

Poly de révision RC

Annales de recherche clinique : toutes les questions, toutes les réponses

Table des matières

Poly de révision RC	3
Annales résolues, méthodes, designs, puissance et réponses types	3
Comment utiliser ce poly	3
1. Carte rapide des annales RC couvertes	3
2. Boîte à outils RC	4
2.1. Toujours commencer par le critère principal	4
2.2. Les trois populations à distinguer	5
2.3. Survie : les briques minimales	5
2.4. Co-primary, composite et hiérarchie	5
2.5. Dépense du risque alpha et analyses intermédiaires	5
2.6. Pourquoi il n'y a souvent pas de p dans les tableaux d'effets indésirables	5
2.7. Quand le modèle n'est ni logistique ni simple Cox	5
2.8. La non-infériorité en une phrase	6
2.9. Approximation de Schoenfeld pour un essai de survie	6
3. Matrice de traçabilité des questions RC	6
4. Comment écrire une bonne réponse RC	7
5. Pièges récurrents	7
6. Annales RC résolues	7
6.1. 2011 RC	7
Article et contexte	7
Questions posées dans l'annale	8
Extrait de l'article à analyser	8
Ce que l'examineur veut vérifier	8
Corrections détaillées	8
6.2. 2012 RC	10
Article et contexte	10
Questions posées dans l'annale	10
Extrait de l'article à analyser	11
Ce que l'examineur veut vérifier	11
Corrections détaillées	11
6.3. 2013 RC	12
Article et contexte	12
Questions posées dans l'annale	13
Extrait de l'article à analyser	13
Corrections détaillées	13
6.4. 2014 RC	14
Article et contexte	14

Questions posées dans l'annale	14
Extrait de l'article à analyser	14
Corrections détaillées	14
6.5. 2016 RC	15
Article et contexte	15
Questions posées dans l'annale	15
Extrait de l'article à analyser	16
Corrections détaillées	16
6.6. 2018 RC	16
Article et contexte	16
Questions posées dans l'annale	17
Extrait de l'article à analyser	17
Corrections détaillées	17
6.7. 2019 RC	18
Article et contexte	18
Questions posées dans l'annale	18
Extrait de l'article à analyser	19
Corrections détaillées	19
6.8. 2020 RC	19
Article et contexte	19
Questions posées dans l'annale	20
Extrait de l'article à analyser	20
Corrections détaillées	20
6.9. 2021 RC	21
Article et contexte	21
Questions posées dans l'annale	21
Extrait de l'article à analyser	21
Corrections détaillées	21
6.10. 2022 RC	22
Article et contexte	22
Questions posées dans l'annale	22
Extrait de l'article à analyser	22
Corrections détaillées	22
6.11. 2023 RC	23
Article et contexte	23
Questions posées dans l'annale	23
Extrait de l'article à analyser	23
Corrections détaillées	23
6.12. 2024 RC	24
Article et contexte	24
Questions posées dans l'annale	24
Extrait de l'article à analyser	24
Corrections détaillées	24
6.13. 2025 RC	25
Article et contexte	25
Questions posées dans l'annale	25
Extrait de l'article à analyser	25
Corrections détaillées	25
6.14. 2026 RC	26
Article et contexte	26
Questions posées dans l'annale	26

Extrait de l'article à analyser	28
Ce que l'examinateur veut vérifier	28
Corrections détaillées	28
7. Réponses prêtes à l'emploi	32
7.1. Quand il faut critiquer un critère principal	32
7.2. Quand il faut expliquer l'absence de p dans un tableau de sécurité	32
7.3. Quand il faut discuter une analyse en sous-groupe	33
7.4. Quand il faut retrouver un NSN approximatif	33
8. Checklist finale	33
9. Conclusion	33

Poly de révision RC

Annales résolues, méthodes, designs, puissance et réponses types

Document de travail orienté examen : **toutes les questions repérées dans les annales RC**, avec corrections didactiques, lecture critique et réponses directement réutilisables en copie.

Ce poly a un objectif simple : **permettre de répondre entièrement aux annales de recherche clinique** du dossier.

Le fil directeur est volontairement le même que dans les autres polys :

1. repartir des **sujets originaux** ;
2. repérer **toutes les questions réellement posées** ;
3. revenir au **papier support** pour comprendre ce qu'il fallait regarder ;
4. produire une **réponse d'examen complète, concise et défendable**.

recherche clinique essais randomisés survie multiplicité annales exhaustives

! Point de méthode

En RC, la bonne réponse n'est pas toujours un calcul de nombre de sujets. Très souvent, la vraie question porte sur le **design**, le **critère principal**, l'**analyse principale**, la **gestion du risque alpha**, l'**interprétation clinique**, ou la **validité de l'aveugle**. Le calcul n'est qu'une partie de la copie.

Comment utiliser ce poly

1. Commencez par identifier le **type d'essai** et la **nature du critère principal**.
2. Regardez ensuite ce que l'examinateur veut vraiment tester : design, statistique, clinimétrie, sécurité, ou lecture critique.
3. Si un calcul est demandé, refaites-le au brouillon avec les données du papier.
4. Si le calcul exact n'est pas reconstituable, donnez l'**ordre de grandeur**, puis expliquez **pourquoi**.
5. Utilisez les **réponses types** comme matrice de copie, pas comme texte à apprendre mot pour mot.

1. Carte rapide des annales RC couvertes

Année	Article support	Angle dominant	Nombre de questions RC
2011	James et al., NEJM 2010	run-in, composite, Cox, ITT, recrutement	7
2012	AIM-HIGH, NEJM 2011	randomisation, endpoints hiérarchisés, event-driven, futility	7
2013	Ryan et al., NEJM 2013	population d'analyse, analyses intermédiaires, arrêt précoce	4
2014	Doody et al., NEJM 2013	co-primary outcomes, MMRM, futilité, arrêt pour nocivité	4
2016	Kappos et al., NEJM 2015	analyse principale, absence de p, sécurité, interims	4
2018	Pavord et al., NEJM 2017 + supplément	critère principal, multiplicité, binomial négatif, sous-groupes	4
2019	Parker et al., Lancet 2018	minimisation, MAMS, FFS, sous-groupes	7
2020	Schnitzer et al., JAMA 2019	co-primary endpoints, NSN, AE	6
2021	Roy et al., PLoS Med 2020	essai en cluster, critère principal, qualitatif	5
2022	Bacharier et al., NEJM 2021	critère principal, binomial négatif, multiplicité, pédiatrie	5
2023	Edwardson et al., BMJ 2022	cluster, complete case, modèle principal	5
2024	Moncrieff et al., Lancet Psychiatry 2023	aveugle, SFS, non-infériorité, design hybride	5
2025	Hautzinger et al., JAMA Psychiatry 2024	score de propension, psychothérapies, log-rank	5
2026	Fang et al., NEJM 2026 + sujet exhaustif	design, méthodes, efficacité, sécurité, critique	35

2. Boîte à outils RC

2.1. Toujours commencer par le critère principal

Une copie solide dit d'abord :

- quel est le critère principal ;

- s'il est **clinique, surrogat, continu, binaire, temps jusqu'à événement**, ou **comptage** ;
- pourquoi il est adapté ou discutable ;
- et quelle conséquence cela a sur l'analyse statistique.

2.2. Les trois populations à distinguer

ITT tous les sujets randomisés, analysés selon leur groupe d'affectation.

Per protocol sujets ayant suffisamment adhéré au protocole ; utile en sensibilité, rarement idéal en principal.

Population de sécurité tous les sujets ayant reçu au moins une dose du traitement.

2.3. Survie : les briques minimales

Kaplan–Meier décrit les courbes de survie, les médianes, les proportions survivantes à un temps donné.

Log–rank teste une différence globale entre courbes.

Cox estime un HR, éventuellement ajusté, sous hypothèse de proportionnalité des risques.

2.4. Co-primary, composite et hiérarchie

- un **critère composite** est positif si un de ses composants survient ;
- des **co-primary endpoints** exigent souvent que plusieurs dimensions soient toutes positives ;
- des **primary / secondary / tertiary endpoints** organisent surtout la hiérarchie interprétative et la dépense du risque alpha.

2.5. Dépense du risque alpha et analyses intermédiaires

- O'Brien–Fleming : très conservateur tôt, plus permissif tard ;
- Lan–DeMets : version flexible de l'alpha-spending ;
- une analyse intermédiaire peut viser l'**efficacité**, la **futilité**, ou la **sécurité**.

2.6. Pourquoi il n'y a souvent pas de p dans les tableaux d'effets indésirables

- la sécurité est en général décrite de façon **exploratoire** ;
- il y a **beaucoup de comparaisons** ;
- une absence de p ne signifie pas que la sécurité est ignorée ;
- il faut regarder les **fréquences**, les **gradients de gravité**, la **nature clinique** des événements, et les **arrêts de traitement**.

2.7. Quand le modèle n'est ni logistique ni simple Cox

Binomial négatif utile pour des comptages récurrents avec surdispersion, comme les exacerbations.

