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 4 – L'analyse finale et les analyses intermédiaires

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 moment de l'analyse d'un essai doit être parfaitement bien défini a priori pour éviter que l'essai soit poursuivi ou arrêté en fonction des résultats du moment.

> *Si le moment de l'analyse n'est pas préfixé, l'essai sera alors analysé a un moment arbitraire qui peut dépendre des résultats (des analyses sont répétées régulièrement jusqu'à ce que le résultat s'avère satisfaisant si cela arrive).*

Cette analyse survient, après l'inclusion de tous les patients nécessaires, soit à une date de point prédéfini, soit lorsque la durée de suivi voulue (mortalité à 1an par exemple) a été atteinte pour tous les patients, soit, le plus souvent actuellement, lorsque le nombre d'évènements nécessaires a été atteint (tous groupes confondus).

> « We estimated that 288 events would be required to detect a hazard ratio for death of 0.675 with an alpha level of 0.05" [10.1056/NEJMoa1412690]

> ***Calibration, calcul d'effectif***
>
> *Dans les essais modernes, la calibration de l'étude, afin de lui garantir une puissance élevée, ne s'effectue plus en termes d'effectifs (nombre de sujets à inclure), mais en termes de nombres d'évènements nécessaires.*
>
> *Un effectif arbitraire est déterminé en fonction de la fréquence attendue des évènements et de la durée de suivi voulu. Par exemple, si 300 évènements sont nécessaires, dont la fréquence est estimée à environ 5%/an, un effectif de 2000 patients sera nécessaire pour obtenir ces 300 évènements au bout de 3 ans de suivi. Si un suivi de 2 ans est envisagé, il faudra 3000 patients.*
>
> *L'intérêt de cette approche est d'éviter qu'un essai, bien qu'arrêté à la date prévue, s'avère non concluant, non pas parce que l'effet du traitement est moindre qu'attendu, mais parce que sa puissance statistique a été réduite en raison d'une fréquence des évènements plus faible qu'attendu. Il aurait été nécessaire que l'essai dure plus longtemps pour atteindre la fréquence nécessaire permettant d'obtenir la puissance souhaitée avec l'effectif inclus. L'ajustement de la date de fin de l'essai sur le nombre d'évènements effectivement obtenus évite cette situation. Cette approche a l'inconvénient de devoir tenir compte de la variabilité de la durée de suivi ou du nombre de sujet à inclure en cas de durée de suivi constant dans l'estimation budgétaire initiale. Elle nécessite par conséquent de définir aussi précisément que possible la fréquence des évènements attendus, ce qui se révèle souvent difficile en dehors de registres prospectifs sur les populations étudiées dans les pays concernés par les essais.*

Lors de cette analyse finale de l'essai, le bénéfice du traitement est recherché en comparant le critère de jugement entre les 2 groupes et la signification statistique de la différence observée est appréciée en calculant le p.

Parfois d'autres analyses sont réalisées avant cette analyse finale. Il s'agit des analyses intermédiaires (AI) qui sont en général au nombre d'une ou deux. Ces analyses ont aussi pour but de mettre en évidence le bénéfice du traitement (analyse d'efficacité) si les résultats le permettent et reposent donc sur une comparaison statistique des 2 groupes.

Si le bénéfice du traitement est démontré à une analyse intermédiaire, l'objectif de l'essai est atteint et il n'ait plus nécessaire de le poursuivre (pour cet objectif, mais parfois l'étude se poursuit pour répondre à un autre objectif, sur un autre co primary endpoint par exemple). On dit que l'essai a été arrêté prématurément pour démonstration anticipée de l'efficacité.

Cependant si une analyse intermédiaire ne permet pas de conclure au bénéfice du traitement, l'essai se poursuit jusqu'à la prochaine analyse intermédiaire ou jusqu'à l'analyse finale.

Le but de ces analyses intermédiaires est triple :

1 Le premier est de pouvoir détecter au plus tôt le bénéfice du traitement afin d'éviter de continuer à traiter des patients par un traitement inférieur (placebo par exemple) alors que les données amassées sont suffisantes pour conclure à l'efficacité du traitement étudié (arrêt pour efficacité). De plus, la confirmation au plus tôt du bénéfice apporté par un traitement permet d'accélérer sa mise à disposition pour tous les patients. La décision d'arrêt de l'étude pour efficacité doit cependant prendre en compte la notion de durée d'exposition pour les traitements chroniques. Par exemple, un essai prévu pour exposer des patients pendant 5 ans à un médicament arrêté après 2 ans d'exposition et de suivi en moyenne permettra de conclure à une efficacité uniquement pour 2 ans d'exposition.

2 Le deuxième objectif est de se donner les moyens de détecter au plus tôt un éventuel effet délétère afin de limiter le nombre de patients exposés au risque (arrêt pour toxicité).

3 Le troisième objectif est d'arrêter une étude dont on peut prédire avec une certitude raisonnable qu'elle ne pourra pas aboutir (arrêt pour futilité). L'arrêt précoce permettra de diriger les ressources vers le test de nouvelles hypothèses.

La réalisation des AI entraîne une répétition potentielle des comparaisons statistiques cherchant à conclure au bénéfice du traitement. Il y a donc potentiellement une inflation du risque alpha global de l'essai.

Les AI sont réalisées à l'aide de méthode statistique adaptée (O'Brien et Flemming, Peto Haybittle, etc.) qui ajuste le seuil de la signification statistique.

Pour pouvoir conclure à une analyse intermédiaire, il faut que le p (nominal) soit inférieur au seuil ajusté calculé par la méthode statistique (on dit alors que la frontière de la signification a été franchie). Le seuil ajusté est en général assez faible (0.0025 par exemple) et il est calculé en fonction du nombre d'évènements observé au moment de l'analyse. Il est rapporté dans la publication. Il peut être intégré dans une analyse hiérarchique.

> "At the data-cutoff date of April 17, 2014, the interim analysis was performed after 222 events had occurred. For the overall survival analysis, 100 patients (28%) in the combination-therapy group and 122 (35%) in the vemurafenib group had died (hazard ratio for death in the combination-therapy group, 0.69; 95% confidence interval [CI], 0.53 to 0.89; P=0.005) (Figure 1A). The prespecified stopping boundary (P<0.0214) was crossed, and the study was stopped for efficacy on July 14, 2014" [10.1056/NEJMoa1412690]

Si l'essai n'est pas arrêté lors des analyses intermédiaires et arrive à l'analyse finale, le seuil de la signification est aussi ajusté à la baisse pour prendre en compte le risque alpha « consommé » lors des

analyses intermédiaires (répartition du risque alpha global entre les différentes analyses). Une exception est l'analyse hiérarchique dans laquelle analyse intermédiaire et finale sont liées de façon anticipée (fallback procedure avec conservation d'un risque alpha global), dans le cas où l'analyse intermédiaire est positive.

> During the course of the trial, two interim analyses were conducted after 50% and 75%, respectively, of the target number of 1,400 participants had experienced a primary cardiovascular endpoint. To conserve alpha for the final analysis and to limit the possibility of a chance positive interim finding, each interim analysis followed the same closed testing procedure, with a one-sided significance level of 0.01% allotted to the first efficacy interim analysis, and a one sided significance level of 0.04% allotted to the second efficacy interim analysis, and thus a one-sided significance level of 2.45% retained for the final analysis. [10.1056/NEJMoa1707914 supplement]

Des spins de conclusion sont fréquemment observés quand l'analyse intermédiaire ne permet pas de conclure formellement, car le p nominal n'est pas inférieur au seuil ajusté, mais qu'il est cependant inférieur à 0.05.

