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 5 – Les analyses en sous-groupes

19 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

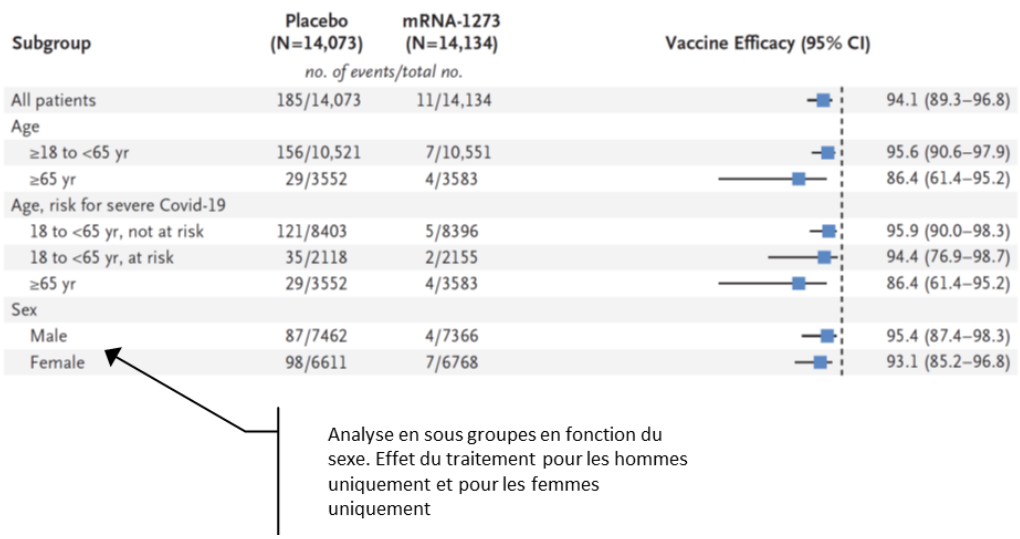
Matthieu Roustit

Table des matières

1 Principe des analyses en sous-groupes

Les analyses en sous-groupes consistent à chercher l'effet du traitement dans des sous-groupes de patients de l'essai. À cette fin, la population totale de l'étude est divisée en fonction d'une variable, par exemple le sexe. La comparaison groupe traité / groupe contrôle et l'estimation de la taille de l'effet du traitement sont alors effectuées pour chaque modalité de cette variable. Dans cet exemple, cette analyse en sous-groupe produira une estimation de l'effet du traitement chez les hommes et chez les femmes (cf. Figure 1). Cependant ces estimations pourront être trompeuses (dû fait des fluctuations aléatoires d'échantillonnage) induisant un fort risque de tirer des conclusions erronées de ces résultats.

Figure 1 – Exemple d'analyses en sous-groupes réalisées dans un essai de vaccins de la COVID19 [10.1056/NEJMoa2035389].
Le résultat de l'essai (all patient) est rappelé en haut du graphique. Plusieurs analyses en sous-groupes sont représentées. La première est l'analyse en fonction de l'âge avec 2 modalités : entre 18 et 65 et supérieure à 65. L'efficacité du vaccin est estimée spécifiquement pour chacune de ces 2 modalités : 95.6% pour les 18-65 ans et 86.4% pour les 65 et plus. L'efficacité vaccinale est la réduction relative du risque (= 1-risque relatif*100%).



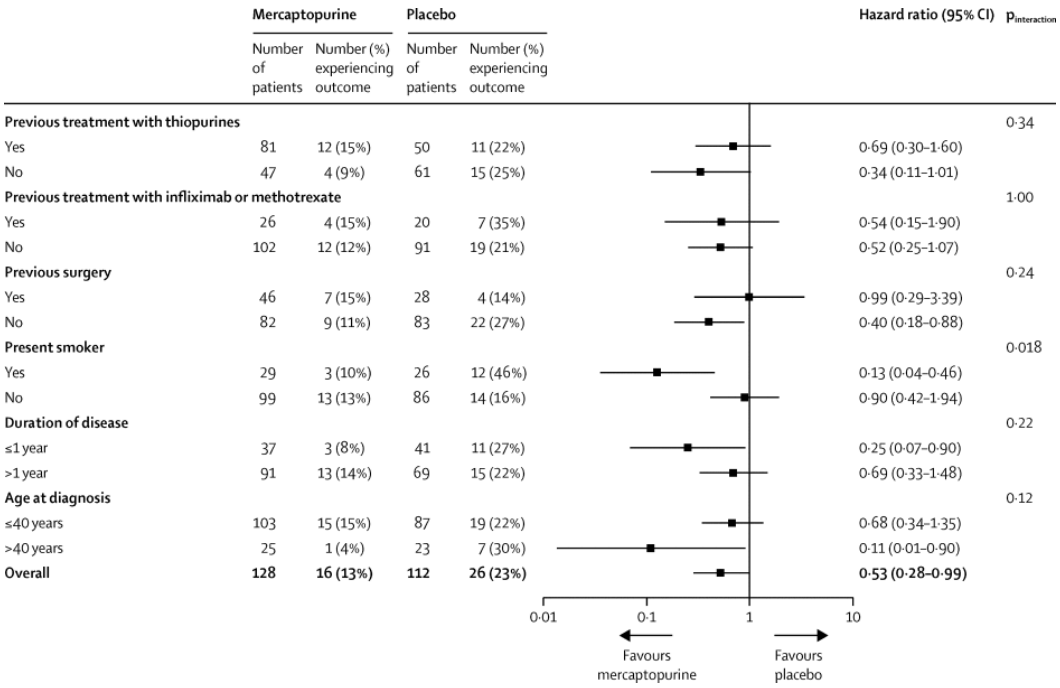
Les analyses en sous-groupes ont des limites statistiques importantes qui empêchent de les utiliser pour conclure à l'effet du traitement ou à son absence au niveau des sous-types de patients.

L'analyse en sous-groupes est typiquement une fausse bonne idée. L'objectif est très pertinent, car on peut imaginer que l'effet d'un traitement (en termes d'efficacité et en termes de sécurité) ne soit pas

identique quelles que soient les caractéristiques des patients. Mais les analyses de sous-groupes réalisées de façon purement exploratoire, sans hypothèse spécifique, ne vont pas permettre d'atteindre ces objectifs et peuvent conduites à établir des fausses vérités.

Au mieux, les analyses en sous-groupes permettent de générer de nouvelles hypothèses, à vérifier dans de nouveaux essais entrepris spécialement.

L'idée sous-jacente est celle de l'existence de facteurs modifiant l'effet du traitement, appelé aussi modificateurs (« modifiers » ou « moderators »). Ces facteurs peuvent être des caractéristiques des patients ou de la maladie[1]. L'identification des modificateurs d'effet exige une méthodologie spécifique allant bien au-delà de la simple analyse en sous-groupe (cf. section 6).



	Mercaptopurine		Placebo		Hazard ratio (95% CI)	$p_{interaction}$
	Number of patients	Number (%) experiencing outcome	Number of patients	Number (%) experiencing outcome		
Previous treatment with thiopurines						0·34
Yes	81	12 (15%)	50	11 (22%)	0·69 (0·30–1·60)	
No	47	4 (9%)	61	15 (25%)	0·34 (0·11–1·01)	
Previous treatment with infliximab or methotrexate						1·00
Yes	26	4 (15%)	20	7 (35%)	0·54 (0·15–1·90)	
No	102	12 (12%)	91	19 (21%)	0·52 (0·25–1·07)	
Previous surgery						0·24
Yes	46	7 (15%)	28	4 (14%)	0·99 (0·29–3·39)	
No	82	9 (11%)	83	22 (27%)	0·40 (0·18–0·88)	
Present smoker						0·018
Yes	29	3 (10%)	26	12 (46%)	0·13 (0·04–0·46)	
No	99	13 (13%)	86	14 (16%)	0·90 (0·42–1·94)	
Duration of disease						0·22
≤1 year	37	3 (8%)	41	11 (27%)	0·25 (0·07–0·90)	
>1 year	91	13 (14%)	69	15 (22%)	0·69 (0·33–1·48)	
Age at diagnosis						0·12
≤40 years	103	15 (15%)	87	19 (22%)	0·68 (0·34–1·35)	
>40 years	25	1 (4%)	23	7 (30%)	0·11 (0·01–0·90)	
Overall	**128**	**16 (13%)**	**112**	**26 (23%)**	**0·53 (0·28–0·99)**	

Figure 2 - Exemple de représentation graphique en forest plot des analyses en sous-groupe d'un essai thérapeutique.

[1] L'enjeu de la médecine de précision est de trouver ces modificateurs.

2 Les limites des analyses en sous-groupes

La première problématique majeure des analyses en sous-groupes est le risque important de fausses découvertes qu'elles font courir du fait de leur multiplicité et des conséquences de cette multiplicité en termes de risque alpha et beta [1, 2].

Des résultats très contrastés entre les sous-groupes peuvent être observés uniquement du fait des fluctuations aléatoires d'échantillonnage. De ce fait, il est impossible de distinguer un résultat lié à une réelle modification de l'effet du traitement d'un résultat artéfactuel, lié à une fausse découverte due au hasard. Le risque de prendre des décisions à tort est non contrôlés.

Il est extrêmement risqué de conclure à l'efficacité du traitement ou à son absence pour certains sous-types de patients à partir des analyses en sous-groupes ordinaires

---

2.1 Cas d'un essai non concluant (« négatif »)

Dans le cas d'un essai « négatif » (non concluant en raison d'un résultat non-significatif), il peut être tentant de conclure quand même à l'intérêt du traitement évalué, mais restreint à un type particulier de patients, car un résultat statistiquement significatif est obtenu dans le sous-groupe de ces patients. L'idée générale est de dire que le traitement ne « marche » pas chez tous les patients, mais seulement chez certains patients qui peuvent ainsi en bénéficier[2].

La problématique est que la multiplication des recherches d'effet et de signification statistique entraine une inflation considérable du risque alpha global. Les données sont découpées dans « tous les sens » jusqu'à l'obtention d'une différence significative.

Le fait que les sous-groupes soient prévus a priori ne change rien à la problématique, car il y a toujours multiplicité des comparaisons, certes moindre qu'en cas de « data dregging » extensif comme évoqué précédemment, mais qui induit toujours une inflation du risque alpha global. De plus il n'est jamais spécifié à priori dans quel sens la modification de l'effet dans ces sous-groupes est attendue. Or les fluctuations aléatoires peuvent conduire à une fausse découverte soit dans le sens de la supériorité soit dans celui de l'infériorité. Comme nous le verrons plus loin, la solution est d'intégrer les analyses en sous-groupes dans un le plan de contrôle strict du risque alpha global (cf. section 6).

La Figure 3 illustre cette problématique. Un essai de 2400 patients simulé avec un traitement réellement sans effet (ratio des risques de 1), est divisé en 12 parties distinctes de manière purement aléatoire. Il n'y a donc aucune raison déterministe que l'effet du traitement soit différent entre ces sous-groupes. Le résultat global de l'essai reflète bien la réalité de l'effet avec un risque ratio observé de 0.96, non significatif. Mais il est possible de conclure à la supériorité du traitement avec le résultat du sous-groupe 10 et aussi à un effet délétère pour le sous-groupe 5 (pour les effets délétères, la

[2] Cependant, lors de la conception de l'étude, les patients à inclure dans un essai ont été soigneusement déterminés en considérant qu'ils étaient tous susceptibles de bénéficier du traitement, et cela en se basant sur ce qui est connu de la physiopathologie et de la pharmacologie. Après avoir obtenu les résultats, partir à tâtons dans les analyses en sous-groupe pour celui des patients bénéficiant du traitement sera donc une démarche complètement exploratoire, non basée sur des hypothèses anticipées et post-hoc.

signification statistique n'est pas obligatoire pour conclure, car le raisonnement se base sur le principe de précaution, cf. section 8).