Mixed-model repeated measures utile pour les mesures répétées continues avec données manquantes supposées manquantes à hasard conditionnel.

Modèle à effet aléatoire de cluster nécessaire quand on randomise des centres mais qu'on analyse des sujets.

2.8. La non-infériorité en une phrase

On ne cherche pas à montrer que le nouveau traitement est meilleur, mais qu'il n'est **pas trop moins bon** qu'un comparateur actif au-delà d'une **marge** jugée cliniquement acceptable.

Formule simplifiée à deux proportions :

$$N \approx \frac{2(z_\alpha + z_\beta)^2 [p_C(1 - p_C) + p_T(1 - p_T)]}{\Delta^2}$$

où Δ est la marge.

2.9. Approximation de Schoenfeld pour un essai de survie

Si l'allocation est 1:1 :

$$D \approx \frac{4(z_{1-\alpha} + z_{1-\beta})^2}{[\log(HR)]^2}$$

D est ici le nombre d'événements requis.

3. Matrice de traçabilité des questions RC

Année	Fichier(s) original(aux) consulté(s)	Nombre de questions RC repérées	Présence dans ce poly
2011	exam_2011_Epi_RC.d Exam_2011_RC.pdf	7	Oui
2012	exam_2012_Epi_RC.d Exam_2012_RC.pdf	7	Oui
2013	exam_2013_Epi_RC.d Exam_2013_RC.pdf	4	Oui
2014	exam_2014_Epi_RC.d exam_2014_RC.pdf	4	Oui
2016	exam_2016_Epi_RC.d exa- men_RC_20152016.pdf	4	Oui
2018	exam_2018_Epi_RC.d Exam_RC_2018_annex	4	Oui
2019	exam_2019_Epi_RC_a M2MSR_Exam_RC_2019 M2MSR_Exam_RC_2019	7	Oui
2020	exam_2020_Epi_RC_a Exam_RC_2020_artic	6	Oui
2021	exam_2021_Epi_RC_a Exam_RC_2021.pdf	5	Oui
2022	exam_2022_Epi_RC_c Su- jet_exam_RC_mars_21 Su- jet_exam_RC_mars_21	5	Oui

Année	Fichier(s) original(aux) consulté(s)	Nombre de questions RC repérées	Présence dans ce poly
2023	exam_2023_Epi_RC_c Ar- ticle_Exam_RC_2023	5	Oui
2024	exam_2024_Epi_RC_c Ar- ticle_Exam_RC_2024	5	Oui
2025	exam_2025_Epi_RC.t Ar- ticle_RC2025.pdf	5	Oui
2026	Examen_RC_M2MSR_20 copie.pdf, ar- ticle_RC_2026.pdf	35	Oui

4. Comment écrire une bonne réponse RC

Une bonne réponse d'examen RC suit souvent le même mouvement :

1. rappeler le **design** et le **critère principal** ;
2. dire ce que fait la méthode statistique et pourquoi elle a été choisie ;
3. discuter les **forces** puis les **limites** ;
4. conclure par un jugement nuancé : approprié, discutable, insuffisant, ou acceptable sous conditions.

5. Pièges récurrents

- Confondre **arrêt de traitement** et **perte de suivi**.
- Confondre **critère composite** et **co-primary endpoints**.
- Critiquer une table de baseline parce qu'il n'y a pas de p, alors qu'en RCT ce n'est pas le point.
- Vouloir absolument un calcul de NSN alors que la bonne question porte en réalité sur le **design**.
- Penser qu'une analyse en sous-groupe positive suffit à refaire toute la conclusion.

6. Annales RC résolues

6.1. 2011 RC

Article et contexte

Article support : **James et al., NEJM 2010**, essai SCOUT sur la **sibutramine** chez des patients obèses ou en surpoids à haut risque cardiovasculaire.

- 10 744 sujets inclus ;
- phase de run-in simple aveugle de 6 semaines avec sibutramine pour tous ;
- 9804 sujets ensuite randomisés ;
- essai en double aveugle sibutramine versus placebo ;
- critère principal : temps jusqu'au premier événement cardiovasculaire majeur du composite.

Questions posées dans l'annale

1. Il est question d'une « lead in period ». Quel en est l'intérêt ? Quel(s) problème(s) pose-t-elle ?
2. Quels sont les problèmes clinimétriques potentiels liés aux mesures anthropométriques réalisées ici ?
3. Quel est le critère principal ? Quel(s) problème(s) pose-t-il ? Comment le(s) gérer ?
4. L'analyse principale du critère principal est faite à l'aide d'un modèle de Cox ajusté. Qu'en pensez-vous ?
5. Dans le texte on trouve : "Secondary analyses of interactions between treatment and cardiovascular-risk group were evaluated with the use of Cox models." A quoi servent ces analyses ?
6. L'analyse est réalisée en ITT et le nombre de sujets nécessaire est calculé en tenant compte d'un nombre de perdus de vue possible. Qu'en pensez-vous ?
7. Il y a eu des difficultés de recrutement. Quelle(s) mesure(s) ont été prise(s) et quelle(s) implication(s) cela peut-il avoir sur l'interprétation des résultats ?

Extrait de l'article à analyser

Ce que l'examineur veut vérifier

- distinction entre **validité interne** et **validité externe** ;
- compréhension d'un **run-in** ;
- lecture critique d'un **composite** et d'un **Cox ajusté** ;
- capacité à ne pas confondre **ITT**, **discontinuation**, et **perdus de vue**.

Corrections détaillées

Question 1

Il est question d'une « lead in period ». Quel en est l'intérêt ? Quel(s) problème(s) pose-t-elle ?

La phase de run-in sert ici à deux choses : vérifier la **tolérance** initiale à la sibutramine et éliminer avant la randomisation les patients chez qui apparaissent rapidement une hausse de pression artérielle, du pouls, une intolérance, ou un défaut d'adhésion majeur. Cela sécurise l'essai et limite les interruptions immédiates après randomisation.

Le problème méthodologique est que la randomisation ne porte plus sur la population cible entière, mais sur une **population sélectionnée** de sujets ayant déjà passé ce filtre. La validité interne de la comparaison randomisée reste bonne, mais la **généralisabilité** est réduite : le résultat concerne surtout des patients capables de tolérer un premier mois et demi de sibutramine.

Réponse type

La lead-in period sert ici à sélectionner des patients capables de tolérer la sibutramine et à éliminer ceux qui ont d'emblée des effets hémodynamiques ou une mauvaise observance. C'est utile sur le plan pratique et sécuritaire, mais cela sélectionne une population particulière avant la randomisation. La comparaison randomisée reste valide en interne, mais les résultats sont moins généralisables à l'ensemble des patients obèses à haut risque cardiovasculaire.

Question 2

Quels sont les problèmes clinimétriques potentiels liés aux mesures anthropométriques réalisées ici ?

Le papier repose sur des mesures comme le **poids**, le **BMI** et le **tour de taille**. Ces variables sont utiles, mais elles ont plusieurs limites clinimétriques :

- elles sont sensibles aux conditions de mesure : vêtements, heure de la journée, rétention hydrosodée, opérateur ;
- le tour de taille est notoirement moins reproductible qu'un poids ;
- BMI et tour de taille ne mesurent pas directement la composition corporelle ni l'adiposité viscérale ;
- une petite différence moyenne peut être statistiquement nette sans être cliniquement majeure.

Dans un essai multicentrique, il faut donc standardiser les procédures et interpréter ces variations comme des **proxies** imparfaits du risque métabolique.

Question 3

Quel est le critère principal ? Quel(s) problème(s) pose-t-il ? Comment le(s) gérer ?

Le critère principal est un **composite de temps jusqu'au premier événement** : infarctus non fatal, AVC non fatal, arrêt cardiaque récupéré, ou décès cardiovasculaire.

Ses atouts :

- il augmente le nombre d'événements ;
- il est pertinent dans un essai cardiovasculaire ;
- il se prête bien à une analyse de survie.

Ses limites :

- tous les composants n'ont pas la même gravité clinique ;
- le traitement peut agir différemment sur chaque composant ;
- le premier événement masque ce qui se passe ensuite.

Il faut donc toujours regarder les **composants séparés** et vérifier si l'effet global n'est pas porté surtout par des composants moins centraux. Ici, l'augmentation observée venait surtout des événements non fatals.

Question 4

L'analyse principale du critère principal est faite à l'aide d'un modèle de Cox ajusté. Qu'en pensez-vous ?

Le choix d'un modèle de Cox est logique puisque le critère principal est un **temps jusqu'à événement**. L'ajustement n'est pas là pour "réparer" la randomisation, mais pour **gagner en précision** à condition que les covariables soient préspecifiées et pronostiques. En revanche, si l'ajustement est opportuniste ou trop riche, on complique inutilement l'interprétation.

Donc la bonne réponse est : **oui, le Cox ajusté est acceptable**, voire classique, mais seulement s'il reste fidèle au protocole et s'il ne remplace pas l'analyse ITT brute.

Question 5

Dans le texte on trouve : "Secondary analyses of interactions between treatment and cardiovascular-risk group were evaluated with the use of Cox models." A quoi servent ces analyses ?

