Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

De la nécessité de la méthodologie dans l'évaluation des médicaments

Document compagnon

Dossier 3 – Le contrôle du risque alpha global

14 février 2022

Comité de rédaction et relecture (par ordre alphabétique)

Jean Luc Cracowski

Michel Cucherat

Dominique Deplanque

Behrouz Kassai

Charles Khouri

Silvy Laporte

Clara Locher

Florian Naudet

Edouard Ollier

Matthieu Roustit

Table des matières

1 Introduction

Le contrôle strict du risque d'erreur statistique alpha est un élément fondamental pour obtenir un degré de certitude suffisant pour baser un éventuel changement de pratique à partir d'un résultat. Sans ce contrôle strict, tout essai, quelle que soit la réelle efficacité du médicament testé, pourrait donner à tort des résultats apparemment « statistiquement significatifs ». Le contrôle strict signifie que l'on connait parfaitement le risque de conclure à tort (généralement 5% bilatéral) à l'issue de l'essai clinique face à un résultat significatif. Cela signifie également que l'on refuse de conclure lorsque ce risque n'est pas maitrisé, connu, ou bien supérieur à 5%.

2 Risque alpha (*type I error rate*)

La détermination qu'un traitement à un effet sur un critère de jugement s'effectue en comparant la valeur du critère de jugement entre les 2 bras de l'essai pour chercher s'il y a une différence en faveur d'un effet bénéfique du traitement (moins de décès dans le groupe traité que dans le groupe contrôlé par exemple).

Une telle différence peut néanmoins survenir du fait des fluctuations aléatoires d'échantillonnage (liées purement au hasard) même si le traitement n'a aucun effet en réalité sur le critère de jugement. Si l'on ne prenait pas ce risque en compte, on pourrait conclure à tort à l'existence d'une différence dans les cas où le traitement n'a pas d'effet. C'est l'erreur statistique alpha (aussi appelé de première espèce ou de type 1, *type I error*) : conclure à tort à une différence qui n'existe pas en réalité.

Cette erreur statistique a de lourdes conséquences dans le cadre d'un essai clinique, car elle conduit à conclure à tort que le nouveau traitement apporte un bénéfice aux patients et donc conduit à le recommander et à l'utiliser indument en pratique.

Pour limiter au maximum cette possibilité catastrophique, on utilise un test statistique qui va limiter le risque de commettre une erreur alpha en deçà d'une valeur faible (on parle de contrôle du risque alpha), 5% en bilatéral en général. Ainsi dans le cas où le traitement n'a pas d'effet sur le critère considéré, on ne conclura à une différence que dans 5% des cas.

> ***N.B.** : En absence de prise en considération du risque d'erreur alpha, on serait amené à utiliser tous les traitements qui n'apportent pas de bénéfice car, du fait des fluctuations aléatoires d'échantillonnage, la moitié des résultats produits serait peu ou prou en faveur de la supériorité du nouveau traitement.*
>
> *En travaillant avec un seuil de la signification statistique bilatéral à 5%, on n'accepte plus que 2.5% des traitements sans effet (car, sous l'hypothèse nulle, seule la moitié des différences dues au hasard sont statistiquement significativement en faveur du nouveau traitement et conduisent à l'utiliser ; l'autre moitié suggérant une infériorité du nouveau traitement).*

L'utilisation de la signification statistique permet de réduire le risque de conclure à tort du fait du hasard, mais ne le réduit pas à zéro. Avec un seuil de signification à 5% bilatéral, on admet encore 2.5% des traitements sans effet.

Un résultat significatif ne signifie *pas qu'il est démontré avec certitude.*

2.1 Signification statistique et pertinence clinique

La signification statistique n'implique pas la pertinence clinique. Une réelle différence aussi petite soit-elle (et donc sans intérêt clinique) peut être rendue statistiquement significative en augmentant l'effectif de l'essai.

Un p très petit (« très significatif ») ne signifie pas que le traitement apporte un grand bénéfice en taille. Le paramètre conditionnant la pertinence clinique de l'importance du bénéfice est la taille de l'effet (*effect size, effect magnitude*).

> L'essai DAIS [10.1016/S0140-6736(00)04209-4] a évalué le fénofibrate en prévention secondaire chez des patients suivi pour une sténose d'une artère coronarienne. Le critère de jugement était l'évolution du diamètre de la sténose.
> Un résultat statistiquement significatif a été observé « a significantly smaller decrease in minimum lumen diameter (−0·06 [0·016] vs −0·10 [0·016] mm, p=0.02) ». La taille de l'effet est cependant très faible et correspond à une différence intergroupe de 0.04 mm. Cette taille d'effet est sans intérêt clinique et, qui plus, a été obtenue sur un critère intermédiaire, lui-même sans pertinence clinique.

2.2 Test unilatéral

Par habitude, des tests bilatéraux (*two-sided*) sont utilisés avec un seuil de la signification à 5%. Ces tests permettent de conclure, quel que soit le sens de la différence : supériorité ou infériorité du nouveau traitement par rapport au contrôle.

Or une seule de ces différences permettra de conclure à l'intérêt du nouveau traitement : la supériorité. Ainsi le risque alpha de conclure à tort à l'intérêt du nouveau traitement n'est que de 2.5%.

Des tests unilatéraux (*one-sided*), qui permettent uniquement de conclure à la supériorité, sont parfois utilisés. Dans ce cas, le seuil de la signification doit être impérativement à **2.5%** pour rester conforme à ce qui se passe avec un test classique bilatéral à 5% sur l'hypothèse de supériorité.

3 L'exploitation de l'erreur alpha pour obtenir à coup sûr des résultats positifs

L'erreur statistique alpha est couramment exploitée pour apporter des « apparentes » démonstrations de bénéfice avec des produits avec peu ou sans efficacité dans les domaines où il n'y a pas d'exigence réglementaire de démonstration formelle du bénéfice (compléments alimentaires, médecines alternatives, dispositifs médicaux, alicament, cosmétiques, homéopathie, etc.). Les comparaisons sont multipliées afin d'augmenter la « chance » d'obtenir un résultat faussement positif qui sera alors mis en avant comme preuve de l'efficacité. En effet, en l'absence de réelle efficacité du traitement, avec un seuil bilatéral à 5%, un résultat statistiquement significatif est attendu en moyenne tous les 20 critères de jugement indépendants, dont la moitié seront en faveur d'un bénéfice du traitement.

Il y a quelques années un yaourt revendiquait une action de renforcement des défenses immunitaires et une protection contre les infections saisonnières. Les publicités étaient accompagnées de la mention « prouvé clinique » et une référence bibliographique était mentionnée.[1].
Cette étude était un essai randomisé dans lequel 67 paramètres biologiques liés à l'immunité étaient comparés entre un groupe recevant le yaourt et un groupe contrôle : cellules immunocompétentes, cytokines, interférons, etc.

	Change over study[a]						
	Control group			Treatment group			Significance of difference between control and treatment groups
	Mean	N	SE	Mean	n	SE	
Leucocytes	154.74	57	271.49	554.14	70	263.30	> 0.05
Lymphocytes	−38.42	57	116.11	369.57	70	112.17	< 0.05
Monocytes	18.42	57	20.35	16.71	70	23.51	> 0.05
Granulocytes	174.56	57	228.99	162.71	70	198.09	> 0.05

[a] Change was calculated as study end minus baseline

Cells/mm³	Change over study[a]				
	Control group, n = 53		Treatment group, n = 69		Significance of difference between control and treatment groups
	Mean	SE	Mean	SE	
CD56	−51.97	21.33	17.29	17.27	< 0.05
CD2	89.43	91.94	271.46	76.46	> 0.05
CD3	130.78	83.33	196.66	66.47	> 0.05
CD4	146.70	57.52	158.67	42.01	> 0.05
CD8	−60.28	34.90	9.83	27.65	> 0.05
CD19	29.76	17.14	67.31	14.90	> 0.05

Pg/mL	Change over study						
	Control group			Treatment group			Significance of difference between control and treatment groups
	Mean	N	SE	Mean	N	SE	
IL-2	108.22	52	93.06	274.69	62	166.12	> 0.05
IFN-γ	10,263.29	52	4,587.60	13,274.14	62	5,520.00	> 0.05
TNF-α	−141.24	52	377.24	−38.59	63	414.44	> 0.05
IL-4	103.92	52	27.26	133.59	63	63.45	> 0.05
IL-5	110.67	52	34.60	149.47	63	95.75	> 0.05
IL-10	99.28	52	246.53	236.07	62	144.23	> 0.05

Aucun critère de jugement principal n'avait été défini à priori et l'obtention d'une différence statistiquement significative sur deux de ces 67 comparaisons a conduit les auteurs à faire la conclusion d'un bénéfice du yahourt.

