L'essai adaptatif en recherche clinique

Jan C. Schuller
EORTC, European Organisation for Research and Treatment of Cancer, Bruxelles, Belgique

Résumé

Le développement de la biologie moléculaire a entraîné une série de nouveaux aspects dans les traitements anticancéreux. De nombreuses substances nouvelles et potentiellement efficaces sont apparues sur le marché (telles que les *biologics*, produits de fabrication biotechnologique), et les récents procédés diagnostiques (tels que par biomarqueurs) montrent que de nombreuses sous-espèces de cancer demandent des thérapies ciblées *(targeted therapies)*. On aboutit alors à des dilemmes à divers égards dans les plans d'études cliniques, car on part du principe que le traitement s'applique à un collectif de patients plus ou moins homogène, alors qu'il faut répondre à un besoin de «médecine personnalisée».
Les essais expérimentaux «classiques» demandent de formuler *a priori* une série d'hypothèses sur les propriétés du traitement à tester. Dans les options thérapeutiques toutes récentes, les données existantes sont souvent insuffisantes pour le faire. C'est là qu'interviennent les essais adaptatifs. Ils sont destinés à accélérer le processus de développement des nouveaux traitements du cancer tout en respectant les exigences posées par la médecine personnalisée.
Cet article présente une introduction à la méthode des essais cliniques adaptatifs et une appréciation critique reflétant le point de vue du département de statistique de l'EORTC.

Mesure du signal

Le signal mesuré lors d'un essai clinique est la différence entre l'efficacité du nouveau traitement et celle du traitement témoin. On peut choisir comme témoin un bras de traitement par placebo, un bras de traitement standard ou un bras historique. Une part non négligeable de la tâche du statisticien consiste à établir un plan d'études pouvant mesurer le signal recherché avec une probabilité élevée (dans le cas où le nouveau traitement est effectivement meilleur que le témoin). On dira dans ce cas que l'étude a une puissance suffisante pour mesurer un effet déterminé. Il n'est pas difficile d'imaginer que des signaux faibles requièrent des instruments de mesure plus précis. En statistique, on augmente la précision en choisissant un échantillon plus grand.
La taille de l'échantillon est un facteur essentiel du coût: il faut donc lui accorder une attention particulière: si le nombre de patients inclus dans un essai est trop faible, nous courons le risque de ne pas pouvoir mesurer le signal (puissance insuffisante), et s'il est trop élevé, nous heurtons les principes d'éthique et l'essai devient trop coûteux.

La règle de décision «classique»

Le procédé classique veut que l'on formule les hypothèses, alpha et bêta (erreurs de type I et de type II) et que l'on fixe la taille de l'échantillon avant de commencer les observations. Une fois l'étude et la collecte de données terminées, le plan d'analyse statistique fournit une règle de décision claire permettant de déterminer si le nouveau traitement exerce ou non l'effet recherché. La décision n'est prise qu'après avoir analysé les données de tous les patients de l'essai. Pour maintenir l'erreur de type I (rejet erroné de l'hypothèse nulle), on renonce à toute évaluation intermédiaire.

Méthodes séquentielles

Le mathématicien Abraham Wald, de l'Université Columbia, modifia ce procédé en 1943 lorsqu'il présenta au *National Defense Research Committee* un rapport intitulé «Sequential analysis of statistical data: theory». Son rapport traitait entre autres du contrôle de qualité de processus industriels impliquant une production en flux continu. Il s'agissait de déterminer les taux de pièces défectueuses parmi les pièces finies. La règle de décision classique demandait d'attendre que l'on produise un certain nombre d'unités, de dénombrer ensuite les unités défectueuses et de s'appuyer sur ces données pour l'évaluation statistique.
Mais combien de pièces faut-il fabriquer pour pouvoir procéder à l'évaluation statistique? La taille de l'échantillon nécessaire dépend du taux de pièces défectueuses: elle peut être petite si le taux est élevé, mais elle doit être grande s'il est faible. Si l'on ne connaît pas le taux d'unités défectueuses auquel il faut s'attendre, le procédé classique peut s'avérer trop rigide. Prenons un exemple extrême: supposons qu'à la suite d'une erreur passée inaperçue, presque toutes les unités soient défectueuses, mais que le contrôle de qualité ait été dé-

Ce travail a été soutenu par le *National Cancer Institute* (NCI, Bethesda, Maryland, USA, Grant N° 5U10 CA011488-40), et par le Fonds Cancer (FOCA, Belgique).
L'auteur se déclare responsable du contenu de cet article, qui ne reproduit pas nécessairement le point de vue officiel du NCI.

cidé sur une base de n = 1000 unités. Si l'on ne produit que 100 unités par jour, il faut attendre 10 jours pour détecter l'erreur. Dans un tel cas, il ne faut pas être statisticien pour affirmer au bout de 10 ou 20 pièces défectueuses qu'il y a déjà une erreur dans le procédé de fabrication. L'analyse séquentielle fournit alors un nouvel élément important: toute unité produite au sein d'un processus est immédiatement contrôlée et donne lieu à une évaluation. En fonction de résultat, la fabrication s'arrête ou non.