Ces analyses cherchent une **modification d'effet** : le traitement a-t-il un effet différent selon qu'on est dans le groupe DM only, CV only, ou CV + DM ?

Elles sont utiles pour :

- explorer l'hétérogénéité d'effet ;
- générer des hypothèses cliniques ;
- vérifier si le résultat global cache un profil particulier.

Mais elles sont rarement très puissantes, et elles ne doivent pas renverser une conclusion principale si elles ne sont pas fortement étayées.

Question 6

L'analyse est réalisée en ITT et le nombre de sujets nécessaire est calculé en tenant compte d'un nombre de perdus de vue possible. Qu'en pensez-vous ?

Il faut distinguer deux choses :

- **arrêt du traitement** ;
- **perte de vue** vraie, c'est-à-dire perte de l'information sur le critère principal.

En ITT, un sujet qui arrête la sibutramine reste analysable si l'on continue à suivre les événements cardiovasculaires. Le surdimensionnement pour pertes de vue est donc légitime seulement si l'on anticipe une perte d'information sur l'issue, pas la simple non-observance. Ici, l'essai a beaucoup de discontinuations mais peu de vraies pertes de suivi, donc il ne faut surtout pas confondre les deux.

Question 7

Il y a eu des difficultés de recrutement. Quelle(s) mesure(s) ont été prise(s) et quelle(s) implication(s) cela peut-il avoir sur l'interprétation des résultats ?

L'article explique qu'en raison d'un taux d'événements plus faible que prévu et d'un recrutement difficile, les auteurs ont restreint secondairement le recrutement vers les profils les plus à risque, ont prolongé la période de recrutement, puis le suivi. C'est défendable dans un essai event-driven, mais cela a plusieurs conséquences :

- la population finale devient plus riche en sujets très à risque ;
- la généralisabilité se déplace ;
- l'interprétation doit tenir compte de ce changement de composition.

Ce n'est pas forcément un biais fatal, mais il faut le signaler explicitement.

6.2. 2012 RC

Article et contexte

Article support : **AIM-HIGH Investigators, NEJM 2011**, essai comparant **niacine LP** versus placebo sur fond de traitement intensif par statine.

- 3414 patients randomisés ;
- essai event-driven ;
- critère principal composite cardiovasculaire ;
- arrêt après environ 3 ans pour absence d'efficacité.

Questions posées dans l'annale

1. Que pensez-vous de la façon dont est présentée la randomization ?

2. Dans la section “end points” il est question de “Primary”, “secondary”, “tertiary end points”. Quelle logique est derrière cette nomenclature ?
3. Il est indiqué que : “Because the rate of the primary end point was lower than projected, the protocol was amended to change the end point of “high-risk acute coronary syndrome” to include...”. Est-ce que cela pose problème sur un plan statistique ?
4. Essayez de retrouver approximativement le nombre de sujet nécessaire.
5. Des analyses intermédiaires étaient prévues. Quelle est la règle d'arrêt ?
6. Que pensez-vous du tableau 1 ?
7. Quelle est l'analyse principale du critère principal ?

Extrait de l'article à analyser

Ce que l'examineur veut vérifier

- compréhension d'une **randomisation précédée d'un run-in** ;
- lecture critique d'une **hiérarchie d'end points** ;
- jugement sur un **amendement de critère principal** ;
- capacité à retrouver le **NSN** et la **règle d'arrêt**.

Corrections détaillées

Question 1

Que pensez-vous de la façon dont est présentée la randomization ?

La randomisation en elle-même est correcte : allocation 1 : 1, traitement masqué, placebo construit pour mimer le flushing. Mais elle intervient **après une phase ouverte** avec simvastatine et niacine, ce qui signifie que l'essai randomise en fait des patients déjà sélectionnés sur la tolérance et l'adhésion. Cela protège la faisabilité, mais réduit la généralisabilité et doit être explicité.

Question 2

Dans la section “end points” il est question de “Primary”, “secondary”, “tertiary end points”. Quelle logique est derrière cette nomenclature ?

Cette nomenclature correspond à une **hiérarchie de confirmation** :

- le primary endpoint porte l'hypothèse principale ;
- les secondary endpoints complètent l'interprétation, parfois dans un ordre hiérarchique ;
- les tertiary endpoints sont encore plus exploratoires.

Autrement dit, plus on descend, moins l'interprétation est confirmatoire et plus la question de la multiplicité devient importante.

Question 3

Il est indiqué que : “Because the rate of the primary end point was lower than projected, the protocol was amended ...”. Est-ce que cela pose problème sur un plan statistique ?

Oui, potentiellement, car modifier un critère principal en cours d'essai est toujours délicat. Le risque est d'introduire de la flexibilité opportuniste et donc d'altérer le contrôle du risque alpha. Ici, l'argument des auteurs est que la décision a été prise sur une **information agrégée et en aveugle** parce que le taux

d'événements était trop faible. Cela rend la démarche plus défendable, mais elle reste méthodologiquement sensible et doit être signalée comme une faiblesse.

Question 4

Essayez de retrouver approximativement le nombre de sujet nécessaire.

Le papier est un essai **event-driven**. Les auteurs visent 800 événements avec une puissance de 85%, un alpha unilatéral de 2,5%, et un effet attendu de l'ordre de 25% de réduction relative. En ramenant cela à un rendement plausible d'événements sur la durée de suivi, on retrouve un besoin de l'ordre de **plusieurs milliers de patients**, cohérent avec les 3414 effectivement randomisés.

Question 5

Des analyses intermédiaires étaient prévues. Quelle est la règle d'arrêt ?

Il y avait une règle de **futilité**. L'article précise qu'un HR observé de 1,02 ou plus, associé à une très petite probabilité de renversement favorable ($P < 0,001$ pour futilité), pouvait justifier l'arrêt. C'est une logique classique : arrêter tôt si l'essai n'a pratiquement plus de chance d'atteindre son objectif principal.

Question 6

Que pensez-vous du tableau 1 ?

Le tableau 1 doit être lu comme une **description** des groupes randomisés, pas comme une batterie de tests d'hypothèses. Le point important est l'équilibre global des caractéristiques pronostiques. Dans un RCT, on ne reproche pas l'absence de p au tableau de baseline ; au contraire, les p y sont souvent peu utiles.

Question 7

Quelle est l'analyse principale du critère principal ?

Le critère principal est un **temps jusqu'au premier événement** du composite cardiovasculaire. L'analyse principale est donc une analyse de survie, avec comparaison par **log-rank** et estimation d'un **hazard ratio** par modèle de Cox. Il faut l'indiquer clairement, en précisant qu'il s'agit d'une analyse ITT.

6.3. 2013 RC

Article et contexte

Article support : **Ryan et al., NEJM 2013**, essai en double aveugle comparant **abiratérone + prednisone** versus **placebo + prednisone** dans le cancer de la prostate métastatique résistant à la castration sans chimiothérapie préalable.

- 1088 patients randomisés ;
- 2 co-primary endpoints : radiographic progression-free survival et overall survival ;
- plusieurs analyses intermédiaires de survie prévues ;
- essai démasqué après une analyse intermédiaire.

Questions posées dans l'annale

1. Quelle est la moyenne d'âge des hommes inclus dans l'étude ? Que pensez-vous de cela ?
2. Dans l'analyse des résultats d'un essai randomisé on parle souvent de population « per protocol » ou « intent to treat ». Quelle est la population d'analyse ici selon-vous ?
3. Décrivez la procédure d'analyse intermédiaire.
4. L'essai a été interrompu lors d'une analyse intermédiaire. Que pensez-vous de cette décision, pesez le pour et le contre ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Quelle est la moyenne d'âge des hommes inclus dans l'étude ? Que pensez-vous de cela ?

L'âge des patients est d'environ **70 ans** dans ce type de population de cancer de la prostate métastatique résistant à la castration. C'est cohérent avec l'épidémiologie de la maladie. La bonne remarque n'est donc pas "ils sont vieux", mais plutôt : la population incluse est globalement pertinente, tout en restant assez sélectionnée sur le plan fonctionnel (ECOG 0–1, peu symptomatiques), donc un peu plus favorable que certains patients vus en routine.

Question 2

... Quelle est la population d'analyse ici selon-vous ?

La population d'analyse principale est l'**ITT** : tous les patients randomisés sont analysés selon leur groupe d'affectation. C'est la bonne stratégie pour les endpoints de survie d'un essai confirmatoire. Une population per protocol peut exister à titre de sensibilité, mais elle ne doit pas remplacer l'ITT ici.

Question 3

Décrivez la procédure d'analyse intermédiaire.

L'essai avait un vrai plan d'analyses intermédiaires sur l'OS, avec plusieurs looks planifiés à des nombres prédéfinis de décès. La dépense du risque alpha suivait une logique **O'Brien-Fleming / Lan-DeMets**, donc très conservatrice aux premiers looks. Les co-primary endpoints avaient en plus une répartition de l'alpha global.

Question 4

L'essai a été interrompu lors d'une analyse intermédiaire. Que pensez-vous de cette décision, pesez le pour et le contre ?