Sur un tel nombre de comparaisons, avec un traitement sans aucun effet, il est quasi certain d'obtenir au moins un $p<0.05$ (le nombre attendu en moyenne est de 3). Ce type d'approche n'a donc aucune valeur scientifique vu qu'elle est en mesure de confirmer l'hypothèse dans tous les cas, quelle que soit la réalité de cette hypothèse.

4 Risque alpha global (overall type I error rate)

Il existe en fait deux niveaux de risque alpha.

Risque alpha nominal	Risque que l'on prend de conclure à tort à l'existence d'un effet du traitement au niveau d'un test particulier, dans le cas où le traitement n'a pas d'effet au niveau de ce test particulier.
Risque alpha global de l'essai	Risque que l'on prend de conclure à tort à un quelconque intérêt du traitement à l'issue de l'essai. C'est l'unique **risque alpha d'intérêt** dans l'essai thérapeutique, qu'il convient de garder strictement inférieur à 5% en bilatéral.

Le risque alpha au niveau d'un test (sur un critère de jugement par exemple) est le risque de conclure à tort à une différence au niveau de ce test particulier. Ce niveau correspond à la présentation classique du risque alpha dans les cours de statistique.

L'autre niveau est celui du risque alpha global de l'essai. Celui-ci est au centre de la problématique statistique de l'essai. Un essai est entrepris pour faire la conclusion de l'intérêt du traitement. Or cette conclusion va reposer sur un test statistique. Elle pourra donc être prise à tort du fait du hasard. C'est le risque alpha global de l'essai qui doit être parfaitement bien contrôlé à moins de 5% en bilatéral (2.5% en unilatéral sur la conclusion à l'intérêt du nouveau traitement).

> "To control the overall type I error, …"
> "to preserve the overall type I error rate at 0.05 (two-sided) after accounting for one interim analysis."

Si la conclusion de l'essai ne peut être faite qu'à partir d'un seul et unique test statistique, le risque de conclure à tort au niveau de l'essai est celui de conclure à tort au niveau du test.

Dans les essais modernes, on souhaite aller au-delà de la contrainte de la conclusion unique et pouvoir conclure à l'intérêt du traitement à partir de plusieurs tests (par exemple à partir de plusieurs critères de jugement, ou en effectuant plusieurs analyses et éventuellement en considérant des sous -groupes de patients). Il est donc nécessaire d'autoriser une multiplicité des comparaisons sans que cela entraine une inflation du risque alpha global.

> Statistical testing in the COMPASS study [10.1056/NEJMoa1709118 supplement] involves multiple testing in 3 main areas:
> 1. Multiple intervention comparisons: Rivaroxaban 2.5 mg bid + aspirin 100 mg od (rivaroxaban plus aspirin) compared to active control aspirin 100 mg od (aspirin); Rivaroxaban 5.0 mg bid (rivaroxaban) compared to active control aspirin 100 mg od (aspirin)
> 2. Multiple outcomes: One primary efficacy outcome and 3 key secondary efficacy outcomes.
> 3. Multiple decision points: A first interim analysis was to be conducted after approximately 50% of the target number of subjects had experienced an unrefuted primary efficacy outcome, a second interim analysis was to be conducted after approximately 75% of the target number of outcomes, and a final analysis was to be conducted after the target number of 2,200 unrefuted primary efficacy outcomes.
>
> Testing multiple hypotheses may increase the Type I error rate and we used a variety of statistical procedures to control the overall Type I error.

Cependant cette multiplicité des tests statistiques, où chacun pourrait conduire à la conclusion de l'intérêt du traitement, augmente le risque alpha global, même si les tests unitaires ont toujours le

même risque alpha nominal de 0.05. On parle d'inflation du risque alpha global (voir section suivante). On pourrait exposer la situation de façon provocante pour aider à la compréhension : je veux montrer que le traitement est efficace, il suffit de faire plusieurs tests et de conclure au premier test statistique significatif que l'on trouve, du fait du hasard.

> *En multipliant les tests, on augmente le risque de trouver au moins un test avec un p<0.05 même si le traitement n'a aucun effet. 5% c'est 1/20. Sous l'hypothèse nulle d'absence d'effet traitement, on s'attend à avoir en moyenne un p<0.05 tous les 20 tests réalisés, simple faux positif uniquement dû au hasard.*

Dans un essai thérapeutique, un résultat statiquement significatif signifie qu'il permet de conclure à l'intérêt du traitement avec un risque alpha global parfaitement bien contrôlé.

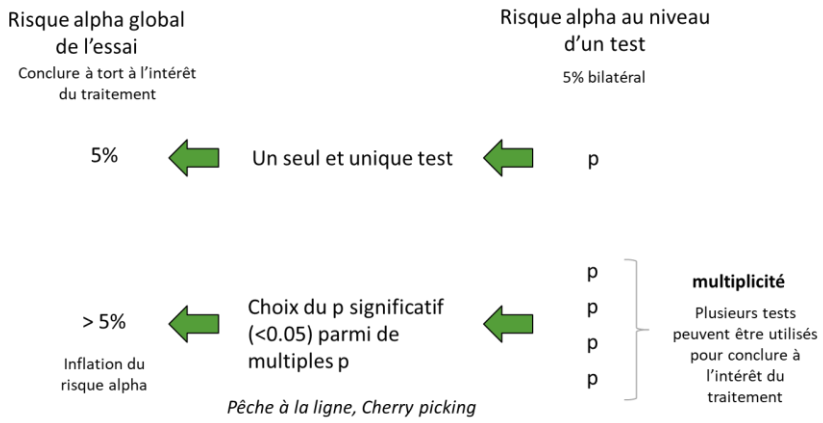
Ainsi des p<0.05 pourront ne pas être statistiquement significatifs, car il ne contrôle pas le risque alpha global[1].

---

[1] on peut dire à la rigueur qu'ils sont nominalement significatifs

Figure 1 – Les deux niveaux de risque alpha et les conséquences de la multiplication des tests pour conclure à l'intérêt du traitement

Le risque alpha global au niveau de l'essai (colonne de gauche) peut conduire à conclure à tort à l'intérêt du traitement et à recommander l'utilisation d'un traitement en réalité sans intérêt. C'est le risque alpha qui doit être parfaitement contrôlé dans l'essai. Quand on dit qu'un résultat est statiquement significatif dans un essai randomisé, cela signifie qu'il permet de conclure à l'intérêt du traitement avec ce risque alpha global parfaitement bien contrôlé. Cependant ce risque alpha global peut augmenter abusivement si cette conclusion est effectuée à partir de multiples tests (partie inférieure du schéma). Il faut donc mettre en œuvre des techniques particulières pour gérer la multiplicité et empêcher cette inflation du risque alpha global (cf. section 6).



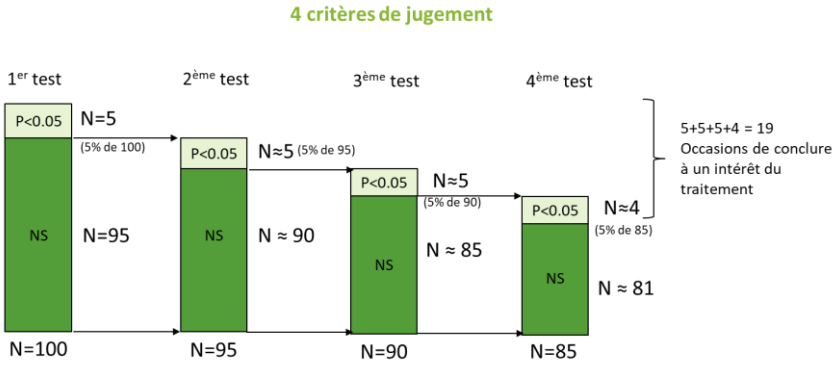
Signification statistique nominale	Contrôle le risque alpha de conclure à tort sur le test considéré	H0 : absence d'effet sur ce test particulier
Signification statistique en termes de risque alpha global de l'essai	Contrôle le risque alpha de conclure à tort à l'intérêt du traitement à l'issue de l'essai	H0 : absence d'intérêt du traitement, c'est-à-dire absence d'effet sur tous les tests qui sont réalisés

Contrairement à ce que l'on imagine couramment, la signification statistique ne renseigne en rien sur la réalité du résultat. La signification statistique ne cherche pas à déterminer quelle est la plausibilité de l'effet du traitement, elle évite seulement de conclure à tort trop fréquemment si le traitement n'a pas d'effet.