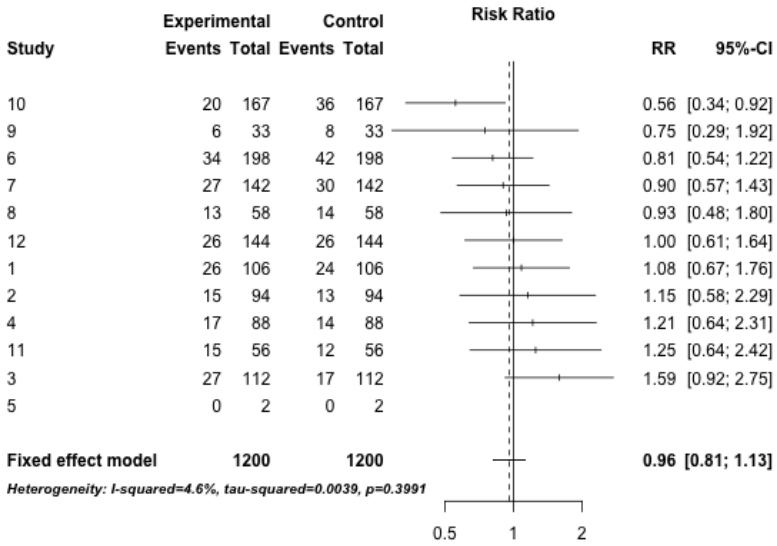


Figure 3 – Visualisation des fluctuations aléatoires et de l'inflation du risque alpha ainsi induit par la subdivision en 12 parties de manière complètement aléatoire d'un essai simulé avec un traitement sans effet

Cette problématique disparait quand les analyses en sous-groupe sont intégrées dans le plan de contrôle strict du risque alpha global et on fait l'objet d'un calcul d'effectif spécifique. Leurs résultats sont alors décisionnels. Sans cette approche, les sous-groupes ne peuvent servir qu'à évaluer la généralisabilité du résultat de l'essai à l'ensemble des sous types de patients inclus dans l'essai avec l'aide exploratoire du test d'interaction.

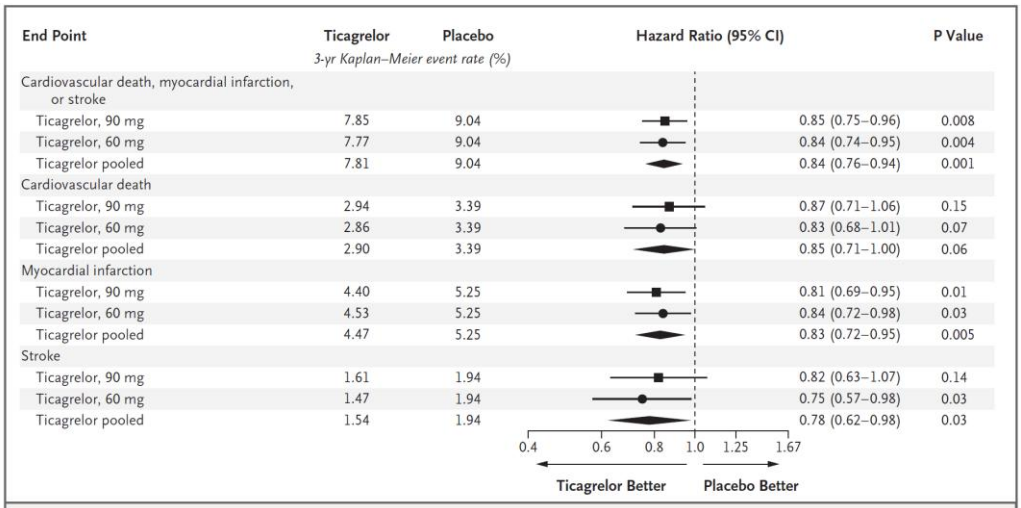
Actuellement, ces limites méthodologiques sont de moins en moins prises en considération en dehors des cercles centrés sur la méthodologique, les statistiques et la régulation [3]. Une grande partie des résultats discutés dans les congrès médicaux par exemple sont issus de sous-groupes, en particulier dans l'optique de chercher à « personnaliser » au mieux les traitements (médecine de précision). Dans ces discussions les conséquences des décisions prises à tort, dû fait du hasard, ne sont en général pas considérées.

Dans un essai « négatif » (non concluant), les sous-groupes ne permettent pas de conclure à l'effet du traitement pour un sous-type de patients particulier, car il existe une inflation du risque alpha global lié à la multiplicité des comparaisons induites par les sous-groupes (souvent plusieurs dizaines, voire centaines).

2.1.1 Étude de cas : Pegasus

L'essai Pegasus [4] a évalué le ticagrelor en prévention secondaire cardiovasculaire 1 an après l'événement ischémique initial. Deux doses de ticagrelor ont été comparées au clopidogrel sur les

événements cardiovasculaires. Aucune différence notable n'a été observée entre les 2 doses, faisant conclure à l'absence de différence d'efficacité entre les 2 doses testées (comme c'est souvent le cas avec les antiagrégants).

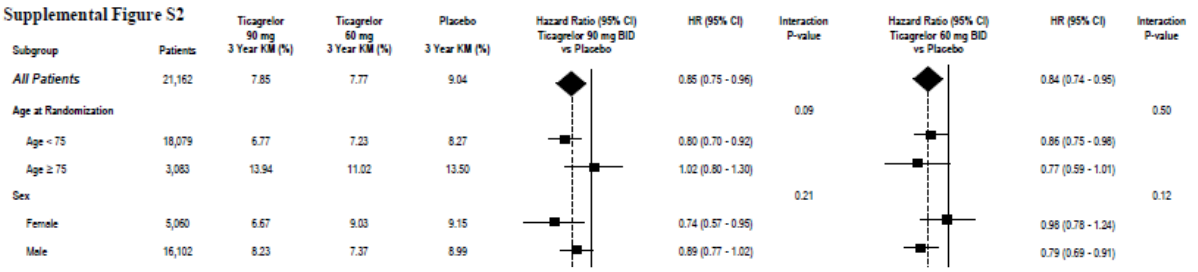


End Point	Ticagrelor	Placebo	Hazard Ratio (95% CI)	P Value
	3-yr Kaplan–Meier event rate (%)			
Cardiovascular death, myocardial infarction, or stroke				
Ticagrelor, 90 mg	7.85	9.04	0.85 (0.75–0.96)	0.008
Ticagrelor, 60 mg	7.77	9.04	0.84 (0.74–0.95)	0.004
Ticagrelor pooled	7.81	9.04	0.84 (0.76–0.94)	0.001
Cardiovascular death				
Ticagrelor, 90 mg	2.94	3.39	0.87 (0.71–1.06)	0.15
Ticagrelor, 60 mg	2.86	3.39	0.83 (0.68–1.01)	0.07
Ticagrelor pooled	2.90	3.39	0.85 (0.71–1.00)	0.06
Myocardial infarction				
Ticagrelor, 90 mg	4.40	5.25	0.81 (0.69–0.95)	0.01
Ticagrelor, 60 mg	4.53	5.25	0.84 (0.72–0.98)	0.03
Ticagrelor pooled	4.47	5.25	0.83 (0.72–0.95)	0.005
Stroke				
Ticagrelor, 90 mg	1.61	1.94	0.82 (0.63–1.07)	0.14
Ticagrelor, 60 mg	1.47	1.94	0.75 (0.57–0.98)	0.03
Ticagrelor pooled	1.54	1.94	0.78 (0.62–0.98)	0.03

Figure 2. Hazard Ratios and Rates of the Primary End Point and Individual Components for Each Dose of Ticagrelor and for the Two Doses Pooled.

The squares and circles reflect the point estimates for the ticagrelor 90-mg group and the ticagrelor 60-mg group, respectively, as compared with placebo. The horizontal lines indicate 95% confidence intervals. The diamonds indicate both the point estimate (center of diamond) and the 95% confidence interval (width) for the two doses pooled versus placebo.

Les analyses en sous-groupes ont été réalisées en fonction de la dose, ce qui augmente la multiplicité et permet aussi d'avoir une réplication de chaque analyse en sous-groupe. Comme il n'existe pas de différence d'effet entre les 2 doses, la réplication de la même analyse en sous-groupe avec l'autre dose devrait conduire à une modification (ou non) d'effet identique. Des interactions opposées sont même observées au niveau de plusieurs sous-groupes (comme l'âge et le sexe) ce qui montre bien l'influence des fluctuations aléatoires. Avec l'âge, pour la dose de 90mg, il pourrait être conclu que l'efficacité du ticagrelor est modifié par l'âge et qu'il n'est pas supérieur au clopidogrel chez les plus de 75 ans. Avec la dose de 60mg, on pourrait faire la conclusion inverse avec une efficacité apparemment supérieure chez les personnes les plus âgées. Une même différence de sens de l'interaction entre les 2 doses est observée avec le sexe.



Supplemental Figure S2

Subgroup	Patients	Ticagrelor 90 mg 3 Year KM (%)	Ticagrelor 60 mg 3 Year KM (%)	Placebo 3 Year KM (%)	Hazard Ratio (95% CI) Ticagrelor 90 mg BID vs Placebo	HR (95% CI)	Interaction P-value	Hazard Ratio (95% CI) Ticagrelor 60 mg BID vs Placebo	HR (95% CI)	Interaction P-value
All Patients	21,162	7.85	7.77	9.04		0.85 (0.75 - 0.96)			0.84 (0.74 - 0.95)	
Age at Randomization							0.09			0.50
Age < 75	18,079	6.77	7.23	8.27		0.80 (0.70 - 0.92)			0.86 (0.75 - 0.98)	
Age ≥ 75	3,083	13.94	11.02	13.50		1.02 (0.80 - 1.30)			0.77 (0.59 - 1.01)	
Sex							0.21			0.12
Female	5,060	6.67	9.03	9.15		0.74 (0.57 - 0.95)			0.98 (0.78 - 1.24)	
Male	16,102	8.23	7.37	8.99		0.89 (0.77 - 1.02)			0.79 (0.69 - 0.91)	

En général aucune p value du test de l'existence de l'effet du traitement n'est rapportée au niveau des sous-groupes. Malgré cela l'intervalle de confiance est souvent abusivement utilisé pour déterminer la signification du résultat.

La préspécification des sous-groupes au protocole ne solutionne pas la problématique de l'inflation du risque alpha liée à la multiplicité.

2.1.2 Etude de cas : Plato

Pour permettre de conclure à l'intérêt particulier du traitement pour un ou des sous-groupes de patients, ce ou ces sous-groupes doivent être inclus dans le plan de contrôle du risque alpha global (par hiérarchisation ou répartition du risque alpha, cf. section 6).