> "Although the difference in overall survival did not cross the prespecified superiority boundary (P<0.0096), continuous lenalidomide–dexamethasone reduced the risk of death, as compared with MPT (hazard ratio, 0.78; 95% CI, 0.64 to 0.96; P=0.02)" [10.1056/NEJMoa1402551]

2 Problématiques et solutions

La réalisation de plusieurs analyses statistiques dans la même expérience, pour chercher à faire la même conclusion, conduit à des comparaisons statistiques multiples. À chaque analyse intermédiaire, un test statistique est réalisé pour chercher à montrer l'intérêt du traitement. Il y a donc répétition de la prise de risque de trouver à tort un argument pour revendiquer l'intérêt du traitement à chaque. In fine le risque alpha global de conclure à tort à l'intérêt du traitement à une quelconque de ces analyses répétées n'est plus de 5% (même si c'est le seuil retenu pour chaque test), mais il est bien supérieur.

L'utilisation de techniques statistiques adaptées est nécessaire pour empêcher cette augmentation du risque alpha, appelée en jargon statistique « inflation du risque alpha ». Le but de ces méthodes est de garantir un risque global, sur l'ensemble des comparaisons effectuées, de conclure à tort à l'efficacité du traitement de 5%. Sur l'ensemble des comparaisons effectuées, le risque d'obtenir au moins un résultat significatif par le fait du hasard est contrôlé et garde sa valeur prédéfinie de 5%.

2.1 Principe de la solution

Plusieurs solutions sont possibles qui sont à la base de différentes méthodes. L'une d'entre elles consiste à diminuer le seuil de signification statistique de chacune des comparaisons intermédiaires, par exemple en divisant le risque alpha global α par le nombre de comparaisons effectuées n. C'est la méthode de Bonferroni. Ainsi malgré l'inflation du risque alpha, le risque final de conclure à tort à l'efficacité restera compris dans les valeurs habituelles.

Avec 3 analyses intermédiaires prévues, le nombre total de comparaisons qui seront effectuées est de 4 : les 3 intermédiaires plus la comparaison finale. Le seuil à utiliser pour chacune de ces analyses est de 5%/4=1,25%. Si un p inférieur à 1,25% est obtenu à l'une des analyses intermédiaires, il est alors possible de conclure et d'arrêter l'essai sans attendre la fin du recrutement prévu.

Cas de figure n°1. Dans la situation dépeinte par le tableau ci-dessous, l'essai peut être arrêté à la 2e analyse intermédiaire. Le p obtenu lors de cette analyse est inférieur au seuil de signification corrigé et l'essai peut donc être arrêté prématurément. Cette situation met en avant tout l'intérêt des analyses intermédiaires.

Analyses intermédiaires			Analyse finale
1	2	3	
p=0.10	p=0.011		

Cas de figure n°2. Dans ce deuxième exemple, le p<5% de la troisième analyse intermédiaire ne permet pas de conclure à une différence significative, car la valeur obtenue reste supérieure au seuil corrigé pour 4 tests (1,25%). L'essai va donc à son terme et lors de l'analyse finale le p devient inférieur au seuil corrigé ce qui donne donc finalement un résultat statistiquement significatif.

Analyses intermédiaires			Analyse finale
1	2	3	

p=0.25	p=0.08	p=0.04	P=0.012

Cas de figure n°3. Le cas suivant peut paraitre déroutant. Aucune analyse intermédiaire ne conduit à interrompre prématurément l'essai. Lors de l'analyse finale, un p de 4% est obtenu. Cette valeur, bien qu'elle soit inférieure à 5% n'autorise pas à conclure à un résultat statistiquement significatif, car elle reste supérieure au seuil corrigé. Il ne peut pas être considéré comme significatif, car du risque alpha a été consommé au cours des analyses précédentes, effritant le contrôle du risque d'erreur de première espèce apporté par un p<5% au niveau d'une comparaison donnée (le coté gênant de ce résultat a conduit au développement d'une méthode qui évite de se retrouver dans cette situation).

Analyses intermédiaires			Analyse finale
1	2	3	
p=0.42	p=0.28	p=0.12	P=0.04

Cas de figure n°4. Dans le dernier cas de figure, aucune analyse n'atteint le seuil corrigé de signification statistique. L'essai n'obtient donc pas de résultat statistiquement significatif.

Analyses intermédiaires			Analyse finale
1	2	3	
p=0.89	p=0.48	p=0.25	P=0.10

2.2 Méthode de Lan et DeMets

La méthode la plus fréquemment utilisée pour réaliser les analyses intermédiaires sans entrainer d'inflation du risque alpha est la méthode de Lan et DeMets [1]. Cette méthode présente l'avantage de permettre des analyses flottantes, en nombre quelconque, y compris des analyses non initialement prévues (contrairement à la méthode de Bonferroni qui impose de préfixer le nombre d'analyses).

Le seuil de la signification statistique à atteindre à une AI (ou à l'AF) pour pouvoir conclure est déterminé à partir de la fraction d'information. La fraction d'information est le rapport du nombre du nombre d'évènements obtenus à l'analyse intermédiaire divisé par le nombre attendu pour l'analyse finale, exprimée en pourcentage. Si 300 évènements sont nécessaires pour l'analyse finale, une AI réalisée à 150 évènements correspond à une fraction d'information de 50%. Cette approche permet aussi de prendre en compte le nombre effectif d'évènements disponible pour l'AI, qui peut être différents du nombre prévu au protocole en raison des inerties de remontée des informations qui existe dans les grands essais thérapeutiques.

Initialement, la méthode repose sur un graphique où sont tracées la limite (boundary) à franchir (to cross) pour pourvoir conclure à l'efficacité lors de l'AI considérée (cf. Figure 1).

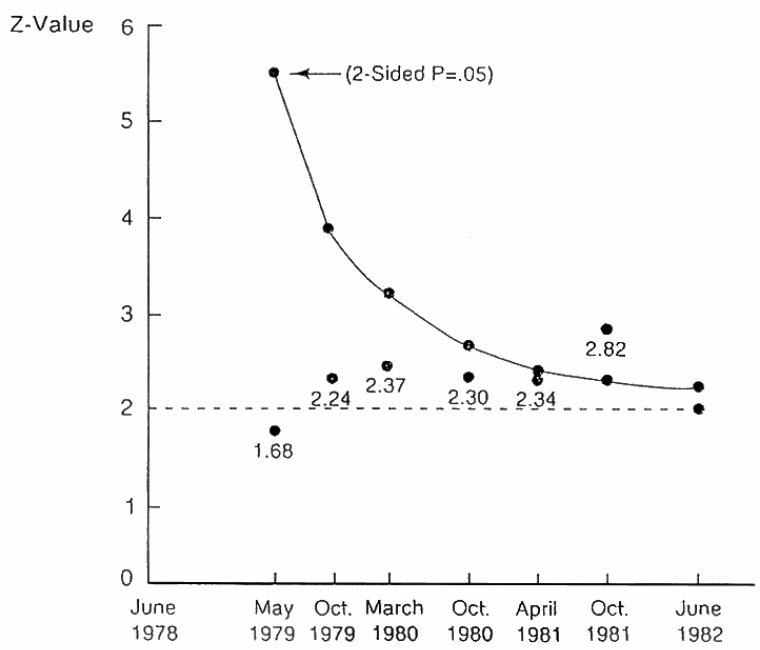


Figure 1 – Exemple de fonctionnement graphique de la méthode de Lan et DeMets. Chaque point numéroté représente le résultat d'une analyse intermédiaire avec en abscisse la fraction d'information correspondante et en ordonnées la différence (standardisée) entre les 2 groupes. La démonstration de l'efficacité est obtenue quand le point représentatif d'une AI se trouve à l'extérieure de la frontière de rejet de l'hypothèse nulle. Cette frontière est une façon de représenter le seuil de signification ajusté. (d'après ref [2])

Cette terminologie « crossed the significance boundary », « did not cross the significance boundary » est encore utilisée même si le raisonnement se base actuellement sur l'obtention d'une p value nominale[1] inférieure ou non au seuil de la signification ajusté.