> *Le test d'hypothèse est en fait un outil bancal qui n'évalue pas la plausibilité que le traitement ait un effet », mais qui indique seulement qu'elle est la plausibilité d'obtenir un tel résultat du fait uniquement du hasard si le traitement n'a pas d'effet. Il existe une autre approche statistique, l'inférence bayésienne, qui répond directement à la question en donnant la probabilité que le traitement soit efficace compte tenu du résultat observé. Mais cette approche est aussi bancale car elle nécessite d'introduire une idée a priori du résultat. Pour éviter que le résultat obtenu ne dépende que de l'idée préconçue de l'investigateur et non pas des données, un a-priori non informatif (qui ne fait aucune hypothèse a priori sur l'efficacité du traitement) doit être impérativement utilisé.*

5 Multiplicité et inflation du risque alpha global

L'inflation du risque alpha gobal peut être illustrée de manière assez simple sans recours à une formule mathématique. Considérons un traitement sans aucun effet et imaginons que 100 essais randomisés versus placebo ont été réalisés. Sur ces 100 essais, on accepte donc de conclure à tort à l'intérêt du traitement dans cinq d'entre eux, mais pas plus ! (Risque alpha à 5%, on ne rentre pas dans la problématique du test bilatéral pour simplifier[2]). Imaginons aussi que 4 critères de jugement complètement différents sont analysés, et qu'il est possible de trouver un intérêt au traitement à partir du moment où un quelconque de ces 4 critères montre un effet du traitement à son niveau.



Parmi ces 100 essais réalisés, 5 d'entre eux auront un p<0.05 sur le 1er critère de jugement et permettront de conclure à l'intérêt du traitement. L'examen du 2ème critère aura lieu pour 95 essais (100-5). Parmi ces 95 essais, 5 auront un p<0.05 (5% de 95 ≈5) et permettront de conclure à l'intérêt du traitement. Cela laisse ≈90 essais qui sont négatifs sur le 1er et le 2ème critère. Parmi ces 90, ≈5 auront un p<0.05 sur le 3ème critère et permettront de conclure à l'intérêt du traitement et finalement il restera ≈85 essais pour lesquels le 4ème critère sera examiné et qui donneront ≈4 nouvelles occasions (5% de 85) de conclure à l'intérêt du traitement. Au total, globalement, il y aura eu 5+5+5+4=19 occasions de conclure à tort à l'intérêt d'un traitement qui en est dépourvu en réalité. Le risque alpha global est donc de 19/100 = 19%, ce qui montre l'importance du processus d'inflation du risque alpha global lorsqu'une multiplicité (*multiplicity*) des comparaisons statistiques est présente.

Maintenant, comment pourrait-on toujours continuer à envisager 4 critères pour déterminer l'intérêt du traitement sans que cela n'induise d'inflation du risque alpha global. Une solution simple est de ne retenir que les tests où le p est inférieur à 5%/4 = 1.25%. Cela conduira à donner lors du 1er test que 1.25 occasion de conclure à l'intérêt du traitement, puis 1.25% * (100-1.25) ≈1.25 nouvelles occasions lors de l'examen du 2ème test, puis encore 1.25 et 1.25 pour le 3ème et 4ème test. Au total il y aura 1.25+1.25+1.25+1.25 = 5 occasions de conclure à l'intérêt du traitement à tort, soit un risque alpha global de 5%.

[2] Dans les démonstrations par l'exemple données dans ce document, des approximations simplificatrices sont susceptibles d'être faites dans un but pédagogique. Toute la complexité mathématique du problème sous-jacent n'est pas abordée pour éviter de noyer le lecteur dans des détails inutiles pour la compréhension générale et l'appropriation des concepts (comme, entre autres, l'indépendance en probabilité des tests multiples, le risque alpha de l'hypothèse de supériorité qui n'est que de 2.5% et non pas 5%, etc.). Le lecteur ayant une expertise en statistique comprendra le pourquoi de ces simplifications compte tenu du public visé.

Diviser le risque alpha global par le nombre de tests induit par la multiplicité (méthode de Bonferroni) permet ainsi d'éviter l'inflation tout en autorisant la multiplicité. Cependant pour cela la règle de décision change ($p<0.0125$) et la signification statistique n'est plus $p<0.05$.

Une autre façon d'éviter l'inflation est de limiter à un seul le nombre de test permettant de faire la conclusion recherchée. C'est le principe du critère de jugement principal unique, mais dont l'usage a été progressivement abandonné depuis 2010. En effet, cette approche est très contraignante car, au cours d'une étude, elle permet seulement de démontrer l'intérêt d'un traitement sur un seul critère ; pourtant, il est des situations où l'on peut espérer davantage qu'un bénéfice unique.

6 Technique de contrôle du risque alpha global gérant la multiplicité

6.1 Répartition

Dans la méthode par répartition, le risque alpha global est réparti entre les différents critères de jugement. Il peut s'agir d'une équi-répartition ou non, cela n'a pas de conséquence.

Cette approche est parfois appelée co primary enpoints pour bien insister sur le fait que les 2 critères auront le statut de critère de jugement principal, c'est-à-dire permettant de faire des démonstrations. La répartition peut se faire entre 2 ou plusieurs critères de jugement

Pour pouvoir conclure sur un critère de manière statistiquement significative, il faudra que la p-value soit inférieure au risque alpha attribué à ce critère.

Ainsi il apparait que la signification statistique en termes de risque alpha global n'est plus du tout synonyme de p inférieur à 0.05.

L'avantage de la méthode par répartition et de pouvoir conclure sur l'un **ou** l'autre des critères de jugement, indépendamment, en fonction de la valeur de p obtenue

> Dans cet essai d'oncologie, 3 critères de jugements (OS, PFS et ORR[3]) étaient envisagés pour chercher un quelconque intérêt au traitement évalué (association nivolumab et ipilumab). Le risque alpha global de trouver un intérêt à tort au nivolumab plus ipilumab a été réparti entre ces 3 critères :
> « *The coprimary end points were overall survival (alpha level, 0.04), objective response rate (alpha level, 0.001), and progression-free survival (alpha level, 0.009) among patients with intermediate or poor prognostic risk*" [10.1056/NEJMoa1712126]
> Pour être significatifs (et permettre de conclure à l'intérêt du nivolumab plus ipilumab), les p obtenus pour chaque critère (p nominal) doivent être inférieur au risque alpha attribué au critère.
> Les résultats obtenus sont : « *overall survival rate was 75% with nivolumab plus ipilimumab and 60% with sunitinib (hazard ratio for death, 0.63; P<0.001). The objective response rate was 42% versus 27% (P<0.001. The median progression-free survival was 11.6 months and 8.4 months, respectively (hazard ratio for disease progression or death, 0.82; P=0.03, not significant per the prespecified 0.009 threshold).*"
> Pour interpréter ce type d'analyse statistique, le plus simple est de faire un tableau du type :

[3] OS : survie globale, PFS : survie sans progression, ORR : réponse objective (réduction de taille de la masse tumorale

Critère	Risque alpha attribué (seuil de signification)	P nominal	Verdict
OS	0.04	P<0.001	Significatif, permet de conclure à l'intérêt du nivolumab plus ipilumab en raison d'un bénéfice statistiquement démontré sur la survie
PFS	0.009	P=0.03	Non significatif
ORR	0.001	P<0.001	Significatif permet de conclure à un effet démontré, mais ce critère n'a que très peu de pertinence clinique. Si la survie n'avait pas été significative, ce résultat démontré n'aurait pas été suffisant pour justifier l'utilisation du traitement.

Dans les essais modernes, la signification statistique n'est plus du tout synonyme de p<0.05

6.2 Hiérarchisation (*closed testing*)

L'approche par hiérarchisation consiste à hiérarchiser dans le protocole les critères de jugement (le 1er, le 2ème, le 3ème, etc.).

> "We used a closed testing procedure, with prespecified hierarchical testing of the primary and secondary outcomes."
> "a hierarchical sequential testing approach of outcomes was used to control for the type 1 error rate, with testing of outcomes as described in the order listed in the Outcomes section, beginning with the CDR-SB score" [10.1056/NEJMoa1812840]
> "Analyses followed a predefined hierarchical hypothesis-testing strategy to adjust for multiplicity to maintain a familywise type I error of 5%. According to this strategy, the statistical significance of each secondary end point could be investigated only if the previous end point was significant (P<0.05 for pooled analyses). The statistical-hierarchy testing order was as follows: ACR20 response, PASI 75, PASI 90, DAS28-CRP, physical component summary of SF-36, HAQ-DI, ACR50, mTSS, dactylitis and enthesitis, and mTSS."
> Adapté à partir de [10.1056/NEJMoa1412679]

Une fois les résultats obtenus, ils sont analysés dans l'ordre de la hiérarchie. Tous les premiers critères de la hiérarchie où p<0.05 sont alors significatifs et permettent de conclure au bénéfice du traitement sur ces critères. Dès l'obtention d'un p>=0.05, l'analyse s'interrompt et tous les autres critères situés en dessous dans la hiérarchie sont non concluants (quelle que soit la valeur du p, y compris si p<0.05).