> L'essai Plato [10.1056/NEJMoa0904327] évaluait le ticagrelor lors d'un syndrome coronarien aigu (SCA). Dans cette pathologie les patients peuvent bénéficier d'une procédure invasive de revascularisation (PCI, stent). Ces patients représentent une population bien particulière par rapport à ceux traités exclusivement de manière médicale. Il convient d'avoir la preuve formelle que le ticagrelor apporte bien un bénéfice chez eux. Il y a donc nécessité de pouvoir conclure sur un sous-groupe de patients (ceux traités invasivement). Ce sous-groupe a pour cela été inclus dans une analyse avec contrôle de risque alpha global par hiérarchisation des tests en 2ème position (cf. section 6.1) :
> « The primary efficacy variable was the time to the first occurrence of composite of death from vascular causes, myocardial infarction, or stroke. … The principal secondary efficacy end point was the primary efficacy variable studied in the subgroup of patients for whom invasive management was planned at randomization. Additional secondary end points (analyzed for the entire study population) were …".
> Sans cette inclusion dans le plan de contrôle du risque alpha global, ce résultat aurait été uniquement exploratoire et n'aurait pas permis de décider si le ticagrelor avait sa place dans la prise en charge des SCA traités invasivement.

2.2 Cas d'un essai concluant (« positif »)

Dans le cas d'un essai concluant (qui a obtenu un résultat statistiquement significatif en faveur de la supériorité du nouveau traitement), les analyses en sous-groupes peuvent être utilisées pour chercher d'éventuels patients chez lesquels le traitement n'apporterait pas, ou trop peu, de bénéfice et qui ne serait pas donc pas à traiter avec ce traitement.

La problématique statistique sous-jacente est triple.

Il existe, tout d'abord, une inflation du risque beta, qui est le risque de ne pas trouver une différence alors que cette différence existe réellement. En multipliant les comparaisons, le risque de conclure à tort à une absence d'effet dans au moins un cas de figure où le traitement est réellement efficace augmente.

La Figure 4 illustre les conséquences en termes d'erreur beta globale des fluctuations aléatoires d'échantillonnages liées à la multiplicité des sous-groupes. Il s'agit d'un essai simulé avec un traitement efficace (vrai risque ratio de 0.50) et qui globalement, sur l'ensemble des patients inclus, amène à faire la bonne conclusion. Cet essai a été divisé de manière purement aléatoire en 12 parties. Comme cette division est purement aléatoire, il n'y a aucune raison mécanistique que l'efficacité du traitement soit modifiée dans ces sous-groupes. Cependant, l'estimation ponctuelle du risque ratio fluctue entre ces sous-groupes purement du fait du hasard. Plusieurs sous-groupes pourraient amener à conclure à l'absence de bénéfice devant un résultat du sous-groupe nominalement non significatif.

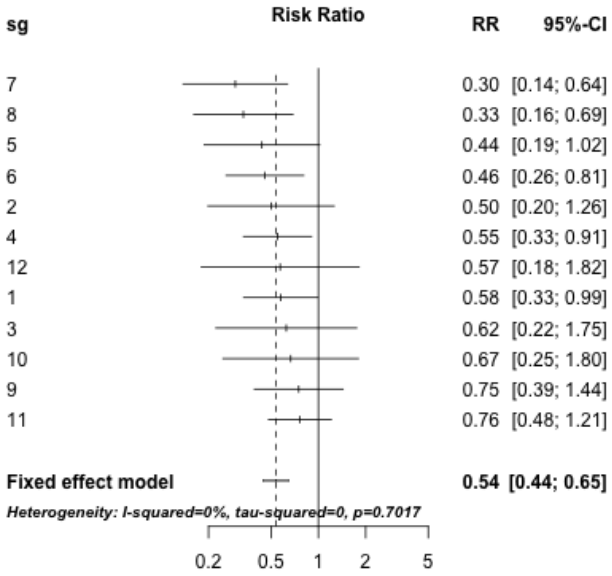


Figure 4 – Illustration des conséquences des fluctuations aléatoires sur les résultats des sous-groupes dans un essai simulé avec un traitement réellement efficace (vrai risque ratio de 0.5). Plusieurs de ces résultats pourraient conduire à conclure à tort à l'absence d'effet.

La deuxième problématique statistique est celle de la réduction d'effectif dans les sous-groupes qui potentialise l'inflation du risque beta. Étant de taille inférieure à l'essai, la précision des estimations est moindre (les intervalles de confiance sont plus larges que celui du résultat de l'essai). Par exemple dans la Figure 4, le sous-groupe n°12 donne la même estimation d'effet traitement que l'essai, mais par réduction de son effectif le résultat n'est plus significatif.

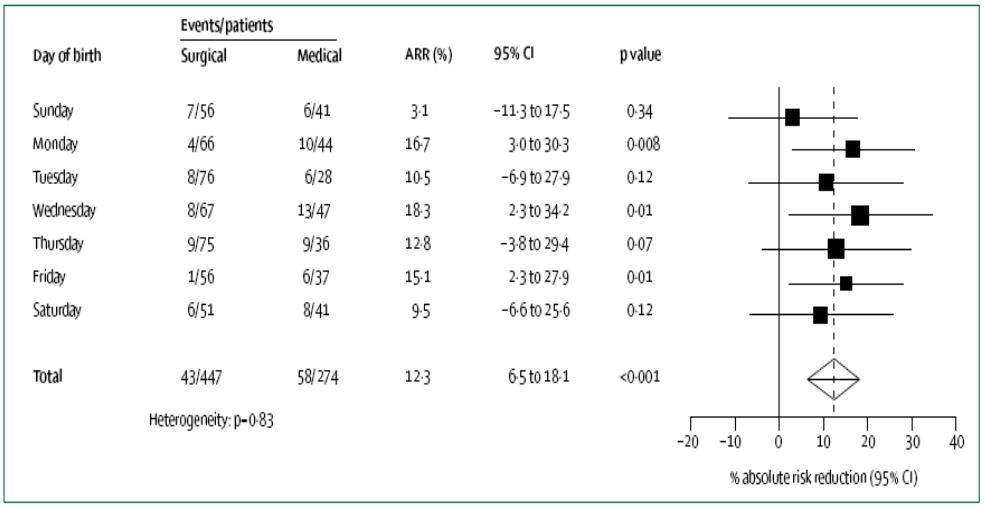
La troisième problématique est celle de conclure à l'absence d'effet devant une différence non significative. Elle est le corolaire de la problématique précédente. En effet il n'est pas possible de conclure à l'absence d'effet devant une différence non significative, car une l'absence de signification peut provenir de 2 phénomènes non distinguables : une réelle absence d'effet ou un manque de puissance. Une conclusion d'absence d'effet devrait se baser sur une approche type essai de non-infériorité pour gérer correctement cette problématique.

Dans un essai concluant (montrant l'intérêt du traitement au niveau global), les analyses en sous-groupe ne permettent pas de conclure à l'absence d'effet pour certains sous-types de patients en raison de : 1) Inflation du risque beta (de ne pas conclure à tort à l'effet du traitement) liée à la multiplicité ; 2) Réduction d'effectif, entrainant une réduction de la précision des estimations (largeur des intervalles de confiance) et de la puissance statistique ; 3) Conclusion à l'absence d'effet à partir d'une différence non significative impossible

---

Ces problématiques ont été illustrées par plusieurs analyses de sous-groupes pédagogiques basées sur des variables insolites (comme les signes du zodiaque [5]).

Dans un essai montrant la supériorité de la chirurgie d'endartériectomie par rapport au traitement médical chez des patients ayant un antécédent d'AVC et une sténose serrée d'une artère carotide. L'essai démontre une réduction absolue de la fréquence de récidive des AVC de 12%. Une analyse en sous-groupe a été réalisée en fonction du jour de la semaine de la naissance[6].



Day of birth	Events/patients Surgical	Medical	ARR (%)	95% CI	p value
Sunday	7/56	6/41	3·1	−11·3 to 17·5	0·34
Monday	4/66	10/44	16·7	3·0 to 30·3	0·008
Tuesday	8/76	6/28	10·5	−6·9 to 27·9	0·12
Wednesday	8/67	13/47	18·3	2·3 to 34·2	0·01
Thursday	9/75	9/36	12·8	−3·8 to 29·4	0·07
Friday	1/56	6/37	15·1	2·3 to 27·9	0·01
Saturday	6/51	8/41	9·5	−6·6 to 25·6	0·12
Total	43/447	58/274	12·3	6·5 to 18·1	<0·001

Heterogeneity: p=0·83

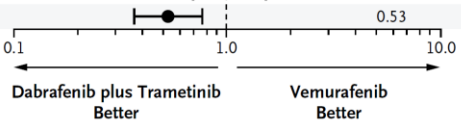
Figure 2: Effect of carotid endarterectomy in patients with ⩾70% symptomatic stenosis in ECST[126] according to day of week on which patients were born

Il est peu probable que le bénéfice de la chirurgie soit modifié, 50-70 après, par le jour de la semaine de la naissance, mais les résultats obtenus pourraient conduire à conclure à l'absence de bénéfice pour ceux nés un dimanche, un mardi, un jeudi ou un samedi. Le but de cette démonstration par l'absurde et de montrer les dangers auxquels exposent les analyses en sous-groupe en cas d'essais positifs. Leurs résultats peuvent être le pur produit du hasard et ne refléter en rien une absence de bénéfice pour certains patients. Bien sûr s'il existe une véritable modification de l'effet, cela va conduire à des différences d'effet entre les sous-groupes (avec de potentielles distorsions du fait du hasard). Mais en pratique, la problématique est que devant des résultats différents entre les sous-groupes, il est impossible de savoir si cela provient que du hasard ou d'un vrai déterminisme.

L'association dabrafenib trametinib a été évaluée en 1er ligne dans le mélanome métastatique versus vemurafenib [7]. L'essai montre une réduction de la mortalité totale avec un hazard ratio de 0.69 95% CI, 0.53 to 0.89; P = 0.005. Cependant le résultat suivant est obtenu dans les analyses en sous-groupe au niveau du performance status ECOG :

B **Overall Survival in Subgroups**

Subgroup	No. of Patients	Hazard Ratio (95% CI)
Age		
<65 yr	538	0.74
≥65 yr	166	0.61
Sex		
Male	388	0.87
Female	316	0.46
BRAF mutation		
V600K	68	0.85
V600E	629	0.70
Tumor stage		
IIIc, IVM1a, or IVM1b	273	0.40
IVM1c	429	0.77
Baseline LDH		
≤ULN	471	0.58
>ULN	232	0.78
No. of disease sites		
≤2	378	0.59
≥3	325	0.68
Baseline ECOG		
1	206	1.03
0	496	0.53



L'analyse en sous-groupe en fonction du performance statuts ECOG suggère l'absence de bénéfice chez les ECOG de 1 avec à la fois un hazard ration très proche de 1 (1.03) et un résultat nominalement non significatif. Aucune décision de restriction de l'utilisation aux patients ECOG de 0 n'a été prise devant ce résultat qui a été considéré comme étant de l'ordre des fluctuations aléatoires.

Il n'est donc pas possible de conclure que certains patients ne sont pas répondeurs au traitement à partir d'analyse en sous-groupes ordinaires sans utilisation d'une méthode statistique adaptée.

---

3 L'interaction statistique

Les analyses en sous-groupes peuvent être abordées d'une façon complètement différente, bien au-delà de la simple recherche de la démonstration statistique de l'effet à l'intérieur des sous-groupes (qui comme nous venons de le voir dans la section précédente est très problématique).

