Attention aux analyses de sous-groupes!

Peter Kleist
GlaxoSmithKline AG, Münchenbuchsee

Quintessence

- Même si leur signification ne doit pas être surestimée, les analyses de sous-groupes effectuées sur les données des essais cliniques peuvent parfois servir de référence pour les décisions thérapeutiques du médecin.
- De petits échantillons pris au hasard, un déséquilibre entre les sous-groupes et un cumul de tests statistiques augmentent la probabilité d'un résultat relevant du hasard.
- L'évaluation de la valeur d'une analyse de sous-groupes se fonde sur un certain nombre de critères minimaux qui doivent être remplis: la spécification d'un faible nombre d'analyses de sous-groupes dans le protocole d'étude, la réalisation de tests d'interactions et une correction statistique tenant compte de l'existence de tests multiples.
- Le principal objectif des analyses de sous-groupes dans les essais cliniques est de recueillir des informations sur la consistance d'un effet thérapeutique. Les différences d'efficacité ou de sécurité d'une intervention dans des sous-groupes déterminés n'ont la plupart du temps qu'une valeur indicative, doivent être interprétées avec prudence et nécessitent confirmation par d'autres études.
- Le fait de triturer les données après coup («data dredging») dans le but de trouver des réponses significatives et même parfois de «sauver» une «étude négative» relève d'une mauvaise pratique scientifique.
- La plupart des publications sur des analyses de sous-groupes ne remplissent cependant pas ces standards, mêmes dans les journaux médicaux de haut niveau.

Summary

Subgroup analyses in clinical studies: handle with care

- *While analyses of treatment effects in subgroups of clinical studies may be of importance for therapeutic decision-making in individual patients, the importance of such analyses should not be overestimated.*
- *Small patient numbers, imbalances between subgroups and the growing range of statistical tests increase the probability of chance findings.*
- *The validity of subgroup analyses depends on compliance with basic requirements: prior specification of a small number of subgroup analyses in the protocol, conduct of treatment interaction tests and statistical adjustment for multiple testing.*
- *The main objective of subgroup analyses is to obtain information on the uniformity of the overall treatment effect. Subgroup differences in the efficacy or safety of an intervention are mainly of an exploratory character, should be interpreted with caution and need confirmation by subsequent clinical studies.*
- *Data dredging and fishing for statistical significances, sometimes for the purpose of "rescuing a negative study", are not sound scientific practices.*
- *Most publications containing subgroup analyses, even in high-ranking medical journals, fail to comply with current standards.*

La controverse

«Tous ceux qui prennent ce remède guérissent en un rien de temps – sauf ceux chez qui il ne marche pas et qui meurent tous. Il est donc certain que ce remède n'échoue que dans les cas inguérissables.»

Galenos de Pergamon (Galen), IIe siècle après J.-C.

«Subgroup analyses in clinical trials – fun to look at, but don't believe them.»

Professeur Peter Sleight, Oxford, XXe siècle après J.-C.

Les médecins et les statisticiens ont peut-être appris à vivre ensemble. Il n'en demeure pas moins que leurs relations ne sont pas toujours amicales. La controverse au sujet du (non)sens des analyses de sous-groupes dans les essais cliniques en est un bon exemple. Ce n'est pas pour rien que l'on parle dans le monde anglosaxon de «clinicostatistical tragedy» [1].

D'un côté, les statisticiens lancent une mise en garde de principe face aux pièges que constituent les analyses de sous-groupes: leur usage disproportionné, les trouvailles basées exclusivement sur le hasard et/ou une surinterprétation des résultats peuvent donner lieu à des recommandations thérapeutiques erronées et priver certains patients d'un traitement efficace ou au contraire leur prescrire un médicament totalement inefficace. Un bon exemple est celui de la limitation, basée sur des analyses de sous-groupes, de la thrombolyse et du traitement bêtabloquant aux patients avec infarctus antérieur ou du tamoxifène dans le cancer du sein aux patientes de plus de 50 ans – des limitations qui se sont avérées par la suite sans fondement [2]. Guère étonnant que les détracteurs des analyses de sous-groupes se soient écriés: «subgroups may kill people» [3].

