contexte

Article support : **Schnabel et al., Lancet 2015**, sur cinquante ans de tendances de fibrillation atriale dans Framingham.

- cohorte prospective sur 50 ans ;
- 9511 participants ;
- 202 417 personnes-années ;
- 1544 cas incidents ;
- analyses par décennies et par sexe.

Extrait de l'article à analyser

Question 1

> Il s'agit d'une étude de cohorte. Est-ce intéressant ici ? expliquez.

Réponse type

Oui, le design de cohorte est très pertinent ici car il permet de suivre l'apparition **incidentielle** de la fibrillation atriale, de relier les facteurs de risque mesurés avant l'événement à l'issue, et de décrire des **tendances temporelles** sur une longue durée. Une étude transversale n'aurait pas permis de distinguer incidence et prévalence, et une étude cas-témoins aurait été moins naturelle pour étudier l'évolution conjointe de l'incidence, des facteurs de risque et de la mortalité.

Question 2

> Quel est le paramètre présenté table 4 ? comment interpréter les résultats ?

Réponse type

Le tableau 4 présente des **risques attribuables populationnels** (*population-attributable risks*). Il ne s'agit pas d'un risque individuel ni d'un hazard ratio, mais d'une estimation de la part de l'incidence de fibrillation atriale attribuable, au niveau de la population, à chaque facteur de risque. Ce paramètre dépend à la fois de la **force de l'association** et de la **fréquence** du facteur dans la population. Il faut donc l'interpréter comme un indicateur de poids populationnel potentiel, pas comme une preuve qu'en supprimant entièrement un facteur on obtiendrait exactement cette baisse d'incidence.

Question 3

> Comment aborder la question de la puissance statistique dans une telle étude ?

Réponse type

Il ne faut pas chercher un NSN fictif. La bonne discussion porte sur le **nombre d'événements** et leur répartition. Globalement, 1544 cas incidents donnent une étude très informative. En revanche, lorsque l'on découpe ces données en cinq décennies puis par sexe, certaines strates deviennent petites, surtout dans les périodes anciennes, avec des erreurs standard plus larges. La bonne conclusion est donc : puissance suffisante au niveau global, mais puissance **hétérogène** et parfois limitée pour les analyses fines de tendance.

Question 4

> Comment ont été gérées les données manquantes ? Donnez votre point de vue sur la question.

Réponse type

Le papier ne met pas en avant une stratégie sophistiquée d'imputation multiple ; il repose plutôt sur les examens de référence disponibles et exclut certaines observations lorsqu'il manque un examen index éligible pour les analyses de facteurs de risque. Cette approche est simple mais revient essentiellement à une logique de **complete case analysis** sur certaines parties du papier. Elle est acceptable si les données manquantes sont peu nombreuses et peu informatives, mais elle peut introduire un biais si l'absence de mesures est liée à l'état de santé ou à la période historique.

6.6. 2018 EPI

Article et contexte

Article support : **Autier et al., BMJ 2017**, étude populationnelle sur efficacité et surdiagnostic du dépistage mammographique aux Pays-Bas.

- programme national de dépistage biennal ;
- analyse d'incidence par stades ;
- comparaison des groupes d'âge invités et non invités ;
- évaluation simultanée de la mortalité et du surdiagnostic.

Extrait de l'article à analyser

Question 1

> Quelle est la définition de l'overdiagnosis, expliquez (brièvement) la logique ?

Réponse type

Le surdiagnostic correspond à des cancers détectés par le dépistage qui **n'auraient jamais donné de symptômes ni de décès** pendant la vie de la personne si le dépistage n'avait pas existé. La logique empirique est la suivante : si le dépistage augmente fortement les cancers précoces sans faire baisser en parallèle les cancers avancés dans les mêmes proportions, l'excès de cas dépistés ne peut pas être interprété uniquement comme un gain d'avance diagnostique. Une partie de cet excès correspond alors probablement à du surdiagnostic.

Question 2

Dans cet article les auteurs utilisent des « joinpoint regressions », de quoi s'agit-il ?

Réponse type

Une joinpoint regression est un modèle de tendance par segments qui cherche des **points de rupture** dans l'évolution temporelle d'un taux. On ajuste plusieurs droites reliées entre elles et l'algorithme identifie les années où la pente change significativement. Ici, cela permet de voir si l'introduction ou l'extension du dépistage s'accompagne d'une rupture dans les tendances d'incidence ou de mortalité.

Question 3

Dans l'article il est question de modèle « âge-période-cohorte », de quoi s'agit-il ?

Réponse type

Un modèle âge-période-cohorte essaie de décomposer une tendance observée en trois composantes : effet de l'**âge** des femmes, effet de la **période calendaire** et effet de la **cohorte de naissance**. C'est utile quand on veut distinguer ce qui vient du vieillissement, des changements de contexte historique et des différences générationnelles. La difficulté classique est l'**identifiabilité** : âge, période et cohorte sont mathématiquement liés, donc l'interprétation dépend toujours d'hypothèses ou de contraintes supplémentaires.

Question 4

Décrivez la stratégie utilisée pour évaluer le surdiagnostic.

Réponse type

La stratégie consiste à comparer l'augmentation des cancers in situ et de stade précoce chez les femmes invitées au dépistage avec l'évolution des cancers avancés et avec les tendances dans les groupes d'âge non invités. Les auteurs retranchent ensuite ce qui peut être expliqué par l'**avance au diagnostic** (*lead time*) et par les tendances séculaires. Le surdiagnostic correspond au résidu persistant d'incidence excédentaire après ces corrections. C'est une stratégie indirecte, raisonnable à l'échelle populationnelle, mais dépendante des hypothèses retenues sur les tendances de fond.

Question 5

Comment des essais randomisés ont-ils pu conduire à une surestimation du bénéfice du dépistage alors qu'ils étaient randomisés ?

Réponse type

La randomisation protège contre la confusion de base, mais elle ne protège pas contre tous les biais. Des essais de dépistage peuvent surestimer le bénéfice si les traitements de référence deviennent obsolètes, s'il existe de la **contamination** du groupe contrôle, de la **non-adhésion** dans le groupe invité, une classification discutable de la cause du décès, ou un suivi insuffisant pour bien mesurer les dommages du surdiagnostic. En plus, beaucoup d'essais historiques ont été conduits dans des contextes thérapeutiques très différents d'aujourd'hui. Donc "randomisé" ne veut pas dire "impossible à surestimer".

6.7. 2019 EPI

Article et contexte

Article support : **Shan et al., BMJ 2018**, cohorte prospective sur travail de nuit tournant, mode de vie et risque de diabète de type 2 chez des infirmières américaines.

- `143 410` femmes ;
- suivi de `22` à `24` ans ;
- `10 915` cas incidents de diabète ;
- exposition fréquente et effet estimé avec précision.

Extrait de l'article à analyser

Question 1

> Dans une cohorte, l'attrition est un problème essentiel. Comment sont gérés ici les perdus de vue ?

Réponse type

Les perdus de vue sont essentiellement gérés comme des **censures** dans le cadre du modèle de Cox. C'est la stratégie standard, mais elle repose sur l'hypothèse que la censure est **non informative** conditionnellement aux variables du modèle. Le papier bénéficie d'un suivi de très bonne qualité, mais la bonne réponse d'examen doit rappeler que cette hypothèse n'est jamais complètement vérifiable et que l'attrition aurait mérité d'être davantage discutée dans l'article.

Question 2

> L'article étudie l'interaction entre deux expositions et ce sous deux aspects : interaction multiplicative, interaction additive. A quoi correspondent ces deux types d'interactions ?

Réponse type

L'interaction multiplicative est celle que l'on teste naturellement dans des modèles comme le Cox ou la régression logistique : on regarde si l'effet combiné s'écarte du **produit** des effets séparés. L'interaction additive regarde si le surplus de risque absolu associé à l'exposition conjointe dépasse la **somme** des surplus de risque individuels. En santé publique, l'échelle additive est souvent plus parlante, car elle répond mieux à la question : "combien de cas en plus sont liés à la combinaison des deux expositions ?"

Question 3

> Dans une analyse de sensibilité, il y a un ajustement sur l'hypertension et l'hyperlipidémie. Quels sont les avantages et inconvénients d'un tel ajustement ?