Il s'agit de la recherche d'une interaction, c'est-à-dire de savoir si la variable, en fonction de laquelle sont définis les sous-groupes (âge, sexe, etc.), modifie l'intensité de l'effet du traitement. Par exemple si un hazard ratio de 0.45 est observé dans le sous-groupe des hommes et 0.80 dans celui des femmes, la question est de savoir si ces 2 valeurs sont statistiquement différentes, compte tenu de leur incertitude statistique (matérialisée, par exemple, par leur intervalle de confiance respectif).

La réponse à cette question est apportée par le test d'interaction qui est significatif ($p<0.05$) lorsque les effets traitements sont significativement différents les uns des autres. Ce test permet de savoir s'il y a autre chose que le hasard derrière les différences observées entre les sous-groupes. Ce test n'a donc aucune relation avec le test d'existence d'un effet non nul qui peut être fait au sein de chaque sous-groupe (et qui a les limitations que nous venons d'exposer dans la section précédente).

La Figure 5 illustre trois situations différentes d'interaction. En A, le même hazard ratio a été observée dans les deux sous-groupes. Il y a absence d'interaction avec une p velue du test d'interaction proche de 1. En B, le hazard ratio n'est pas le même dans les 2 sous-groupes, mais les intervalles de confiance se chevauchent largement. Ces 2 estimations ne sont pas vraiment différentes, compte tenu de l'incertitude entourant ces estimations. Le test d'interaction n'est pas significatif. En C cependant, les résultats des sous-groupes sont significativement différents comme le témoigne le test d'interaction avec une p value <0.05. Les intervalles de confiance ne se chevauchent plus[3]. L'effet du traitement est donc vraisemblablement différent entre ces 2 sous-groupes.

On constate qu'il ne s'agit pas de rechercher un bénéfice du traitement dans un des sous-groupes, l'effet traitement étant observé pour les deux modalités du sous-groupe. De ce fait, l'existence d'un test d'interaction significatif ne signifie pas que le traitement est efficace dans un sous-groupe et sans effet dans un autre sous-groupe comme le montre la Figure 5-C. L'interaction signifie simplement que les effets sont différents quantitativement et cela ne préjuge pas de l'existence ou non d'un effet dans un sous-groupe.

[3] Il y a interaction à partir du moment où l'estimation ponctuelle d'un résultat n'est pas comprise dans l'intervalle de confiance de l'autre résultat. Il n'est pas nécessaire que les 2 intervalles soient complètement disjoints.

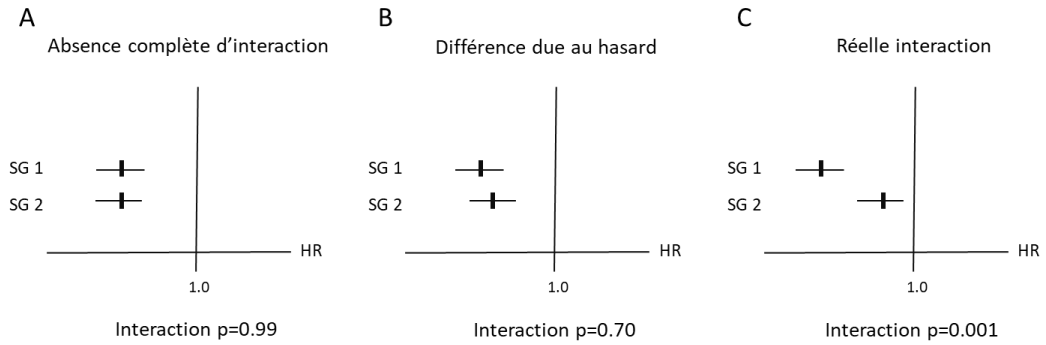


Figure 5 – Différents cas de figure d'interaction

De ce fait l'interaction ne permet pas de conclure à l'existence de l'efficacité au sein des sous-groupes. Son utilisation ne pose pas les problèmes statistiques évoqués plus haut. L'interaction à a une vocation uniquement exploratoire, pour documenter et non pas pour décider de changement dans la pratique. Cette exploration n'a donc pas d'enjeu décisionnel. Mais le risque de découverte de fausse interaction existe bel et bien exposant au risque de discussion fondée sur des artefacts statistiques induits par la multiplicité des comparaisons

La Figure 7 montre un exemple de graphique des résultats des sous-groupes.

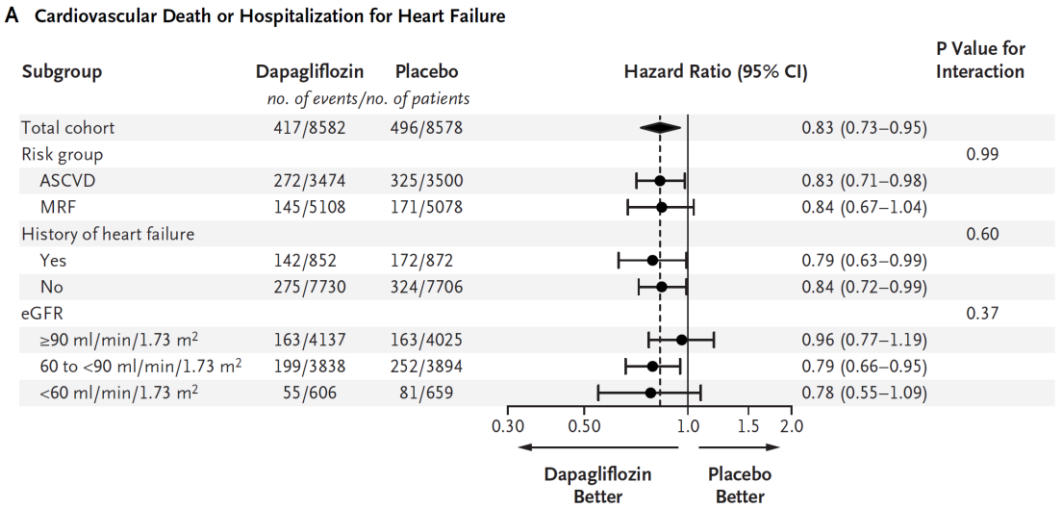


Figure 6 – Exemple d'analyse en sous-groupes avec test d'interaction.
Pour l'analyse en sous-groupe en fonction des antécédents d'insuffisance cardiaque, le p du test d'interaction est de 0.60, ne permettant pas de conclure qu'il existe une différence statistiquement significative entre l'effet du traitement chez les patients ayant un antécédent (HR=0.79) par rapport à l'effet chez les patients sans antécédents (HR=0.84). Compte tenu de l'incertitude entourant ces 2 estimations, il n'est pas possible de conclure que ces 2 hazard ratio (0.79 et 0.84) sont différents.
[10.1056/NEJMoa1812389]

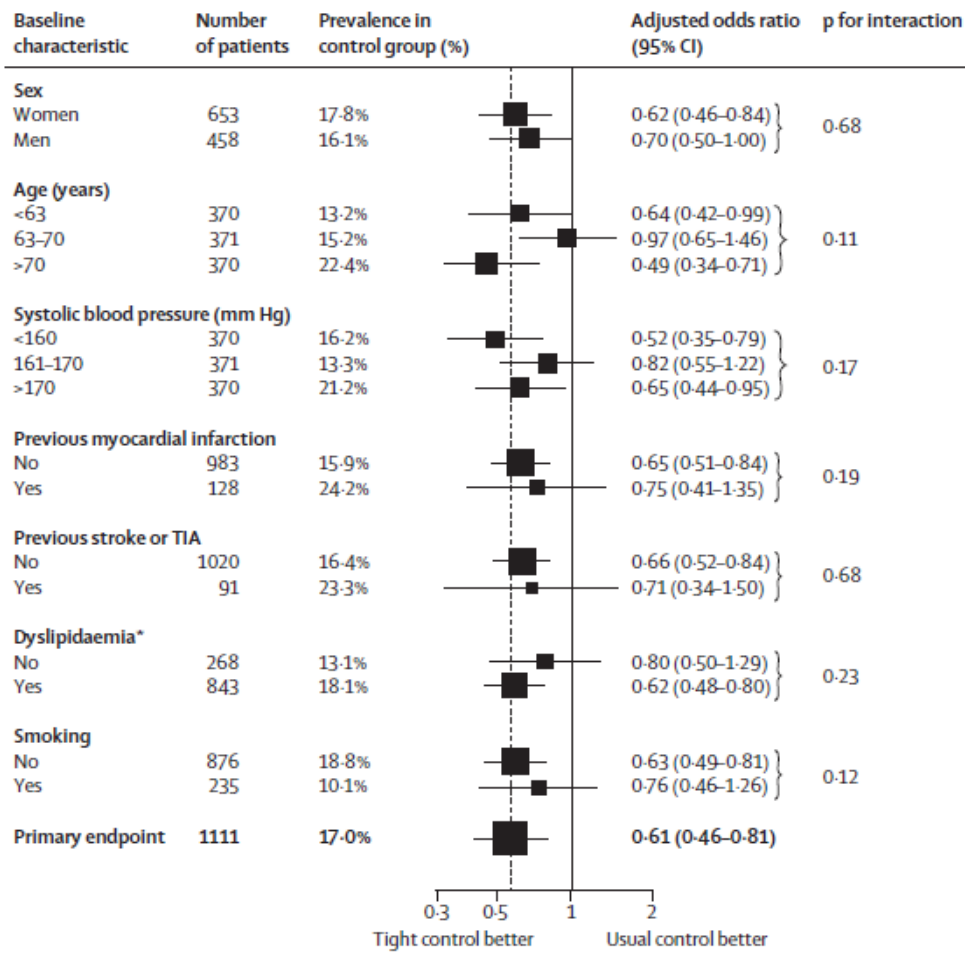


Figure 7 – Exemple d'analyses en sous-groupe présentées avec le test d'interaction.

La multiplicité des analyses en sous-groupes expose à une inflation du risque alpha au niveau des tests d'interaction [8], mais comme ces tests ne contribuent pas directement à la décision d'utiliser ou non le traitement, cette inflation n'a pas de conséquences sérieuses. Cependant la multiplicité expose au risque de trouver une interaction uniquement du fait du hasard parmi tous les tests réalisés au niveau de l'essai. La Figure 8 montre un exemple d'une telle interaction avec une variable qui certainement ne modifie en rien l'efficacité de la chirurgie d'endartériectomie dans la prévention de la récidive de l'AVC [6].

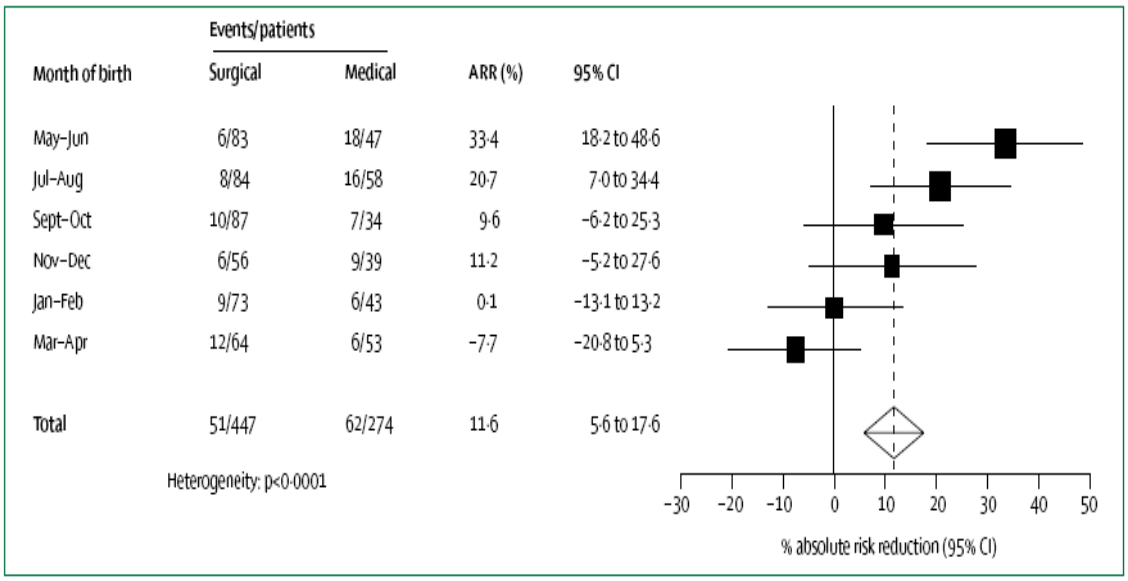


Figure 3: Effect of carotid endarterectomy in patients with ≥70% symptomatic stenosis in ECST[126] according to month of birth in six 2 month periods

Figure 8 – Illustration du risque de trouver une interaction uniquement du fait du hasard [6].