De l'autre côté, le médecin est confronté au problème du transfert des résultats des grandes études randomisées, dont les collectifs sont souvent très hétérogènes, à la pratique clinique de tous les jours et à ses différents patients. L'évaluation du rapport risque/bénéfice d'un traitement peut en effet différer en fonction du profil de risque initial ou du degré de sévérité de la maladie. On en veut pour preuves le problème de

Vous trouverez les questions à choix multiple concernant cet article à la page 791 ou sur internet sous www.smf-cme.ch.

l'anticoagulation dans les formes non compliquées et dans les formes à risque de fibrillation auriculaire ou la thrombendartérectomie carotidienne dans les sténoses de degrés variables. S'il faut considérer qu'un traitement peut être plus ou moins efficace ou même dangereux pour certains groupes de patients, il existe un devoir scientifique et éthique d'identifier ces sous-groupes [4]. Les données relatives à certains sous-groupes de patients constituent donc un point de référence pour la décision thérapeutique du médecin.

Ces deux points de vue sont-ils vraiment inconciliables? L'article qui suit part du principe que le recours à certains sous-groupes pour répondre à des questions spécifiques est essentiel dans les grands essais cliniques randomisés. Les besoins de la pratique clinique et les exigences méthodologiques doivent cependant être mis en relation pour donner aux données des analyses de sous-groupe tout leur sens et leur conférer une réelle valeur scientifique, donc un véritable intérêt pour la clinique. Après quelques exemples initiaux incitant à la prudence, nous préciserons les possibilités et les limites de la réalisation et de l'interprétation des analyses de sous-groupes.

Problèmes inhérents aux analyses de sous-groupes

«Le hasard travaille bien mieux, si on lui donne un petit coup de pouce.»

Henryk Bereska (1926–2005), poète et traducteur de littérature polonaise

Le seuil conventionnel de signification statistique situé à 5% indique que la probabilité qu'un résultat positif soit dû au hasard est de 5%. Cela semble peu à première vue. Mais attention: la probabilité est ici près de deux fois plus élevée que celle d'obtenir deux «six» lors d'un lancer des deux dés simultanément. Le fait de procéder – en plus de la vérification de l'hypothèse primaire – à plusieurs analyses de sous-groupes revient à lancer plusieurs fois les dés: plus le nombre de tests statistiques est élevé, plus la probabilité de parvenir à un résultat dû au hasard est grande. En d'autres termes, la probabilité d'*un* résultat fortuit (faux positif) atteint à ce moment déjà 23%. Pour dix analyses, elle sera de 40%, pour 20 de 64% et pour 50 de 92%! Il n'est donc pas étonnant que l'étude Betablocker-Heart-Attack-Study [5], au cours de laquelle pas moins de 146 sous-groupes ont été évalués, ait donné lieu à quelques résultats «positifs» aléatoires. On remarquera de plus que les échantillons pris au hasard dans les sous-groupes sont relativement petits et que la puissance du test (power) diminue donc rapidement dans ces analyses. La probabilité de découvrir une vraie différence entre les sous-groupes est par conséquent plutôt faible. Si un traitement est efficace et que le résultat global d'une étude clinique est significatif, il faut forcément s'attendre à trouver des sous-groupes dans lesquels le traitement semble inopérant (par ex. les bêtabloquants dans l'infarctus postérieur) – surtout lorsque ces sous-groupes sont petits et que des groupes de patients y sont sous-représentés, par exemple les femmes ou les sujets âgés. Des simulations statistiques ont montré que la probabilité de l'existence d'un sous-groupe avec des résultats non significatifs dans l'étude peut aller jusqu'à 66% et ceci pour une seule question testée dans le sous-groupe donné. Si le résultat global d'une étude est négatif, la probabilité de trouver un résultat positif dans un sous-groupe atteint tout de même encore 21% [6].