Les essais séquentiels par groupes forment une variante de ce concept, et sont appliqués en recherche clinique depuis longtemps déjà. Au cours d'un essai, on procède à une ou plusieurs analyses intermédiaires. Si un nouveau traitement est extrêmement efficace ou extrêmement inefficace, cette méthode permet éventuellement de le détecter avant que tout le collectif de patients prévu dans l'étude n'ait eu à subir le traitement. C'est ici que l'adaptation intervient dans les essais cliniques. Après une analyse intermédiaire des données, on peut interrompre prématurément un essai si l'efficacité du traitement est estimée soit très élevée, soit très faible. Ce procédé répond aux impératifs éthiques et économiques.

Méthodes bayésiennes

Ces méthodes portent le nom du mathématicien et pasteur presbytérien Thomas Bayes (env. 1702–1761). Selon la règle de Bayes, la probabilité d'un événement n'est pas uniquement déterminée par les résultats expérimentaux nouveaux, mais aussi par des résultats antérieurs (probabilité *a priori*). De nombreux chercheurs apprécient particulièrement ces méthodes, car elles correspondent bien au raisonnement par inférence utilisé en sciences naturelles et en médecine. Elles sont également appliquées au quotidien dans les prises de décision: nous voyons le soleil se lever le premier jour, le deuxième et tous les autres jours, et pour finir nous nous attendons à le voir se lever demain également. Le statisticien bayésien dirait: «Le lever du soleil de demain suit une probabilité *a priori* élevée.»

La médecine réserve une grande place aux tests d'hypothèses statistiques. On cherche à obtenir des résultats significatifs. Mais il est bien connu qu'en soi, «$p < 0{,}05$» informe peu et mène fréquemment à des interprétations erronées. Les méthodes bayésiennes permettent d'évaluer les effets de traitements sans imposer la recherche de valeurs significatives.

Définition et exemples d'essais adaptatifs

Un essai adaptatif permet de modifier le plan d'étude et/ou changer les paramètres statistiques en cours d'essai sans en affecter pour autant la validité ou l'intégrité de l'essai. L'effet recherché est l'amélioration de l'efficacité, de la rapidité et de la flexibilité des essais cliniques.

La suite de l'article contiendra les principaux exemples d'essais adaptatifs, avec un accent particulier sur les plans d'étude en phases II et III. Les essais séquentiels par groupes évoqués plus haut sont déjà utilisés depuis longtemps en pratique et ils occupent une place quelque peu particulière, car les adaptations possibles sont très limitées (il s'agit presque toujours d'abandon d'essais en raison de l'inefficacité ou du manque de sécurité du traitement). Par contre, les essais adaptatifs d'un type plus récent, dit flexible, impliquent les adaptations les plus diverses, comme la modification de la taille d'échantillon, la randomisation adaptative, les plans d'étude combinés *(seamless)* de phase II/III, l'exclusion de certains bras de traitement, la modification du critère d'évaluation ainsi que le changement de catégorie dans les bras de traitement, les faisant passer de «supérieurs» à «non inférieurs».

Essai séquentiel par groupes et arrêt précoce

Dans ce modèle, l'essai est interrompu prématurément si l'analyse intermédiaire montre que l'efficacité du traitement est nettement plus élevée ou nettement plus faible que prévu. C'est le type d'essai adaptatif le plus répandu, notamment dans les études de phases II et III.

Essai combiné de phase II/III *(seamless phase II/III design)*

Ce type d'essai réunit les objectifs de la phase II (déterminer dans quelle mesure le traitement antitumoral est efficace) et de la phase III (évaluer l'efficacité du nouveau traitement par rapport au traitement témoin) au sein d'une seule étude.

Supposons que l'on cherche parmi trois nouveaux traitements à déterminer le meilleur. Dans un premier temps, il suffit de tester ces trois traitements sur un nombre restreint de patients par un essai contrôlé de phase II (le groupe de contrôle peut être historique), avec un petit nombre de patients et donc une puissance statistique réduite. On sélectionne le traitement qui donne les meilleurs résultats (et qui laisse espérer une amélioration par rapport au témoin). Ce traitement sera ensuite étudié contre un traitement témoin dans un essai contrôlé de phase III suffisamment puissant. Il peut s'écouler entre plusieurs mois et quelques années entre les phases II et III, car il faut du temps pour analyser les données de la phase II et planifier l'essai de phase III.