> For the primary and key secondary outcomes only, the type I error was controlled by a hierarchical gate-keeping procedure, wherein each successive outcome was tested only if the preceding comparison was significant at a two-sided P value of 0.05. [10.1056/NEJMoa1714631]

> L'essai DAPA-HF [10.1056/NEJMoa1911303] a utilisé une hiérarchisation pour contrôler le risque alpha global sur plusieurs critères :
> « We used a closed testing procedure, with prespecified hierarchical testing of the primary and secondary outcomes. The type I error was controlled at a two-sided alpha level of 0.0499 for multiple comparisons across primary and secondary outcomes, with one interim efficacy analysis taken into account. »
> NB : Le seuil de signification dans la hiérarchie n'est pas 0.05, mais 0.0499, car une partie du risque alpha global a été attribué à une analyse intermédiaire (cf. section 2.1).
> Pour interpréter les résultats, il convient en premier d'identifier les critères inclus dans la hiérarchie et leur position respective :

"The primary outcome was 1 a composite of worsening heart failure or death from cardiovascular causes. …. A key secondary outcome was 2 a composite of hospitalization for heart failure or cardiovascular death. The additional secondary outcomes were the total number of 3 hospitalizations for heart failure and cardiovascular deaths; 4 the change from baseline to 8 months in the total symptom score on the Kansas City Cardiomyopathy Questionnaire; 5 a composite of worsening renal function; and 6 death from any cause"

Le tableau des résultats se lit alors dans l'ordre de cette hiérarchie. Le p pour les critères 1 à 5 est inférieur au seuil de 0.0499 et permet donc de conclure au bénéfice du traitement sur ces critères. Dans ce tableau (cf. note de bas de tableau) le sigle NA est utilisé pour les tests qui ne permettent pas de conclure. Il s'avère donc que le p du critère n° 5 ne permettait pas de conclure (il n'est pas rapporté, mais c'est le premier NA de la hiérarchie). De ce fait le p du critère n° 6 n'est pas rapporté (cf. section 6.4).

Table 2. Primary and Secondary Cardiovascular Outcomes and Adverse Events of Special Interest.*

	Variable	Dapagliflozin (N=2373)		Placebo (N=2371)		Hazard or Rate Ratio or Difference (95% CI)	P Value
			events/100 patient-yr		events/100 patient-yr		
	Efficacy outcomes						
1	Primary composite outcome — no. (%)†	386 (16.3)	11.6	502 (21.2)	15.6	0.74 (0.65 to 0.85)	<0.001
	Hospitalization or an urgent visit for heart failure	237 (10.0)	7.1	326 (13.7)	10.1	0.70 (0.59 to 0.83)	NA
	Hospitalization for heart failure	231 (9.7)	6.9	318 (13.4)	9.8	0.70 (0.59 to 0.83)	NA
	Urgent heart-failure visit	10 (0.4)	0.3	23 (1.0)	0.7	0.43 (0.20 to 0.90)	NA
	Cardiovascular death	227 (9.6)	6.5	273 (11.5)	7.9	0.82 (0.69 to 0.98)	NA
	Secondary outcomes						
2	Cardiovascular death or heart-failure hospitalization — no. (%)	382 (16.1)	11.4	495 (20.9)	15.3	0.75 (0.65 to 0.85)	<0.001
3	Total no. of hospitalizations for heart failure and cardiovascular deaths‡	567	—	742	—	0.75 (0.65 to 0.88)	<0.001
4	Change in KCCQ total symptom score at 8 mo§	6.1±18.6	—	3.3±19.2	—	1.18 (1.11 to 1.26)	<0.001
5	Worsening renal function — no. (%)¶	28 (1.2)	0.8	39 (1.6)	1.2	0.71 (0.44 to 1.16)	NA STOP
6	Death from any cause — no. (%)	276 (11.6)	7.9	329 (13.9)	9.5	0.83 (0.71 to 0.97)	NA

* Plus–minus values are means ±SD. NA denotes not applicable because P values for efficacy outcomes are reported only for outcomes that were included in the hierarchical-testing strategy.

Dans cet exemple, il ne faut surtout pas tomber dans le piège de conclure à un résultat significatif pour les décès de toute cause (critère n° 6) en se basant sur l'intervalle de confiance. En effet il s'agirait d'une signification nominale qui n'a rien à voir avec la signification en termes de risque alpha global. Sur ce critère aucune conclusion ne peut être portée, car le test de la hiérarchie s'arrête au-dessus (au niveau du critère n° 5).

Il ne faut pas déduire la signification statistique de l'intervalle de confiance quand le p n'est pas rapporté

Les p<0.05 pour des critères situés dans la hiérarchie en dessous du premier « non significatif » ne doivent pas être considérés et ne permettent pas de conclure au bénéfice du traitement.

Les résultats de l'essai Odyssey Outcome ont d'abord été présentés à un congrès de cardiologie avec la diapositive suivante :

ACC.18

Main Secondary Efficacy Endpoints:
Hierarchical Testing

Endpoint, n (%)	Alirocumab (N=9462)	Placebo (N=9462)	HR (95% CI)	Log-rank P-value
CHD event	1199 (12.7)	1349 (14.3)	0.88 (0.81, 0.95)	0.001
Major CHD event	793 (8.4)	899 (9.5)	0.88 (0.80, 0.96)	0.006
CV event	1301 (13.7)	1474 (15.6)	0.87 (0.81, 0.94)	0.0003
Death, MI, ischemic stroke	973 (10.3)	1126 (11.9)	0.86 (0.79, 0.93)	0.0003
CHD death	205 (2.2)	222 (2.3)	0.92 (0.76, 1.11)	0.38
CV death	240 (2.5)	271 (2.9)	0.88 (0.74, 1.05)	0.15
All-cause death	334 (3.5)	392 (4.1)	0.85 (0.73, 0.98)	0.026*

*Nominal P-value

ODYSSEY OUTCOMES 33

http://clinicaltrialresults.org/Slides/ACC2018/ODYSSEY_Steg.pdf

Un bénéfice de l'alirocumab a été montré sur les 4 premiers critères de jugement secondaires. La valeur du p sur le premier critère de mortalité (CHD, coronary heart disease death) interrompt la hiérarchie et il est donc impossible de conclure sur les 3 derniers critères, y compris sur les décès de toute cause, même si son p nominal est inférieur à 0.05. Cette subtilité statistique n'a cependant pas été perçue par tout le monde et ce résultat de mortalité a ensuite été largement repris dans des sources secondaires[4] et en communication promotionnelles pour mettre en avant une réduction de la mortalité de toute cause comme le montre les titres suivants :

http://www.news.sanofi.us/2018-11-11-ODYSSEY-OUTCOMES-investigators-highlight-at-AHA-that-Praluent-R-alirocumab-Injection-was-associated-with-fewer-deaths-from-any-cause

Dans la publication dans le NEJM [10.1056/NEJMoa1801174], aucune p value n'est rapportée pour les décès de toutes causes conformément aux pratiques de ce journal (cf. section 6.4) afin de prévenir ce genre de surinterprétation des résultats et les spins de conclusions qui pourraient être engendrés.

[4] Les sources secondaires sont des revues journalistiques ou des revues professionnelles, très nombreuses et souvent distribuées gratuitement aux médecins. Elles sont souvent des revues promotionnelles (publirédactionnel).

Table 2. **Composite Primary End Point and Secondary End Points (Intention-to-Treat Population).**

End Point	Alirocumab (N=9462)	Placebo (N=9462)	Hazard Ratio (95% CI)	P Value
	number of patients (percent)			
Primary end point: composite of death from coronary heart disease, nonfatal myocardial infarction, fatal or non-fatal ischemic stroke, or unstable angina requiring hospitalization	903 (9.5)	1052 (11.1)	0.85 (0.78–0.93)	<0.001
Major secondary end points, in order of hierarchical testing				
Any coronary heart disease event*	1199 (12.7)	1349 (14.3)	0.88 (0.81–0.95)	0.001
Major coronary heart disease event†	793 (8.4)	899 (9.5)	0.88 (0.80–0.96)	0.006
Any cardiovascular event‡	1301 (13.7)	1474 (15.6)	0.87 (0.81–0.94)	<0.001
Composite of death from any cause, nonfatal myocardial infarction, or nonfatal ischemic stroke§	973 (10.3)	1126 (11.9)	0.86 (0.79–0.93)	<0.001
Death from coronary heart disease	205 (2.2)	222 (2.3)	0.92 (0.76–1.11)	0.38¶
Death from cardiovascular causes	240 (2.5)	271 (2.9)	0.88 (0.74–1.05)	
Death from any cause	334 (3.5)	392 (4.1)	0.85 (0.73–0.98)	

Cet exemple illustre bien les **dangers des sources secondaires et de la communication promotionnelle** et montre l'intérêt de pouvoir interpréter par soi-même les résultats des essais pour se forger sa propre opinion sur le réel intérêt clinique d'un nouveau traitement, en toute indépendance.