L'exemple suivant illustre le rôle joué par le hasard. L'étude ISIS-2 [7], qui a porté sur plus de 17 000 patients, a montré que l'acide acétylsalicylique (AAS) diminue de façon remarquable la mortalité post-infarctus du myocarde. Une analyse de sous-groupe basée sur le signe du zodiaque des patients a cependant laissé entendre que l'AAS ne serait pas efficace chez les sujets nés sous le signe de la Balance et sous celui des Gémeaux (tab. 1) [8]. Un autre exemple est celui de l'efficacité de la thrombendartérectomie carotidienne pour la prévention des accidents vasculaires cérébraux dans les sténoses symptomatiques de haut grade: chez les sujets nés en mai et en juin, l'intervention chirurgicale semble particulièrement efficace, tandis que chez ceux nés en mars et en avril, elle semblerait carrément dangereuse (fig. 1) [2, 9]. Ceci prête évidemment à sourire, car il ne viendrait à l'idée de personne (si ce n'est peut-être de quelques astrologues) que le bénéfice clinique des deux interventions mentionnées ci-dessus dépend du signe du zodiaque ou du mois de naissance des patients. Pourtant, si nous étions face à une relation de ce

Tableau 1. Influence de l'acide acétylsalicylique (AAS) sur la mortalité après infarctus du myocarde (analyse de sous-groupe de l'étude ISIS-2, basée sur le signe du zodiaque des participants [7, 8]).

Signe du zodiaque	Décès d'origine vasculaire après un mois		Niveau de signification
	ASS	Placebo	
Balance ou Gémeaux	150 (11,1%)	147 (10,2%)	p = 0,5
Tous les autres signes	654 (9,0%)	869 (12,1%)	p <0,0001
Tous les patients (tous les signes)	804 (9,4%)	1016 (11,8%)	p <0,0001

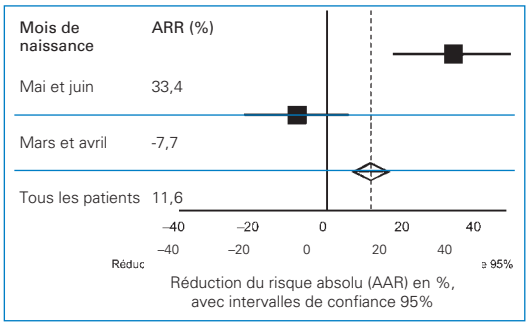

Figure 1
Effets de la thrombendartérectomie chez les patients avec sténose ≥70% symptomatique dans l'étude ECS [9]. Présentation des résultats sélectionnés d'un sous-groupe basé sur le mois de naissance des patients (d'après [2]).

type plausible, par exemple l'âge des patients ou leur tension artérielle, nous serions facilement tentés d'attribuer à ces résultats une véritable signification clinique.

Il n'est à vrai dire même pas nécessaire que des résultats soient plausibles médicalement pour que certains résultats d'analyses de sous-groupes soient pris pour argent comptant. Dans l'étude PRAISE-1 [10], l'antagoniste du calcium amlodipine a été comparé au placebo chez plus de 1100 patients souffrant d'insuffisance cardiaque. Aucune différence entre l'amlodipine et le placebo n'a été mise en évidence pour l'endpoint primaire (combinaison mortalité globale et morbidité cardiovasculaire). Comme les investigateurs étaient partis de l'idée que les patients ayant une insuffisance cardiaque secondaire à une ischémie profiteraient davantage du traitement, le protocole prévoyait une inclusion stratifiée des patients et une analyse du sous-groupe correspondant en fonction de la cause de l'insuffisance cardiaque (ischémique versus non ischémique). Contre toute attente, les résultats ont mis en évidence une diminution de 31% du risque sous amlodipine versus placebo dans le groupe de patients atteints d'une cardiomyopathie dilatée non ischémique, tandis que chez les patients atteints d'une cardiomyopathie ischémique, on ne voyait pas de différence. Pour «confirmer» cet effet de sous-groupe, il a été procédé à une étude complémentaire, l'étude PRAISE-2 [11], auprès des patients insuffisants cardiaques avec angiographie coronaire normale. Les résultats de cette seconde étude n'ont pas confirmé les observations faites durant la première – au contraire, ils ont même révélé une tendance en faveur d'une augmentation de la mortalité sous l'antagoniste du calcium (tab. 2 ←). Le résultat de l'analyse de sous-groupe de PRAISE-1 était donc simplement dû au hasard.