Les arguments **pour** :

- bénéfice déjà convaincant sur plusieurs dimensions ;
- éthique de ne pas maintenir inutilement des patients dans le bras contrôle ;
- cohérence avec un plan intermédiaire prévu à l'avance.

Les arguments **contre** :

- un arrêt précoce peut surestimer la taille d'effet ;
- il rend l'OS plus immature ;
- il réduit la précision à long terme sur sécurité et durabilité.

La bonne réponse est donc nuancée : **décision défendable si elle suit le plan prévu, mais qui doit être interprétée avec la prudence habituelle des essais arrêtés tôt.**

6.4. 2014 RC

Article et contexte

Article support : **Doody et al., NEJM 2013**, essai de phase 3 sur le **semagacestat** dans la maladie d'Alzheimer probable.

- 1537 patients randomisés ;
- 3 bras : placebo, 100 mg, 140 mg ;
- 2 co-primary outcomes à 76 semaines : ADAS-cog et ADCS-ADL ;
- analyse principale par mixed-model repeated-measures ;
- arrêt précoce sur recommandation du DSMB pour nocivité et futilité.

Questions posées dans l'annale

1. Discutez la pertinence clinique des (co)critères principaux d'efficacité. Parmi l'ensemble des critères présentés (principaux et secondaires), pensez-vous qu'il manque quelque chose ?
2. Décrire précisément la procédure d'analyse statistique du (co)critère principal d'efficacité
3. Que pensez-vous de la gestion du risque de première espèce dans cet essai.
4. Pensez-vous que l'essai a été arrêté à tort ou qu'il aurait dû être arrêté plus tôt ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Discutez la pertinence clinique des (co)critères principaux d'efficacité...

Le choix d'associer ADAS-cog et ADCS-ADL est logique : on veut mesurer à la fois la **cognition** et le **retentissement fonctionnel**, pas seulement un score neuropsychologique. C'est défendable dans Alzheimer.

Les limites sont toutefois nettes :

- ces scores évoluent lentement et sont parfois peu sensibles à court terme ;
- leur interprétation clinique concrète n'est pas toujours simple ;
- en ouvert, ils seraient très vulnérables au biais, même si ici l'essai est en double aveugle.

On peut dire qu'il manque idéalement une dimension plus directement perceptible pour le patient et l'entourage : **qualité de vie, charge de l'aidant**, ou un critère global clinique fortement interprétable.

Question 2

Décrire précisément la procédure d'analyse statistique du (co)critère principal d'efficacité

L'analyse principale compare les changements de scores du baseline à la semaine 76 au moyen d'un **mixed-model repeated-measures**. Ce type de modèle :

- exploite toutes les mesures répétées ;
- ajuste sur des covariables de base ;
- tient compte de la corrélation intra-sujet ;
- évite de recourir à une imputation simple de type LOCF.

La bonne justification est que, pour des scores répétés en cours de suivi avec attrition, le MMRM est plus propre qu'une comparaison ponctuelle finale naïve.

Question 3

Que pensez-vous de la gestion du risque de première espèce dans cet essai.

Il y a plusieurs sources de multiplicité : deux doses actives, deux co-primary endpoints, plusieurs secondaires. Cela rend le contrôle du risque alpha non trivial. On attend donc une hiérarchie ou une stratégie claire. Ici, le vrai problème de l'essai n'est pas tant d'avoir produit un faux positif que l'inverse : les résultats vont vers la **nocivité** et la **futilité**. Il faut donc dire que le contrôle du risque alpha reste important en théorie, mais que la lecture clinique finale est surtout dominée par l'absence de bénéfice et les signaux défavorables.

Question 4

Pensez-vous que l'essai a été arrêté à tort ou qu'il aurait dû être arrêté plus tôt ?

L'arrêt paraît justifié : il existait à la fois un **signal de nocivité** et une **futilité** importante sur les co-primary outcomes. On peut toujours discuter s'il aurait fallu arrêter quelques mois plus tôt, mais rétrospectivement la décision du DSMB paraît raisonnable et prudente. La bonne position est donc : **arrêt légitime, et probablement plus défendable qu'un maintien prolongé.**

6.5. 2016 RC

Article et contexte

Article support : **Kappos et al., NEJM 2015**, essai de phase 3 comparant **daclizumab HYP** versus **interféron bêta-1a** dans la sclérose en plaques rémittente-récurrente.

- 1841 patients randomisés ;
- essai randomisé, en double aveugle, comparateur actif ;
- critère principal : **annualized relapse rate** ;
- analyse principale par **régression binomiale négative**.

Questions posées dans l'annale

1. Explicitez, expliquez et justifiez l'analyse principale du critère principal
2. Dans l'article il est dit "The estimated percentage of patients who were free from relapse at week 144 was 67% ... why n'y a-t-il pas de "p" ?"
3. Pourquoi n'y a-t-il pas de « p » dans les comparaisons de survenue des effets indésirables ?
4. Dans cette étude il n'est pas fait mention d'analyses intermédiaires. Pensez-vous que ça ait été un bon choix ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Explicitez, expliquez et justifiez l'analyse principale du critère principal

Le critère principal est le **taux annualisé de rechutes**, donc un **comptage récurrent par unité de temps**, pas un simple oui/non. Une régression binomiale négative est adaptée car :

- elle modélise un nombre d'événements ;
- elle peut intégrer le temps d'observation ;
- elle gère la **surdispersion**, fréquente pour les rechutes de SEP.

La justification attendue est donc : le modèle est cohérent avec la nature des données et préférable à un Poisson trop rigide ou à un simple test binaire sur "au moins une rechute".

Question 2

... pourquoi n'y a-t-il pas de "p" accompagnant la phrase ?

Parce que cette phrase rapporte une **estimation descriptive dérivée** de la courbe de survie ou du modèle, pas nécessairement le test principal. Le p central porte déjà sur l'analyse principale. Répéter des p partout n'apporte rien et alourdit l'interprétation. Une proportion estimée à un temps donné se lit très bien avec son contexte et, idéalement, son intervalle de confiance.

Question 3

Pourquoi n'y a-t-il pas de « p » dans les comparaisons de survenue des effets indésirables ?

Comme dans beaucoup d'essais, la sécurité est surtout décrite **de manière exploratoire**. Les tableaux d'EI comportent énormément de comparaisons, souvent corrélées entre elles. Tester systématiquement chaque ligne donnerait une inflation de faux positifs et détournerait l'attention de la vraie question clinique : **quels EI sont graves, fréquents, plausibles et asymétriques entre bras ?**

Question 4

... pas d'analyses intermédiaires. Pensez-vous que ça ait été un bon choix ?

On peut défendre ce choix pour l'efficacité : dans une SEP rémittente, on n'attend pas forcément un signal très précoce justifiant une dépense alpha complexe. En revanche, l'absence d'analyses intermédiaires formelles d'efficacité ne dispense pas d'un **monitoring de sécurité** étroit, surtout avec un immunomodulateur. Donc : choix globalement acceptable pour l'efficacité, à condition qu'un comité indépendant surveille bien la sécurité.

6.6. 2018 RC

Article et contexte

Article support : **Pavord et al., NEJM 2017** et supplément, sur le **mepolizumab** dans la BPCO éosinophilique, avec les essais METREX et METRE0.

- critère principal : **annual rate of moderate/severe exacerbations** ;
- population enrichie ou stratifiée par phénotype éosinophilique ;
- analyses fondées sur un **modèle binomial négatif** ;
- gestion de multiplicité détaillée dans le supplément.

Questions posées dans l'annale

1. Quels sont les aspects positifs et négatifs du critère principal utilisé dans ces études ?
2. Décrire et justifier les modalités de dépense du risque de première espèce dans les deux études. Proposez une alternative. Dans METREX il y a 22200 patients, pourquoi est-il alors discutable de proposer les seuils de significativité de 4% et 1% dans chaque sous population (avec ou sans hyperéosinophilie) ?
3. L'analyse principale du critère principal repose sur un modèle binomial négatif. Une alternative aurait-elle été possible ?
4. ... "A greater effect ... among patients with higher blood eosinophil counts ..." Quels arguments les auteurs peuvent-ils utiliser pour étayer cette affirmation ? Qu'en pensez-vous ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Quels sont les aspects positifs et négatifs du critère principal utilisé dans ces études ?

Le critère principal est intéressant car il porte sur les **exacerbations** de BPCO, donc un événement cliniquement important, fréquent, et central pour le patient et pour le système de soins. Il utilise l'information répétée dans le temps.

Ses limites :

- une exacerbation "modérée" n'a pas la même gravité qu'une "sévère" ;
- la définition peut dépendre des pratiques de prescription ou d'hospitalisation ;
- le critère reste composite ;
- il ne résume pas à lui seul le vécu global, la qualité de vie ou la fonction respiratoire.

Question 2

Décrire et justifier les modalités de dépense du risque de première espèce ...