Dans les méthodes hiérarchiques, le seuil de la signification n'est pas toujours 0.05. Il peut être plus petit en raison, par exemple, de la réalisation d'analyses intermédiaires ou d'une répartition en amont pour gérer plusieurs doses de traitement (cf. section 0).

> We used a closed testing procedure, with prespecified hierarchical testing of the primary and secondary outcomes. The type I error was controlled at a two-sided alpha level of 0.0499 for multiple comparisons across primary and secondary outcomes, with one interim efficacy analysis taken into account.

La hiérarchisation permet de valider des bénéfices supplémentaires contrairement à la répartition qui permet d'aménager plusieurs possibilités pour démontrer que le traitement à un intérêt (même si ce n'est pas sur un critère au moins sur l'autre éventuellement).

> *Ces 2 situations, montrer qu'un traitement apporte plusieurs bénéfices et pouvoir conclure à l'intérêt d'un traitement sur un critère ou un autre, entrainent toutes les deux une multiplicité, mais leur finalité n'est pas la même. La répartition permet de conclure sur l'un* ***ou*** *l'autre des tests impliqués (donne plus de flexibilité pour obtenir au moins un résultat pour justifier l'enregistrement ou l'utilisation du traitement). La hiérarchisation permet de montrer qu'éventuellement un traitement apporte un 1er bénéfice et un 2ème et un 3éme, etc. Cependant, si le 1er critère ne permet pas de conclure, les autres tests prévus dans la hiérarchie ne sont d'aucun secours (contrairement à la répartition où, en cas de non-significativité sur un critère, le ou les autres peuvent éventuellement rattraper le coup !).*

6.2.1 Comment cela marche ?

Cette méthode permet un contrôle de l'inflation du risque alpha de la manière suivante.

Classiquement l'inflation du risque alpha était évitée par l'utilisation d'un critère de jugement principal unique défini à priori. Seule une conclusion sur ce critère est possible, les autres résultats obtenus sur les autres critères (critères secondaires) ont seulement une valeur exploratoire (ou explicative en cas de résultat concluant sur le critère principal) et ne permettent pas de justifier une utilisation ou une AMM. Les résultats sur les critères secondaires suggèrent des effets et ne les démontrent pas. Seul le

résultat sur le critère principal est susceptible d'apporte une démonstration formelle de l'effet du traitement. En effet, un essai clinique a pour finalité de permettre de décider, en se basant sur les faits, si le traitement évalué à un intérêt ou non (et ainsi s'il doit être mis sur le marché et recommandé pour la pratique). Dans cet exercice, on ne souhaite pas courir un risque de faire cette conclusion et recommandation à tort (risque de considérer que le traitement apporte un bénéfice sous hypothèse nulle) supérieur à certain niveau, classiquement de l'ordre de 2.5% pour un essai (car la conclusion est unilatérale et les tests à 5% bilatéraux), voire plus faible sur un dossier d'enregistrement où l'on demande classiquement au moins 2 essais concluants.

Pour que ce risque de conclusion à tort reste au niveau voulu –disons pour simplifier 5%, il est nécessaire de n'examiner qu'un seul et unique test statistique. En effet, s'il devient possible de conclure à l'intérêt du traitement à partir de la réalisation de plusieurs tests, le risque de conclure à tort à l'intérêt du traitement n'est plus de 5%, mais il est bien plus important, car chaque test envisagé apporte un risque de 5% qu'il soit significatif par hasard. Ainsi sous l'hypothèse nulle (le traitement n'apporte aucun bénéfice sur aucun plan) si on examine 100 critères de jugement indépendants, 5 seront significatifs (c'est le reflet du risque alpha consentie de 5%). Pour éviter cela, la décision de reconnaître un intérêt au traitement évalué se base sur un seul et unique test statistique, choisi à priori indépendamment des résultats : celui du critère de jugement principal. Ainsi l'approche du critère de jugement principal permet de contrôler parfaitement le risque de recommander (mettre sur le marché) un traitement qui en réalité n'apporte aucun bénéfice aux patients.

Souvent lorsque le critère de jugement principal est significatif, on scrute les critères de jugement secondaires à la recherche d'effets supplémentaires du traitement qui permettraient ainsi de dire que ce traitement présente d'autres avantages que son effet sur le critère principal (par exemple, une réduction de la mortalité totale en plus de la réduction des évènements cardiovasculaire). Aucune précaution statistique n'est en général mise en œuvre, ce qui conduit à faire de la pêche à la ligne et à courir un risque de fausse découverte important.

Prenons l'exemple d'un traitement qui n'a aucun autre effet que son effet sur le critère principal, plus on scrute de critères de jugement secondaires (où par hypothèse on a un risque de 5% d'avoir un résultat significatif) plus la probabilité de découvrir à tort un autre avantage du traitement augmente.

Souvent cette analyse des critères de jugement secondaires débouche sur la mise en concurrence de 2 ou plusieurs traitements ayant le même effet primaire pour voir si l'un ne se distinguerait pas en ayant un autre avantage que n'auraient pas les autres. Cette approche n'a souvent pas plus de valeur que de faire cette recherche de facteurs discriminants à l'aide de la « roue de la fortune » !

Le principal intérêt de l'approche séquentielle hiérarchique est de parfaitement contrôler le risque de fausse découverte dans cette recherche d'autres avantages.

En effet au niveau du premier critère testé, tout se passe comme avec l'approche du critère principal. La prise de risque sur la reconnaissance à tort d'un intérêt au traitement est parfaitement contrôlée (limitée à 5% pour faire simple). Si ce test est non significatif, tout s'arrête et ce traitement ne sera jamais, sur la base de cet essai, considéré comme utile.

Si ce test est significatif, se pose alors la question de l'existence d'au moins un autre bénéfice supplémentaire. Comme le critère à utiliser pour faire cette recherche d'intérêt complémentaire est pré fixé, aucune pêche à la ligne n'est effectuée. L'hypothèse que le traitement a un premier (au moins un) avantage supplémentaire est testée avec un risque de fausse découverte parfaitement contrôlé de 5%. Si ce test est significatif, on peut conclure que l'on a démontré que le traitement avait un intérêt (1er test) et qu'il a aussi un avantage supplémentaire (le 2ème test). En faisant cette conclusion, on

court un risque parfaitement maitrisé de recommander à tort, dans l'absolu, l'utilisation de ce traitement (le 1er test) et aussi de promouvoir à tort ce traitement en disant qu'il apporte un bénéfice supplémentaire alors qu'en réalité il n'apporte que le bénéfice obtenu sur le 1er critère.

Et ainsi de suite... Avec un traitement qui a démontré son bénéfice primaire et un premier bénéfice supplémentaire, se pose alors la question de savoir s'il n'aurait pas un second bénéfice supplémentaire.

Pour éviter une pêche à la ligne sur tous les critères secondaires restants, la méthode séquentielle hiérarchique a parfaitement défini le critère unique qui devait être examiné pour chercher ce second avantage supplémentaire et ainsi il n'y a pas d'inflation du risque alpha sur cette recherche. Et ainsi de suite jusqu'à la fin de la hiérarchie ou jusqu'au premier test non significatif. Aucune conclusion n'est alors possible en deçà de ce 1er test non significatif, même en cas de résultat statistiquement significatif.

6.3 Combination des deux approches

Les deux approches de la hiérarchisation et de la répartition sont assez fréquemment combinées. Par exemple, lorsque deux doses du nouveau traitement sont évaluées dans un essai (essai à 3 bras) le risque alpha global est d'abord réparti entre les 2 doses (2.5% pour chaque dose par exemple) et ensuite, pour chaque dose, une série de critères de jugement est testée de manière séquentielle.

> A Bonferroni approach (splitting the overall α between the two dose levels of verubecestat) in conjunction with a hierarchical sequential testing approach of outcomes was used to control for the type 1 error rate, with testing of outcomes as described in the order listed in the Outcomes section.
> Separately for each dose level, if significant superiority was not shown, all subsequent outcomes were assumed not to have differed significantly between the groups [10.1056/NEJMoa1812840]

Dans d'autres cas, les 2 doses peuvent être testées de manière complètement hiérarchique

> Hypotheses were tested in the following order: the 20-mg cannabidiol group, followed by the 10-mg cannabidiol group, was compared with the placebo group with respect to the primary outcome; the 20-mg cannabidiol group was then compared with the placebo group with respect to each key secondary outcome in the order listed above, and then the 10-mg cannabidiol group was compared with the placebo group with respect to each key secondary outcome in the same order [10.1056/NEJMoa1714631]

6.4 Réallocation, recyclage du risque alpha

Une amélioration du principe de répartition du risque alpha peut être apportée en procédant à la réallocation du risque alpha. Cela consiste, quand un test est significatif sur un *co-primary endpoints,* à réallouer le risque alpha de ce test sur l'autre co-primary.