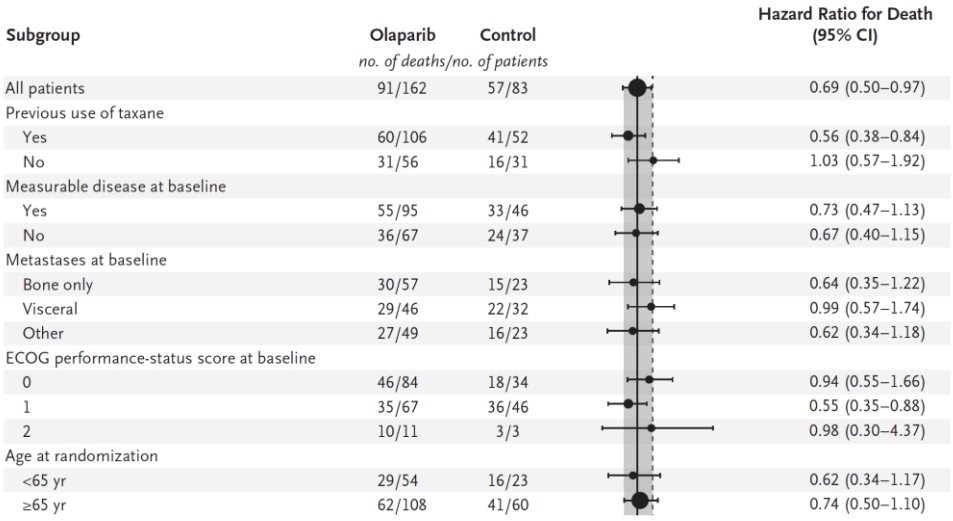
4 Vérification de la généralisabilité du résultat

Les analyses en sous-groupes ont pour seul but de s'assurer de la généralisabilité du résultat global de l'essai à l'ensemble des sous-types de patients inclus dans l'essai, c'est-à-dire de s'assurer que, l'hypothèse qui a été faite lors de l'élaboration des critères d'éligibilité du protocole que tous les patients inclus bénéficieraient de la même façon du traitement, n'est pas remise en cause par les résultats des analyses en sous-groupes.

Les analyses en sous-groupes standards[4] ne servent qu'à juger de la généralisabilité du résultat de l'étude aux différents types de patients qui ont été inclus dans l'essai.

Lors de la conception de l'étude, les critères d'éligibilité sont définis afin de recruter des patients que l'on pense tous en mesure de tirer le même bénéfice du traitement. L'analyse de la généralisabilité du résultat est donc une vérification, a posteriori, qu'il était justifié de regrouper ces différents types de patients au sein d'une même étude.

Cette analyse s'effectue graphiquement. Les analyses en sous-groupes sont le plus souvent représentées à l'aide d'un forest plot où le résultat de l'étude est projeté sur les sous-groupes par un trait vertical.



Subgroup	Olaparib	Control	Hazard Ratio for Death (95% CI)
	no. of deaths/no. of patients		
All patients	91/162	57/83	0.69 (0.50–0.97)
Previous use of taxane			
Yes	60/106	41/52	0.56 (0.38–0.84)
No	31/56	16/31	1.03 (0.57–1.92)
Measurable disease at baseline			
Yes	55/95	33/46	0.73 (0.47–1.13)
No	36/67	24/37	0.67 (0.40–1.15)
Metastases at baseline			
Bone only	30/57	15/23	0.64 (0.35–1.22)
Visceral	29/46	22/32	0.99 (0.57–1.74)
Other	27/49	16/23	0.62 (0.34–1.18)
ECOG performance-status score at baseline			
0	46/84	18/34	0.94 (0.55–1.66)
1	35/67	36/46	0.55 (0.35–0.88)
2	10/11	3/3	0.98 (0.30–4.37)
Age at randomization			
<65 yr	29/54	16/23	0.62 (0.34–1.17)
≥65 yr	62/108	41/60	0.74 (0.50–1.10)

Figure 9 – Utilisation des analyses en sous-groupes pour s'assurer de la généralisabilité du résultat de l'essai à tous les sous-types de patients inclus dans l'essai.
La bande grise verticale représente la projection de l'intervalle de confiance du résultat de l'essai (All patients) sur les résultats des analyses en sous-groupes. Les intervalles de confiances des sous-groupes ont tous des parties communes avec l'intervalle de confiance du résultat de l'essai. Il n'existe pas de cas où le résultat obtenu par un sous-groupe serait discordant avec le résultat global compte tenu de l'incertitude des estimations. Le résultat de l'essai est donc généralisable à l'ensemble des catégories de patients inclus. [10.1056/NEJMoa2022485]

[4] c'est-à-dire non inclus dans un plan de contrôle du risque alpha global

En cas de constatation de modifications de l'effet dans certains sous-groupes, les difficultés d'interprétation commencent. Le premier élément de discussion va être une interaction due au hasard. Compte tenu de la multiplicité des sous-groupes la probabilité d'avoir au moins une interaction sur le nombre du fait du hasard est importante.

Ensuite naturellement arrive la question de restreindre la population cible du traitement (si le résultat global de l'essai est en faveur du bénéfice). Comme nous l'avons vu, la présence d'une interaction ne signifie pas l'absence d'effet dans un sous-groupe. Elle signifie simplement que l'effet du traitement change entre les sous-groupes. Ainsi la difficulté de conclure à l'absence d'intérêt d'un traitement dans un sous-groupe reste identique.

Utilisation **erronée** des sous-groupes pour conclure vis-à-vis de l'effet du traitement	Analyse de la signification statistique (nominale) pour chaque sous-groupe, par exemple chez les hommes, chez les femmes	Correspond au même objectif que l'essai, entraine donc une inflation du risque alpha global de conclure à tort à un quelconque intérêt du traitement
Utilisation **appropriée** des sous-groupes pour rechercher si un facteur modifie la taille de l'effet du traitement (interaction)	Comparer la taille de l'effet entre les sous-groupes (par exemple entre les hommes et les femmes pour explorer si le sexe est un facteur modifiant l'effet du traitement)	Ne cherche pas à conclure à l'intérêt du traitement Résultat purement exploratoire, cognitif. Ne permettant pas de faire des conclusions sur l'intérêt du traitement, cette analyse n'entraine pas d'inflation du risque alpha global de l'essai

Il faut cependant noter, qu'en pratique, hors du champ d'une interprétation rigoureuse, les analyses en sous-groupe sont largement surinterprétées, exposant à des prises de décisions potentiellement erronées. La littérature regorge d'exemples de résultats de sous-groupes qui ont été invalidés par les études de vérification ultérieures.

La nouvelle politique concernant les p-values du NEJM se traduit par la disparition des p-values des tests d'existence de l'effet du traitement au niveau des résultats de sous-groupes. La Figure 10 montre un exemple de l'application de cette nouvelle politique éditoriale.

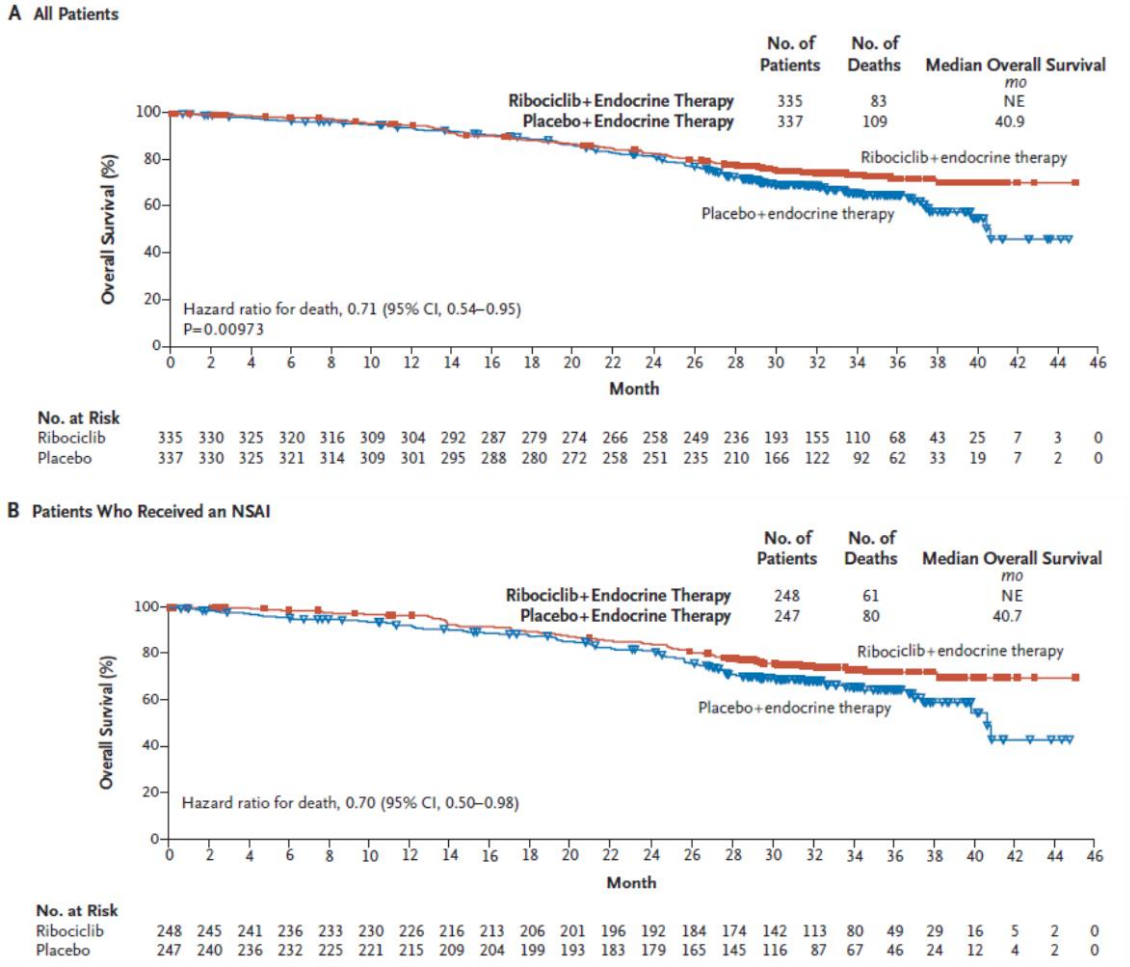


Figure 10 - Conséquence de la nouvelle politique concernant les p values du NEJM sur la présentation des résultats de sous -groupe.