La forme de la limite à franchir peut-être variable. Elle correspond à ce qui s'appelle la fonction de consommation du risque alpha (« alpha spending function »). Elle est préfixée au protocole. En général les fonctions utilisées (comme celle de O'Brien et Flemming) consomment peu de risque alpha au début afin de réserver la quasi-totalité de ce risque pour l'analyse finale. Pour un risque global de 5% bilatéral à dépenser sur l'ensemble des analyses à réaliser (AI+AF), le seuil de la signification (en termes de p value) sera donc très faible initialement (très inférieur à 0.05), par exemple 0.0002. À l'analyse finale sera alors proche de 0.05, 0.0496 par exemple.

[1] La p value nominale est identique à la p value habituelle. Le terme nominal est seulement là pour insister sur le fait que cette p value doit être comparée à un seuil ajusté et non pas à la valeur de 0.05. Nominal insiste sur le fait que cette valeur est seulement un instrument et ne permet pas en soit, sans l'algorithme de détermination de la signification, de conclure si elle est ou non significative.

3 Le principe général des analyses intermédiaires

Compte tenu des éléments mis en place dans la section précédente, la réalisation des analyses intermédiaires (AI) d'efficacité s'effectue de la manière illustrée par la Figure 2.

Le calcul de « l'effectif » détermine le nombre d'évènements (par exemple décès, ou évènements cardiovasculaires) nécessaire pour réaliser l'analyse finale. C'est le nombre d'évènements nécessaire pour assurer la puissance voulue. Dans cet exemple il est de 500.

Ensuite le nombre d'analyses intermédiaires est déterminé de manière arbitraire. Ces analyses sont aussi définies en nombre d'événement. Ces nombres d'évènements définissent alors les fractions d'information auxquelles s'effectueront les AI ; fraction d'information qui permettent à leur tour de calculer le seuil de la signification à croiser lors de l'AI ou de l'analyse finale pour pouvoir conclure à l'existence de l'effet du traitement. Dans cet exemple la première analyse intermédiaire a été prévue à 50 évènements, ce qui correspond à une fraction d'information de 50/500, soit 10%. Pour une fraction d'information de 10% la méthode statistique donne un seuil ajusté de 0.001.

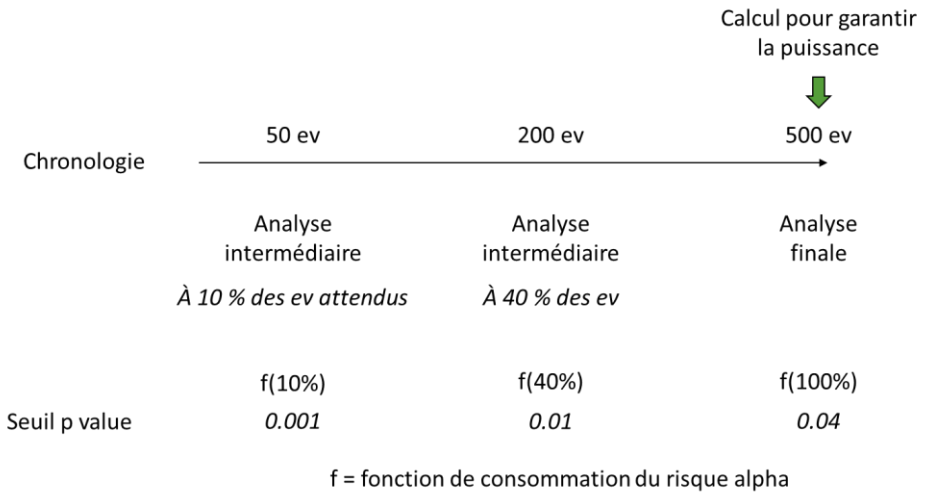


Figure 2 – Illustration du processus des analyses intermédiaires dans les essais modernes

La première analyse intermédiaire sera déclenchée quand le centre de coordination de l'étude aura connaissance de la survenue d'au moins 50 évènements, mais il se peut très bien que cela survienne avec en fait 55 évènements (si le taux de survenues des évènements est important par exemple) et que lorsque l'on extrait le fichier pour l'analyse statistique de la base de données de l'étude, il y a 57 évènements (en raison du petit délai technique entre prise de décision et réalisation de l'extraction). Une nouvelle fraction d'information effective est alors calculée, soit avec cet exemple 57/500 = 11.4%, ce qui conduit aussi à un nouveau seuil ajusté. La p value obtenue par l'analyse statistique de ces données (p value nominale) est comparée à ce seuil. Si elle est inférieure, l'analyse intermédiaire démontre alors (avec un risque alpha global contrôlé à moins de 5% bilatéral) l'efficacité du traitement. Si ce n'est pas le cas, l'essai se poursuit vers l'analyse prévue suivante.

> **Exemple** - The O'Brien-Fleming type of boundary with Lan-DeMets alpha spending function will be used for the interim efficacy analyses to assess superiority. As currently planned, two interim efficacy analyses are to be expected approximately at 1/3 and 2/3 of information time (i.e., approximately 803 and 1607

patients, respectively, with a primary events of CV mortality or HF hospitalization). The interim efficacy analysis with the boundary will spend approximately an alpha of 0.0001 (one-sided) at the first interim analysis and 0.00605 (one-sided) at the second interim analysis. The actual alpha to be spent for the interim efficacy analyses will be precisely determined based on the Lan-DeMets alpha spending function using the actual number of patients who have experienced a primary events at the interim efficacy analyses.

Exemple - During the course of the trial, two interim analyses were conducted after 50% and 75%, respectively, of the target number of 1,400 participants had experienced a primary cardiovascular endpoint. To conserve alpha for the final analysis and to limit the possibility of a chance positive interim finding, each interim analysis followed the same closed testing procedure, with a one-sided significance level of 0.01% allotted to the first efficacy interim analysis, and a one sided significance level of 0.04% allotted to the second efficacy interim analysis, and thus a one-sided significance level of 2.45% retained for the final analysis.

En cas de démonstration de l'efficacité à une analyse intermédiaire et si cette démonstration est la seule (ou l'ultime) recherchée par l'essai, celui-ci s'interrompt alors prématurément. Il a atteint son but et il se termine comme s'il était arrivé à l'analyse finale.