Dans METREX, l'hypothèse forte portait sur le **phénotype éosinophilique**, d'où l'attribution de 4% de l'alpha à cette comparaison et 1% à l'analyse globale mITT-ALL. L'idée est de mettre la plus grande part de l'erreur de type I là où la plausibilité biologique est la plus forte.

Pourquoi est-ce discutable ? Parce que les populations ne sont pas totalement indépendantes : la population éosinophilique est incluse dans la population globale, et les mêmes patients contribuent donc à plusieurs contrastes. Ce n'est pas une répartition aussi propre que s'il s'agissait de deux essais séparés.

Une alternative raisonnable aurait été :

- une hiérarchie stricte eosinophilic first, puis overall seulement si la première passe ;
- ou une stratégie unifiée de type gatekeeping / interaction biologique d'abord.

Question 3

L'analyse principale ... repose sur un modèle binomial négatif. Une alternative aurait-elle été possible ?

Oui, mais avec des limites :

- un **modèle de Poisson** avec correction robuste : plus simple, mais moins bon si la surdispersion est importante ;
- une analyse de **temps jusqu'à première exacerbation** : plus lisible, mais elle jette une partie de l'information ;
- un modèle de survie récurrente type **Andersen-Gill / frailty**.

Le binomial négatif reste ici un bon choix parce que le critère est un **nombre d'exacerbations** et qu'on sait que ces comptages sont souvent surdispersés.

Question 4

... "A greater effect ... among patients with higher blood eosinophil counts ..." Quels arguments ... ? Qu'en pensez-vous ?

Les auteurs peuvent invoquer :

- la plausibilité biologique ;
- des analyses préspecifiées par niveau d'éosinophiles ;
- une relation dose-réponse ou gradient de l'effet ;
- la cohérence avec l'asthme éosinophilique.

Mais il faut rester prudent : dès qu'on discute finement des sous-groupes et des seuils biologiques, la multiplicité et le risque de surinterprétation remontent. La bonne réponse est donc : **affirmation plausible et étayée, mais qui reste en partie dépendante d'analyses de sous-groupes**.

6.7. 2019 RC

Article et contexte

Article support : **Parker et al., Lancet 2018**, essai STAMPEDE sur la **radiothérapie de la tumeur primitive** chez des hommes avec cancer de la prostate métastatique nouvellement diagnostiqué.

- essai de phase 3, ouvert, multi-bras / multi-stades ;
- 2061 patients dans la comparaison SOC versus SOC + RT ;
- critère principal : overall survival ;
- analyses intermédiaires de failure-free survival prévues dans le cadre MAMS.

Questions posées dans l'annale

1. Une randomisation par minimisation est utilisée, à quoi cela correspond-il ?
2. L'essai n'est pas en aveugle, est-ce un problème ?
3. Pourquoi les auteurs utilisent-ils des « flexible parametric models »
4. Le protocole relève de l'approche MAMS. A quoi correspond cette approche, qu'en pensez-vous ?
5. Dans l'article les auteurs mentionnent « three intermediate analyses of failure free survival ». Qu'elle est la logique de ces analyses ?
6. Que pensez-vous du critère « Failure free survival » ?
7. La conclusion de l'étude repose avant tout sur une analyse en sous-groupe. Qu'en pensez-vous ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

La minimisation est une randomisation dynamique qui cherche à équilibrer les bras sur plusieurs facteurs pronostiques. Ici elle inclut un **élément aléatoire de 20%**, ce qui limite la prévisibilité de l'allocation.

Question 2

L'absence d'aveugle n'est pas idéale, mais le critère principal OS est objectif. Le risque majeur serait surtout sur des critères plus subjectifs ou sur certaines décisions cliniques intermédiaires.

Question 3

Les flexible parametric models permettent une modélisation plus souple de la fonction de risque et facilitent l'estimation de médianes, de moyennes de survie ou de survies à temps fixes. Ils complètent utilement l'analyse de Cox.

Question 4

MAMS signifie Multi-Arm Multi-Stage. C'est un design adaptatif efficace qui permet d'évaluer plusieurs stratégies dans une même plateforme avec des arrêts intermédiaires pour futilité ou activité. Intéressant, mais méthodologiquement lourd.

Question 5

Les trois analyses intermédiaires de FFS ont une logique de **tri précoce** : arrêter rapidement des options peu actives sans attendre l'OS finale. C'est une logique classique dans les plateformes de type MAMS.

Question 6

Le Failure-free survival est un critère acceptable comme critère d'**activité** et de pilotage d'un design adaptatif. Il est beaucoup plus discuté comme critère final centré patient.

Question 7

Un sous-groupe, même préspecifié, ne doit pas effacer un résultat global négatif. Ici, le signal dans le sous-groupe à faible burden métastatique est plausible et intéressant, mais il faut rester prudent dans la formulation des conclusions.

6.8. 2020 RC

Article et contexte

Article support : **Schnitzer et al., JAMA 2019**, essai randomisé en double aveugle sur le **tanezumab** dans l'arthrose de hanche ou de genou.

- 698 patients randomisés;
- 3 bras;
- 3 co-primary endpoints continus à 16 semaines.

Questions posées dans l'annale

1. Quels sont les co-critères principaux d'efficacité de l'étude ? Quels sont leurs forces et leurs faiblesses ?
2. Dans l'étude il est calculé un nombre de sujets à inclure égal à 3×230 . Proposez de votre côté un calcul de ce paramètre.
3. Si ces endpoints étaient à analyser séparément, on aurait pensé qu'un nombre de 165 sujets/bras aurait pu garantir la puissance de 90% à chaque fois. Mais les auteurs ont calculé un NSN de 230/bras. Pourquoi ?
4. ... "Tanezumab treatment groups were considered more effective than placebo if all 3 co-primary end points were statistically significant.". Cette approche est-elle habituelle ? Quels sont ses intérêts et inconvénients ?
5. Quelles sont les différences d'efficacité observées entre les bras actifs et le bras placebo pour chacun des co-critères principaux d'efficacité ? Interprétez l'importance de ces différences.
6. Dans la table 4 ... il n'y a pas de « p ». A votre avis pourquoi ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Les trois co-critères sont l'évolution du WOMAC Pain, du WOMAC Physical Function, et du PGA-0A. C'est pertinent cliniquement car on couvre douleur, fonction et perception globale. Les limites sont la subjectivité, la faible sensibilité de certaines échelles et l'absence de marqueur plus objectif.

Question 2

Si l'on fait un calcul naïf à deux moyennes sur chaque endpoint séparément, on tombe sur un besoin d'environ 140 à 174 sujets par bras selon le critère, donc autour de 165 par bras. C'est la bonne base de raisonnement.

Question 3

Le NSN est plus grand car l'essai n'est positif que si **les trois co-primary endpoints** sont tous significatifs. La corrélation entre critères doit donc être prise en compte. D'où un effectif plus élevé, proche de 230 par bras.

Question 4

Cette approche est moins fréquente qu'un critère composite, mais elle a une vraie logique : exiger une amélioration concordante sur plusieurs dimensions importantes. L'avantage est la cohérence clinique ; l'inconvénient est le coût en puissance et la complexité.

Question 5

Les écarts vs placebo sont statistiquement positifs mais modestes : environ $-0,6$ à $-0,73$ sur la douleur, $-0,66$ à $-0,89$ sur la fonction, et $-0,22$ à $-0,25$ sur l'évaluation globale. Il faut donc parler d'un bénéfice probablement **discret** sur le plan clinique.

Question 6

La sécurité est présentée de manière descriptive. Multiplier des tests dans une table d'effets indésirables n'est généralement pas la bonne stratégie ; on regarde surtout la fréquence, la gravité et la plausibilité clinique.

6.9. 2021 RC

Article et contexte

Article support : **Roy et al., PLoS Medicine 2020**, essai **cluster randomisé apparié** sur les **adherence clubs** dans des cliniques VIH en Zambie.

- 10 cliniques ;
- critère principal : temps jusqu'au premier retrait tardif (>7 jours) ;
- composante qualitative intégrée.

Questions posées dans l'annale

1. Dans un essai, la question de la pertinence clinique du critère principal d'efficacité est essentielle. Quel est ce critère ici et qu'en pensez-vous ?
2. Si le design avait été celui d'un essai randomisé standard, combien de patients aurait-il fallu inclure pour obtenir une puissance statistique « convenable » ?
3. Quel est l'intérêt, ici, de faire un essai randomisé en cluster ?
4. ... pourquoi les auteurs parlent-ils de facteurs de confusion potentiels alors que l'essai est randomisé ?
5. Décrivez de façon synthétique les principales étapes de la méthodologie utilisée pour la partie qualitative de l'étude ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Le critère principal est le **temps jusqu'au premier retard de retrait du traitement de plus de 7 jours**. C'est un très bon critère de mise en œuvre et d'observance, proche de la logique de l'intervention, mais il reste indirect vis-à-vis du bénéfice clinique ultime.

Question 2

En oubliant le cluster et en ramenant le problème à un essai individuel à deux proportions avec un risque contrôle autour de 60% et un effet attendu de division par deux, on obtient **un peu moins de 50 sujets par bras**.