Réponse type

L'avantage est de tester la robustesse des résultats à un ajustement plus poussé, notamment si l'on pense que ces variables sont des marqueurs de risque cardiovasculaire ou de mode de vie global. Le problème est qu'elles peuvent aussi être **sur le chemin causal** entre exposition et diabète, ou être des variables intermédiaires liées au mode de vie. Les ajuster risque alors de conduire à un **surajustement** et à une atténuation artificielle de l'effet. C'est précisément pour cela qu'un tel ajustement est plus raisonnable en **analyse de sensibilité** qu'au cœur du modèle principal.

Question 4

Interprétez la taille d'effet des résultats trouvés.

Réponse type

L'effet du travail de nuit est **modeste** : `HR = 1,31` par tranche de cinq ans de travail tournant. Cela n'est pas négligeable, surtout si l'exposition est fréquente, mais on n'est pas face à un effet massif. À l'inverse, le score de mode de vie défavorable a un effet beaucoup plus fort, autour de `HR = 2,30` par incrément. La lecture correcte est donc double : il existe un signal compatible avec un rôle du travail de nuit, mais l'importance pratique semble plus marquée pour les facteurs de mode de vie, et l'interprétation causale du travail de nuit reste sensible à une confusion résiduelle.

Question 5

Dans une cohorte de cette taille, comment se discute la question de la puissance statistique ?

Réponse type

Dans une cohorte de cette ampleur, la puissance n'est plus le vrai sujet. Avec `143 410` participantes, une exposition fréquente, plus de `10 000` cas incidents et des `IC95%` étroits, l'étude est très largement capable de détecter des effets modestes. La discussion doit donc porter surtout sur la **taille d'effet**, la **pertinence clinique** et les **biais résiduels**, pas sur l'existence d'une p-value petite. En pratique, ici, on devrait presque regarder les `IC95%` avant de regarder les tests.

6.8. 2020 EPI

Article et contexte

Article support : **Das-Munshi et al., Lancet Psychiatry 2019**, sur densité ethnique, fragmentation sociale et mortalité chez les patients avec maladie mentale sévère dans un contexte urbain britannique.

- cohorte rétrospective de registres cliniques et administratifs ;
- patients avec **severe mental illness** ;
- modèle principal : **multi-level Poisson regression** ;
- question centrale : quel rôle jouent le quartier, la densité ethnique et les caractéristiques contextuelles sur la mortalité ?

Extrait de l'article à analyser

Question 1

Dans la cohorte étudiée, comment les sujets ont-ils été classés en cas (malade psychiatrique) et en contrôle (non malade psychiatrique) ? Ce classement est-il fiable ?

Réponse type

Le point important est qu'à proprement parler la cohorte ne contient pas un groupe témoin individuel recruté comme tel ; elle contient surtout des **patients avec maladie mentale sévère**, identifiés à partir des dossiers cliniques, des diagnostics `ICD-10` et d'outils d'extraction automatique dans les textes libres. La comparaison avec les "non malades" est essentiellement externe ou implicite, via les taux attendus dans la population générale. Le classement des cas paraît plutôt **spécifique** : les patients identifiés ont de fortes chances d'être de vrais cas. En revanche, la sensibilité peut être imparfaite et certains diagnostics peuvent être mal documentés.

Question 2

> Le modèle statistique utilisé dans l'article est un « multi-level Poisson regression » à quoi correspond ce type de modèle ?

Réponse type

Il s'agit d'une régression de Poisson adaptée à des **taux** ou des **comptages**, ici avec une structure **multiniveaux**. Cela signifie que le modèle combine des variables individuelles et des variables de quartier, et tient compte du fait que plusieurs sujets partagent un même contexte géographique. En pratique, cela permet d'estimer des rapports de taux ajustés tout en modélisant l'hétérogénéité entre aires géographiques par un effet aléatoire ou une structure hiérarchique.

Question 3

> Contrairement aux résultats trouvés dans l'article, d'anciennes études ont mis en évidence la présence d'une association entre la fragmentation sociale et la mortalité. Dans la discussion les auteurs suggèrent que ces études étaient possiblement entachées d'un biais écologique. Qu'est-ce qu'un biais écologique ?

Réponse type

Le biais écologique apparaît lorsqu'on attribue au niveau **individuel** une association observée sur des données **agrégées**. Par exemple, un quartier très fragmenté peut avoir un taux de mortalité élevé sans que l'on puisse conclure que chaque personne isolée du quartier a, individuellement, ce sur-risque pour cette raison. Le papier `2020` essaie justement de réduire ce problème en combinant des données individuelles de patients avec des indicateurs contextuels de quartier, plutôt que de raisonner uniquement à l'échelle agrégée.

Question 4

> Quel est le résultat principal de l'étude ? Comment l'interpréter ?

Réponse type

Le résultat principal est que, chez les patients ayant une maladie mentale sévère, vivre dans une zone à forte densité ethnique semble être associé à une **mortalité plus faible pour les minorités ethniques** relativement au groupe blanc britannique, alors qu'urbanicité, fragmentation sociale et privation n'émergent pas comme déterminants robustes de mortalité après ajustement. Il faut interpréter cela comme un **effet de contexte potentiel** ou un rôle protecteur de l'environnement communautaire, mais pas comme une preuve causale définitive : des différences non mesurées de soutien social, de recours aux soins ou de sévérité clinique peuvent encore expliquer une partie du signal.

Question 5

> Une étude qualitative pourrait être réalisée pour approfondir les résultats. Quel protocole proposeriez-vous en quelques lignes ?

Réponse type

On peut proposer une étude qualitative par **entretiens semi-structurés** ou **théorisation ancrée**, avec échantillonnage raisonné de patients SMI issus de quartiers contrastés sur la densité ethnique et la fragmentation sociale. L'idée serait d'explorer le soutien social, le vécu communautaire, les discriminations, l'accès aux soins somatiques et psychiatriques, et les trajectoires de suivi. Une vingtaine à une trentaine d'entretiens, analysés de façon thématique par au moins deux chercheurs, permettraient de tester des mécanismes plausibles derrière l'association observée.

6.9. 2021 EPI

Article et contexte

Article support : **Srinivasan et al., Lancet Psychiatry 2020**, étude longitudinale sur symptômes dépressifs périnataux maternels et expériences psychotiques à 18 ans chez l'enfant.

- cohorte de naissance ALSPAC ;
- exposition : score EPDS anténatal et postnatal ;
- outcome principal : expériences psychotiques à 18 ans ;
- enjeu central : confusion, médiation, génétique et temporalité.

Extrait de l'article à analyser

Question 1

> Dans une étude d'épidémiologie analytique il y a généralement une « variable à expliquer ». Qu'elle est-elle ici ? Comment est-elle mesurée ?

Réponse type

La variable à expliquer principale est la survenue d'**expériences psychotiques** chez l'enfant à l'âge de 18 ans. Elle est mesurée par le **PLIKSi** (*Psychosis-Like Symptom Interview*), c'est-à-dire un entretien semi-structuré bien plus solide qu'un simple auto-questionnaire. C'est un point fort important du papier, car l'outcome n'est pas seulement auto-déclaré mais recueilli avec une procédure clinique plus rigoureuse.

Question 2

> Concernant la variable d'exposition d'intérêt … que pensez-vous de la stratégie proposée ici par les auteurs ?

Réponse type

La stratégie est globalement bonne. Utiliser le score EPDS en **continu** permet de conserver l'information et d'éviter la perte de puissance liée à une dichotomisation arbitraire. Vérifier ensuite les résultats avec un seuil clinique (> 12) est utile pour la lisibilité clinique et la robustesse. La limite est qu'une dichotomisation reste méthodologiquement moins élégante et qu'un effet éventuellement non linéaire aurait pu aussi être exploré par splines ou catégories ordonnées.

Question 3

> Concernant la construction du modèle … Justifiez cette stratégie de sélection de variables.

Réponse type

La stratégie est **theory-driven** : on ajuste sur des variables connues pour être associées à la fois à la dépression périnatale et au risque d'expériences psychotiques, mais on évite d'ajuster sur des facteurs susceptibles d'être des **médiateurs** situés après la naissance, comme certaines expériences d'abus ou de harcèlement. C'est exactement la logique d'un ajustement causal raisonné. Le principal risque serait d'oublier un vrai confondeur important ou de mal classer une variable comme médiateur alors qu'elle aurait aussi un rôle de confusion.

Question 4

> Discutez l'existence et la prise en compte dans l'article de possibles facteurs de confusion génétiques.