4 Gestion des autres types de multiplicité lors des analyses intermédiaires

Dans les essais modernes, le plan de contrôle du risque alpha global gère d'autres multiplicités que celle engendrée par les analyses intermédiaires, par exemple celle liée à plusieurs critères de jugement décisionnel par répartition du risque alpha ou par hiérarchisation. À ce moment les méthodes dédiées aux analyses intermédiaires se greffent sur celles utilisées en amont pour la multiplicité des critères. Comme évoqué précédemment, la hiérarchisation peut intégrer les analyses intermédiaires aux différents critères de jugement

Dans les essais en oncologie moderne, les analyses intermédiaires sont quasi systématiques. La multiplicité des comparaisons qu'elles entrainent s'ajoute alors à celle engendrée par l'utilisation de plusieurs critères de jugement décisionnel (comme l'OS, la PFS et l'ORR). Le contrôle du risque alpha global prend alors en compte ces 2 sources de multiplicités.

L'intrication entre analyse intermédiaire et critères de jugement multiple conduit à des schémas d'étude similaire à celui représenté sur la Figure 3.

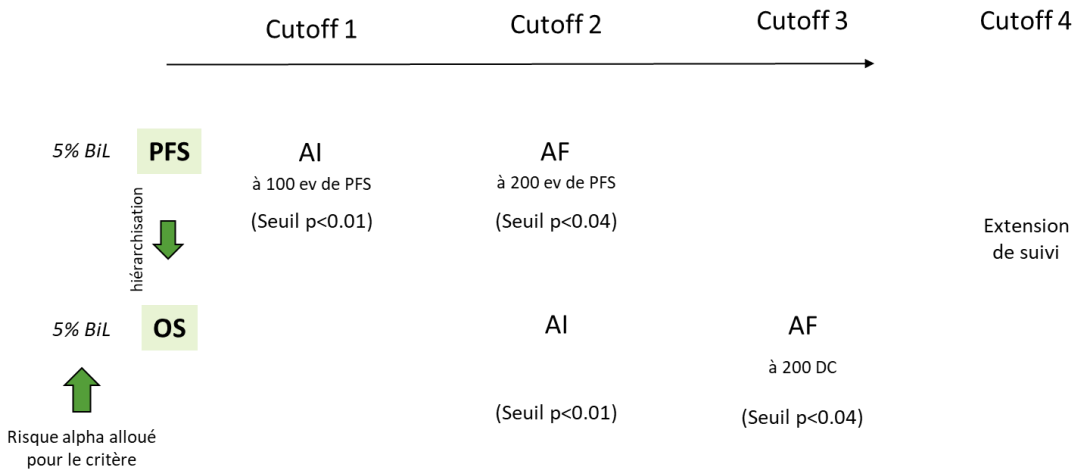


Figure 3 – Schéma typique des analyses intermédiaires dans un essai moderne d'oncologie.
L'objectif de la hiérarchie décrite ici est de rediriger l'intégralité du risque alpha à 0.05 vers l'OS, ceci ne serait possible que si le bénéfice sur la survie sans progression était démontré au cutoff 1 et 2. Comme nous le verrons dans l'exemple suivant, d'autres approches encore plus complexes, idéalement illustrées sur un plan graphique sont utilisées

Dans cet essai la présence 2 critères de jugement décisionnels (PFS et OS) induit deux séquences indépendantes analyses intermédiaires / analyse finale. Une analyse intermédiaire (AI) est prévue pour éventuellement permettre de démontrer de manière anticipée le bénéfice du traitement sur la PFS. Cette AI de PFS est prévue lorsque 100 évènements de PFS (progression ou décès) auront été accumulés. Le moment où ce nombre est atteint définit alors la date de point (cut-off) de l'analyse

intermédiaire[2]. Le seuil de la signification statistique à atteindre pour pouvoir faire la conclusion anticipée de bénéfice sur la PFS est de 0.01. S'il n'est pas possible de faire cette conclusion anticipée (car le p nominal obtenu est supérieur au seuil), le bénéfice du traitement sur la PFS pourra être à nouveau testé lors de l'analyse finale de PFS qui est prévue à 200 évènements de PFS. Pour pouvoir conclure à ce bénéfice, à cette analyse, le p nominal devra être inférieur à 0.04.

Pour l'OS, les analyses intermédiaires et finales vont être décalées dans le temps, car les évènements d'OS (les décès) dont plus rare que ceux de PFS. Il faut donc un suivi plus important pour obtenir le même nombre d'évènements par rapport la PFS. En général il n'est pas effectué d'analyse intermédiaire d'OS au moment de l'AI de PFS, car le nombre de décès sera très faible, rendant improbable la possibilité de démontrer avec si peu d'événement un bénéfice du traitement.

La première AI d'OS survient donc en même temps que l'AF de PFS. Cette AI d'OS n'est donc pas définie en termes de nombre d'évènements comme les autres, mais est simplement synchrone de l'analyse de PFS[3]. Le seuil à atteindre pour conclure à un bénéfice sur l'OS est de 0.01. Si le p nominal n'est pas inférieur à ce seuil, aucune conclusion anticipée n'est possible et il convient d'attendre 200 décès pour déclencher l'AF à la recherche d'un éventuel bénéfice du traitement sur l'OS avec un seuil de la signification de 0.04.

Comme la multiplicité des critères de jugement a été gérée dans cette étude par une hiérarchisation (PFS puis OS), ces 2 séquences d'AI et d'AF sont à considérer l'une après l'autre pour respecter cette hiérarchie (et ainsi avoir un contrôle du risque alpha global).

En d'autres termes, lorsque l'hypothèse nulle est rejetée au cutoff 2 pour la survie sans progression et uniquement dans ce cas, l'analyse de la survie globale peut être considérée. Ceci implique que l'analyse de la survie globale ne pourra pas être testée si un bénéfice sur la survie sans progression n'est pas démontrée. Si l'analyse de survie globale ne permet pas de démontrer un bénéfice, l'essai n'est alors pas négatif mais n'aura démontré un bénéfice que sur la survie sans progression

> *Ainsi l'étude ne pourra démontrer un bénéfice en OS que si auparavant il a été possible de démontrer un bénéfice de PFS lors de l'AI de PFS ou de l'analyse finale de PFS. Si à l'analyse finale de PFS le p nominal n'est pas inférieur au seuil ajusté, aucun effet du traitement sur la PFS n'aura été démontré et l'étude s'arrête à ce stade. Le résultat obtenu sur l'OS sera purement exploratoire, quelle que soit sa valeur par rapport à son seuil ajusté (principe de la hiérarchisation des critères de jugement).*
>
> *Si la PFS est démontré (en AI ou en AF), l'OS pourra être envisagée à but de démonstration. Si lors de l'AI d'OS le p nominal n'est pas inférieur au seuil ajusté, aucune conclusion n'est possible et l'essai se poursuit vers l'analyse finale. Ainsi à ce stade un p nominal non inférieur au seuil ne veut pas dire que l'essai est négatif. Aucune conclusion n'est possible puisque l'essai se poursuit vers l'AF. Cependant, si à l'analyse finale, le p nominal obtenu ne franchit pas la limite du seuil de la signification, l'essai sera non concluant pour l'OS. Dans l'autre cas, il aura démontré le bénéfice du traitement sur la survie.*

[2] Cette date de point de l'analyse n'est donc pas fixée, mais dépends de la vitesse avec laquelle survient effectivement les évènements dans l'essai lui-même. Cette date ne peut donc pas être anticipée au protocole de l'étude. Dans ce protocole, seul figure le nombre d'événement définissant l'AI.