Question 3

Le cluster est pertinent ici car l'intervention est **organisationnelle et collective**. Une randomisation individuelle exposerait à une forte contamination et ne refléterait pas la vraie mise en œuvre du soin.

Question 4

La randomisation porte sur les centres, pas sur les patients. Avec peu de clusters, des déséquilibres au niveau individuel peuvent persister ; parler de facteurs de confusion potentiels reste donc défendable.

Question 5

La partie qualitative repose sur un échantillonnage raisonné, des **focus groups** et des **in-depth interviews**, puis un codage et une analyse selon un **cadre thématique** pour documenter acceptabilité, faisabilité et appropriation.

6.10. 2022 RC

Article et contexte

Article support : **Bacharier et al., NEJM 2021**, essai randomisé en double aveugle de **dupilumab** chez des enfants avec asthme modéré à sévère non contrôlé.

- critère principal : taux annualisé d'exacerbations sévères ;
- modèle principal : **binomial négatif** ;
- deux populations principales d'efficacité.

Questions posées dans l'annale

1. Quel est le critère principal d'efficacité ? Quelles sont ses force(s) et faiblesses(s) ?
2. Quel autre ou quels autres critère(s) d'efficacité aurai(en)t-pu être envisagé(s) ?
3. Expliquez les modalités de calcul du nombre de sujets nécessaires
4. ... les intervalles de confiance sont donnés à 95%. Est-ce que cela vous choque ?
5. ... Quel type de design serait envisageable pour alléger l'étude chez l'enfant ? expliquez brièvement.

Extrait de l'article à analyser

Corrections détaillées

Question 1

Le critère principal est le **taux annualisé d'exacerbations sévères**. C'est clinique, fréquent et très utilisé. Ses limites sont la possible hétérogénéité des définitions et le fait qu'il ne résume pas tout l'impact sur la vie quotidienne.

Question 2

On pouvait discuter la qualité de vie de l'enfant, le fardeau parental, les jours d'école manqués, les jours sans asthme, voire certains paramètres respiratoires.

Question 3

Le NSN repose ici sur un **modèle de comptage avec surdispersion**. Il faut donc citer les ingrédients : α , puissance, taux de base d'exacerbation, effet relatif attendu, allocation 2 : 1, durée d'exposition moyenne, et paramètre de dispersion.

Question 4

Ce n'est pas choquant. Les tests confirmatoires hiérarchisés et les intervalles de confiance 95% n'ont pas exactement le même statut ni le même but de communication.

Question 5

Une option raisonnable serait un **emprunt bayésien** des données adulte/adolescent, ou un dispositif réglementaire plus léger avec fort ancrage pharmacologique et suivi post-autorisation.

6.11. 2023 RC

Article et contexte

Article support : **Edwardson et al., BMJ 2022**, essai **cluster randomisé à trois bras** chez des salariés de bureau.

- 78 clusters ;
- 756 participants ;
- critère principal : temps assis quotidien à 12 mois ;
- modèle principal linéaire mixte.

Questions posées dans l'annale

1. Expliquez brièvement les avantages et les inconvénients d'un essai randomisé en cluster
2. Si l'essai avait randomisé des sujets et non des clusters, quel aurait été le nombre de sujets à inclure ?
3. L'analyse principale a été faite « on a complete case basis ». Discutez cette option.
4. Détaillez et justifiez le modèle utilisé pour l'analyse principale du critère principal
5. Dans l'abstract les auteurs concluent ... Qu'en pensez-vous ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Les avantages : moins de contamination, meilleure faisabilité, cohérence avec l'intervention réelle. Les inconvénients : perte d'efficacité, analyse plus complexe, besoin de modèles mixtes, difficulté de l'ITT au niveau individuel.

Question 2

En retirant l'effet cluster, on obtient environ 50 sujets par bras, puis 150 au total pour 3 bras, enfin environ 195 après majoration de 30%. Il faut répondre **environ 200 sujets au total**.

Question 3

L'analyse en complete case simplifie mais expose à un biais si les données manquantes ne sont pas ignorables. Elle est plus défendable ici que dans un RCT individuel pur, mais reste une faiblesse.

Question 4

Le modèle principal est un **modèle linéaire mixte** avec groupe en effet fixe, cluster en effet aléatoire, variables de stratification, valeur baseline du critère et variables liées au port du dispositif. C'est cohérent avec le design.

Question 5

La conclusion de l'abstract est trop affirmée : les différences sont petites, multiples, et non protégées de manière très convaincante. Il faut parler de **surinterprétation**.

6.12. 2024 RC

Article et contexte

Article support : **Moncrieff et al., Lancet Psychiatry 2023**, essai RADAR comparant réduction/discontinuation d'antipsychotiques versus maintien.

- essai ouvert, groupes parallèles ;
- évaluateurs autant que possible masqués ;
- critère principal : Social Functioning Scale ;
- raisonnement secondaire de non-infériorité sur les rechutes sévères.

Questions posées dans l'annale

1. Dans les essais randomisés en psychiatrie la question de l'aveugle est souvent considérée comme particulièrement importante. Comment est gérer cette question ici ?
2. Le critère principal est la « Social Functioning Scale ». êtes-vous satisfait de ce choix ? ...
3. Pensez-vous qu'il aurait été souhaitable de prévoir des analyses intermédiaires ?
4. Essayez, dans la mesure du possible, de retrouver le « Nombre de Sujets Nécessaires » ... de non-infériorité.
5. Le design de cette étude est assez inhabituel ... qu'est ce que cette étude était susceptible d'apporter comme conclusion ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

L'aveugle complet était pratiquement impossible. Les auteurs ont donc masqué ce qui pouvait l'être : évaluateurs, adjudication et analyse. C'est imparfait mais probablement le mieux faisable.

Question 2

Le SFS a un intérêt réel car la question clinique porte aussi sur le fonctionnement social. On peut toutefois discuter sa sensibilité au changement et sa lisibilité clinique ; un critère de rechute aurait été plus directement parlant.

Question 3

Des analyses intermédiaires d'efficacité avaient peu d'intérêt. En revanche, des analyses de **sécurité** centrées sur le risque de rechute auraient été défendables.

Question 4

En utilisant la formule simplifiée de non-infériorité avec une marge de 10%, on retrouve un ordre de grandeur autour de **360 à 372 sujets**, cohérent avec le papier.

Question 5

Le design pouvait théoriquement conclure à une amélioration fonctionnelle sans excès cliniquement inacceptable de rechutes sévères. Mais même dans le meilleur des cas, l'arbitrage bénéfice/risque restait délicat.

6.13. 2025 RC

Article et contexte

Article support : **Hautzinger et al., JAMA Psychiatry 2024**, essai randomisé comparant deux psychothérapies adjuvantes de prévention de rechute dans le trouble bipolaire euthymique.

- 305 participants ;
- SEKT versus FEST ;
- critère principal : temps jusqu'à rechute ;
- log-rank et Cox ;
- analyse complémentaire avec propensity score matching.

Questions posées dans l'annale

1. Cet article porte sur un essai randomisé. Les auteurs proposent pourtant une analyse faisant appel à un score de propension. Pourquoi et qu'en pensez-vous ?
2. L'étude compare deux stratégies thérapeutiques : « SEKT » et « FEST ». Quel est la rationnel à l'origine de cette comparaison ?
3. L'étude porte sur deux traitements non pharmacologiques pour lesquels l'effet thérapeute peut être important. Comment les investigateurs le gèrent-ils ? Qu'en pensez-vous ?
4. A l'aide des données disponibles dans l'article, essayez de retrouver au moins approximativement une estimation du nombre de sujets à inclure dans l'essai.
5. Dans la conclusion ... "SEKT and FEST were equally effective ..." Qu'en pensez-vous ?

Extrait de l'article à analyser

Corrections détaillées

Question 1

Les auteurs utilisent un score de propension non pas pour remplacer la randomisation, mais pour **tester la robustesse** des résultats si certains déséquilibres résiduels existent ou si l'analyse porte sur des sous-échantillons moins bien équilibrés. Ce n'est pas l'analyse principale la plus naturelle dans un RCT, mais comme **analyse de sensibilité** ce n'est pas absurde.

Question 2

Le rationnel est d'opposer une thérapie **structurée, orientée compétences et matériel** (SEKT) à une thérapie **soutenante, centrée patient et émotion** (FEST). Ce sont deux logiques psychothérapeutiques plausibles, pas simplement actif versus placebo.

Question 3

L'effet thérapeute est géré par des manuels, une formation des thérapeutes aux deux approches, de la supervision, et une évaluation de l'adhérence aux protocoles. C'est bien pensé, mais cela n'annule jamais totalement l'effet thérapeute.

Question 4

En ramenant l'hypothèse 50% versus 70% de patients sans rechute à un calcul binaire naïf à un an, on trouve environ 93 sujets par bras, puis environ 106 après correction de 12% de dropout. Comme le papier est en réalité dimensionné en **survie/log-rank** avec recrutement étalé et censure, le bon ordre de grandeur final reste **environ 150 par bras**, soit 300 au total.