La répartition du risque alpha sert à éviter l'inflation du risque alpha lorsque l'on est obligé de regarder le 2ème co-primary endpoint après n'avoir pas pu conclure à la signification statistique sur le 1er test. Dans ce cas il faut que la nouvelle prise de risque de conclure à tort à un intérêt du traitement avec ce 2ème critère soit contrôlée avec un seuil abaissé, car il s'agit d'une deuxième tentative de trouver un intérêt au traitement. Mais si le premier test permet de conclure à l'intérêt du traitement, la réalisation du deuxième n'a plus d'enjeu pour conclure à cet intérêt (cela est déjà acquis avec le 1er). Il n'y a donc plus lieu de corriger le seuil de ce test, car ce n'est plus un plan de secours (pour récupérer un échec

sur le 1er test). Il est donc possible de tester ce 2ème critique au seuil de 5% (c'est-à-dire après avoir réalloué le risque alpha initialement attribué au premier test).

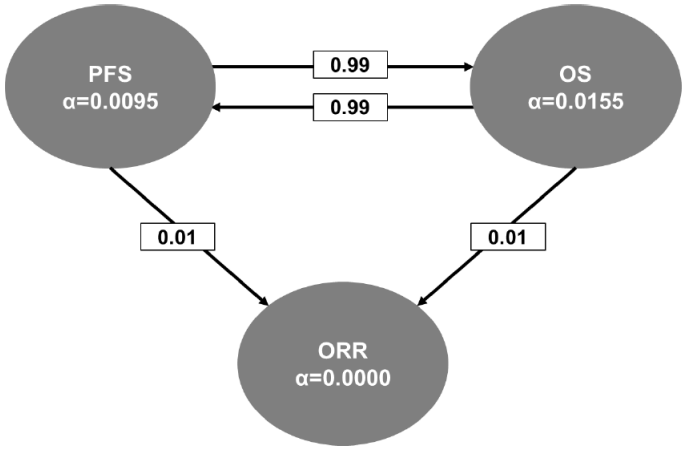
Si par exemple le risque alpha est équiréparti entre 2 co-primary endpoints A et B, le risque alpha attribué à chaque critère est de 2.5%. Avec la réallocation, le résultat sera statistiquement significatif sur le critère A si p<2.5% ou si p<5% quand la value de l'autre critère est p<2.5%, et vice versa.

> *La réallocation du risque alpha est le principe de plusieurs anciennes méthodes comme la méthode de Holm ou de Hochberg. Ce concept a été ensuite généralisé récemment et devient incontournable dans les essais modernes.*

6.4.1 Exemple 1

L'essai Keynote-189 a évalué le pembrolizumab dans le cancer du poumon non à petites cellules métastatique [2]. Une répartition du risque alpha global avec réallocation a été effectuée entre les 3 critères classiques de l'oncologie : l'OS, la PFS et l'ORR (réponse objective) suivant le schéma suivant disponible dans le supplément[5]:

Figure S1. Type I Error Reallocation Strategy.



> *"The weights for reallocation from each hypothesis to the others are represented in the boxes on the lines connecting the hypotheses. The progression-free survival and overall survival hypotheses were to be tested at α=0.0095 and α=0.0155, respectively. If the overall survival test is significant, the progression-free survival hypothesis may be tested at α=0.025, and if the progression-free survival test is significant, the overall survival hypothesis may be tested at α=0.025. If both the overall and progression-free survival tests are significant, the objective response hypothesis will be tested at α=0.025."*

Initialement, avant la disponibilité des résultats, le risque alpha global unilatéral de 2.5% a été réparti entre la PFS (0.95%) et l'OS (1.55%). Aucun risque alpha n'a été attribué initialement à l'ORR.

Lorsque les résultats sont disponibles, la p-value (unilatérale) de la PFS est comparée au seuil de 0.95%. Si elle est inférieure à ce seuil, le résultat de PFS est statistiquement significatif et il permet de conclure à l'intérêt du traitement. À ce moment, l'examen du résultat d'OS n'a plus une occasion supplémentaire de faire cette conclusion, étant donné qu'elle est déjà faite. Il n'y a donc plus de risque

[5] La lecture de ces papiers nécessite de plus en plus de se référer aux suppléments électroniques ou au protocole pour trouver les détails cruciaux de la méthode.

d'inflation du risque (comme cela aurait été le cas si le résultat sur la PFS n'était pas significatif, et il s'agirait alors d'une 2eme tentative de trouver un intérêt au traitement). L'examen de l'OS peut donc se faire avec le seuil à 2.5% unilatéral, car le risque alpha de la PFS est recyclable, en fait 2.4905%, car seulement 99% (le poids qui est au milieu de la flèche) de l'alpha de la PFS est recyclé vers l'OS, 1% étant attribué à l'ORR soit 0.0095% d'alpha. Cependant, si le résultat de PFS n'est pas significatif, l'examen de l'OS est une deuxième tentative et le seuil doit donc être le seuil ajusté. Si l'OS sort avec ce seuil il est alors possible de recycler 99% de l'alpha initialement attribué à l'OS vers la PFS (ce qui fait un seuil de PFS de 2.4845%). Il se peut alors que la PFS devienne significative avec ce nouveau seuil (un recyclage bidirectionnel avait été choisi, cf. la figure). On ne peut donc conclure que lorsque la totalité du réseau de réallocation a été parcourue.

Si PFS ou l'OS ou les deux sont significatifs, l'ORR devient testable (elle a récupéré de l'alpha par réallocation) au seuil de 0.0095% si seul la PFS est significative, au seuil de 0.0155% si seul l'OS est significatif ou au seuil de 0.025% si les 2 sont significatifs.

6.4.2 Exemple 2

L'essai DECLARE [3] a évalué la sécurité et l'efficacité de la dapagliflozine sur les évènements cardiovasculaires dans le diabète de type 2.

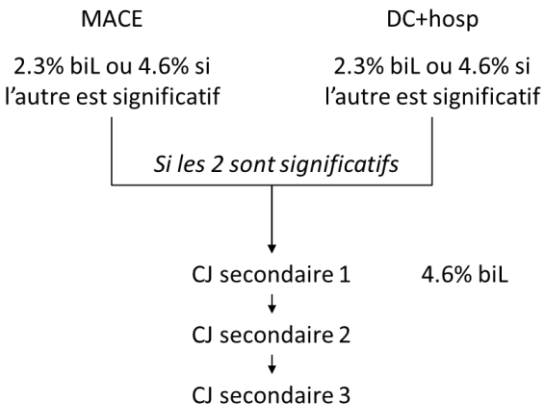
Au total, 5 critères de jugement étaient potentiellement décisionnels :

> *« The two primary efficacy outcomes were MACE and a composite of cardiovascular death or hospitalization for heart failure. Two secondary efficacy outcomes were prespecified. The first was a renal composite outcome, defined as a sustained decrease of 40% or more in estimated glomerular filtration rate (eGFR) —calculated by means of the Chronic Kidney Disease Epidemiology Collaboration equation22 —to less than 60 ml per minute per 1.73 m2 of body-surface area, new end-stage renal disease, or death from renal or cardiovascular causes. The other secondary outcome was death from any cause. A prespecified additional renal composite outcome included all the criteria described for the secondary renal outcome except for cardiovascular death. »*

Cette multiplicité a été gérée par un mélange de hiérarchisation et de répartition avec réallocation :

> *« the two efficacy outcomes of MACE and the composite of cardiovascular death or hospitalization for heart failure were to be tested in parallel, each at a two-sided alpha level of 0.023. If either was significant, the alpha value could be recycled24 to test the other efficacy outcome at a two-sided alpha level of 0.046. If after this procedure both efficacy outcomes were significant, the secondary outcomes were to be tested, at a two-sided alpha level of 0.046, in a hierarchical fashion."*

L'approche choisie correspond donc au schéma suivant :



Les résultats obtenus sont les suivants :