Dans un essai combiné de phase II/III, le recrutement des patients entre les deux phases ne sera pas interrompu. En contrepartie, plusieurs analyses intermédiaires seront effectuées dans le but de rejeter le(s) bras de traitement le(s) moins efficace(s). Ce modèle inclut les patients de la phase II dans l'analyse finale de la phase III.

En théorie, le nombre nécessaire de patients est moins élevé dans un essai combiné que dans des essais séparés et les résultats de la phase III sont obtenus plus rapidement.

Dans la pratique, on se heurte au problème de recrutement trop rapide *(overrun)*: pendant l'étape d'analyse intermédiaire, on continue à inclure de nouveaux patients dans l'étude, et si cet afflux est trop élevé, il arrive que tous les patients aient déjà reçu le traitement et par conséquent, la décision d'adaptation arrive trop tard. Les essais de phase II sont très instructifs par rap-

port à la nouvelle méthode de traitement, et leurs résultats exercent une grande influence sur la planification et la réalisation de la phase III. L'essai combiné de phase II/III est à cet égard désavantageux: sa planification s'effectue d'un seul tenant, et les résultats de la phase II ne peuvent être pris en compte qu'avec réserves dans la phase III.

Adaptation de la taille de l'échantillon
(sample size re-estimation)
Les résultats intermédiaires permettent d'adapter la taille de l'échantillon de façon à obtenir la puissance statistique nécessaire pour déterminer l'efficacité du traitement. L'expérience montre qu'il faut augmenter la taille de l'échantillon lorsque la puissance conditionnelle résultant de l'analyse intermédiaire se situe entre 30 et 90% (la puissance conditionnelle étant la probabilité de détecter un effet, s'il existe, sur la base des données présentes à cette étape). Contrairement à la puissance fixée en début d'essai, par ex. 80%, la puissance conditionnelle dépend des résultats intermédiaires. Si elle est inférieure à 30%, l'essai a peu de chances de produire un résultat positif, même si l'on augmente la taille de l'échantillon; si elle dépasse 90%, il est probable que l'étude montre l'existence d'un effet sans nécessité d'adaptation.
La méthode adaptative est susceptible de réduire les coûts en début d'essai. Cependant s'il faut augmenter la taille de l'échantillon, le budget s'accroît en conséquence. Cela pose parfois problème, car les décisions concernant l'adaptation de la taille de l'échantillon sont généralement prises par un comité indépendant de suivi des données *(Independent Data Monitoring Committee, IDMC)* alors que la compétence budgétaire appartient à une autre instance (par ex. une entreprise industrielle).
Dans un essai, il faut toujours définir avec soin le niveau de signification de l'efficacité clinique d'un traitement. Or si l'on adapte la taille de l'échantillon aux données intermédiaires, cela revient à réévaluer ce niveau, ce qui rend le procédé discutable.

Randomisation adaptative
Normalement, la randomisation s'effectue selon un rapport constant (par ex. 1:1 ou 1:2). Lorsqu'elle est adaptative, elle est ajustée à chaque étape d'analyse intermédiaire (méthode bayésienne) afin d'augmenter la probabilité pour un patient d'être randomisé vers un bras de traitement efficace. Ce procédé, auquel on recourt fréquemment dans les essais adaptatifs basés sur un marqueur biologique, pose toutefois un problème: il introduit un élément déterministe lors de la distribution vers les différents bras de traitement de l'essai (biais opérationnel) et ne respecte donc pas l'idée fondamentale de la randomisation.

Essai basé sur la sélection du gagnant
(pick the winner design)
Il s'agit le plus souvent d'un essai de phase II avec plusieurs bras de traitement et une ou plusieurs analyses intermédiaires (comme l'exemple présenté dans la section sur l'essai combiné de phase II/III). On utilise les résultats des analyses intermédiaires pour sélectionner les bras de traitement efficaces et exclure les bras inefficaces.

Essai adaptatif basé sur l'utilisation d'un marqueur biologique *(biomarker-adaptive design)*
On adapte le déroulement de l'essai en fonction de la réponse à différents biomarqueurs associés à la pathologie traitée. Ce type d'essai sert à déterminer, avec plus de précision, la population de patients pour laquelle un nouveau traitement est efficace, et il constitue un pas en avant vers la médecine personnalisée et les thérapies ciblées. Mais attention: il ne suffit pas d'associer un marqueur biologique et un critère d'évaluation clinique pour construire un modèle prédictif.

Pilotage du recrutement
(ramping-up of the patient accrual)
Pour éviter d'inclure trop de patients dans une étude (*overrun*, voir plus haut), on peut mettre à profit la méthode adaptative et ralentir le recrutement en maintenant délibérément l'échantillon à une taille faible, pour augmenter ensuite la vitesse de recrutement après adaptation de l'essai.