La figure du haut correspond à la population totale de l'étude. La p-value est donnée. La figure du bas est un sous-groupe, aucune p-value n'est rapportée, car ce résultat ne peut pas être considéré comme inférentiel, permettant une conclusion sur ce sous-groupe.

5 Méta-épidémiologie

Les risques de fausses découvertes dans les analyses en sous-groupe sont parfaitement bien illustrés par toute une série d'exemple où des résultats initiaux de sous-groupes n'ont pas été confirmés ultérieurement lorsqu'un essai spécifique a été mis en place. Rothwell et al. donnent une liste de tels exemples.

Observation	Refutation
Aspirin is ineffective in secondary prevention of stroke in women[29,30]	31
Antihypertensive treatment for primary prevention is ineffective in women[32,33]	34
Antihypertensive treatment is ineffective or harmful in elderly people[35]	36
Angiotensin-converting enzyme inhibitors do not reduce mortality and hospital admission in patients with heart failure who are also taking aspirin[37]	38
β blockers are ineffective after acute myocardial infarction in elderly people,[39] and in patients with inferior myocardial infarction[41]	40
Thrombolysis is ineffective >6 hours after acute myocardial infarction[42]	43
Thrombolysis for acute myocardial infarction is ineffective or harmful in patients with a previous myocardial infarction[42]	44
Tamoxifen citrate is ineffective in women with breast cancer aged <50 years[45]	46
Benefit from carotid endarterectomy for symptomatic stenosis is reduced in patients taking only low-dose aspirin due to an increased operative risk[47]	48
Amlodipine reduces mortality in patients with chronic heart failure due to non-ischaemic cardiomyopathy but not in patients with ischaemic cardiomyopathy[49]	50

Table 1: **Examples of subgroup analyses that have shown apparently clinically important heterogeneity of treatment effect which has subsequently been shown to be false**

Figure 11 – D'après Rothwell PM, Lancet 2005[6]

Plusieurs études méta-épidemiologiques ont décrit les pratiques récentes concernant les sous-groupes dans les essais thérapeutiques et montrent une utilisation et une interprétation excessive de leurs résultats [9, 10, 11, 12, 13].

6 La gestion du risque alpha global

Les limitations statistiques des analyses en sous-groupes ordinaires (inflation des risques alpha et beta) peuvent être contournées en intégrant les sous-groupes d'intérêt dans le plan de contrôle strict du risque alpha global, soit par hiérarchisation, soit par répartition avec un calcul d'effectif pour chaque sous-groupe. Dans ce cas les démonstrations peuvent être obtenues au niveau de sous-groupe et les réserves habituellement faites sur les résultats de sous-groupes disparaissent.

6.1 Exemple avec hiérarchisation

L'essai Keynote-010 [14] a évalué 2 doses de pembrolizumab dans le cancer du poumon avancé en première ligne versus docetaxel. Le modèle pharmacologique conduit à supposer que le pembrolizumab n'apporte un bénéfice que lorsque les cellules tumorales expriment le ligand L1. De ce fait la logique voudrait que l'essai n'inclue que des patients exprimant le ligand, les autres étant non-répondeurs. Cependant les difficultés d'apprécier le niveau d'expression de ce ligand et la possible existence d'autres mécanismes d'action font qu'il est aussi concevable que le pembrolizumab apporte un bénéfice chez tous les patients.

Avec une approche classique, l'essai aurait inclus tous les patients et ceux qui exprimeraient le PDL1 à plus de 50% auraient été un sous-groupe. En cas d'échec de l'essai à montrer un bénéfice (probable si l'hypothèse pharmacologique initiale est correcte), il n'aurait pas été possible de conclure pour les patients exprimant le PDL1>50%.

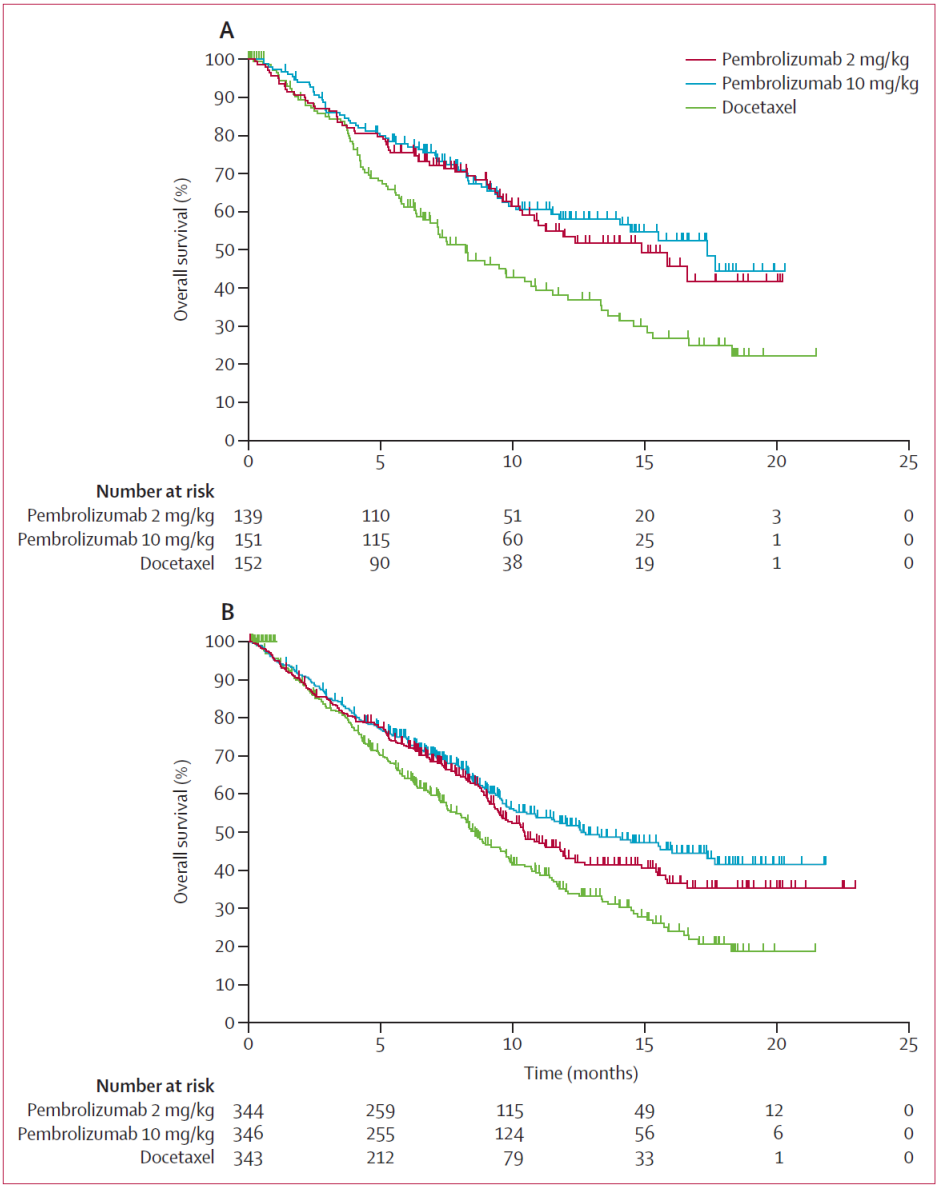


Figure 2: **Kaplan-Meier analysis of overall survival**
(A) For patients with a PD-L1 tumour proportion score of 50% or greater. (B) For all patients.

La solution retenue a été de hiérarchiser les populations de patients en commençant par la recherche du bénéfice du pembrolizumab chez les patients PD-L1>50% puis chez tous les patients. Ainsi, si l'hypothèse pharmacologique est exacte seul le premier test sera concluant et pas le 2ème, mais le bénéfice sera formellement démontré dans la sous-population concernée. Si les 2 tests sont concluants, une démonstration sera apportée pour tous les patients, quel que soit leur niveau d'expression du PD-L1. En revanche ce design ne permet pas de démontrer spécifiquement le bénéfice chez les moins de 49%.

Les résultats permettent de conclure pour les 2 populations de patients :

> *In the total population, … overall survival was significantly longer for pembrolizumab 2 mg/kg versus docetaxel (hazard ratio [HR] 0·71, 95% CI 0·58–0·88; p=0·0008) and for pembrolizumab 10 mg/kg versus docetaxel (0·61, 0·49–0·75; p<0·0001). … Among patients with at least 50% of tumour cells expressing PD-L1, overall survival was significantly longer with pembrolizumab 2 mg/kg than*

with docetaxel (… HR 0·54, 95% CI 0·38–0·77; p=0·0002) and with pembrolizumab 10 mg/kg than with docetaxel (… 0·50, 0·36–0·70; p<0·0001).

En revanche ce design ne permet pas de démontrer spécifiquement le bénéfice chez les moins de 49%. Ce point pourrait être gênant si le traitement n'apportait aucun bénéfice chez ces patients. Malgré cette absence de bénéfice, le résultat sur la population entière pourrait être toujours en faveur du traitement évalué si le bénéfice dans l'autre sous-groupe est important et/ou si ce sous-groupe représente la majeure partie des patients. Ainsi on traiterait des patients, car ils sont inclus dans la population globale, mais sans qu'ils en bénéficient. Ce point peut être exploré par le résultat du sous-groupe complémentaire, même si celui n'est pas décisionnel. Ce point représente la limite de la gestion des sous-groupes par hiérarchisation et fait que l 'approche par répartition est certainement plus appropriée (mais rarement mise en œuvre).

6.2 Exemple avec répartition du risque alpha

L'essai SATURN a évalué l'erlotinib comme traitement de maintenance dans le cancer du poumon avancé non à petite cellule [15]. Les patients EGFR positifs représentent un sous-groupe d'intérêt. Afin de donner un statut décisionnel à ce sous-groupe tout en permettant aussi de pouvoir conclure sur la population totale (quel que soit le statut EGFR) une répartition du risque alpha global a été effectuée avec 3% de risque alpha pour la totalité des patients et 2% pour le sous-groupe EGFR+ :

The co-primary endpoints were PFS in all analysable patients irrespective of EGFR status, and PFS in patients with EGFR immunohistochemistry-positive tumours. … The alpha level of 5% was split between the two co-primary endpoints: 3% for all patients and 2% for patients with EGFR immunohistochemistry-positive tumours.

Le terme "co-primary endpoint" est utilisé ici non pas pour désigner 2 critères de jugements principaux différents, mais plutôt dans une acception de « co-objectif ». Un résultat statistiquement significatif a été obtenu aussi bien pour le sous-groupe EGFR+ (HR 0·69, 0·58–0·82; p<0·0001) que pour tous les patients (HR 0·71, 95% CI 0·62–0·82; p<0·0001). Comme dans l'exemple précédent de la Keynote-010, cette approche statistique ne permet pas de démontrer le bénéfice dans le sous-groupe complémentaire.

7 Analyse en sous-groupe et prise de décision

Les problématiques statistiques des sous-groupes et le risque inhérent de fausses découvertes font qu'en théorie les résultats des sous-groupes[5] ne permettent pas de modifier la pratique.