Question 5

Dire que SEKT et FEST sont "equally effective" est un peu trop fort si l'essai n'est pas formellement un essai d'équivalence. La formulation plus prudente est : **aucune différence claire n'a été mise en évidence** entre les deux approches dans ce cadre précis.

6.14. 2026 RC**Article et contexte**

Article support : **Fang et al., NEJM 2026**, essai de phase 3 OptiTROP-Lung04 comparant **sacituzumab tirumotecan** à une chimiothérapie standard dans le NSCLC avancé muté EGFR résistant aux TKI.

- phase 3, multicentrique, randomisé 1:1, ouvert ;
- 376 patients randomisés (188 par bras) ;
- critère principal : progression-free survival (PFS) évaluée par BICR ;
- critère secondaire hiérarchisé clé : overall survival (OS) ;
- essai dimensionné pour 249 événements de PFS et environ 356 sujets ;
- sujet d'examen exhaustif avec 35 questions.

Questions posées dans l'annale

1. Décrivez le design de l'essai OptiTROP-Lung04. Justifiez chaque choix méthodologique (phase, randomisation, ouverture, contrôle).
2. Quelles sont les populations d'analyse utilisées dans l'essai ? Distinguez ITT, PP, et population de sécurité. Discutez leur pertinence.
3. Les patients atteints de métastases cérébrales sont inclus dans l'essai. Discutez les implications méthodologiques et cliniques de ce choix.
4. Cet essai est conduit uniquement en Chine avec une population 100% asiatique. Quelles sont les conséquences pour la généralisabilité des résultats ?
5. Expliquez le rôle du BICR (Blinded Independent Central Review) dans cet essai ouvert. En quoi est-il indispensable ?

6. Retrouvez et justifiez le calcul du Nombre de Sujets Nécessaires (NSN) présenté dans l'article. Vérifiez si les hypothèses sont cohérentes.
7. Expliquez la stratégie de dépense du risque de 1ère espèce (alpha-spending) utilisée dans cet essai. Justifiez les seuils α retenus.
8. Les résultats de la SSP sont présentés par la méthode de Kaplan-Meier et le modèle de Cox. Expliquez ce que chacun apporte et leurs hypothèses.
9. Pour le taux de réponse objective (ORR), les auteurs utilisent un intervalle de confiance de Clopper-Pearson et le test de Cochran-Mantel-Haenszel (CMH). Expliquez ces choix.
10. Les données manquantes pour les critères de survie sont gérées sous l'hypothèse de "censure non informative". Discutez cette hypothèse dans le contexte de cet essai.
11. Quel est le critère principal d'efficacité ? Quelles sont ses forces et faiblesses dans le contexte d'un essai oncologique de phase 3 ?
12. Auriez-vous préféré un autre critère principal ? Discutez notamment la survie globale (OS) comme critère principal alternatif.
13. Commentez la significativité statistique versus la pertinence clinique des résultats de SSP et SG obtenus dans cet essai.
14. Qu'est-ce que le critère de réponse RECIST 1.1 ? Détaillez les catégories de réponse et comment l'ORR est calculé.
15. Analysez les résultats des sous-groupes prédéfinis de SSP et SG. Comment interpréter un forest plot de sous-groupes ?
16. Les résultats d'OS sont présentés à une analyse intermédiaire. Que pensez-vous de rapporter des résultats d'OS à une analyse intermédiaire plutôt qu'à l'analyse finale ?
17. Interprétez le Hazard Ratio de 0,49 (IC 95% : 0,39–0,62) pour la SSP. Qu'est-ce qu'un Hazard Ratio ? Quelles sont ses limites d'interprétation ?
18. Commentez les données du tableau des caractéristiques initiales (Table 1). Y a-t-il des déséquilibres entre les deux groupes ?
19. Interprétez les résultats de survie globale (OS). Pourquoi la SG médiane n'est-elle pas atteinte dans le bras sac-TMT ?
20. Analysez les résultats de qualité de vie (PRO). Leur intégration est-elle satisfaisante dans cet essai ?
21. Dans la conclusion, les auteurs déclarent que "sac-TMT could be considered a favorable treatment option before pemetrexed plus platinum-based chemotherapy". Qu'en pensez-vous ?
22. Analysez le profil de sécurité du sac-TMT. Y a-t-il des différences qualitatives importantes avec la chimiothérapie ?
23. Pourquoi n'y a-t-il pas de valeur p dans les tableaux comparant les effets indésirables entre les groupes ?
24. Comment les auteurs gèrent-ils la stomatite liée au sac-TMT ? Discutez l'impact des réductions de doses sur l'interprétation des résultats.
25. Comment comparer le profil de sécurité de sac-TMT avec celui d'autres anticorps-médicament conjugués (ADC) dans le NSCLC ?
26. Décrivez le mécanisme d'action du sacituzumab tirumotecan en tant qu'ADC. Pourquoi cible-t-il spécifiquement les cellules NSCLC EGFR-mutées résistantes ?
27. Expliquez le rationnel des mutations EGFR dans le NSCLC. Quelles mutations sont incluses et pourquoi ?
28. Comparez sac-TMT avec les autres traitements de 2e ligne disponibles pour les NSCLC EGFR-mutés résistants aux TKI (HARMONi-A, ORIENT-31, MARIPOSA-2).
29. Discutez les aspects éthiques de cet essai : consentement, comité indépendant, Déclaration de Helsinki.
30. L'essai est financé par le fabricant (Sichuan Kelun-Biotech). Quels sont les risques et les mécanismes de contrôle ?
31. Discutez la question de la généralisation éthique des résultats : est-il éthique d'utiliser sac-TMT

- pour des patients non-asiatiques sur la base de cet essai ?
32. Quelles sont les implications de ces résultats pour la politique de santé publique en oncologie pulmonaire ?
 33. Discutez la place du sac-TMT dans l'algorithme décisionnel de traitement du NSCLC EGFR-muté. Comment s'intègre-t-il dans le parcours patient ?
 34. Quelles sont les principales forces et limites méthodologiques de l'essai OptiTROP-Lung04 ?
 35. Si vous deviez améliorer la conception de cet essai, quelles modifications proposeriez-vous ? Quelles études futures seraient nécessaires ?

Extrait de l'article à analyser

Ce que l'examinateur veut vérifier

- lecture d'un essai de phase 3 oncologique complet ;
- maîtrise des bases de survie, de multiplicité et de sécurité ;
- capacité à articuler **statistique, clinique, pharmacologie, éthique et santé publique**.

Corrections détaillées

Question 1

L'essai est un **phase 3**, multicentrique, randomisé 1:1, ouvert, à comparateur actif. Chaque choix est cohérent : phase 3 pour confirmer un signal antérieur, comparateur actif car la chimiothérapie reste standard, ouvert car les schémas d'administration diffèrent fortement, randomisation centralisée et stratifiée pour équilibrer les facteurs pronostiques majeurs.

Question 2

La population d'**efficacité** est l'ITT : tous les patients randomisés. La **population de sécurité** comprend les patients ayant reçu au moins une dose. Une population PP peut être explorée, mais n'est pas prioritaire ici. En oncologie confirmatoire, ce choix est classique et robuste.

Question 3

Inclure des patients avec métastases cérébrales stabilisées améliore la **représentativité clinique**. En contrepartie, cela ajoute de l'hétérogénéité pronostique et nécessite de bons facteurs de stratification ou des analyses de sous-groupes.

Question 4

La validité interne n'est pas remise en cause, mais la **validité externe** est limitée : pharmacogénétique, habitudes de prise en charge, ligne thérapeutique précédente et profil de patients peuvent différer hors Chine. Il faut donc éviter de sur-généraliser.

Question 5

Le BICR est indispensable parce que le critère principal (PFS) dépend d'une lecture d'imagerie et qu'un essai ouvert expose à un biais d'évaluation. Une revue centrale indépendante et en aveugle réduit fortement ce risque.

Question 6

Le calcul repose sur une logique de Schoenfeld. Avec :

$$HR = 0,67, \quad \alpha = 0,025 \text{ unilatéral}, \quad \text{puissance} = 88\%$$

on obtient environ 249 événements. Sous une hypothèse de rendement d'événements d'environ 70%, cela donne environ 356 sujets. Les chiffres de l'article sont cohérents.

Question 7

L'essai utilise une dépense alpha de type **Lan-DeMets avec frontière O'Brien-Fleming**. L'idée est de préserver le risque global de première espèce malgré les analyses intermédiaires, en étant très conservateur tôt et moins conservateur tard.

Question 8

Kaplan-Meier décrit les courbes, les médianes et les proportions survivantes. Le modèle de Cox résume l'effet par un HR, éventuellement ajusté, sous hypothèse de proportionnalité des risques. Les deux sont complémentaires, pas interchangeables.

Question 9

Le Clopper-Pearson est un intervalle exact pour une proportion binaire comme l'ORR. Le test CMH est pertinent quand on veut comparer des proportions en tenant compte de la **stratification** utilisée à la randomisation.