Réponse type

La question génétique est centrale : une vulnérabilité familiale pourrait expliquer à la fois la dépression maternelle et le risque ultérieur de symptômes psychotiques chez l'enfant. Les auteurs essaient de la prendre en compte par les antécédents familiaux et par certaines analyses de sensibilité incluant le risque génétique de schizophrénie. C'est utile, mais probablement **incomplet**. La bonne réponse est donc de reconnaître l'effort méthodologique tout en disant que la confusion génétique résiduelle reste une limite sérieuse du papier.

Question 5

> Qu'entendent les auteurs par « effect modifier » ? Comment sont-ils recherchés statistiquement ?

Réponse type

Un **effect modifier** est une variable qui modifie la force ou parfois la direction de l'association entre exposition et outcome. Ici, l'idée est de savoir si l'effet de la dépression anténatale maternelle sur les expériences psychotiques de l'enfant dépend de l'existence d'épisodes dépressifs maternels ultérieurs. Statistiquement, cela se recherche classiquement par un **terme d'interaction** dans le modèle, ou par des analyses stratifiées si l'on veut présenter les résultats plus simplement.

6.10. 2022 EPI

Article et contexte

Article support : **Niederkrotenthaler et al., BMJ 2021**, sur l'association entre la chanson de Logic "1-800-273-8255", les appels à la Lifeline et les suicides aux États-Unis.

- série temporelle interrompue ;
- données écologiques agrégées ;
- modèle principal : `SARIMA` ;
- enjeu central : validité causale d'un signal médiatique en santé publique.

Extrait de l'article à analyser

Question 1

> Expliquez brièvement ce qu'est un "Seasonal autoregressive integrated moving average models".

Réponse type

Un modèle SARIMA est un modèle de séries temporelles qui combine quatre briques : une composante **autorégressive** (AR), une composante de **moyenne mobile** (MA), une éventuelle **différenciation** pour rendre la série plus stationnaire (I), et une composante **saisonnière** (S). Il sert à capter les tendances, les autocorrélations et les variations périodiques d'une série avant d'évaluer l'effet d'une intervention ou d'un choc exogène.

Question 2

> Une section de l'article est intitulée « Possible confounding exogenous events ». Les facteurs discutés dans cette section sont-ils des facteurs de confusion ?

Réponse type

Au sens strict, ce ne sont pas toujours des facteurs de confusion au sens individuel classique. Dans une série temporelle interrompue, il s'agit plutôt d'**événements concomitants** ou de co-interventions susceptibles d'expliquer une rupture de la série au même moment que l'exposition d'intérêt. La bonne idée à exprimer est donc : ce sont des facteurs qui peuvent **biaiser l'attribution causale** de la rupture observée, même si le mot "confusion" est ici utilisé dans un sens un peu élargi.

Question 3

> Dans la partie supérieure de la figure 1 il y a de petites flèches noires. A quoi correspondent-elles ? Comment ont-elles été déterminées ?

Réponse type

Ces flèches matérialisent les **périodes de forte exposition médiatique** ou les **changepoints** identifiés autour de la chanson, de ses performances publiques et du bruit médiatique associé. Elles servent à délimiter les fenêtres où un effet sur les appels ou les suicides est plausible. Elles sont déterminées à partir des données de diffusion médiatique, notamment l'activité sur les réseaux sociaux, avec un appui d'algorithmes de détection de ruptures et d'une lecture du contexte.

Question 4

> Dans la discussion, les auteurs indiquent que « The observational nature of this research does not allow us to establish causality ». Qu'en pensez-vous ?

Réponse type

La phrase est juste et méthodologiquement saine. Une ITS est un design puissant pour renforcer une inférence causale quand il existe une rupture nette, un bon calage temporel et peu d'interventions concurrentes, mais ici on reste dans des données **écologiques** et **observationnelles**. La plausibilité psychologique et le calendrier sont intéressants, mais ils ne suffisent pas à "prouver" la causalité. La bonne position d'examen est donc de dire : il existe des arguments en faveur d'un effet causal, mais pas une démonstration définitive.

Question 5

> A propos de la baisse numérique du nombre de suicides, discutez la façon dont les auteurs présentent leurs résultats.

Réponse type

Les auteurs ont raison de mentionner la **baisse numérique**, mais ils doivent le faire avec prudence. Les intervalles de confiance restent larges et la robustesse statistique est limitée. Il faut donc éviter un ton triomphaliste. En revanche, on peut aussi reconnaître qu'en santé publique, un signal même modeste et incertain peut rester intéressant s'il est peu coûteux et potentiellement bénéfique. La bonne réponse équilibre ces deux idées : prudence statistique, mais pas indifférence clinique ou sociétale.

6.11. 2023 EPI

Article et contexte

Article support : **Brynge et al., Lancet Psychiatry 2022**, cohorte suédoise sur infection maternelle pendant la grossesse, autisme et déficience intellectuelle, avec analyses de contrôle négatif et intra-fratrie.

- registres nationaux suédois ;
- très grand effectif ;
- outcome : autisme et déficience intellectuelle ;
- stratégie forte de triangulation : modèles emboîtés, contrôle négatif, sibling comparison.

> **Numérotation du sujet**
>
> Dans le fichier source consulté, la partie EPI `2023` comporte **quatre questions** numérotées `1`, `3`, `4` et `5`. Aucune question `2` n'apparaît dans la version fournie du sujet.

Extrait de l'article à analyser

Question 1

> Les données de cette étude proviennent d'un registre, quels sont les avantages et les inconvénients d'un tel type de données.

Réponse type

Les registres offrent de grands effectifs, une forte puissance descriptive, un coût marginal faible et une excellente profondeur temporelle. Ils sont très adaptés pour étudier des événements rares ou des sous-groupes. Leur limite majeure est que les variables n'ont pas été recueillies spécifiquement pour la question posée : qualité imparfaite du codage, manque de certaines variables essentielles, exposition parfois grossière, et risque de confusion résiduelle. La bonne réponse doit donc être équilibrée : **très forte faisabilité et puissance, mais qualité de mesure pas toujours optimale**.

Question 3

> Dans leur analyse, les auteurs utilisent une série de 3 modèles emboités. Quel est l'intérêt d'une telle stratégie ?

Réponse type

Les modèles emboîtés permettent de montrer comment l'association évolue quand on ajoute progressivement des covariables. Ce n'est pas seulement une coquetterie statistique : cela permet de voir si l'effet brut disparaît, s'atténue ou persiste, donc de raconter une **histoire analytique**. En examen, il faut dire que cette stratégie aide à juger la robustesse de l'association et l'importance des ajustements, même si elle ne prouve pas à elle seule la causalité.

Question 4

> En épidémiologie les possibles facteurs de confusion non mesurés sont une hantise. Quelle stratégie les auteurs ont-ils utilisés pour contourner cette difficulté (compte double) ?

Réponse type

Les auteurs utilisent deux stratégies élégantes et complémentaires. Premièrement, un **contrôle négatif d'exposition** : l'infection maternelle dans l'année précédant la grossesse, qui partage beaucoup de facteurs de confusion potentiels mais ne devrait pas avoir le même mécanisme causal direct. Deuxièmement, une **comparaison intra-fratrie** (*sibling comparison*), qui réduit une partie des facteurs familiaux partagés, génétiques ou environnementaux. Cette "double" stratégie ne supprime pas toute confusion, mais elle renforce beaucoup la crédibilité de l'interprétation.

Question 5

> Les auteurs construisent leur conclusion à partir d'une mise en contraste de résultats positifs … avec des résultats négatifs … Qu'en pensez-vous ?

Réponse type

Cette façon de conclure est intéressante parce qu'elle s'appuie sur une logique de **triangulation** : un signal positif gagne en crédibilité si des analyses voisines censées être nulles restent effectivement faibles ou nulles. Mais il faut rester prudent : un "résultat négatif" n'est pas synonyme de "preuve d'absence d'effet", surtout si la précision ou la puissance de cette analyse secondaire sont limitées. La bonne réponse d'examen consiste donc à saluer l'élégance du raisonnement tout en rappelant que les résultats négatifs doivent eux aussi être lus à la lumière de leurs `IC95%`.

6.12. 2024 EPI

Article et contexte

Article support : **Hasin et al., Lancet Psychiatry 2023**, étude transversale répétée sur douleur chronique, légalisation du cannabis et trouble de l'usage du cannabis chez les vétérans américains.