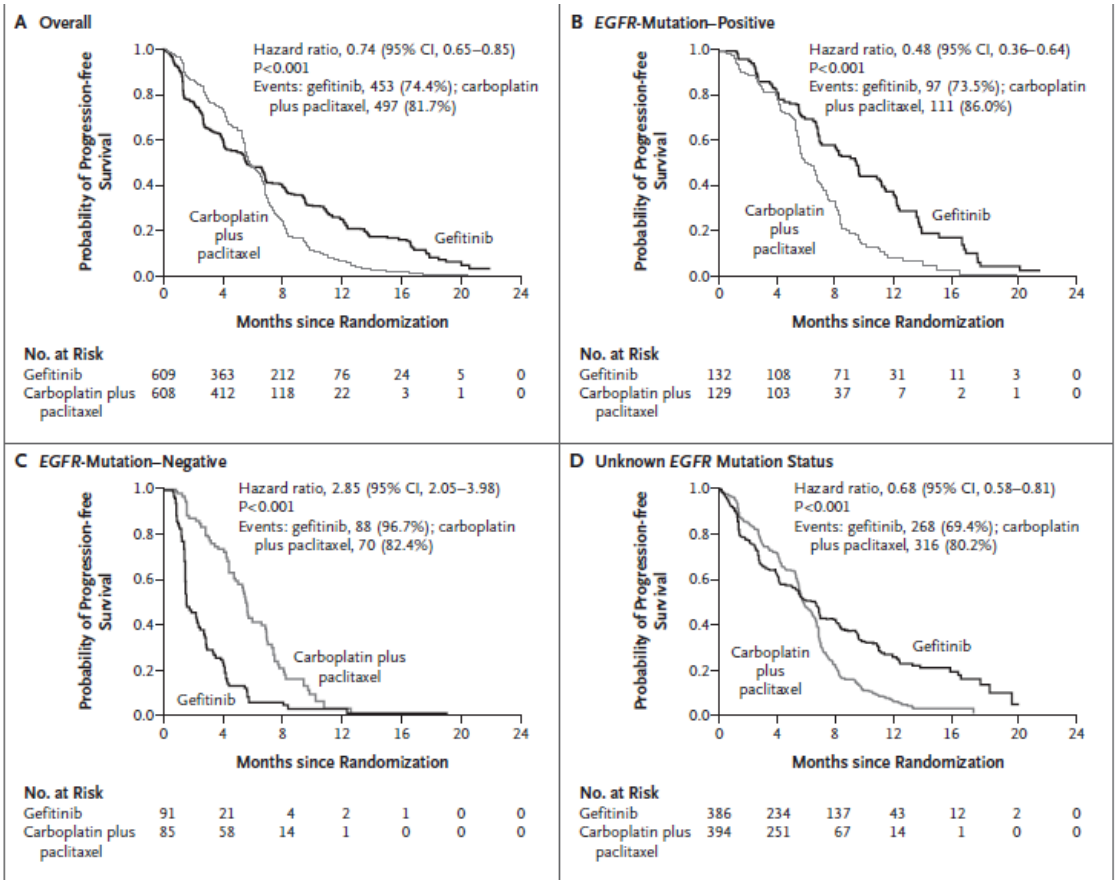
Ce principe est en pratique assez bien respecté lorsque l'essai est négatif. Il n'y a pas d'exemple récent où un traitement a été adopté uniquement sur la base d'un résultat de sous-groupe alors que l'essai était négatif. Les tentatives sont nombreuses (quasiment systématique avec tous les essais négatifs [16]), mais ne débouchant sur aucune prise de décision en faveur du traitement.

Cependant, la situation est plus complexe quand il s'agit d'un sous-groupe suggérant l'absence de bénéfice dans un essai concluant par lui-même. En effet, dans le cas d'un essai concluant, la décision d'exclure de l'indication d'un traitement certains patients, à la suite d'un résultat de sous-groupe non en faveur du bénéfice, est lourde de conséquences. Compte tenu des limites des sous-groupes cette exclusion peut être faite à tort et conduit à une perte de change pour les patients. En fait il existe plusieurs cas de figure.

> Un premier essai du gefitinib dans le cancer du poumon [17] a obtenu un résultat statistiquement significatif, mais avec des courbes de survie qui se croisent. Les croisements dans les courbes de survie sont entre autres évocateurs d'un mélange de population répondant de manière opposée au traitement (interaction forte), une recherche d'explication a été entreprise par des analyses en sous-groupe. Dans l'analyse en fonction de la mutation du récepteur EGFR[6], un résultat en défaveur du gefitinib a été trouvé chez les patients sans mutation de l'EGFR (HR 2.85), tandis que le bénéfice du produit était retrouvé chez les patients mutés (HR 0.48).

[5] ordinaires

[6] La valeur prédicatrice de l'efficacité de la mutation du récepteur n'était pas connue à l'époque. C'est cette étude qui l'a suggérée.



Comme cette analyse était purement exploratoire et qu'un risque de fausse découverte n'était pas exclu, il était difficile de recommander le gefinitib chez les patients mutés sans plus de preuves. Pour confirmer cette nouvelle hypothèse (le gefitinib n'est efficace que chez les EGFR mutés) un autre essai a été entrepris et a obtenu un résultat positif [18].

La démarche a été vertueuse et en accord avec les principes habituels d'utilisation des résultats de sous-groupe : l'analyse en sous-groupe a bien était considérée comme exploratoire. Elle a généré une nouvelle hypothèse qui a été confirmée dans de nouvelles études avant la prise de décision.

Le premier cas de figure est le plus simple. Un hazard ratio proche de 1, non nominalement significatif est observée dans un sous-groupe, mais aucune interaction significative n'est détectée. Il ne peut donc pas être considéré qu'il y a modification de l'effet traitement. Il n'y a pas lieu de considérer que l'effet du traitement dans ce sous-groupe est différent de l'effet global. Comme cet effet est démontré il n'y a pas lieu de suivre le résultat du sous-groupe et d'exclure ces patients de l'indication.

En revanche si une forte interaction est détectée il n'est plus possible de récuser le résultat du sous-groupe aussi clairement. Il n'est pas possible n'ont plus d'affirmer que le traitement ne marche pas dans ce sous-groupe, car les problématiques statistiques liées à la multiplicité : inflation du risque beta, diminution de la puissance et risque de fausse découverte persistent.

Le fait de trouver, a posteriori, une explication biologique ou pharmacologique à ce résultat n'est pas non plus une preuve en soit, car cette explication est élaborée de manière post hoc. Or compte tenu de la complexité des mécanismes sous-jacents il est presque toujours possible d'expliquer tout et son contraire a posteriori (cf. dossier 4).

Ce point illustre l'importance que les sous-groupes reposent sur une hypothèse formulée a priori de modification de l'effet du traitement, formulée à partir des mécanismes ou d'autres connaissances

(résultats exploratoires d'études précédentes) disponibles a priori. Cette hypothèse permettant aussi de prévoir le sens de l'interaction [2, 19]. Dans cette situation, les résultats des sous-groupes s'inscriront dans un cadre hypothético-déductif et auront une valeur certaine, même s'ils ne sont pas inclus dans le plan de contrôle du risque alpha global. Cependant si un tel raisonnement a priori débouche sur la possibilité que le traitement ne puisse pas apporter de bénéfice à certains patients, il est probable que ces patients ne seront pas inclus dans l'essai. Ce point montre bien le côté complètement inattendu des résultats des sous-groupes et explique aussi pourquoi, il faut considérer, a priori, les absences de bénéfice dans un sous-groupe comme une découverte fortuite artéfactuelle. Les prendre en considération reposera sur un raisonnement purement inductif.

La prise de décision de réaliser un nouvel essai pour démontrer l'hypothèse générée par ce résultat de sous-groupe n'est pas simple non plus : il s'agirait de faire un nouvel essai pour démontrer l'absence de bénéfice !

De nombreuses techniques avancées [20] ont aussi été proposées dans le cadre de la médecine personnalisée en remplacement des analyses en sous-groupes traditionnelles pour rechercher les facteurs d'hétérogénéité des effets traitements . Ces approches ne sont pas encore utilisées dans les essais de phase 3. Elles sont donc, pour cette raison, en dehors du champ de ce document.

8 Points divers

8.1 Confusion

Nonobstant les problématiques statistiques liées au risque de fausses découvertes, les analyses en sous-groupes sont aussi limitées dans leur conclusion en termes de risque de biais de confusion. Bien que réalisées dans le cadre d'un essai randomisé, les analyses en sous-groupes sont de nature purement observationnelle. Ce sont des analyses univariées sans aucun ajustement sur de potentiels facteurs de confusion. Par exemple, une modification d'effet observée avec la présence en comorbidité d'une altération cognitive peut simplement être le reflet d'une mauvaise observance. Le réel facteur modifiant l'effet du traitement peut ainsi être un tout facteur que celui utilisé pour faire l'analyse en sous-groupe.

8.2 Sous-groupes stratifiés

Il est aussi souvent objecté à des analyses en sous-groupes que celles-ci de sont pas stratifiées (c'est-à-dire réalisées avec un facteur de stratification de la randomisation). Bien qu'encore débattu, cet aspect ne change en rien les limites des analyses en sous-groupes. La stratification de la randomisation sur la variable servant à faire les sous-groupes ne lève pas les limitations des résultats obtenus [21]. Elle apporte seulement un gain de précision et limite un problème très technique, la non collapsibilité de certains estimateurs de l'effet traitement qui ne sera pas abordé dans ce document.

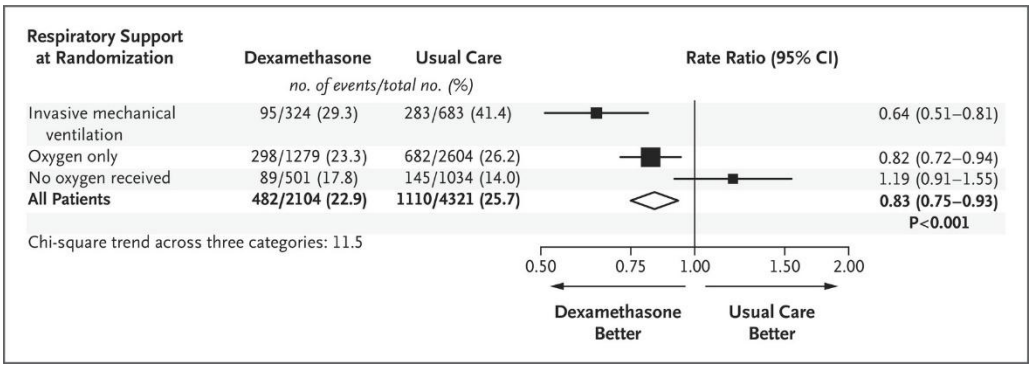
8.3 Cas particulier où le sous-groupe suggère un effet délétère

Le seul cas où la décision est relativement simple est celui où le sous-groupe suggère un effet délétère. En effet la problématique n'est plus une question d'efficacité, mais de sécurité où les décisions se prennent suivant une autre logique. L'exclusion de ces patients de l'indication est alors justifiable par le principe de précaution qui régit les décisions de sécurité.

En effet dans ces décisions de sécurité, la crainte n'est plus de se faire abuser par un résultat faussement positif (faisant croire à tort à l'efficacité avec un produit sans intérêt), mais par un résultat faussement négatif (faisant croire à tort à l'absence d'événement indésirable). De ce fait tout argument suggérant un problème, même s'il n'est pas parfaitement démontré, débouche sur la décision de non-utilisation, retrait du médicament.

Ce cas de figure est d'ailleurs prévu par un guideline européen de l'EMA [1].