[3] Cette AI d'OS est donc indirectement conditionnée par le nombre d'évènements de PFS nécessaires à l'AF de PFS.

Remarque terminologique

Souvent il y a confusion dans le langage courant entre date de point (cut-off) et d'analyse intermédiaires. On parle ainsi de 1ère , 2ème analyse intermédiaire pour désigner en fait les cut-off.

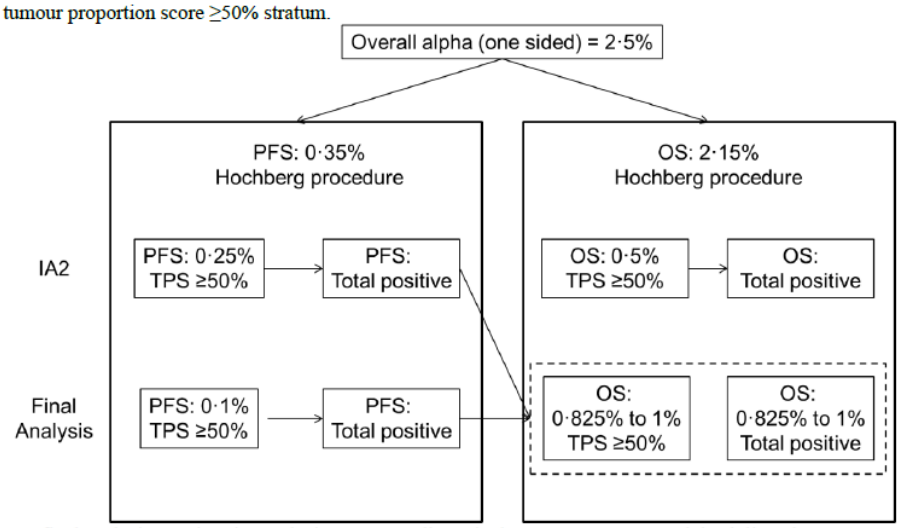
Si un seul critère de jugement est utilisé dans l'essai, ces appellations ne sont pas ambigües.

Par contre, en cas d'utilisation de plusieurs critères, comme ci-dessus, parler de la 2ème analyse intermédiaire pour désigner le 2ème cut-off n'est pas souhaitable, car, à cette date, une analyse finale de PFS et une analyse intermédiaire d'OS sont réalisées. Il est bien plus clair de dire qu'au 2ème cut-off, analyse finale de PFS et la première analyse intermédiaire d'OS sont réalisées.

Dans un autre cas de figure, la multiplicité des critères de jugement pourrait être gérée par une répartition du risque alpha. Dans ce cas, les 2 séquences d'AI/AF deviennent complètement indépendantes. Cependant les seuils ajustés de la signification sont calculés à partir d'un capital initial d'alpha, non plus de 5% (comme avec la hiérarchisation), mais de 2.5% (en cas d'équirépartition). Ces seuils ajustés seront plus petits avec la répartition qu'avec la hiérarchisation. L'intérêt de la répartition sera de pouvoir conclure sur l'OS même si aucun bénéfice sur la PFS n'a pu être démontré (en AI ou en AF) car les deux analyses sont strictement indépendantes

L'essai keynote-010 a évalué le pembrolizumab à 2 doses en 2eme ligne du cancer du poumon non à petites cellules [3]. La multiplicité provenait de 2 critères de jugement décisionnel (PFS et OS) et deux populations de patients (patients ayant un niveau d'expression du PD-L1 >= 50% et totalité des patients positifs au PD-L1 quel que soit le degré d'expression). Une analyse intermédiaire (IA2) était prévue. Le plan de contrôle du risque alpha fut le suivant :

Figure S2. Multiplicity control strategy. "Total" refers to the total population. "TPS ≥50%" refers to the PD-L1 tumour proportion score ≥50% stratum.



FA=final analysis; IA=interim analysis; OS=overall survival; PD-L1=programmed death ligand 1; PFS=progression-free survival; TPS=tumour proportion score.

Le risque alpha global total de l'étude (2.5% exprimé en unilatéral, ce qui correspond bien à 5% bilatéral) a été réparti entre la PFS 0.35% et l'OS 2.15%. Les populations de patients ont été hiérarchisées (d'abord les >=50% puis la population totale).
L'analyse intermédiaire (IA2) débutait par le test de la PFS chez les TPS>=50% avec une consommation de 0.25% des 0.35% d'alpha attribué à ce critère. Si ce test était « significatif » la PFS était testé dans la population totale avec le même risque alpha (hiérarchisation). Pour l'analyse finale, il restait 0.1% d'alpha (sur les 0.35% alloué à la PFS) avec le même test hiérarchisé en fonction des populations de patients.

Pour l'OS, l'analyse intermédiaire consommait 0.5% d'alpha par rapport au 2.15% attribué à ce critère (avec la même hiérarchisation pour les populations de patients que pour la PFS). Pour l'analyse finale d'OS, le risque alpha était déterminé avec un processus de réallocation (et divisé en 2 en raison des 2 doses[4]). Il pouvait donc aller de 0.825% (qui est 2.15% attribué initialement à l'OS moins les 0.5% consommés à l'analyse intermédiaire divisée par 2) à 1% (2.15% moins 0.5% plus 0.25% et 0.1% de réallocation en provenance de l'IA2 et de l'AF de la PFS si elles sont significatives et le tout divisé par 2).

[4] cf. le texte du protocole de l'essai pour la signification de l'encadré en pointillé non explicité dans la légende : « *At the final analysis, OS in the strongly positive PD-L1 stratum and the overall positive PD-L1 population will be tested simultaneously, with available alpha split evenly between the two tests* »

5 Autres objectifs possibles pour les analyses intermédiaires

5.1 Analyse de sécurité (safety)

À côté de l'objectif de mise en évidence anticipée de l'efficacité, l'autre grand objectif des analyses intermédiaires est celui de la surveillance de la sécurité (safety) du traitement. Le but est de détecter au plutôt un effet indésirable du nouveau produit afin de limiter l'exposition des patients à un produit qui pourrait être dangereux ou avoir une balance bénéfice risque défavorable.

Au niveau statistique ces analyses ne posent aucun problème en termes de risque alpha, car leur objectif n'est pas de faire admettre l'intérêt du produit, mais, au contraire, de conclure à l'absence d'intérêt. De plus, l'évaluation de la sécurité se fait suivant le principe de précaution et non pas suivant celui de la recherche de la preuve formelle. Une présomption suffisamment forte d'un effet indésirable gênant est suffisante pour faire abandonner le produit sans qu'il y ait besoin d'avoir une démonstration formelle de l'existence de cet effet indésirable.

Ainsi les analyses intermédiaires de sécurité s'effectuent sans recours à une méthode statistique particulière, souvent par une simple comparaison naïve des fréquences des évènements entre les 2 groupes (et par rapport à la fréquence attendue) et par une analyse type pharmacovigilance à la recherche d'évènements inattendus ou atypiques.

Par exemple le développement d'un des premiers anticoagulants directs oraux, le ximelagratran, a été interrompu par la mise en en évidence lors d'une analyse intermédiaire d'un essai de phase 3 dans la FA d'un problème d'hépatotoxicité [4].

> **Exemple de description des analyses intermédiaires de saf**ety - Interim safety assessments are planned to be performed twice a year. In order to identify potential safety signals earlier, a review of selected safety assessments is planned to be performed when 200 patients have completed the first 4 weeks of double-blind randomized treatment. Additionally, reviews of selected safety assessments are planned when the first 100, 300, and 600 patients completing the run-in period. Summary of the selected safety assessments during the run-in period will be provided for both treatments separately. No further alpha adjustment will be made due to these interim safety assessments.