- 15 vagues transversales répétées ;
- système de santé des vétérans ;
- design quasi-expérimental de type **staggered DiD** ;
- outcome : prévalence du trouble de l'usage du cannabis.

Extrait de l'article à analyser

Question 1

> Il s'agit d'une étude transversale répétée. Quels sont les avantages et inconvénients par rapport à une étude de cohorte ?

Réponse type

Une étude transversale répétée est très utile pour décrire des **tendances de prévalence** dans le temps, avec un coût souvent plus faible qu'une cohorte longue et sans problème d'attrition individuelle. En revanche, elle ne suit pas les mêmes personnes au fil du temps ; elle donne donc une force bien plus faible pour la **temporalité individuelle** et les mécanismes causaux. On peut suivre l'évolution d'un phénomène collectif, mais pas raconter des trajectoires individuelles.

Question 2

> Le design de l'étude … relève d'une approche dite de "staggered difference-in-difference model". De quoi s'agit-il ?

Réponse type

Le principe d'une différence-en-différences est de comparer l'évolution du critère dans les unités exposées à une politique avec l'évolution du même critère dans des unités non encore exposées. Le qualificatif **staggered** signifie que les États n'adoptent pas la loi tous au même moment. Le modèle exploite donc cette adoption échelonnée dans le temps. L'hypothèse clé est celle des **tendances parallèles** en l'absence d'intervention ; si elle est violée, l'interprétation causale devient fragile.

Question 3

> Dans l'article, les auteurs indiquent "our model was on the additive, not multiplicative scale …". Expliquez ce qu'ils veulent dire.

Réponse type

Ils veulent dire qu'ils modélisent directement des **différences absolues de prévalence** plutôt que des rapports de cotes, des rapports de risques ou des log-risques. Sur une échelle additive, un coefficient se lit comme un surplus absolu de cas, ce qui est souvent plus parlant en santé publique. L'avantage est l'interprétation directe. La difficulté est que ces modèles peuvent être plus délicats statistiquement et moins standard que les modèles multiplicatifs habituels.

Question 4

> Que pensez-vous de la conclusion de l'article telle que présentée dans l'abstract ?

Réponse type

La conclusion de l'abstract est factuellement défendable, mais elle peut paraître un peu **péremptoire** au regard de la complexité du design, de l'hétérogénéité entre États et des limites de généralisabilité. Le bon commentaire d'examen est de dire qu'il existe bien un signal moyen, mais qu'il reste **modeste**, hétérogène et dépendant d'hypothèses fortes. L'abstract simplifie donc un peu trop un résultat plus nuancé quand on lit le papier en détail.

Question 5

> Quelle est, selon vous, la principale limite méthodologique de l'étude ?

Réponse type

Deux limites fortes sont recevables. La première est la **validité externe** : la population étudiée est celle des vétérans américains, majoritairement masculine et très spécifique. La seconde est la **validité du**

staggered DiD en présence d'hétérogénéité des effets selon les États et de possibles violations des tendances parallèles. Si je devais en choisir une seule, je dirais que la plus structurante est la fragilité de l'interprétation causale dans un contexte d'adoption législative hétérogène.

6.13. 2025 EPI

Article et contexte

Article support : **Salvatore et al., American Journal of Psychiatry 2024**, sur les effets sociaux génétiques des pairs sur le risque de troubles liés à l'usage de substances, dépression majeure et trouble anxieux.

- très grand échantillon national suédois ;
- expositions : scores génétiques agrégés des pairs (*peer FGRS*) ;
- outcome : enregistrements de troubles psychiatriques et addictifs ;
- modèles `Cox`, analyses de contrôle négatif, modèles additifs pour interactions.

Extrait de l'article à analyser

Question 1

> Proposez et justifiez un Directed Acyclic Graph (DAG) du modèle A.

Réponse type

Le DAG minimal doit faire apparaître : `peer FGRS –> environnement social des pairs –> risque du sujet`, mais aussi `FGRS propre du sujet –> risque du sujet`, `caractéristiques familiales/SES –> composition du groupe de pairs` et `caractéristiques familiales/SES –> risque du sujet`. Il faut aussi intégrer le **contexte scolaire** ou de filière, qui influence à la fois la structure des pairs et le risque ultérieur. L'intérêt du DAG est de montrer qu'un effet des pairs est plausible, mais qu'il existe plusieurs chemins de confusion, notamment par sélection sociale et génétique dans les groupes.

Question 2

> Les auteurs indiquent que « To evaluate potential bias, we also conducted a negative control analysis ». Expliquez en quoi consiste et quel est le rôle de cette analyse.

Réponse type

Une analyse de contrôle négatif consiste à tester une association censée être **non causale** afin de détecter une confusion résiduelle ou une structure de biais. Ici, les auteurs utilisent la **taille** comme outcome de contrôle négatif dans un sous-échantillon d'hommes. Si le `peer FGRS` présentait une association analogue avec cette variable biologiquement peu plausible dans ce contexte social, cela affaiblirait l'interprétation causale des effets observés sur les troubles psychiatriques. C'est donc un test de crédibilité du raisonnement causal.

Question 3

> Les auteurs indiquent également « To focus on interactions on an additive scale, we used a linear probability model ». Qu'est-ce que cela signifie et pourquoi prennent-ils cette option ?

Réponse type

Cela signifie qu'ils modélisent directement une **différence absolue de probabilité** d'issue selon les niveaux croisés de `peer FGRS` et de `FGRS` individuel, au lieu de modéliser un log-risque ou un logit. Ce choix facilite l'étude des **interactions additives**, souvent plus parlantes en santé publique : on peut quantifier un excès absolu de risque dû à la combinaison des susceptibilités. L'inconvénient est que le modèle de probabilité linéaire est moins standard et peut être statistiquement moins confortable que les modèles multiplicatifs usuels.

Question 4

> Dans une étude épidémiologique, l'importance clinique ou de santé publique des effets mis en évidence est au moins aussi importante que la significativité statistique de ces derniers. Interprétez cette taille d'effet pour les principaux résultats rapportés dans l'article.

Réponse type

Le papier montre des effets **petits à modérés**. Pour les pairs définis géographiquement, les HR sont très proches de 1, donc probablement modestes en pratique. Pour certaines définitions scolaires des pairs, les effets montent davantage et deviennent plus intéressants, surtout pour les troubles liés à l'usage de substances. La bonne lecture est donc : les résultats sont statistiquement convaincants dans un très grand échantillon, mais leur importance clinique individuelle est souvent limitée ; leur intérêt se situe davantage au niveau **populationnel** et mécanistique.

6.14. 2026 EPI

Article et contexte

Article support : **Rosenström et al., Lancet Psychiatry 2025**, comparaison naturaliste finlandaise entre **internet-delivered CBT** guidée (`iCBT`) et **face-to-face CBT** (`fCBT`) pour la dépression.

- cohorte rétrospective sur registres de soins ;
- `5834` patients analysés, dont environ `5446` iCBT et `388` fCBT ;
- outcome principal : changement de score `PHQ-9` ;
- estimateur principal : **ATE par TMLE** ;
- estimation rapportée : `ATE = 0,745`, `IC95% 0,156 à 1,334` ;
- point critique : étude observationnelle non pré-enregistrée, très riche analytiquement, mais sans randomisation.

> **i** Source utilisée pour cette section
>
> Le sujet `2026 EPI` disponible dans le dossier est déjà un **sujet exhaustif corrigé** organisé en `35` questions. Cette section réintègre ces `35` questions dans le format du poly, avec des réponses reformulées pour rester utilisables comme support de révision.

Extrait de l'article à analyser

Thème 1. Design de l'étude

Question 1

> Il s'agit d'une cohorte rétrospective utilisant des registres naturalistes. Quels sont les avantages et les inconvénients de ce type de données par rapport à un essai contrôlé randomisé ?

Réponse type

Les avantages sont la **grande taille d'échantillon**, la validité externe élevée, le faible coût marginal et la possibilité d'étudier l'efficacité en conditions réelles. Les inconvénients sont l'absence de randomisation, donc la **confusion résiduelle**, le manque de certaines variables clés comme l'affinité numérique ou certaines préférences thérapeutiques, et le fait que les groupes reçoivent en réalité des "paquets" de soins différents. Une bonne copie doit opposer `effectiveness` naturaliste et `efficacy` expérimentale.

Question 2

> Expliquez le flowchart : comment les 32 343 entrées ont-elles été réduites à 5 834 patients analysés ? Que révèle ce processus sur la représentativité de l'échantillon ?