Si des analyses de sous-groupes sont effectuées après coup, c'est-à-dire après que les auteurs aient eu connaissance des résultats de l'étude (analyses dites post-hoc), la probabilité de résultats fortuits est encore plus grande. Nous avons déjà vu cela dans les analyses post-hoc sur l'influence des signes du zodiaque et du mois de naissance. L'exemple de la Canadian Cooperative Study, une étude pionnière des années 1970, montre le danger que peuvent représenter les analyses post-hoc. Cette étude avait porté sur les effets de l'acide acétylsalicylique chez les patients victimes d'attaques ischémiques transitoires [12]: comme l'effet de l'AAS n'avait atteint le seuil de signification statistique que chez les hommes (jusqu'à ce que d'autres travaux prouvent le contraire), la substance n'a été pendant longtemps admise aux Etats-Unis que dans la prévention de l'accident vasculaire cérébral chez les hommes, ce qui a eu pour conséquence de priver de nombreuses femmes d'un traitement susceptible de leur sauver la vie [8].

Critères de validité des analyses de sous-groupes

Les critères de validité des analyses de sous-groupes se déduisent aisément des exemples décrits ci-dessus et le tableau 3 ←, tout en ne mentionnant que l'essentiel, en donne un aperçu. Tous ces critères ont en commun un contrôle aussi efficace que possible des effets du hasard, pour conférer aux analyses de sous-groupes un degré acceptable de fiabilité et de crédibilité. Ceci implique en règle générale qu'il convient de se limiter à un petit nombre de questions cliniquement importantes avec une justification biologique claire [2, 13] – ce qui est difficile pour de nombreux directeurs d'étude, en raison de l'immense

Tableau 2. Amlodipine vs placebo chez des patients avec insuffisance cardiaque (IC). Résultats d'une analyse de sous-groupe chez des patients sans cardiopathie ischémique PRAISE-1 [10] et d'un essai ultérieur, PRAISE-2 [11], chez des patients avec insuffisance cardiaque et coronarographie normale.

	PRAISE-1		PRAISE-2		PRAISE-1 + PRAISE-2 combinées	
	Sous-groupe Uniquement des patients avec IC non ischémique		*Tous les patients de l'étude* Coronarographie normale			
Traitement	Placebo	Amlodipine	Placebo	Amlodipine	Placebo	Amlodipine
Nombre de patients	212	209	826	826	1408	1397
Mortalité	34,9%	21,5%	31,7%	33,7%	34,0%	33,4%

Tableau 3. Dix critères idéaux pour la réalisation correcte d'analyses de sous-groupes. Les principaux critères (de l'avis de l'auteur) figurent en gras dans le texte.

1.	**Mention des analyses de sous-groupes dans le protocole de l'étude (ou dans un plan d'analyses séparé).** Formulation d'hypothèses et description des effets (direction et importance)
2.	Présentation de la justification de l'analyse de sous-groupes (effet biologique attendu)
3.	**Définition précise des différents sous-groupes à l'avance** (par ex. quartiles ou limites supérieures/inférieures des variables mesurées)
4.	**Limitation de l'analyse à un petit nombre de sous-groupes importants** (les effets de sous-groupes ne devraient être examinés qu'en relation avec l'endpoint primaire de l'étude)
5.	Prise en compte lors de la planification des cas, si des effets de sous-groupe importants sont attendus (pour assurer une puissance de test suffisante
6.	Envisager une randomisation stratifiée
7.	**Limitation de l'analyse de sous-groupes aux variables de la ligne de base**
8.	**Réalisation d'un test d'interactions comme première étape d'évaluation** (seul un test d'interactions significatif permet une analyse subséquente séparée de sous-groupes)
9.	**Correction statistique pour tests répétés (multiples)**
10.	Mention dans la publication de toutes les analyses de sous-groupes effectuées

masse de données de patients recueillie dans les grandes études. Mieux vaut donc s'en tenir à l'adage «peu mais bien». L'inclusion de questions moins importantes augmente le nombre de tests statistiques et donc la probabilité d'effets du hasard, ce qui limite involontairement la portée des questions véritablement essentielles.