Cependant, l'absence de tests statistiques ne résout en rien le problème de la multiplicité des analyses de la sécurité en ce qui concerne l'affirmation d'une plus grand iatrogénie. En effet, l'analyse de la sécurité consiste en l'observation de tableaux de multiples évènements indésirables entre deux groupes, fréquemment plusieurs centaines. Elle expose donc à un risque de disproportion liée au hasard, sans conséquence si il y a moins de proportion d'évènements sous traitement, mais problématique si il y a moins de proportion d'évènements sous placebo. On recourt alors à une analyse qualitative, facile lorsque la disproportion est majeure ou comportant un nombre d'évènements immortants, quand l'évènement est attendu, ou typique d'un lien de cause à effet. Elle est cependant beaucoup plus compliquée à analyser dans le cas où l'évènement est courant en population générale et que la disproportion est basée sur quelques évènements. Un exemple typique est la disproportion de paralysies faciales dans l'essai pivot du vaccin tozinameran contre la covid (4 cas de paralysie faciale dans le groupe vaciné versus 1 dans le groupe contrôle [5]) ayant conduit à noter cet effet indésirable

dans les résumés des caractéristiques du produit alors qu'une étude populationnelle a ultérieurement infirmé ce risque [6].

5.2 Autres types d'objectifs

Les analyses intermédiaires s'intègrent dans un processus global de surveillance des essais [7]. À côté de la recherche anticipée d'un effet du traitement et de la protection des personnes incluses dans l'essai, cette surveillance a pour objectif de vérifier le bon déroulement de l'essai. Il s'agit d'éviter des dérives dans la réalisation de l'essai, qui, si elles n'étaient détectées qu'à la fin, rendraient l'essai inutilisable en raison de défauts de qualité rédhibitoires.

- Les éléments à surveiller sont les suivants :
- Le taux d'écart au protocole : l'essai est-il de qualité ?
- Le taux d'inclusion : est-ce que l'essai pourra être réalisé dans un délai acceptable ?

Les caractéristiques des patients inclus : le risque de base des patients effectivement inclus correspond-il à celui initialement prévu et utilisé dans le calcul du nombre de sujets nécessaire ? Les patients recrutés correspondent-ils à la population cible de l'essai?

Cette surveillance permet de prendre au plus tôt des mesures correctrices. Les centres investigateurs ayant des difficultés à suivre le protocole pourront rectifier le tir. En cas de taux de recrutement insuffisant, d'autres centres investigateurs pourront être recrutés afin d'éviter qu'un essai dure trop longtemps. En effet, une durée excessive limite l'intérêt d'un essai.

Cette surveillance par analyse des données amassées se superpose à la surveillance « de terrain » de l'essai (appelé parfois « monitorage ») qui est focalisée sur le contrôle de qualité des données (visite de centres, contrôles des données, audit).

6 Les analyses intermédiaires en pratique

La réalisation d'une analyse statistique implique la levée de l'insu. Une organisation particulière est donc nécessaire pour éviter que la réalisation d'analyses intermédiaire perturbe la réalisation de l'essai en double insu et conduise à l'introduction de biais. En particulier le résultat de ces analyses doit rester inconnu de toutes les personnes impliquées dans la réalisation de l'essai : investigateurs, personnels de coordinations, promoteur. En effet, la divulgation des résultats des analyses intermédiaires pourrait avoir de nombreuses conséquences délétères pour l'essai : arrêt des inclusions en cas de tendance favorable et utilisation en pratique d'un traitement sans que la démonstration de l'efficacité n'ait pu être obtenue.

En pratique, les analyses intermédiaires sont réalisées par une structure indépendante de la coordination de l'essai. Les résultats de l'analyse sont communiqués à un comité de surveillance (« safety committee , Independent Data Monitoring Committee (IDMC), Data and Safety Monitoring Board (DSMB), Monitoring Committee, Data Monitoring Committee ») composé de personnes indépendantes. Ce comité émettra au vu des résultats des analyses statistiques une recommandation destinée au comité directeur de l'essai. Cette recommandation peut être de poursuivre le recrutement, d'interrompre l'essai à ce stade, de modifier le protocole.

6.1 Justification de la non-divulgation des résultats intermédiaires d'efficacité et/ou de sécurité en dehors du DSMB

La connaissance des résultats intermédiaires (véritables analyses intermédiaires ou résultats tabulés par groupe d'efficacité et/ou de sécurité) d'un essai par l'investigateur principal peut conduire à des modifications de l'étude qui vont favoriser l'obtention d'un résultat positif (et cela peut-être à tort) ou arrêter à tort un essai pour futilité. Ainsi l'essai ne sera plus un test loyal de l'hypothèse, où seule la réalité des données conditionnera la confirmation ou la réfutation de l'hypothèse thérapeutique. La connaissance des résultats intermédiaires au sens large par le *steering committee*, l'investigateur principal, ou les investigateurs remets ainsi en cause l'intégrité scientifique de l'étude et entraine une méconduite.

Encadré – FAME 2

L'essai FAME 2 évaluait la mesure de la réserve coronarienne pour guider la revascularisation coronarienne chez des patients ayant un angor stable [8]. L'abstract relate que l'essai a été interrompu prématurément :

> *Recruitment was halted prematurely after enrollment of 1220 patients (888 who underwent randomization and 332 enrolled in the registry) because of a significant between-group difference in the percentage of patients who had a primary end-point event:*

Mais le protocole précise qu'aucune analyse intermédiaire n'était planifiée (page 26) :

> **13.3.2. Interim Analysis**
> No interim analysis is planned.

La justification de l'arrêt prématuré de l'essai est l'existence d'une différence significative. Mais cette différence n'aurait jamais dû être connue de l'investigateur principal étant donné l'absence de méthode de prise en charge de la multiplicité engendrée par d'éventuelles analyses intermédiaires. Ainsi cette

décision d'arrêt est entièrement drivée par les résultats et rien ne permet d'écarter une sacralisation en tant que résultat démontré d'une simple fluctuation aléatoire particulièrement favorable, mais ne reflétant pas la réalité de l'efficacité du traitement (cf. encadré marche aléatoire de résultats ci-dessous). Malgré cette limite cet essai est à l'origine de l'adoption de cette technique, bien qu'un autre essai, FUTURE, toujours non publié, a débouché sur un surcroit de mortalité à une analyse intermédiaire[5].

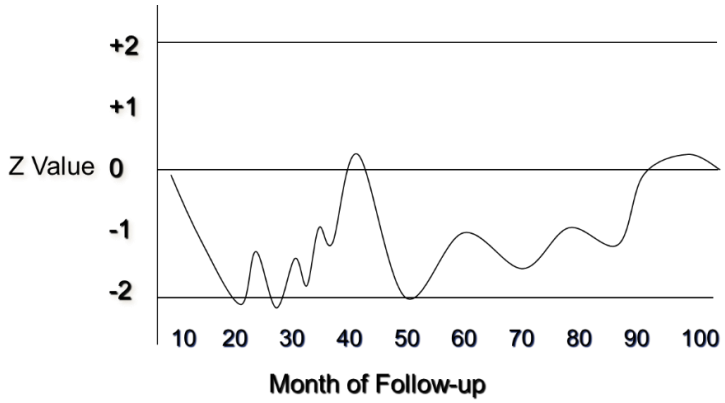
Les mauvaises pratiques à ce niveau vont invalider complètement la force de conviction des résultats obtenus et font courir le risque qu'en cas d'obtention d'un résultat positif, celui-ci ne puisse pas convaincre la communauté médicale et/ou les agences de régulation de changer les pratiques, car il y aura toujours un doute sur la fiabilité du résultat et donc sa réalité.