Réponse type

Le flowchart élimine successivement les sujets sans bonne indication diagnostique, sans PHQ–9 de base, ayant reçu les deux modalités, ou commencés trop tard pour permettre un suivi minimal. Cela conduit à une très forte réduction, surtout dans le bras `iCBT`. Le message méthodologique est que l'échantillon final représente **les patients traités avec données utilisables**, pas tous les patients initialement orientés vers ces prises en charge. On garde donc une bonne validité clinique, mais au prix d'un possible biais de sélection.

Question 3

> Expliquez la distinction entre "efficacy" et "effectiveness" telle qu'elle est utilisée dans cet article. En quoi cette étude apporte-t-elle des preuves sur les deux ?

Réponse type

L'`effectiveness` est l'efficacité en conditions réelles de soin ; l'`efficacy` renvoie à l'effet dans des conditions idéalement comparables. Les données naturalistes brutes documentent surtout l'`effectiveness`. Le recours au TMLE permet ensuite d'approcher une comparaison contrefactuelle plus proche d'une logique d'`efficacy`, sans atteindre toutefois le niveau de preuve d'un essai randomisé. La bonne réponse est donc : **oui pour effectiveness, partiellement pour efficacy, mais pas équivalent à un ECR**.

Question 4

> L'étude n'a pas été pré-enregistrée. Quelles sont les implications méthodologiques de ce choix et comment les auteurs tentent-ils d'y remédier ?

Réponse type

Le non-pré-enregistrement expose au p–hacking, au HARKing et à une plus grande liberté dans les choix analytiques a posteriori. Les auteurs tentent de limiter ce problème par la transparence du code, la multiplicité d'analyses de sensibilité et la cohérence globale des résultats obtenus par des méthodes différentes. Cela atténue la critique, mais ne l'efface pas totalement.

Question 5

> Quelles sont les différences entre le groupe iCBT et le groupe fCBT au niveau du traitement lui-même ? En quoi ces différences compliquent-elles l'interprétation des résultats ?

Réponse type

L'étude ne compare pas seulement "internet" contre "présentiel". Elle compare aussi des différences de **standardisation**, de **durée**, d'**intensité de contact**, de **type de thérapeute** et probablement de sélection des patients. Cela complique l'interprétation car l'effet attribué au mode de délivrance peut en réalité refléter un mélange de différences organisationnelles, cliniques et thérapeutiques. La bonne conclusion est qu'on compare des **modalités de prise en charge réelles**, pas un facteur unique parfaitement isolé.

Thème 2. Méthodes statistiques

Question 6

> Expliquez ce qu'est le TMLE et pourquoi il est supérieur à une régression multivariée classique ou à un score de propension simple pour estimer l'ATE.

Réponse type

Le `TMLE` (*Targeted Maximum Likelihood Estimator*) combine un modèle d'outcome et un modèle d'assignation au traitement, puis "cible" l'estimation vers le paramètre causal d'intérêt, ici l'`ATE`. Sa force est sa **double robustesse** et son aptitude à travailler avec des outils de machine learning sans perdre de vue le paramètre causal. Par rapport à une simple régression ajustée ou à un score de propension isolé, il est moins dépendant d'une seule spécification de modèle et exploite mieux l'information disponible, sous réserve que les hypothèses causales restent plausibles.

Question 7

> Qu'est-ce qu'un Super Learner ? Expliquez le principe de validation croisée utilisé et son intérêt.

Réponse type

Un `Super Learner` est un **ensemble learning** : plusieurs algorithmes candidats sont mis en compétition et combinés selon leurs performances prédictives en validation croisée. L'intérêt est d'éviter de parier sur un seul modèle potentiellement mal spécifié. La validation croisée protège contre le surapprentissage et permet de sélectionner ou pondérer les apprenants de façon plus robuste qu'un choix arbitraire unique.

Question 8

> Dans l'article, les auteurs mentionnent des analyses de sensibilité incluant G-computation, AIPTW, simplification du TMLE et restriction à Uusimaa. Quel est l'intérêt de cette stratégie ?

Réponse type

L'intérêt est de vérifier que le résultat principal n'est pas l'artefact d'une seule méthode ou d'une seule sous-population. Si `G-computation`, `AIPTW`, versions simplifiées du `TMLE` et analyses restreintes conduisent toutes à un message cohérent, la crédibilité du résultat augmente. Cela ne prouve pas la causalité, mais cela renforce la **robustesse analytique**.

Question 9

> Les auteurs utilisent la validation croisée imbriquée (nested cross-validation). Pourquoi est-ce important ?

Réponse type

La validation croisée imbriquée sépare mieux la phase de **choix/tuning** des modèles et la phase d'**évaluation**. Elle réduit le risque d'optimisme artificiel dû à la réutilisation des mêmes données pour sélectionner et juger un apprenant. Dans un contexte de machine learning causal, c'est un garde-fou important contre le surajustement et la "bonne surprise" purement algorithmique.

Question 10

> Comment l'imputation multiple (superMICE) a-t-elle été utilisée dans cette étude ? Quels sont ses intérêts et ses limites ?

Réponse type

L'imputation multiple sert ici à explorer des scénarios plus favorables que la règle principale, qui assimilait certains patients sans seconde mesure à une absence d'amélioration. `superMICE` combine l'esprit des équations en chaîne avec des outils de prédiction plus souples. L'intérêt est de tester la sensibilité des résultats à l'hypothèse faite sur les données manquantes. La limite est que tout cela repose encore sur une hypothèse de type **MAR** et qu'aucune imputation ne peut réparer des données manquantes **non aléatoirement** liées à l'engagement thérapeutique.

Thème 3. Inférence causale et DAG

Question 11

> Proposez et justifiez un DAG représentant les relations causales plausibles dans l'article.

Réponse type

Le DAG doit au minimum contenir : âge, `sexe`, `gravité initiale`, `comorbidités`, `région`, `statut social`, `affinité numérique` et `type de thérapeute` comme causes plausibles du **choix de traitement** et du **résultat final**. Le traitement influence ensuite le changement de `PHQ-9`, avec le `dropout` comme variable post-traitement intermédiaire possible. Le vrai intérêt du DAG est de rendre visible la masse de **chemins de confusion** que le `TMLE` essaie de bloquer, sans garantir qu'ils le soient tous.

Question 12

> Expliquez les hypothèses causales nécessaires au TMLE dans ce contexte : consistance, échangeabilité, positivité.

Réponse type

La **consistance** suppose que le traitement observé correspond bien à l'intervention étudiée et que "recevoir iCBT" ou "recevoir fCBT" a un sens stable. L'**échangeabilité** suppose qu'après ajustement il n'existe plus de confondeurs non mesurés. La **positivité** impose qu'à profils comparables, chaque patient ait une probabilité non nulle de recevoir chacune des deux options. Dans ce papier, la positivité est discutable à cause du fort déséquilibre des groupes et des contraintes régionales ; l'échangeabilité est la grande hypothèse fragile.

Question 13

> Qu'est-ce qu'une analyse de contrôle négatif ? Comment aurait-elle pu être utile ici ?

Réponse type

Une analyse de contrôle négatif consiste à choisir une exposition ou un outcome qui ne devrait pas être causalement lié au mécanisme étudié, pour tester la présence d'un biais caché. Le papier n'en fait pas un usage central, mais un contrôle négatif aurait pu être utile pour vérifier si certaines différences entre `iCBT` et `fCBT` reflètent seulement une sélection sociale ou organisationnelle. L'intérêt théorique est de détecter une **confusion résiduelle** que même un modèle causal sophistiqué ne voit pas.

Question 14

> Qu'entend-on par « triangulation causale » dans cet article et dans la littérature ?

Réponse type

La triangulation causale consiste à confronter plusieurs approches imparfaites, mais biaisées différemment, pour voir si elles convergent vers le même message. Ici, on peut parler de triangulation entre données naturalistes, méthodes causales modernes (`TMLE`), analyses de sensibilité, et résultats antérieurs d'essais ou de littérature comparative. L'idée n'est pas qu'une seule méthode donne la vérité, mais qu'un faisceau d'indices convergents est plus convaincant qu'une analyse isolée.

Question 15

> La question contrefactuelle posée dans l'article est : « quel serait l'ATE si on pouvait allouer tous les patients à iCBT ou tous à fCBT ? ». Que faut-il comprendre exactement ?