> L'essai RECOVERY[22] a évalué la dexaméthasone dans la COVID-19 chez des patients hospitalisés. Un bénéfice sur la mortalité totale est démontré par l'essai. Mais dans le sous-groupe des patients ne nécessitant pas d'oxygène à l'entrée une tendance, non significative, à une surmortalité est observée.



Bien que ce résultat puisse simplement être une fluctuation aléatoire, la plupart des recommandations pour la pratique ne recommandent pas l'usage des corticoïdes chez ces patients ne nécessitant pas d'oxygénothérapie. L'idéal serait la réalisation de nouveaux essais thérapeutiques spécifiquement chez ces patients pour statuer définitivement. En effet, il s'agit, à la date de la rédaction de ce document, du seul produit dans la COVID-19 ayant démontré une réduction de mortalité chez tous les patients hospitalisés. Comme il persiste un doute que ce résultat ne soit qu'une fausse découverte due au hasard, l'équipoise reste entière et autorise la randomisation.

8.4 Le paradoxe de Stein

La taille de l'effet dans une sous-population peut être un paramètre d'intérêt dans certaines circonstances (comme pour paramétrer un modèle médico-économique). Il est alors tentant de prendre le résultat obtenu dans le sous-groupe correspondant. Cette estimation n'est cependant pas forcément la meilleure estimation à utiliser. Dans certaines circonstances la meilleure estimation pour une sous population peut être l'estimation globale de l'essai et non pas le résultat du sous-groupe correspondant. Ce phénomène est connu sous le nom de paradoxe de Stein [23]. Tout dépend du degré d'interaction dans l'analyse en sous-groupe d'intérêt (Figure 12).

Si aucune interaction n'existe, la meilleure estimation est l'estimation globale. En effet il n'y a pas de différence d'effet entre les sous-groupes, et l'estimation d'un sous-groupe particulier peut être éloignée de la vraie valeur du fait d'une fluctuation aléatoire. L'estimation globale, en portant sur plus de patients, est sujette à des fluctuations aléatoires moindres.

En revanche s'il existe une forte interaction, l'effet est différent entre les sous-groupes et l'estimation du sous-groupe d'intérêt est probablement plus proche de la vraie valeur recherchée que l'estimation globale, car, dans ce cas, la différence entre les vrais effets est plus grande que les fluctuations aléatoires affectant les résultats du sous-groupe (sinon le test d'interaction ne serait pas significatif).

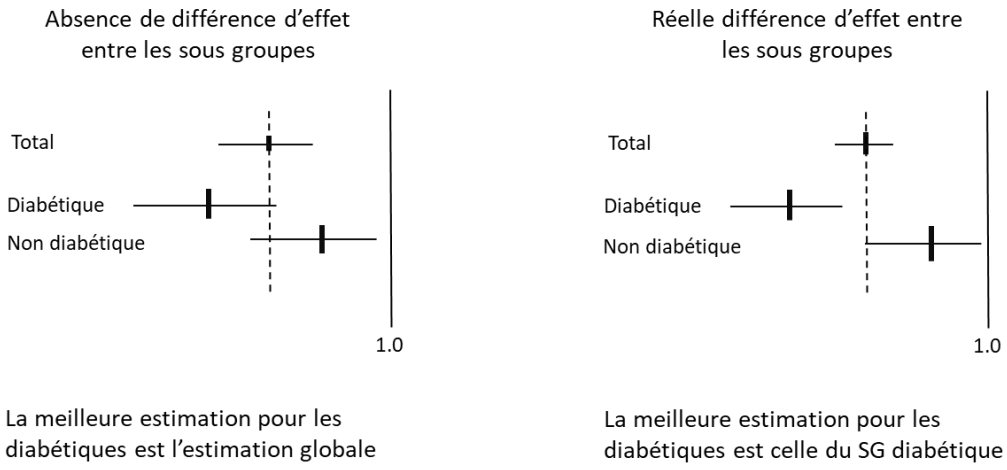


Figure 12 – Illustration du principe de Stein

En dehors de ces cas extrêmes, la meilleure estimation pour l'effet dans un sous-groupe peut être déterminée en appliquant un coefficient de « shrinkage ». Cette estimation s'apparente à une interpolation entre l'estimation globale et l'estimation obtenue dans le sous-groupe, au prorata du degré d'interaction (Figure 13). Son intérêt est de relativiser la taille de l'effet qui semble plus important dans un sous-groupe qu'avec la totalité des patients de l'essai et de tempérer ainsi le sur-enthousiasme lié à ce résultat.

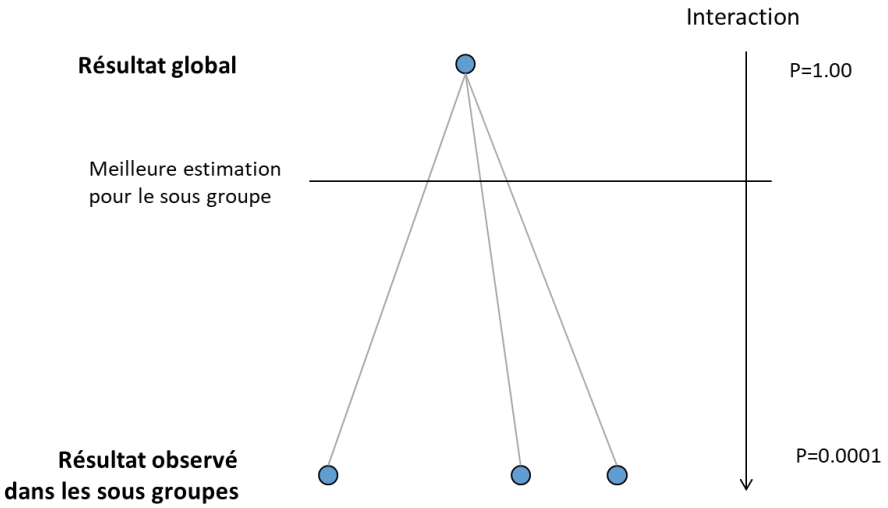


Figure 13 – Illustration du « shrinkage » des estimations en fonction du degré d'interaction

9 Conclusion

Au total, leurs limites statistiques font qu'il est extrêmement risqué de prendre des décisions concernant l'efficacité ou la non-efficacité d'un traitement à partir d'analyses en sous-groupes [9, 24].

Le guideline ICH E9 précise :

> *"In most cases…subgroup or interaction analyses are exploratory and should be clearly identified as such;…these analyses should be interpreted cautiously;…any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses are unlikely to be accepted." [5.7 Subgroups, Interactions and Covariates]*

Que peut-on faire avec des résultats de sous-groupes	• Générer une nouvelle hypothèse à tester prospectivement dans un nouvel essai
	• Contrindiquer un traitement chez certains patients en cas d'effet délétère par principe de précaution
	
	• Prendre une décision si l'hypothèse concernant le SG a été faite de manière EXPLICITE :
	  ○ A priori, avec fixation du sens de l'interaction
	  ○ Avec un contrôle du risque alpha (méthode hiérarchique)
	  ○ Un calcul du NSN

Ce que l'on ne peut vraiment pas faire	Récupérer un essai négatif à l'aide d'un résultat de sous-groupe
Ce que l'on fait avec une prise de risque non contrôlée	Mais que l'on ne devrait pas faire, car la prise de risque est souvent considérable • Conclure que le traitement n'est pas ou insuffisamment efficace pour certains sous-groupes de patients • Conclure que le traitement est bien plus efficace pour certains sous-groupes de patients • Conclure à une généralisabilité du résultat à tous les patients inclus (en considérant que l'absence d'interaction significative montre l'absence de différence)

Only one thing is worse than doing subgroup analyses---believing the results
Richard Peto [25]

Références

1 European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials ;

2 Assmann SF, Pocock SJ, Enos LE, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064–69 ;

3 Fan J, Song F, Bachmann MO. Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials. *J Clin Epidemiol* 2019;108:17–25 doi:10.1016/j.jclinepi.2018.12.009; PMID:30557676;

4 Bonaca MP, Bhatt DL, Cohen M, et al. Long-term use of ticagrelor in patients with prior myocardial infarction. *N Engl J Med* 2015;372:1791–800 doi:10.1056/NEJMoa1500857; PMID:25773268;

5 Horton R. From star signs to trial guidelines. *The Lancet* 2000;355:1033–34 doi:10.1016/S0140-6736(00)02031-6; PMID:10744086;

6 Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* 2005;365:176–86 doi:10.1016/S0140-6736(05)17709-5;

7 Robert C, Karaszewska B, Schachter J, et al. Improved overall survival in melanoma with combined dabrafenib and trametinib. *N Engl J Med* 2015;372:30–39 doi:10.1056/NEJMoa1412690; PMID:25399551;

8 Brookes ST, Whitely E, Egger M, et al. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004;57:229–36 ;

9 Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;5:1–56 doi:10.3310/hta5330; PMID:11701102;

10 Hernández AV, Boersma E, Murray GD, et al. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J* 2006;151:257–64 doi:10.1016/j.ahj.2005.04.020; PMID:16442886;

11 Kasenda B, Schandelmaier S, Sun X, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. *BMJ-BRITISH MEDICAL JOURNAL* 2014;349:g4539 doi:10.1136/bmj.g4539; PMID:25030633;

12 Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. *Am Heart J* 2000;139:952–61 doi:10.1067/mhj.2000.106610; PMID:10827374;

13 Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 2012;344:e1553 doi:10.1136/bmj.e1553; PMID:22422832;

14 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;

15 Cappuzzo F, Ciuleanu T, Stelmakh L, et al. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study. *The Lancet Oncology* 2010;11:521–29 doi:10.1016/S1470-2045(10)70112-1;

16 Naggara O, Raymond J, Guilbert F, et al. The problem of subgroup analyses: an example from a trial on ruptured intracranial aneurysms. *AJNR Am J Neuroradiol* 2011;32:633–36 doi:10.3174/ajnr.A2442; PMID:21436333;

17 Mok TS, Wu Y-L, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009;361:947–57 doi:10.1056/NEJMoa0810699; PMID:19692680;

18 Maemondo M, Inoue A, Kobayashi K, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* 2010;362:2380–88 doi:10.1056/NEJMoa0909530; PMID:20573926;

19 Tanniou J, van der Tweel I, Teerenstra S, et al. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Med Res Methodol* 2016;16:20 doi:10.1186/s12874-016-0122-6; PMID:26891992;

20 Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Ann. Intern. Med.* 2020;172:35–45 doi:10.7326/M18-3667; PMID:31711134;

21 Kaiser LD. Stratification of randomization is not required for a pre-specified subgroup analysis. *Pharm Stat* 2013;12:43–47 doi:10.1002/pst.1550; PMID:23281052;

22 Horby P, Lim WS, Emberson JR, et al. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. *N Engl J Med* 2020 doi:10.1056/NEJMoa2021436; PMID:32678530;

23 Lipsky AM, Gausche-Hill M, Vienna M, et al. The importance of "shrinkage" in subgroup analyses. *Annals of emergency medicine* 2010;55:544-552.e3 doi:10.1016/j.annemergmed.2010.01.002; PMID:20138396;

24 Aronson D. Subgroup analyses with special reference to the effect of antiplatelet agents in acute coronary syndromes. *Thromb Haemost* 2014;112:16–25 doi:10.1160/TH13-09-0801; PMID:24599493;

25 Sleight P. Subgroup analyses in clinical trials - fun to look at, but don't believe them! *Curr Control Trials Cardiovasc Med* 2000;1:25–27 ;