Question 10

La censure non informative suppose que le fait d'être censuré n'est pas lié au risque futur conditionnel d'événement. C'est plausible dans bien des cas, mais pas garanti dans un essai ouvert si l'arrêt ou le changement de prise en charge sont liés à l'évolution clinique.

Question 11

Le critère principal est la PFS évaluée par BICR. Ses forces : rapidité, puissance, sensibilité au signal antitumoral, usage standard en oncologie. Ses limites : critère de substitution, interprétation clinique moins forte que l'OS, sensibilité à la fréquence des évaluations.

Question 12

On peut défendre l'OS comme critère principal plus "dur" et plus parlant cliniquement. Mais en seconde ligne oncologique, attendre l'OS demande plus de temps et de décès, avec risque de dilution par traitements ultérieurs. La PFS est donc un compromis défendable.

Question 13

Les résultats sont à la fois **statistiquement très significatifs** et **cliniquement importants** : HR 0,49 sur la PFS et HR 0,60 sur l'OS sont de grosses tailles d'effet dans ce contexte. La bonne copie doit citer les deux dimensions, pas seulement la p-value.

Question 14

RECIST 1.1 classe les réponses en CR, PR, SD, PD. L'ORR correspond à :

$$ORR = \frac{CR + PR}{\text{patients évaluable}}$$

Le DCR ajoute la stabilité (SD), ce qui en fait une mesure plus large mais moins spécifique.

Question 15

Un forest plot de sous-groupes s'interprète d'abord par la **cohérence visuelle des effets**, puis par la crédibilité biologique et éventuellement les tests d'interaction. Il ne faut pas surinterpréter une différence apparente entre sous-groupes si les intervalles sont larges et si l'interaction n'est pas robuste.

Question 16

Rapporter une OS à une analyse intermédiaire est acceptable si cela était **prévu** et si la dépense alpha est correctement contrôlée. La limite est la maturité encore incomplète des données, mais l'information peut être cliniquement très utile.

Question 17

Un HR = 0,49 signifie qu'à chaque instant, le risque de progression ou décès est environ réduit de moitié dans le bras sac-TMT. Limites : ce n'est pas un ratio de médianes, cela suppose une certaine stabilité de l'effet dans le temps, et le HR reste parfois abstrait pour le clinicien.

Question 18

Le tableau 1 semble globalement équilibré. C'est ce qu'on attend d'une randomisation bien menée. Il faut le lire de façon descriptive et vérifier l'absence de déséquilibre clinique massif, pas rechercher des p-values.

Question 19

La médiane d'OS n'est pas atteinte dans le bras sac-TMT parce qu'au moment de l'analyse intermédiaire, moins de 50% des patients de ce bras étaient décédés. Cela reflète un bénéfice réel possible, mais aussi l'immaturation relative des données de survie globale.

Question 20

L'intégration des PRO est plutôt satisfaisante : instruments standardisés, taux d'évaluabilité élevés, lecture complémentaire à l'efficacité. Mais ces résultats restent souvent secondaires et plus vulnérables aux biais dans un essai ouvert.

Question 21

La conclusion des auteurs est globalement cohérente avec les résultats, mais elle doit être tempérée par trois réserves : population exclusivement chinoise, essai ouvert, et besoin de confirmation de la place exacte dans la séquence thérapeutique globale.

Question 22

Le profil de sécurité de sac-TMT n'est pas simplement "plus toxique" ou "moins toxique" que la chimiothérapie; il est **différent**. Certains EI sont comparables ou plus fréquents, d'autres sont moins sévères, et le profil qualitatif d'un ADC doit être décrit séparément.

Question 23

Il n'y a pas de p dans les tableaux de sécurité parce que la sécurité est généralement décrite de façon **exploratoire**. Tester chaque EI un par un serait peu informatif et exposerait à une inflation majeure des faux positifs.

Question 24

La stomatite est gérée par adaptation de dose et mesures de soutien. Les réductions de dose compliquent l'interprétation causale stricte de "dose pleine", mais elles font partie intégrante de l'évaluation pragmatique d'un traitement en oncologie. L'important est de raisonner en ITT puis de commenter la faisabilité.

Question 25

Comparer avec d'autres ADC impose de raisonner en **profil de classe** et non en simple fréquence brute. Certains ADC exposent surtout à la pneumopathie interstitielle, d'autres à des toxicités hématologiques ou muqueuses. La comparaison doit rester prudente si elle n'est qu'indirecte.

Question 26

Le sacituzumab tirumotecan est un **anticorps-médicament conjugué** ciblant Trop-2, qui délivre un dérivé inhibiteur de topoisomérase I dans la cellule tumorale après internalisation. Le rationnel est renforcé dans les NSCLC mutés EGFR résistants, où Trop-2 est fortement exprimé.

Question 27

Les mutations incluses sont les mutations sensibilisantes usuelles, surtout exon 19 deletion et L858R. Ce sont les altérations qui définissent le sous-groupe clinique traité par EGFR-TKI et donc concerné par la résistance secondaire étudiée ici.

Question 28

Par rapport à d'autres options comme HARMONi-A, ORIENT-31 ou MARIP0SA-2, sac-TMT se positionne comme une option potentielle supplémentaire après échec des EGFR-TKI. Les comparaisons restent indirectes; il faut éviter toute hiérarchisation trop affirmative sans tête-à-tête.

Question 29

Les éléments éthiques essentiels sont là : consentement écrit, comité indépendant de monitoring, approbations réglementaires, respect de la Déclaration d'Helsinki. Le comparateur actif est particulièrement important sur le plan éthique.

Question 30

Le financement industriel expose à un risque de biais de conception, d'analyse et de communication. Les garde-fous sont la transparence, l'accès aux données, un IDMC indépendant, un protocole public, et une lecture critique attentive des formulations de conclusion.

Question 31

Il n'est pas "interdit" d'utiliser ces résultats pour des patients non asiatiques, mais ce ne serait pas rigoureux de les généraliser sans discussion. Éthiquement, cela plaide pour des données complémentaires dans d'autres populations, ou au minimum pour une grande prudence de transposition.

Question 32

Les implications de santé publique touchent l'accès, le remboursement, le coût-efficacité, et l'organisation des lignes de traitement. Un bénéfice important sur PFS et OS peut modifier les référentiels, mais le coût et la généralisation hors Chine resteront déterminants.

Question 33

Le sac-TMT pourrait s'insérer avant ou à la place de certaines chimiothérapies standards dans la séquence après EGFR-TKI. Sa place exacte dépendra des comparaisons indirectes, de la toxicité, de la disponibilité et des standards locaux.

Question 34

Forces : randomisation, comparateur actif, BICR, cohérence des résultats, hiérarchie statistique. Limites : essai ouvert, population exclusivement chinoise, comparaisons indirectes avec d'autres stratégies, maturité relative de certaines données secondaires.

Question 35

On pourrait améliorer l'essai par :

- une extension internationale ;
- un aveugle partiel si techniquement possible sur certains volets ;
- davantage de données PRO et de qualité de vie ;
- des comparaisons plus directes avec d'autres standards contemporains ;
- un suivi plus long de OS et de sécurité.

7. Réponses prêtes à l'emploi

7.1. Quand il faut critiquer un critère principal

Le critère principal est pertinent car il est proche de la question clinique, mais il présente plusieurs limites : hétérogénéité de ses composantes, possible sensibilité aux pratiques locales, et interprétation parfois moins directe qu'un critère dur comme la survie globale.

7.2. Quand il faut expliquer l'absence de p dans un tableau de sécurité

Les tableaux de sécurité sont en général descriptifs. L'absence de p n'est pas choquante, car il existe de très nombreuses comparaisons et la question principale est la fréquence, la gravité, la plausibilité et la cohérence clinique des événements indésirables.

7.3. Quand il faut discuter une analyse en sous-groupe

Une analyse en sous-groupe peut être intéressante si elle était préspecifiée et biologiquement plausible, mais elle doit rester prudente. Elle ne doit pas, à elle seule, annuler ou renverser un résultat global.

7.4. Quand il faut retrouver un NSN approximatif

Le calcul exact n'est pas toujours parfaitement reconstituable à partir du seul article. En revanche, on peut souvent retrouver un ordre de grandeur cohérent en identifiant α , puissance, effet attendu, critère principal, variabilité ou taux de base, puis les ajustements liés au design.

8. Checklist finale

- Ai-je identifié le **vrai** critère principal ?
- Ai-je dit pourquoi le modèle statistique choisi était adapté ou discutable ?
- Ai-je distingué **validité interne** et **validité externe** ?
- Si un calcul était demandé, ai-je donné au moins un **ordre de grandeur défendable** ?
- Ai-je évité de confondre **absence de preuve** et **preuve d'absence** ?
- Ai-je conclu en langage d'examen et pas seulement en jargon statistique ?

9. Conclusion

Les annales RC récompensent surtout les copies qui savent articuler **design, statistique, clinique** et **lecture critique**. Une bonne réponse n'est pas forcément la plus longue ; c'est celle qui dit clairement ce que fait l'essai, pourquoi il le fait ainsi, ce que cela apporte, et ce qui reste discutable.