Réponse type

Il faut comprendre qu'on ne compare pas seulement deux groupes observés tels quels. On cherche à estimer la différence moyenne qu'on aurait obtenue **si les mêmes patients**, en moyenne, avaient tous reçu `iCBT` puis tous reçu `fCBT`. C'est donc une question **contrefactuelle** classique d'inférence causale, bien plus ambitieuse qu'une simple comparaison brute des moyennes observées.

Thème 4. Résultats et taille d'effet

Question 16

> L'ATE estimé est de 0,745 point PHQ-9 (IC95% : 0,156–1,334). Interprétez ce résultat sous l'angle statistique et clinique.

Réponse type

Statistiquement, l'intervalle de confiance exclut `0`, donc il existe un signal compatible avec une meilleure réduction symptomatique sous `iCBT` que sous `fCBT` selon la convention de signe du papier. Cliniquement, l'effet reste **modeste** : moins d'un point de `PHQ`−9 en moyenne. La bonne interprétation est donc : résultat statistiquement crédible, mais bénéfice clinique moyen probablement limité à l'échelle individuelle.

Question 17

> Expliquez la différence entre l'ATE estimé (0,745) et la différence observée brute (1,120).

Réponse type

La différence brute mélange l'effet du traitement et les différences initiales entre groupes. L'ATE estimé par TMLE corrige partiellement ces écarts observés en essayant de reconstituer une comparaison plus équilibrée. Le fait que l'estimation causale soit plus petite que la différence brute suggère qu'une partie de l'écart brut venait probablement de **déséquilibres de base** et pas uniquement de l'effet du traitement.

Question 18

> Discutez le problème du dropout dans cette étude. Comment les auteurs le gèrent-ils et dans quelle mesure est-ce convaincant ?

Réponse type

Le dropout est un enjeu majeur, car l'absence de seconde mesure de PHQ−9 peut refléter un arrêt précoce, un désengagement ou un autre profil clinique. Dans l'analyse principale, certains patients sans suivi sont considérés comme n'ayant pas changé, ce qui est conservateur mais discutable. Les auteurs complètent cela par des analyses d'imputation multiple. C'est sérieux, mais pas totalement convaincant : si les patients manquants diffèrent fortement pour des raisons non mesurées, l'incertitude persiste.

Question 19

> Les tailles d'effet intra-groupe sont importantes. Pourquoi ne faut-il pas les confondre avec l'effet comparatif principal ?

Réponse type

Des tailles d'effet intra-groupe élevées peuvent simplement refléter l'amélioration spontanée, la régression vers la moyenne, l'effet thérapeutique commun aux deux modalités ou la sélection des patients ayant complété. Elles ne disent donc rien à elles seules sur la **supériorité relative** de iCBT par rapport à fCBT. Le critère comparatif principal reste l'ATE entre modalités, pas la baisse moyenne observée au sein d'un seul groupe.

Question 20

> Pourquoi les auteurs ont-ils choisi d'exclure le PHQ-9 baseline de leurs modèles de prédiction principaux ?

Réponse type

Comme l'outcome est le **changement** de PHQ−9, inclure systématiquement le score de base dans tous les modèles peut créer une dépendance mathématique artificielle entre la valeur initiale et la variation observée. Les auteurs ont voulu éviter cette "coupling" excessive. Ils montrent d'ailleurs en sensibilité que l'inclusion du PHQ−9 de base atténue l'effet estimé, ce qui rappelle que ce choix n'est pas neutre.

Thème 5. Biais et facteurs de confusion

Question 21

> En épidémiologie les facteurs de confusion non mesurés sont une hantise. Quelle stratégie générale les auteurs utilisent-ils pour limiter ce problème ?

Réponse type

Ils utilisent une stratégie cumulative : covariables riches, TMLE, super learner, restriction régionale, imputations, analyses alternatives. Cela réduit probablement une partie importante de la confusion observée. Mais aucune de ces techniques n'élimine des confondeurs **non mesurés** comme certaines préférences de traitement, l'affinité numérique, la motivation ou la finesse du jugement clinique initial.

Question 22

> Les patients fCBT présentaient un PHQ-9 baseline plus bas que les patients iCBT. Pourquoi est-ce problématique et comment les auteurs l'abordent-ils ?

Réponse type

Ce déséquilibre suggère une **confusion par indication** : les groupes ne sont pas comparables au départ. Il peut aussi amplifier artificiellement certaines différences de changement par régression vers la moyenne. Les auteurs l'abordent via l'ajustement causal et des analyses de sensibilité, notamment celles incluant explicitement le PHQ–9 initial. Le bon message d'examen est que ce déséquilibre affaiblit la lecture naïve des différences brutes.

Question 23

> Expliquez le concept de biais écologique et pourquoi il est absent de cette étude.

Réponse type

Le biais écologique survient quand on infère des relations individuelles à partir de données agrégées. Ici, le papier travaille au **niveau individuel** avec des patients identifiés un par un, des covariables individuelles et un outcome individuel. Il n'y a donc pas de biais écologique au sens classique. Cela ne veut pas dire qu'il n'y a pas de biais, mais simplement que le problème principal n'est pas une agrégation inadéquate des données.

Question 24

> Les facteurs génétiques pourraient être des confondeurs de l'association traitement-résultat. Qu'en pensez-vous ?

Réponse type

Oui, c'est plausible, mais probablement indirect. Des vulnérabilités génétiques peuvent influencer la sévérité dépressive, la comorbidité, le style cognitif, voire la probabilité d'adhérer à une thérapie numérique ou présentielle. Comme l'article ne dispose pas directement de ces mesures, une part de confusion génétique ou familiale résiduelle peut subsister. Cela renforce l'idée qu'un très bon ajustement statistique ne remplace pas complètement la randomisation.

Question 25

> Dans la discussion, les auteurs mentionnent que "digital affinity" pourrait être un facteur important. Pourquoi ?

Réponse type

L'affinité numérique peut agir comme un **confondeur non mesuré** : elle peut orienter le choix vers iCBT, faciliter l'engagement dans le programme, et améliorer indépendamment l'issue via une meilleure

observance ou une meilleure adéquation avec le format. Si elle n'est pas mesurée, le TMLE ne peut pas l'ajuster. C'est donc un bon exemple de limite substantielle qui reste visible même dans un papier méthodologiquement sophistiqué.

Thème 6. Mesures et variables

Question 26

> Qu'est-ce que le PHQ-9 ? Justifiez son choix comme critère de jugement principal dans cette étude.

Réponse type

Le PHQ−9 est une échelle courte, validée et largement utilisée pour la sévérité des symptômes dépressifs. Son principal intérêt ici est d'être disponible en routine, répétable, simple à interpréter et cohérente avec un suivi naturaliste de très grande taille. Sa limite est d'être un **self-report** symptomatique : il ne capture pas à lui seul le fonctionnement global ni la qualité de vie.

Question 27

> Le critère de jugement est le changement PHQ-9 (variable continue). Discutez l'intérêt d'utiliser une variable continue plutôt qu'un critère dichotomique.

Réponse type

Une variable continue conserve beaucoup plus d'information et donne en général plus de puissance statistique qu'une dichotomisation en "répondeur / non répondeur". Elle évite aussi un seuil arbitraire. En revanche, les cliniciens trouvent parfois plus intuitive une définition dichotomique de rémission. Le bon compromis est souvent de garder la variable continue pour l'analyse principale et de réserver les seuils cliniques à des analyses secondaires.

Question 28

> Le score OASIS est utilisé comme prédicteur. Qu'apporte-t-il dans l'analyse ?

Réponse type

OASIS renseigne sur la sévérité anxieuse et le retentissement associé. Dans ce contexte, il sert surtout de **covariable pronostique** utile, car l'anxiété comorbide peut influencer à la fois le choix du traitement et la réponse clinique. Son inclusion aide donc à mieux ajuster la comparaison et à décrire plus finement le profil initial des patients.

Thème 7. Éthique et réglementaire

Question 29

> Les auteurs indiquent que le consentement éclairé n'était pas requis pour cette étude. Qu'en pensez-vous ?

Réponse type

Dans une étude rétrospective sur registres pseudonymisés, avec autorisation éthique et faible risque additionnel pour les patients, l'absence de consentement individuel peut être justifiable sur le plan réglementaire. Mais ce choix doit rester transparent et proportionné : la gouvernance des données, la

confidentialité et la légitimité de l'usage secondaire des données sont cruciales. La bonne réponse est donc nuancée : **acceptable dans ce cadre, mais pas anodin**.