Les directives des autorités de contrôle des médicaments [14, 15] exigent que les analyses de sous-groupes soient planifiées et définies à l'avance dans le protocole d'étude (ou dans un plan d'analyses) ou du moins avant que les résultats ne soient connus. Les analyses post-hoc motivées par les données recueillies ne sont pas admises par les autorités, car la probabilité de résultats faussement positifs y est particulièrement élevée. De plus, il n'est pas possible de vérifier combien de tests ont réellement été effectués et combien de résultats non significatifs ont passé aux oubliettes. Mais les analyses post-hoc sont également considérées d'un œil critique en dehors du domaine des régulations [8, 16, 17]. Le CONSORT Statement [18], sur lequel se basent actuellement les meilleurs journaux biomédicaux pour l'évaluation des manuscrits qui leur sont soumis, exige des indications claires sur les analyses qui avaient été spécifiées avant l'étude et sur celles qui ont été effectuées après coup.

La première étape d'une analyse de sous-groupe consiste fondamentalement à réaliser un test d'interactions (test d'homogénéité de l'effet du traitement). Ce n'est que si le test d'interactions est significatif qu'il sera ensuite possible d'analyser des sous-groupes séparés. La première règle consiste par conséquent à comparer l'importance des effets thérapeutiques entre eux, autrement dit de déterminer les risques relatifs (RR) ou les odds ratios (OR) – et non les valeurs de p des sous-groupes [19]. Qu'est-ce que cela signifie concrètement? Un exemple fictif va l'illustrer: supposons qu'un nouveau traitement de l'ostéoporose diminue de manière significative de 20% (RR: 0,8; IC: 0,71–0,89; p <0,05) l'incidence des fractures non vertébrales par rapport à un traitement de contrôle. Chez les femmes avec antécédents de fractures (sous-groupe 1), le risque relatif est de 0,77 (0,57–0,97; p <0,05) contre «seulement» 0,83 (0,59–1,07, non significatif) chez les femmes sans antécédents de fracture (sous-groupe 2). Cela signifie-t-il que le nouveau traitement n'est efficace que dans le sous-groupe 1, mais pas dans le sous-groupe 2? La vraie question est de savoir si la réduction du risque de 23% observée dans le sous-groupe 1, est différente des 17% constatés dans le sous-groupe 2. Le test d'interactions, qui compare les deux risques relatifs, n'est cependant pas significatif et les deux valeurs de p des sous-groupes pris séparément (significatif dans le sous-groupe 1 et non significatif dans le sous-groupe 2) ne doivent pas faire conclure à une différence d'effets thérapeutiques [20]. Ce genre d'erreur se rencontre pourtant souvent en lisant la littérature. On a par exemple prétendu à tort, dans la publication d'une étude sur les patients post-infarctus, qu'un suivi psychologique complémentaire était associé à un risque accru chez les femmes [21]. Si l'augmentation apparente de la mortalité cardiaque n'a atteint le seuil de signification que chez les femmes, un test d'interactions (qui n'a malheureusement pas été réalisé) aurait montré que l'effet thérapeutique ne différait en réalité pas entre les hommes et les femmes [4].