Plusieurs points sont à l'origine de cette problématique.

Tout d'abord des aspects statistiques. Les résultats que l'on peut être amenés à obtenir en cours d'étude sont sujets à une « marche aléatoire ». Ainsi, alors que l'essai sera négatif en ne montrant aucune différence en fin d'étude, il est tout à fait possible qu'une ou plusieurs différences apparaissent en cours d'étude. Ce point est pris en considération par les règles d'arrêts prématurés (O'Brien et Flemming par exemple).

Encadré - Marche aléatoire de résultats

Une autre façon de voir la problématique de la répétition des analyses à la recherche de l'effet du traitement passe par la représentation de la marche aléatoire des résultats en cours d'étude [9]. La figure ci-dessous obtenue avec l'essai Coronary Drug Project [10] représente l'évolution de la différence entre les 2 groupes au cours des mois de suivi de l'essai. La ligne horizontale d'ordonnée -2 représente la limite de la signification statistique nominale à p<0.05. À son terme (100° mois) l'essai a été négatif avec une absence de différence entre les 2 groupes. Mais rétrospectivement il apparait que par trois fois la différence entre les 2 groupes a été significative en faveur du traitement étudié. Si l'évolution de cette différence avait été scrutée en continu, par l'investigateur, de manière naïve sans méthode statistique, il aurait pu interrompre l'essai assez précocement qui aurait été positif en faveur du traitement.



Un autre exemple illustre cette marche aléatoire. L'analyse intermédiaire de l'efficacité du molupiravir dans le traitement oral de la Covid sur le critère combiné décès + hospitalisations chez 762 patients était de 7,3 % dans le groupe traité versus 14,1% dans le groupe placebo, soit une taille d'effet remarquable.

[5] Rioufol G. FUTURE trial: Treatment strategy in multivessel coronary disease patients based on fractional flow reserve. Presented at: ESC 2018. August 25, 2018. Munich, Germany

> L'analyse finale sur 1408 patients montrait un bénéfice, mais avec une taille d'effet bien moindre, 6,8 % dans le groupe traité versus 9,7 % dans le groupe placebo [11].

Mais beaucoup d'études ne prévoient pas d'arrêter prématurément pour démonstration anticipée de l'efficacité (ce qui n'est absolument pas un problème). Cependant si l'investigateur principal (IP) (ou le *steering committee*) prend connaissance des résultats de l'AI et perçoit que, si l'étude s'arrête là, l'étude serait concluante, cela peut conduire à ce qu'il recherche un moyen d'arrêter l'étude (même si cela n'était pas prévu). Par exemple si l'étude rencontre des difficultés de recrutement, l'IP peut prendre la décision de l'arrêter pour insuffisance de recrutement. L'éventuelle connaissance des résultats va influencer cette prise de décision :

- Si l'investigateur sait que s'il arrête à ce stade son étude est négative, il laissera vraisemblablement se poursuivre l'étude et fera tout pour augmenter le recrutement.
- L'analyse intermédiaire peut montrer une absence potentielle d'effet qui peut changer lors des analyses suivantes, entrainant un arrêt ou manque de motivation de continuer l'étude pour raison de futilité.
- Cependant s'il sait que les résultats sont en faveur de l'effet du traitement, il aura tendance à arrêter l'étude, prétextant l'insuffisance de recrutement. Mais sa décision sera entièrement orientée par le résultat, augmentant ainsi la probabilité que l'étude soit déclarée positive uniquement sur la base d'un résultat positif dû au hasard. De ce fait, l'étude n'est plus un test loyal de l'hypothèse thérapeutique et devient un test biaisé en faveur d'une sur-révélation de résultats positifs.

À côté de cela, la connaissance par l'IP des résultats intermédiaires peut aussi lui permettre de favoriser indument l'obtention d'un résultat positif par son étude, par exemple, en changeant le critère de jugement principal, en excluant du recrutement de certains patient, d'augmenter le nombre de sujets nécessaire, de changer le schéma posologique, etc. Toutes ces modifications font, là aussi, que l'essai ne devient plus un test loyal de l'hypothèse thérapeutique et aura été modifié en cours d'étude pour faire passer comme probante une tendance aléatoire. En cas de divulgation des résultats des analyses intermédiaires il devient aussi impossible de modifier le protocole par amendement, car il sera impossible de justifier que ces modifications ne sont pas drivées par les résultats et qu'elles sont donc entièrement post-hoc. Cela bloque donc toute adaptation de l'étude à un changement de contexte ce qui peut être dommageable pour l'étude.

La connaissance des résultats intermédiaires et de leur tendance par les investigateurs effectuant le recrutement peut aussi être très néfaste pour l'étude. Si les investigateurs ont connaissance que la tendance des résultats favorise le nouveau produit, ils pourront avoir une certaine réticence à continuer d'inclure des patients dans le groupe contrôle, car ils penseront peut-être (à tort, car il n'y 'a pas de preuve de cela et l'équipoise, clause d'ambivalence, reste valable) que ces patients ont une perte de chances. Cela conduit ces investigateurs à arrêter de recruter ou à provoquer de nombreuses sorties d'études. L'essai ne se terminera jamais et la preuve de l'intérêt du nouveau traitement ne pourra jamais être apportée.

Pour toutes ces raisons, il est donc indispensable que les résultats intermédiaires ne soient pas connus des décideurs de l'essai [12, 13, 14]. C'est pour cela que le principe de comité indépendant a été inventé, afin de permettre une étanchéité parfaite entre les personnes/structures qui ont connaissance des résultats intermédiaires et les structures/personnes qui conduisent l'étude :

- circuit indépendant des documents
- formulation sibylline des recommandations de poursuites de l'étude par le DSMB
- mise en aveugle des comités de lecture des évènements et du DSMB dans les études en ouvert
- création de parties ouvertes et fermées dans les réunions du DSMB
- non transmission des CR des réunions du DSMB par le chairman avant la fin de l'étude

Malheureusement pour les firmes introduites en bourses la législation financière les oblige à communiquer à leurs investisseurs toutes informations à leur disposition pouvant avoir un retentissement sur l'évolution des actions. Les résultats des analyses intermédiaires font partie de ces types d'informations et font donc l'objet de communiqué de presse. Mais si les résultats ne sont connus que du DSMB indépendant, l'information détenue par la firme et qu'elle doit rendre publique n'est que la recommandation laconique de poursuivre ou d'arrêter l'essai.

7 Spins de conclusion

Les résultats d'analyses intermédiaires sont à l'origine de nombreux spin de conclusion en raison de la relative complexité des principes d'interprétation et de leur méconnaissance par la majorité des lecteurs [15]. Ces spins vont présenter comme concluants (statistiquement significatif) des résultats qui ne le sont pas, mais avec lesquels la p value nominale est inférieure à 0.05.