Question 30

> Les auteurs précisent qu'il n'y a pas eu d'implication de personnes ayant une expérience vécue dans la conception de l'étude. Que penser de ce point ?

Réponse type

Ce n'est pas une faute méthodologique majeure, mais c'est une limite de **pertinence clinique et sociale**. L'implication d'usagers aurait pu aider à choisir des outcomes plus proches des priorités des patients, à mieux discuter le sens d'une différence faible de PHQ−9, ou à interpréter le dropout. Dans un papier contemporain en santé mentale, ce manque mérite d'être signalé.

Question 31

> Le financement de l'étude est public. En quoi est-ce important ?

Réponse type

Un financement public diminue le soupçon de conflit d'intérêts commercial direct, surtout ici où il n'y a pas d'intervention médicamenteuse propriétaire. Cela n'annule évidemment pas tous les biais possibles, mais renforce l'indépendance perçue du travail. C'est un point favorable, à mentionner sans naïveté.

Thème 8. Santé publique et clinique

Question 32

> Si vous étiez décideur de santé publique en France ou au Liban, comment utiliseriez-vous les résultats de cette étude ?

Réponse type

Je m'en servirais comme argument pour renforcer une offre de **soins gradués** incluant iCBT, surtout là où l'accès à une psychothérapie présentielle est limité. En revanche, je n'utiliserais pas ce papier pour dire que iCBT doit remplacer partout fCBT. Il s'agit plutôt d'un argument en faveur d'une **option supplémentaire scalable**, à intégrer à une stratégie plus large d'accès aux soins.

Question 33

> Discutez le modèle de soins "stepped care" mentionné dans la discussion. En quoi l'iCBT peut-il y trouver sa place ?

Réponse type

Dans un modèle de stepped care, on commence par des interventions moins intensives, plus accessibles et moins coûteuses, puis on intensifie si la réponse est insuffisante. iCBT s'insère très bien dans cette logique comme première ou deuxième marche, à condition de garder des possibilités d'escalade vers fCBT, des soins spécialisés ou des prises en charge plus complexes pour les patients qui en ont besoin.

Thème 9. Critique globale

Question 34

Que pensez-vous de la conclusion de l'article telle que présentée dans l'abstract ?

Réponse type

La conclusion de l'abstract est globalement cohérente avec les résultats, mais elle peut faire oublier que l'étude reste **observationnelle**, non pré-enregistrée, avec fort déséquilibre de groupes et données manquantes non triviales. Le bon commentaire est donc : conclusion intéressante, raisonnablement soutenue, mais qui doit être lue avec davantage de prudence que dans un essai randomisé.

Question 35

Quelle est, selon vous, la principale force méthodologique et la principale limite de l'étude ?

Réponse type

La principale force est la combinaison rare d'un **très grand jeu de données naturalistes** et de méthodes causales modernes très sérieusement déployées. La principale limite est l'absence de randomisation, donc la possibilité persistante de **confusion non mesurée**, en particulier autour du choix de modalité thérapeutique et de l'affinité numérique. Toute bonne conclusion doit tenir ensemble ces deux idées.

Vingt questions fictives supplémentaires inspirées des annales

> **i** Statut de ce bloc
>
> Les vingt questions ci-dessous sont **fictives**. Elles sont construites à partir des motifs récurrents des annales EPI du dossier, mais elles portent toutes sur le même article `2026`.

Extrait de travail pour les questions fictives

Question fictive 1

Quel est le principe général de l'analyse principale de cet article ? Qu'en pensez-vous ?

Réponse type

Le principe général est d'estimer un **ATE contrefactuel** entre `iCBT` et `fCBT` à partir de données observationnelles, plutôt que de comparer simplement deux moyennes observées. C'est une stratégie ambitieuse et méthodologiquement moderne. Elle est intéressante parce qu'elle cherche explicitement une quantité causale, mais elle repose sur des hypothèses fortes d'échangeabilité qui ne sont jamais totalement vérifiables dans un registre naturaliste.

Question fictive 2

Les critères d'inclusion et d'exclusion sont-ils de nature à améliorer ou à dégrader la comparabilité des groupes ?

Réponse type

Ils font les deux. Ils améliorent la cohérence clinique de l'échantillon en ciblant une population dépressive relativement homogène et en excluant certaines situations incompatibles avec `iCBT`. Mais ils dégradent potentiellement la comparabilité entre groupes, car les critères pratiques d'orientation vers `iCBT` et `fCBT` ne sont pas symétriques. En particulier, les exigences techniques et motivationnelles propres à `iCBT` peuvent introduire une sélection spécifique.

Question fictive 3

> Que pensez-vous du fait que `fCBT` provienne principalement d'Uusimaa alors que `iCBT` est plus largement diffusée ?

Réponse type

Cela pose un problème de **confusion géographique** et de comparabilité de l'offre de soins. Une partie des différences observées pourrait tenir aux spécificités régionales, à l'organisation locale des soins ou au profil des patients adressés. La restriction à Uusimaa est donc une bonne analyse de sensibilité, même si elle fait perdre de la puissance et de la généralisation.

Question fictive 4

> Les auteurs parlent d'un estimateur "optimal and robust". Est-ce une formulation acceptable ?

Réponse type

Oui, mais à condition de ne pas la prendre au pied de la lettre. `TMLE` est un estimateur robuste au sens statistique et souvent performant sur le compromis biais-variance, surtout avec super learner. En revanche, il n'est "optimal" que sous les hypothèses du cadre causal adopté et avec des données suffisamment riches. Dans une copie, il faut donc valider l'idée sans tomber dans l'admiration aveugle.

Question fictive 5

> Comment discuteriez-vous le fait que les diagnostics privés ne soient pas disponibles dans les registres utilisés ?

Réponse type

L'absence du secteur privé peut limiter à la fois la **validité externe** et la **qualité de certaines covariables**. Des patients plus favorisés ou avec certains profils cliniques peuvent avoir une partie de leur trajectoire hors du système public. Cela peut appauvrir l'information historique disponible et laisser subsister une confusion résiduelle. C'est une limite crédible, surtout si le recours au privé est socialement différencié.

Question fictive 6

> Le critère principal est le changement de PHQ-9. Quelle est la principale force de ce choix et quelle est sa principale faiblesse ?

Réponse type

Sa force principale est la **simplicité** : c'est une mesure validée, répétée en routine, facile à exploiter dans un grand registre. Sa faiblesse principale est de rester un **score symptomatique auto-rapporté**, qui ne reflète pas à lui seul le fonctionnement social, la qualité de vie ou la durabilité de l'amélioration. Le choix est donc très raisonnable, mais incomplet sur le plan clinique.

Question fictive 7

> Les auteurs pénalisent le dropout en attribuant un changement nul si une seule mesure de PHQ-9 est disponible. Qu'en pensez-vous ?

Réponse type

Cette décision est défendable car elle évite d'ignorer les patients qui sortent précocement du suivi et pénalise les traitements avec plus d'abandons. Elle a donc un aspect conservateur. Mais elle est aussi discutable : un abandon peut correspondre à une aggravation, à une amélioration rapide ou à un problème purement logistique. La règle choisie est simple et transparente, mais elle reste une hypothèse forte sur les données manquantes.

Question fictive 8

> A partir de l'objectif de précision fixé par les auteurs, retrouvez l'erreur standard maximale acceptable.

Résolution guidée

Les auteurs veulent un intervalle de confiance à 95% de largeur au plus égale à `1,7`.

$$\text{largeur}(IC95\%) \approx 3{,}92 \times SE$$

Donc :

$$SE_{\text{max}} \approx \frac{1{,}7}{3{,}92} \approx 0{,}434$$

L'erreur standard maximale acceptable est donc d'environ **0,43**.

Réponse type

Si l'on utilise l'approximation classique `largeur(IC95%) ≈ 3,92 × SE`, un objectif de largeur maximale `1,7` impose :

$$SE \lesssim 0{,}434$$

L'étude atteint donc un niveau de précision satisfaisant dès lors que son SE final est inférieur à cette valeur.

Question fictive 9

> En utilisant l'ATE rapporté (`0,745`) et son SE (`0,300`), calculez un score de Wald grossier et discutez la précision.