Les tests d'interactions limitent le «flot» de résultats dus au hasard, mais ils ne constituent pas à eux seuls une garantie absolue. Un test d'interactions significatif ne protège en effet pas des conclusions erronées; dans l'exemple de la figure 1, le test d'interactions est hautement significatif avec une valeur de p <0,0001. La puissance du test (power) des tests d'interactions est par ailleurs faible, car les études cliniques sont en règle générale conçues avec une puissance définie en fonction de la question primaire; elle passe ainsi par exemple de 80% pour l'effet thérapeutique dans l'essai global à guère plus de 29% pour l'effet d'interactions dans des proportions semblables [22]. L'augmentation d'un facteur quatre du nombre de cas inclus dans l'étude pour conserver la puissance du test est toutefois considérée comme une option irréaliste.

La répétition des tests dans le cadre des analyses de sous-groupes augmente la probabilité de résultats faussement positifs et nécessite une correction statistique. En l'absence de procédures de test hiérarchiques (i.e., le test suivant ne peut être réalisé que si le précédent était significatif), le niveau de signification doit être adapté au nombre de tests. Le tableau 4 donne un aperçu sur la modification du niveau de signification réel d'un test en fonction du nombre de sous-groupes analysés et sur les adaptations qui sont nécessaires pour conserver un niveau de signification

Tableau 4. Niveau de signification en fonction du nombre de sous-groupes examinés.

Nombre de sous-groupes	Niveau de signification réel, sans correction statistique	Niveau de signification du test isolé pour le maintien du niveau global de 0,05
2	0,0975	0,0253
3	0,1426	0,0170
4	0,1855	0,0127
5	0,2262	0,0102
6	0,2649	0,0085
7	0,3017	0,0073
8	0,3366	0,0064
9	0,3598	0,0057
10	0,4013	0,0051

Niveau de signification nominal = 0,05; colonne du milieu: niveau de signification réel sans correction pour tests multiples; colonne de droite: niveau de signification de chaque test pour le maintien d'un niveau de signification global de 0,05.

nominal pour chaque test pris à part. L'exemple suivant illustre le fait que ce type de correction est aussi nécessaire lorsque plusieurs tests d'interactions sont effectués: un essai comparatif entre l'association clopidogrel plus acide acétylsalicylique et la monothérapie d'AAS n'a pas mis en évidence de différence significative entre les deux schémas thérapeutiques en termes de prévention des événements athérothrombotiques [23]. Seule une analyse sur 20 sous-groupes (de patients symptomatiques) a trouvé un avantage significatif en faveur de l'un des traitements (test d'interactions avec p = 0,045). Si on avait corrigé de manière adéquate le niveau de signification du test d'interactions pour les 20 tests effectués (0,05/20 = 0,0025), aucun des résultats des sous-groupes n'aurait eu la moindre chance d'approcher même le seuil de signification statistique [24].

Les analyses de sous-groupe doivent se limiter à des caractéristiques préexistantes au début de l'étude (par ex. âge des patients, degré de sévérité de la maladie ou facteurs de risque). Il ne doit pas s'agir de variables potentiellement influençables par le traitement et éventuellement susceptibles de décider si un patient appartient à un groupe donné ou non (par exemple, des modifications des valeurs de laboratoire ou des symptômes cliniques) [25]. Des comparaisons basées sur des observations post-randomisation comportent un risque de biais important, par exemple lorsque les répondeurs à une statine sont définis par le degré de réduction du taux de cholestérol sous traitement dans une étude contrôlée par placebo [26]. La survenue d'un end-point clinique (par ex. un événement cardiovasculaire) peut être étroitement liée à la réponse biochimique et ne constitue de ce fait plus une variable indépendante. Les évaluations des groupes de répondeurs et de non répondeurs ne sont par conséquent pas valides. De même, les analyses de sous-groupes sur la compliance au traitement posent aussi problème, dans la mesure où les patients du groupe traité sont classés comme peu compliants ou sont exclus de l'essai pour cause d'effets indésirables, alors que ceux du groupe placebo le sont pour manque d'efficacité thérapeutique, autrement dit pour des raisons complètement différentes. Ceci implique que les deux sous-groupes ne sont en fait plus comparables [8]. L'exemple du Coronary Drug Project est édifiant à ce propos [27]: les patients du groupe clofibrate, qui prenaient au moins 80% des doses de médicament de l'étude, avaient de façon hautement significative une mortalité à cinq ans plus basse que les patients peu compliants (15 vs 24,6%; p = 0,0001). La différence était même plus impressionnante dans le groupe placebo (15,1 vs 28,3%; p <0,00001) – apparemment, les patients prenant régulièrement le placebo retireraient donc le plus de bénéfice du traitement.