Adaptation par réevaluation continue de la dose
(adaptive dose finding)
On choisit ce type d'essai en phase I pour déterminer la dose maximale tolérée d'une nouvelle substance: plutôt que de fixer les doses à tester avant le début de l'étude, on peut les ajuster en fonction des résultats intermédiaires (par ex. au moyen de la méthode de réévaluation continue ou *continual re-assessment method*, *CRM*).

Modification du critère d'évaluation primaire
La modification du critère d'évaluation primaire représente une forme extrême d'adaptation, et elle n'est indiquée qu'en de très rares situations. Dans la plupart des cas, il faut se montrer très critique vis-à-vis d'une modification du critère d'évaluation primaire sur la base des résultats intermédiaires.

Conclusion

La randomisation et l'aveugle sont les principales techniques utilisées pour éviter les biais dans une étude. Les méthodes adaptatives représentent un compromis entre ces deux techniques, et le fait de connaître des résultats intermédiaires peut porter préjudice à la valeur probante de l'essai. Dans un essai adaptatif, il existe toujours un risque d'aboutir à des biais difficilement contrôlables. Ce risque est bien connu des autorités de contrôle: l'Agence européenne des médicaments (EMA) a publié des directives soumettant les essais adaptatifs à des exigences très strictes pour avoir valeur de preuve, par ex. dans le cas d'une étude clinique en vue de l'autorisation d'un médicament. On cherche d'abord à contrôler les erreurs de première espèce (conclure à tort qu'un nouveau traitement est efficace). Certaines adaptations, comme celle du critère d'évaluation primaire, ne sont acceptées par l'EMA que dans des cas exceptionnels si elles sont soigneusement justifiées.

Une chose est claire: les méthodes adaptatives sont nettement plus exigeantes que les méthodes classiques sur le plan de la logistique, de la planification et de la durée. Il arrive souvent que les résultats obtenus ne soient pas à la hauteur des moyens investis.

Ne perdons pas de vue que les essais adaptatifs ne font pas de miracle: ils ne sont pas plus sensibles à l'efficacité du traitement que les essais classiques. Finalement, que l'essai soit adaptatif ou non, la règle est la même: lorsque l'effet examiné est faible, il faut augmenter la taille de l'échantillon.

Ces aspects demandent à être examinés soigneusement. Pour chaque cas, il faut toujours réfléchir si un essai standard ne ferait pas tout aussi bien l'affaire qu'un essai adaptatif.

Remerciement
L'auteur tient à remercier le réviseur anonyme d'une version antérieure de cet article pour ses précieuses remarques et Sandra Collette pour les corrections du manuscrit.

Correspondance:
Jan C. Schuller
Senior Biostatistician
EORTC, European Organisation for Research and Treatment of Cancer
Avenue E. Mounierlaan, 83/11
Bruxelles 1200 Brussel
Belgique
jan.schuller@eortc.be

Références recommandées

– Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics. Draft Guidance. U.S. Department of Health and Human Services, Food and Drug Administration. February 2010.
– Emerson Scott S. Issues in the use of adaptive clinical trial designs. Statist. Med. 2006;25:3270–96.
– Reflection paper on methodological issues in confirmatory trials planned with adaptive Design; Committee for medicinal products for human use. EMA, Doc. Ref. CHMP/EWP/2459/02EMA, October 2007.

Association? Une femme de 37 ans se présente avec de la fièvre (39,2 °C) depuis 3 jours, des céphalées, des nausées et une confusion. Elle obtient un score de 11 sur l'échelle de coma de Glasgow. Le médecin constate une raideur de la nuque et l'absence de réflexes au niveau des extenseurs. La ponction lombaire révèle une pression de 48 cm d'eau, un contenu leucocytaire de 394, une concentration de protéines de 1,21 g/l et une concentration de glucose de 0,6 mmol/l. Une IRM cérébrale montre de multiples lésions hyperintenses et une accumulation circulaire du gadolinium. Les cultures du liquide céphalorachidien révèlent la présence de streptocoques. Que se passe-t-il?

(Réponse dans la colonne de droite)

Réponse: *Une **méningite**? Un abcès **cérébral**? Les deux réponses sont correctes, mais quelle est la cause? Les streptocoques mis en évidence appartiennent au groupe C de Lancefield (Streptococcus equi, subspecies zooepidemicus) et il s'avère que la patiente est une cavalière enthousiaste qui s'est récemment fait mordre par son cheval. Les méningites sont principalement provoquées par les streptocoques de la sous-espèce zooepidemicus et elles sont sensibles à la pénicilline. Après 2 ans, la patiente, qui souffre d'une amnésie grave, n'est toujours pas autonome. (Lancet. 2010;376:1194.)*