> At the time of the interim analysis of overall survival, 173 patients in the continuous lenalidomide–dexamethasone group, 192 in the group that received 18 cycles of lenalidomide–dexamethasone, and 209 in the MPT group had died. The overall survival rates at 3 years were 70% with continuous lenalidomide–dexamethasone, 66% with 18 cycles of lenalidomide–dexamethasone, and 62% with MPT; the overall survival rates at 4 years were 59%, 56%, and 51%, respectively. Although the difference in overall survival did not cross the prespecified superiority boundary (P<0.0096), continuous lenalidomide–dexamethasone reduced the risk of death, as compared with MPT (hazard ratio, 0.78; 95% CI, 0.64 to 0.96; P=0.02) (Fig. 1B).

La dernière phrase de cet extrait est un véritable petit bijou d'astuce rédactionnelle, en disant tout et son contraire. En effet, en cas de spin, il est classique que la tournure de rédaction soit suffisamment ambiguë pour il puisse être contre-objecté, en cas de critique de la conclusion, qu'en fait il s'agit d'une maladresse et que l'on voit le mal partout, car il était bien dit que le résultat n'était pas significatif.

La première partie de la phrase mentionne bien (sans être pour autant très explicite) que le résultat n'est pas concluant : « *the difference … did not cross the … superiority boundary* ». La suite de la phrase mentionne une réduction de la mortalité et donne seulement la p value qui, lut sans expertise statistique suffisante, apparait comme significative, bien que cela ne soit pas non plus mentionné de manière explicite.

Dans cet autre exemple (figure suivante), les éléments permettant de comprendre que la p value de 0.04, bien mise en évidence sur la figure, n'est pas significative n'apparaissent que dans la légende de la figure. La tournure utilisée n'est pas explicite et n'est certainement compréhensible qu'avec un certain bagage statistique que beaucoup de lecteurs et décideurs n'ont pas.

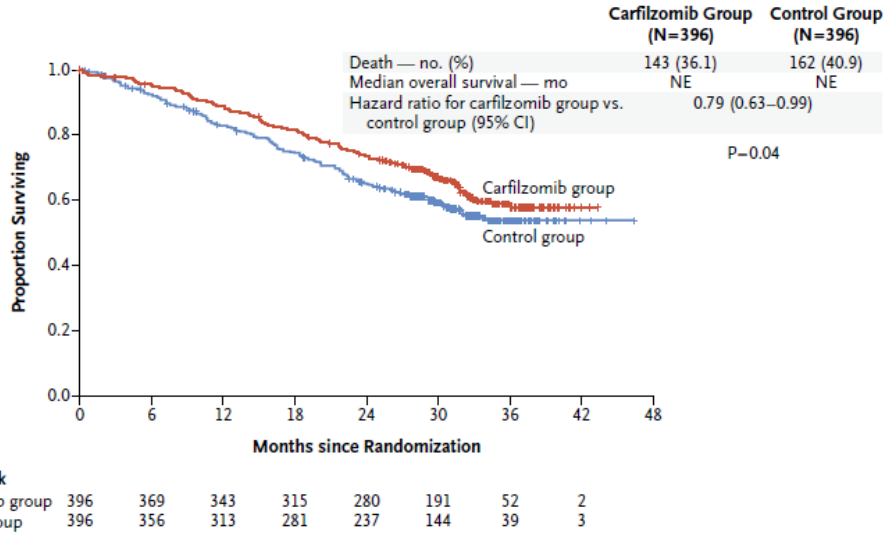


Figure 2. Overall Survival.

Kaplan–Meier estimates of overall survival in the intention-to-treat population are shown. The interim analysis of overall survival was performed after 305 deaths had occurred (60% of the prespecified 510 deaths for the final analysis). The interim analysis of overall survival did not cross the prespecified stopping boundary. NE denotes not estimable.

Pour cette raison, la nouvelle politique du NEJM de présentation des p values (cf. section **Erreur ! Source du renvoi introuvable.**) ne présente plus les p value dans ces situations, prévenant ainsi toute possibilité de spin.

La même ambigüité est présente dans l'abstract de l'article :

RESULTS

Progression-free survival was significantly improved with carfilzomib (median, 26.3 months, vs. 17.6 months in the control group; hazard ratio for progression or death, 0.69; 95% confidence interval [CI], 0.57 to 0.83; P=0.0001). The median overall survival was not reached in either group at the interim analysis. The Kaplan–Meier 24-month overall survival rates were 73.3% and 65.0% in the carfilzomib and control groups, respectively (hazard ratio for death, 0.79; 95% CI, 0.63 to 0.99; P=0.04). The

Références

1 Endometrial cancer and hormone-replacement therapy in the Million Women Study. *The Lancet* 2005;365:1543–51 doi:10.1016/S0140-6736(05)66455-0;

2 DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. *Stat Med* 1994;13:1341-52; discussion 1353-6 doi:10.1002/sim.4780131308; PMID:7973215;

3 Herbst RS, Baas P, Kim D-W, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *The Lancet* 2016;387:1540–50 doi:10.1016/S0140-6736(15)01281-7; PMID:26712084;

4 Boos CJ, Lip GYH. Ximelagatran: an eulogy. *Thrombosis Research* 2006;118:301–04 doi:10.1016/j.thromres.2006.02.012; PMID:16626788;

5 Thomas SJ, Moreira ED, Kitchin N, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine through 6 Months. *N Engl J Med* 2021 doi:10.1056/NEJMoa2110345; PMID:34525277;

6 Barda N, Dagan N, Ben-Shlomo Y, et al. Safety of the BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Setting. *N Engl J Med* 2021;385:1078–90 doi:10.1056/NEJMoa2110475; PMID:34432976;

7 Grant AM, Altman DG, Babiker AB et al. Issues in data monitoring and interim analysis of trials ;

8 Bruyne B de, Pijls NHJ, Kalesan B, et al. Fractional flow reserve-guided PCI versus medical therapy in stable coronary disease. *N Engl J Med* 2012;367:991–1001 doi:10.1056/NEJMoa1205361; PMID:22924638;

9 Jennison C, Turnbull BW. Statistical Approaches to Interim Monitoring of Medical Trials: A Review and Commentary. *Statistical Science* 1990;5:299–317 ;

10 The coronary drug project. Design, methods, and baseline results. *Circulation* 1973;47:I1-50 doi:10.1161/01.cir.47.3s1.i-1; PMID:4570454;

11 Jayk Bernal A, Da Gomes Silva MM, Musungaie DB, et al. Molnupiravir for Oral Treatment of Covid-19 in Nonhospitalized Patients. *N Engl J Med* 2021 doi:10.1056/NEJMoa2116044; PMID:34914868;

12 Anand SS, Wittes J, Yusuf S. What information should a sponsor of a randomized trial receive during its conduct? *Clin Trials* 2011;8:716–19 doi:10.1177/1740774511424421; PMID:22024103;

13 Borer JS, Gordon DJ, Geller NL. When should data and safety monitoring committees share interim results in cardiovascular trials? *JAMA* 2008;299:1710–12 doi:10.1001/jama.299.14.1710; PMID:18398083;

14 Ellenberg SS, Fleming TR, DeMets DL. Data monitoring in clinical trials: A practical perspective. Hoboken NJ: Wiley 2019 ISBN:9781119512653;

15 Wayant C, Vassar M. A comparison of matched interim analysis publications and final analysis publications in oncology clinical trials. *Annals of Oncology* 2018;29:2384–90 doi:10.1093/annonc/mdy447; PMID:30307531;