Résumé de l'évaluation et recommandations

«The best test of the validity of subgroup analyses is not significance but replication.»
Professeur Peter M. Rothwell, Oxford

Les tests de qualité effectués de manière systématique sur les analyses de sous-groupes confirment que la plupart des études publiées – même dans les journaux médicaux de premier plan – ne satisfont pas aux standards évoqués précédemment [4, 13, 22, 28]. Les auteurs ont recours de façon exagérée aux analyses de sous-groupes et les résultats sont souvent surévalués en raison de carences méthodologiques. Un test d'interactions fait souvent défaut, si bien que les résultats représentés se limitent aux valeurs de p (non ajustées) de certains sous-groupes. La probabilité de l'existence de résultats fortuits et de passage à des recommandations thérapeutiques injustifiées est donc très élevée. Les analyses post-hoc invitent à encore plus de scepticisme, surtout lorsque les résultats de l'analyse de sous-groupes sont en contradiction avec l'effet global observé dans l'étude et que l'analyse de sous-

groupe ne constitue qu'une tentative de «sauver une étude négative». Le fait de manipuler les données après coup, à la recherche de significations statistiques («data dredging»), revient à faire des analyses de sous-groupes abusives et doit être qualifié de non scientifique [16].

Les effets significatifs observés dans le test d'interactions ne reposent en général que sur des interactions quantitatives, ce qui veut dire que l'importance de l'effet thérapeutique est plus grande dans un sous-groupe que dans un autre. Ces interactions sont souvent dues au hasard et sans signification clinique. Les interactions qualitatives indiquent que l'un des traitements est nettement supérieur dans l'un des sous-groupes par rapport à l'autre (ou il peut même être dangereux). Ils sont extrêmement rares [28].

Dans les petits essais, il vaut mieux éviter d'une manière générale les analyses de sous-groupes. Les analyses de sous-groupes dans les grandes études multicentriques randomisées ne peuvent être cliniquement significatives que si elles répondent aux critères de validité mentionnés ci-dessus. Si une analyse de sous-groupes revêt un très grand intérêt clinique, la randomisation des patients devrait être stratifiée à l'avance pour prévenir tout biais de randomisation. Mais même dans ce cas, la signification des analyses de sous-groupes ne saurait être surestimée, car l'effet global observé dans l'étude reste plus fiable de par sa puissance élevée que celui constaté dans les différents sous-groupes [8].

Le principal intérêt des analyses de sous-groupes réside dans la confirmation de l'effet global mesuré dans l'étude et dans la vérification de la consistance des résultats dans différents collectifs de patients. Les différences d'efficacité et de sécurité dans quelques groupes de patients spécifiques n'ont qu'un caractère exploratoire et doivent être interprétées avec la plus grande prudence. La présence d'une logique biologique claire, d'informations relatives à une relation dose-effet ou d'une reproductibilité des résultats dans différents échantillons, par exemple sur la base des résultats dans les différents centres d'étude, pourrait renforcer la validité d'une analyse de sous-groupes [25]. Ce type de remarques sur les analyses de sous-groupes ne sert toutefois finalement qu'à la formulation d'hypothèses et devrait plutôt être vérifié dans le cadre d'autres études indépendantes [2, 22]. La réalisation d'une méta-analyse portant sur des sous-groupes de patients provenant de plusieurs études peut être intéressante. Il a par exemple fallu une méta-analyse de plusieurs grandes études pour démontrer que l'utilité de la thrombolyse dans l'infarctus du myocarde est d'autant plus grande que le traitement est instauré rapidement [29].