Outcome	Dapagliflozin (N=8582) no. (%)	Dapagliflozin (N=8582) rate/1000 patient-yr	Placebo (N=8578) no. (%)	Placebo (N=8578) rate/1000 patient-yr	Hazard Ratio (95% CI)	P Value
Cardiovascular death or hospitalization for heart failure	417 (4.9)	12.2	496 (5.8)	14.7	0.83 (0.73–0.95)	0.005
MACE	756 (8.8)	22.6	803 (9.4)	24.2	0.93 (0.84–1.03)	0.17
≥40% decrease in eGFR to <60 ml/min/1.73 m^2, ESRD, or death from renal or cardiovascular cause	370 (4.3)	10.8	480 (5.6)	14.1	0.76 (0.67–0.87)	
Death from any cause	529 (6.2)	15.1	570 (6.6)	16.4	0.93 (0.82–1.04)	
Hospitalization for heart failure	212 (2.5)	6.2	286 (3.3)	8.5	0.73 (0.61–0.88)	
Myocardial infarction	393 (4.6)	11.7	441 (5.1)	13.2	0.89 (0.77–1.01)	
Ischemic stroke	235 (2.7)	6.9	231 (2.7)	6.8	1.01 (0.84–1.21)	
Death from cardiovascular cause	245 (2.9)	7.0	249 (2.9)	7.1	0.98 (0.82–1.17)	
Death from noncardiovascular cause	211 (2.5)	6.0	238 (2.8)	6.8	0.88 (0.73–1.06)	
≥40% decrease in eGFR to <60 ml/min/1.73 m^2, ESRD, or death from renal cause	127 (1.5)	3.7	238 (2.8)	7.0	0.53 (0.43–0.66)	

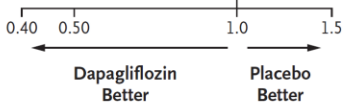


Figure 2. Key Efficacy Outcomes and Their Components.
Two-sided P values are shown for the two primary efficacy outcomes of cardiovascular death or hospitalization for heart failure and MACE. The abbreviation eGFR denotes estimated glomerular filtration rate, and ESRD end-stage renal disease.

Il apparait donc qu'il est possible de conclure à l'intérêt de la dapagliflozine du fait d'une réduction du critère composite décès et hospitalisation, car le p est inférieur au risque alpha alloué à ce critère par la répartition, soit 2.3% bilatéral. Un recyclage est alors possible ce qui donne un seuil de 4.6% bilatéral pour le test du 2eme critère du même niveau, les MACE. Le p obtenu est supérieur à ce seuil (p=0.17) et donc il n'est pas possible de conclure sur ce critère. De plus la hiérarchie s'arrête à ce niveau, car il était prévu que le 2eme niveau de la hiérarchie ne pouvait être testé que si les 2 coprimary endpoints du 1er niveau étaient significatifs. Tous les autres critères sont donc non statistiquement significatifs, quelle que soit leur valeur de p. On notera que ces p-values ne sont d'ailleurs pas rapportées dans la figure suivant la nouvelle politique du NEJM, puisque ces critères deviennent non décisionnels (non inférentiel) du fait de l'arrêt de la hiérarchie avec les MACE.

Il ne faut surtout pas essayer de deviner à partir des intervalles de confiance si le p nominal était inférieur ou non à 5%, car dans ce cas on outrepasserait la méthode de contrôle strict du risque alpha global et les conclusions que l'on pourrait en tirer seraient faites avec un risque alpha non contrôlé. Cela montre que le principe qui a été enseigné par le passé qu'un intervalle de confiance qui

n'englobait pas l'absence d'effet (un ratio de 1 ou une différence de 0) est faux. L'intervalle de confiance ne permet pas de déterminer la signification statistique autre que nominale.

6.4.3 Comment cela marche ?

Cette méthode permet un contrôle de l'inflation du risque alpha comme illustré dans la Figure 2.

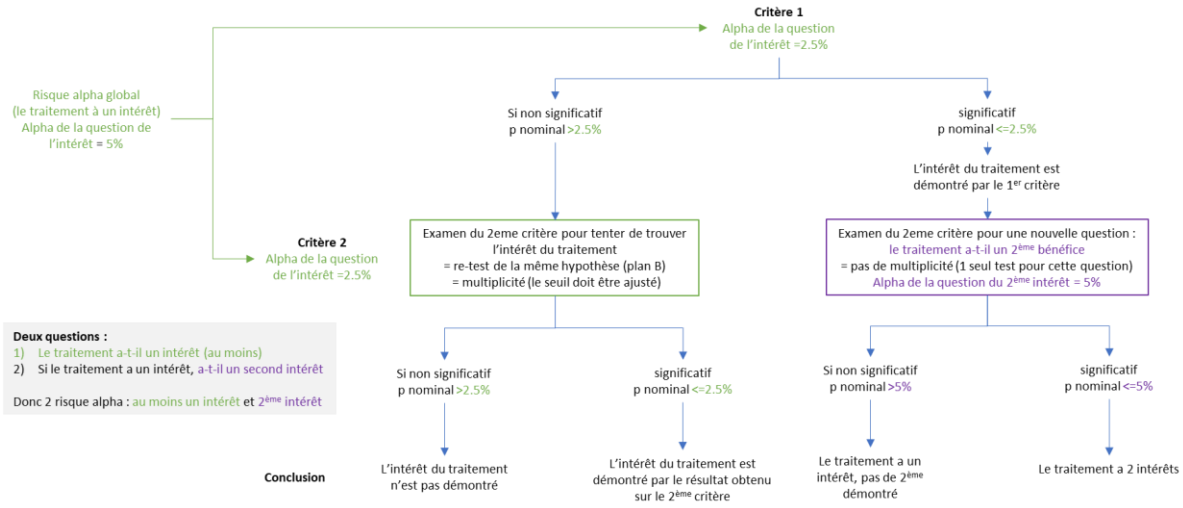


Figure 2 – Illustration du principe du recyclage du risque alpha

La répartition du risque alpha global entre 2 critères aménage la possibilité de pouvoir conclure à l'intérêt du traitement avec l'un ou l'autre. Après échec sur un critère, envisager le second n'augmente pas le risque alpha global puisque celui-ci a été réparti entre les deux critères. Le 2ème critère permet d'avoir un plan B licite en cas d'échec sur le premier.

Cependant si le 1er critère permet de conclure à l'intérêt du traitement, car le p nominal est inférieur au risque alpha attribué à ce critère (seuil ajusté de signification), l'examen du second critère n'a plus pour objectif de conclure à l'intérêt du traitement (cela est déjà acquis). Ce n'est plus l'activation d'un plan B (pour la question le traitement a-t-il un quelconque intérêt, la réponse est déjà obtenue) mais une recherche d'un bénéfice supplémentaire tout simplement. Ce 2ème test sera donc analysé pour une question complètement différente (de celle pour laquelle le 1er test a été utilisé) et il n'y a pas de multiplicité sur cette nouvelle question : le traitement apporte-t-il un second bénéfice ? Seul ce critère sera analysé pour répondre à cette nouvelle question. Il n'y a donc pas de répartition du risque alpha alloué à cette 2èmequestion entre plusieurs critères. Tout le risque alpha que l'on consent pour cette nouvelle question est disponible pour le test de ce second critère.

En revanche, si l'algorithme prévoyait deux critères pour chercher un 2ème bénéfice, dans ce cas, le risque alpha alloué à la question du 2ème bénéfice doit être réparti entre ces 2 critères.

- Il y a un risque alpha par question générique que se pose l'essai
    - Le traitement a t-il au moins un intérêt ?
    - Le traitement a t-il un 2éme intérêt (si le 1er a été démontré) ?
    - Etc.
- Un ou plusieurs critères (tests statistiques) peuvent être prévus pour une même question générique
- Chaque question générique a un risque alpha de 5%,
  qui est ensuite réparti entre les critères correspondant à cette question générique

7 Nouvelle politique de présentation des p value

Depuis 2018, des revues comme le NEJM ne présentent plus dans les articles d'essais cliniques, les p values des tests qui ne permettent pas de conclure à l'intérêt du traitement[6], afin d'éviter à leur lecteur de commettre des surinterprétations abusives de p values nominales inférieures à 0.05, mais qui ne sont pas pour autant statistiquement significatives en termes de risque alpha global [4].