Résolution guidée

$$z \approx \frac{0{,}745}{0{,}300} \approx 2{,}48$$

Ce z est au-dessus de `1,96`, ce qui est cohérent avec l'intervalle de confiance qui exclut `0`.

La précision reste toutefois moyenne :

$$IC95\% = [0{,}156; 1{,}334]$$

L'effet statistique existe, mais l'incertitude sur sa taille exacte reste non négligeable.

Réponse type

On retrouve un `z` d'environ `2,48`, donc un signal statistique compatible avec une différence non nulle entre `iCBT` et `fCBT`. La précision n'est cependant pas exceptionnelle : l'intervalle `0,156` à `1,334` montre que l'effet plausible va d'un bénéfice faible à modéré. La bonne conclusion n'est donc pas "résultat spectaculaire", mais "petit effet compatible avec les données".

Question fictive 10

Que pensez-vous du fait que les auteurs utilisent aussi `OASIS` comme prédicteur de l'issue ?

Réponse type

C'est un bon choix, car l'anxiété comorbide influence probablement la trajectoire dépressive et possiblement le choix de la modalité thérapeutique. Inclure `OASIS` améliore donc la qualité de l'ajustement et la précision prédictive. Il faut simplement veiller à ne pas le traiter comme une variable "neutre" si certaines dimensions qu'il mesure sont aussi liées au processus d'adressage.

Question fictive 11

Pourquoi les auteurs parlent-ils d'ATE plutôt que de simple différence moyenne observée ?

Réponse type

Parce qu'ils ne veulent pas seulement décrire les données observées ; ils cherchent à approcher une **comparaison causale contrefactuelle**. Une différence brute de moyennes ne répond qu'à la question "qu'a-t-on observé dans ces deux groupes ?". L'ATE répond à la question plus ambitieuse : "qu'aurait-on observé si, en moyenne, ces patients avaient tous reçu l'une puis l'autre modalité ?".

Question fictive 12

Dans quelle mesure la très forte dissymétrie d'effectif entre iCBT et fCBT pose-t-elle problème ?

Réponse type

Elle pose au moins trois problèmes : possible violation locale de la **positivité**, plus grande sensibilité du bras `fCBT` à quelques observations atypiques, et difficulté à construire des comparaisons vraiment équilibrées sur certains profils. Le `TMLE` aide, mais il ne supprime pas magiquement les zones du support des données où la comparaison est pauvre. C'est donc une vraie limite.

Question fictive 13

Les analyses de sensibilité sont-elles ici un simple "plus" ou une nécessité ?

Réponse type

Ici, elles sont une **nécessité**. Dans une étude observationnelle non pré-enregistrée avec données manquantes et déséquilibres importants, un seul résultat principal serait insuffisant. Les analyses de sensibilité servent à juger si le message général survit à des choix raisonnablement différents. Ce n'est pas un supplément cosmétique ; c'est une condition de crédibilité.

Question fictive 14

> Que faut-il penser du fait que l'estimation restreinte à Uusimaa devienne compatible avec zéro ?

Réponse type

Il ne faut pas surinterpréter ce passage à un `IC95%` qui touche `0`. Cela peut refléter une perte de puissance liée à la restriction, un contexte plus homogène, ou une vraie hétérogénéité régionale. La lecture correcte est que le signal principal semble moins robuste dans cette sous-population, ce qui incite à rester prudent sur la généralisation et la stabilité exacte de l'effet.

Question fictive 15

> Si l'on force une puissance a posteriori grossière à partir de `z ≈ 2,48`, quel ordre de grandeur obtient-on ? Faut-il beaucoup y croire ?

Résolution guidée

$$\text{puissance} \approx \Phi(z - 1{,}96)$$

Avec `z ≈ 2,48` :

$$\Phi(2{,}48 - 1{,}96) = \Phi(0{,}52) \approx 0{,}70$$

On obtient donc une puissance observée grossière d'environ **70%**.

Réponse type

On retrouve un ordre de grandeur autour de **70%**. Mais il faut peu y croire comme mesure "profonde" de la qualité de l'étude. Ici, la bonne lecture passe surtout par l'intervalle de confiance, la taille d'effet et les hypothèses causales, pas par une puissance a posteriori recalculée après coup.

Question fictive 16

> Les auteurs indiquent que les résultats sont cohérents avec les essais randomisés antérieurs. Est-ce un argument fort ?

Réponse type

Oui, c'est un argument utile dans une logique de **triangulation**. Quand un résultat observationnel va dans le même sens que des essais randomisés antérieurs, cela renforce sa plausibilité. Mais ce n'est pas un argument décisif : la comparaison porte ici sur des contextes de soin réels différents, avec d'autres contraintes et d'autres populations. C'est un renfort, pas une preuve.

Question fictive 17

> Si vous deviez critiquer la validité externe de cette étude, quel serait votre argument principal ?

Réponse type

Mon argument principal serait que l'étude s'inscrit dans un système de soins finlandais particulier, avec une organisation, un remboursement, une culture numérique et une structuration des prises en charge qui ne sont pas automatiquement transposables ailleurs. Ce n'est pas une faiblesse rédhibitoire, mais cela oblige à discuter sérieusement la transférabilité vers d'autres systèmes de santé.

Question fictive 18

> Quel contrôle négatif plausible proposeriez-vous pour cet article ?

Réponse type

On pourrait proposer un outcome peu susceptible d'être influencé par le type de psychothérapie à court terme, mais capté dans les registres, par exemple une pathologie somatique bénigne ou un indicateur administratif sans lien plausible direct avec l'intervention. L'idée serait de vérifier que la comparaison `iCBT` versus `fCBT` ne fabrique pas des associations partout par simple structure de sélection. Le contrôle négatif idéal doit partager les mêmes voies de confusion, sans mécanisme causal crédible.

Question fictive 19

> Si vous deviez proposer une amélioration concrète du design sans perdre le caractère naturaliste de l'étude, laquelle choisiriez-vous ?

Réponse type

Je proposerais un **essai pragmatique randomisé** ou, si cela n'est pas faisable, un design prospectif naturaliste avec collecte standardisée de variables aujourd'hui manquantes : affinité numérique, préférences du patient, attentes, motifs précis d'orientation, qualité/alliance thérapeutique. Cela préserverait une bonne validité externe tout en réduisant une part importante de la confusion non mesurée.

Question fictive 20

> Au total, si vous deviez défendre ou attaquer cet article en dix lignes devant un jury, que diriez-vous ?

Réponse type

Je dirais que c'est un papier **fort**, parce qu'il exploite un très grand registre réel, pose explicitement une question causale, utilise des méthodes modernes sérieuses et multiplie les analyses de sensibilité. J'ajouterais qu'il reste **attaquable** sur ses points les plus classiques : absence de randomisation, déséquilibre massif entre groupes, données manquantes, non-pré-enregistrement, variables clés non mesurées comme l'affinité numérique. Je le défendrais donc comme une bonne pièce de preuve **complémentaire** aux essais randomisés, mais certainement pas comme une preuve causale définitive à lui seul.

7. Réponses prêtes à l'emploi

7.1. Quand la puissance n'est pas le vrai sujet

> Dans cette étude observationnelle, la bonne discussion ne porte pas vraiment sur un nombre de sujets nécessaires a priori, mais plutôt sur la précision des estimations, le nombre d'événements informatifs et le risque de confusion résiduelle.

7.2. Quand il faut discuter une grande cohorte

> Dans une cohorte de cette taille, la significativité statistique n'est plus la question principale. Il faut surtout regarder l'ampleur de l'effet, la largeur des IC95%, et les biais susceptibles d'expliquer l'association observée.

7.3. Quand il faut parler de causalité sans surinterpréter

L'association est compatible avec un effet causal, mais l'étude reste observationnelle. Il faut donc discuter explicitement la temporalité, les ajustements, la plausibilité biologique et la possibilité d'une confusion résiduelle.

8. Checklist finale

— Ai-je bien identifié le design ?
— Ai-je répondu à **toutes** les questions de l'année ?
— Ai-je distingué précision, puissance et pertinence clinique ?
— Ai-je expliqué les biais plausibles ?
— Ai-je conclu de façon prudente et directement utile pour l'examen ?

9. Conclusion

Le point central de ces annales EPI est toujours le même : **ne pas plaquer une réponse statistique standard sur une question qui est en réalité méthodologique**. Les meilleures copies sont celles qui savent reconnaître si le vrai enjeu est la précision, la confusion, le design, l'interprétation de la taille d'effet, ou la validité causale.