Correspondance:
Dr Peter Kleist
Viktoriastrasse 52
CH-3084 Wabern
peter.kleist@bluewin.ch

Littérature recommandée

– Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity. I: clinical trials. Lancet. 2001;357:373–80.
– Counsell CE, Clarke MJ, Slattery J, Sandercock PAG. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ. 1994;309:1677–81.
– Vous trouverez la bibliographie complète [1–29] dans l'édition online de cet article à l'adresse www.medicalforum.ch/pdf/pdf_f/2007-39-210.pdf.

Vorsicht bei Subgruppenanalysen in klinischen Studien

Peter Kleist

Pharma Focus AG, Volketswil

Literatur

1 Feinstein AR. The problem of cogent subgroups: A clinicostatistical tragedy. J Clin Epidemiol 1998;51:297–9.
Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176–86.
2 van Gijn J. Extrapolation of trial data into practice: where is the limit? Cerebrovasc Dis 1995;5:159–62.
3 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917–30.
4 Betablocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA 1982; 247: 1707-14.
5 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.
6 ISIS-2 (Second International Study on Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. Lancet. 1988;2:349–60.
7 Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. Lancet. 2001;357:373–80.
8 European Carotid Surgery Trialists' Collaborative Group. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST). Lancet. 1998;351:1379–87.
9 Packer M, O'Connor CM, Ghali JK, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. N Engl J Med. 1996;335:1107–14.
10 Thackray S, Witte K, Clark AL, Cleland JGF. Clinical trials update: OPTIME-CHF, PRAISE-2, ALL-HAT. Eur J Heart Fail. 2000;2:209–12.
11 The Canadian Cooperative Study Group. A randomised trial of aspirin and sulfinpyrazone in threatened stroke. N Engl J Med. 1978;299:53–9.
12 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355:1064–9.
13 International Conference on Harmonisation (ICH). Guideline E9: Statistical principals for clinical trials. 05 February 1998. www.ich.org/LOB/media/MEDIA485.pdf (Zugriff am 14. Juni 2007).
14 Committee for Proprietary Medicinal Products (CPMP). Points to consider on multiplicity issues in clinical trials. CPMP/EWP/908/99. 19 December 2002. www.emea.europa.eu/pdfs/human/ewp/090899en.pdf (Zugriff am 14. Juni 2007).
15 Lewis SC, Warlow CP. How to spot bias and other potential problems in randomised controlled trials. J Neurol Neurosurg Psychiatry. 2004;75:181–7.
16 Sleight P. Debate: Subgroup analyses in clinical trials – fun to look at, but don't believe them. Curr Control Cardiovasc Med. 2000;1:25–7.
17 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. 2001;134:663–94.
18 Matthews JNS, Altman DG. Interaction 2: compare effect sizes not P values. BMJ. 1996;313:808.
19 Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ. 2003;326:219.
20 Frasure-Smith N, Lespérance F, Prince RH, et al. Randomised trial of home-based psychological nursing intervention for patients recovering from myocardial infarction. Lancet. 1997;350:473–9.
21 Hernandez 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.
22 Bhatt DL, Fox KAA, Hacke W, et al. Clopidogrel and aspirin versus aspirin alone for the prevention of atherothrombotic events. N Engl J Med. 2006;354:1706–17.
23 Lagakos SW. The challenge of subgroup analyses – reporting without distorting. N Engl J Med 2006;354:166–9.
24 Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. MJA 2004;180:289–91.
25 West of Scotland Coronary Prevention Study Group. Influence of pravastatin and plasma lipids on clinical events in the West of Scotland Coronary Prevention Study (WOSCOPS). Circulation 1998;97:1440–5.
26 The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. N Engl J Med. 1980;303:1038–41.
27 Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: Some lessons from major cardiovascu-

Korrespondenz:
Dr. med. Peter Kleist
Chriesbaumstrasse 2
CH-8604 Volketswil
peter.kleist@pfc.ch

lar trials. Am Heart J 2000; 139:952–61.
28 Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Indications for fibrinolytic therapy in suspected myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Lancet. 1994;343:311–22.