Figure 3 – Exemple de nouvelle présentation des p values
Pour les tests qui ne peuvent pas conduire à conclure à l'intérêt du traitement les p values ne sont plus rapportées. Il s'agit des tests non pris en compte dans le plan de contrôle du risque alpha global ou des résultats non significatifs en termes de risque alpha global. [10.1056/NEJMoa1811090]

Table 2. Primary and Secondary End Points in the Modified Intention-to-Treat Population.*

End Point	Rimegepant (N=537)	Placebo (N=535)	Absolute Difference	P Value
	no./total no. (%)†		*percentage points (95% CI)*	
Primary end points				
Freedom from pain 2 hours after the dose	105 (19.6)	64 (12.0)	7.6 (3.3 to 11.9)	<0.001
Freedom from the most bothersome symptom 2 hours after the dose	202 (37.6)	135 (25.2)	12.4 (6.9 to 17.9)	<0.001
Secondary end points				
Freedom from photophobia 2 hours after the dose	183/489 (37.4)	106/477 (22.3)	15.1 (9.4 to 20.8)	<0.001
Freedom from phonophobia 2 hours after the dose	133/362 (36.7)	100/374 (26.8)	9.9 (3.2 to 16.6)	0.004
Pain relief 2 hours after the dose	312 (58.1)	229 (42.8)	15.3 (9.4 to 21.2)	<0.001
Freedom from nausea 2 hours after the dose	171/355 (48.1)	145/336 (43.3)	4.8 (–2.7 to 12.2)	
Use of rescue medication within 24 hr after the dose	113 (21.0)	198 (37.0)	–16.0 (–21.3 to –10.6)	
Sustained freedom from pain 2 to 24 hr after the dose	66 (12.3)	38 (7.1)	5.2 (1.7 to 8.7)	
Sustained pain relief 2 to 24 hr after the dose	229 (42.6)	142 (26.5)	16.1 (10.5 to 21.7)	
Sustained freedom from pain 2 to 48 hr after the dose	53 (9.9)	32 (6.0)	3.9 (0.7 to 7.1)	
Sustained pain relief 2 to 48 hr after the dose	195 (36.3)	121 (22.6)	13.7 (8.3 to 19.1)	
Pain relapse 2 to 48 hr after the dose	52/105 (49.6)	32/64 (50.0)	–0.4 (–15.8 to 15.1)	
Ability to function normally 2 hr after the dose	175 (32.6)	125 (23.4)	9.2 (3.9 to 14.6)	

* The modified intention-to-treat population included patients who underwent randomization, had a migraine attack with pain of moderate or severe intensity, took a dose of rimegepant or placebo, and had at least one efficacy assessment after administration of the dose. To maintain the type I statistical error rate at 0.05, a prespecified hierarchical testing procedure was applied; end points are presented in the sequence in which they were evaluated. Because the incidence of freedom from nausea did not differ significantly between the groups, all statistical tests below this end point in the hierarchy are reported without P values, and no inferences can be made from those results. Percentages,

Dans ces papiers, l'absence de p value **ne doit absolument pas être compensée** en cherchant la signification statistique à partir de l'intervalle de confiance à 95%. Cela conduirait à déterminer une signification nominale, mais qui ne permet pas de conclure à la signification du résultat vis-à-vis du risque alpha global. Si le p n'a été rapporté, c'est que le test ne peut pas être utilisé pour déterminer l'intérêt du traitement (il est en est en dehors du plan de contrôle du risque alpha global ou non significatif).

[6] Ces tests sont parfois appelés non inférentiels, car ils ne permettent pas d'inférer si le traitement à un intérêt ou pas, compte tenu du plan de contrôle du risque alpha de l'essai.

8 Critères de jugement secondaires

8.1 Essai avec un critère de jugement principal unique

Les 2 terminologies « critère de jugement principal » (*primary endpoint ou outcome*) et « critères de jugement secondaires » (*secondary endpoint/outcome*) n'ont vraiment de sens qu'avec les essais utilisant un critère de jugement principal unique.

Dans ce cas, seul ce critère principal unique peut permettre de conclure à l'intérêt du traitement. C'est le seul qui peut être statistiquement significatif en termes de risque alpha global de l'essai et qui peut apporter une démonstration de l'intérêt du traitement.

Dans ce cadre, les critères secondaires ne permettent pas de démontrer et ne peuvent pas être statistiquement significatifs en termes de risque alpha global de l'essai, quelle que soit la valeur nominale de leur p (dans le NEJM ces p ne sont plus rapportés pour cette raison, cf. section 6.4).

Ces critères ne peuvent pas permettre de conclure à l'intérêt du traitement. Tout au plus, ils peuvent faire générer de nouvelles hypothèses à tester dans un nouvel essai.

8.2 Essai gérant la multiplicité par un plan de contrôle du risque global

Dans les essais cliniques modernes les termes critère de jugement principal et critères de jugement secondaires n'ont plus beaucoup d'intérêt et, surtout, l'interprétation ne doit pas s'arrêter aux termes utilisés, mais doit rentrer dans le détail du plan de contrôle du risque alpha global.

Par exemple, en cas d'utilisation d'une hiérarchisation, le terme critère principal désigne le premier de la hiérarchie. Les suivants sont souvent encore appelés critères de jugement secondaires, mais en précisant hiérarchisés ou clés (*key secondary endpoints, main secondary endpoints*, etc.).

Ainsi dans ce cas, des « critères de jugement secondaires » peuvent permettre de conclure à l'intérêt du traitement, mais parce qu'ils sont hiérarchisés (entre eux et avec le critère principal). On voit ainsi toute **l'ambigüité actuelle du terme** critère secondaire et la nécessité de ne pas s'arrêter aux termes et de bien disséquer la méthode.

De plus dans ces études avec hiérarchisation, il peut aussi exister des critères de jugement secondaires qui ne sont pas dans le plan de contrôle du risque alpha et qui n'ont donc qu'une valeur exploratoire. Ces différents critères en fonction des approches sont présentés dans le tableau ci-dessous.

	Essai avec un critère de jugement principal unique (Espèce en voie d'extinction !)	**Essai gérant la multiplicité avec un plan de contrôle du risque alpha global**
Critères **pouvant permettre de conclure** à l'intérêt du traitement **Démonstration possible**	• Critère de jugement principal	• Critère de jugement principal • Critères de jugement secondaires hiérarchisés • Co primary endpoints (répartition)

Critères **ne pouvant pas permettant de conclure** à l'intérêt du traitement = critère exploratoire **Pas de démonstration possible**	• Critères de jugement secondaires • Critères tertiaires • Critères exploratoires	• Critères de jugement secondaires non hiérarchisés (non inclus dans le plan de contrôle du risque alpha global) • Critères tertiaires • Critères exploratoires

Figure 4 – Exemple d'un critère de jugement secondaire avec contrôles du risque alpha global et un critère de jugement secondaire ordinaire [10.1056/NEJMoa1714631]

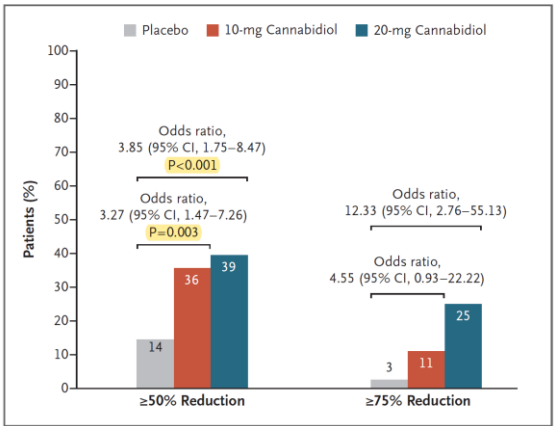


Figure 3. Reductions of at Least 50% and 75% from Baseline in Drop-Seizure Frequency during the Treatment Period.

The odds ratios are for the comparisons between each cannabidiol group and the placebo group. The P values were calculated with the use of a Cochran–Mantel–Haenszel test, with stratification according to age group (2 to 5 years, 6 to 11 years, 12 to 17 years, and 18 to 55 years). P values for reductions of at least 75% from baseline in drop-seizure frequency are not shown because this was not a key secondary outcome, and type 1 error was not controlled.

Dans cet exemple, il est bien précisé que la valeur de p ne peut pas être utilisé pour conclure à l'efficacité du traitement sur la fréquence des crises de goutte, mais c'est rarement notifié comme aussi clairement. Il convient d'être prudent quant à l'interprétation des p valeurs nominales.

Références

1 Marcos A, Wärnberg J, Nova E, et al. The effect of milk fermented by yogurt cultures plus Lactobacillus casei DN-114001 on the immune response of subjects under academic examination stress. *Eur J Nutr* 2004;43:381–89 doi:10.1007/s00394-004-0517-8; PMID:15309418;

2 Gandhi L, Rodríguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *N Engl J Med* 2018;378:2078–92 doi:10.1056/NEJMoa1801005; PMID:29658856;

3 Wiviott SD, Raz I, Bonaca MP, et al. Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. *N Engl J Med* 2019;380:347–57 doi:10.1056/NEJMoa1812389; PMID:30415602;

4 Harrington D, D'Agostino RB, Gatsonis C, et al. New Guidelines for Statistical Reporting in the Journal. *N Engl J Med* 2019;381:285–86 doi:10.1056/NEJMe1906559; PMID:31314974;