REVUE DE STATISTIQUE APPLIQUÉE

A. KRAMAR
C. BASCOUL-MOLLEVI
S. GOURGOU-BOURGADE

Règles d'arrêt précoce dans les essais cliniques basées sur une surveillance séquentielle des événements indésirables graves

Revue de statistique appliquée, tome 53, n° 4 (2005), p. 49-59
<http://www.numdam.org/item?id=RSA_2005__53_4_49_0>

Rev. Statistique Appliquée, 2005, **LIII** (4), 49-59

RÈGLES D'ARRÊT PRÉCOCE DANS LES ESSAIS CLINIQUES BASÉES SUR UNE SURVEILLANCE SÉQUENTIELLE DES ÉVÉNEMENTS INDÉSIRABLES GRAVES

A. KRAMAR, C. BASCOUL-MOLLEVI, S. GOURGOU-BOURGADE

CRLC Val d'Aurelle-Paul Lamarque, Unité de Biostatistiques,
Parc Euromédecine, 34298 Montpellier cedex 5, FRANCE

RÉSUMÉ

Plusieurs méthodes statistiques ont été développées pour définir des règles d'arrêt en terme d'efficacité dans la planification des essais cliniques de phase II et III, mais elles reposent sur une première analyse après l'inclusion d'un nombre fixe de patients. Cependant, ces méthodes ne sont pas adéquates dans le cas de l'évaluation d'événements indésirables graves (EIG) qui peuvent se produire assez tôt dans l'essai. Leur survenue peut nécessiter un arrêt précoce de l'essai si leur nombre est trop important. Les méthodes développées ici définissent des règles d'arrêt séquentielles après chaque survenue d'un EIG en comparant le nombre total de patients inclus au nombre maximal de patients qui vont satisfaire des critères d'EIG maximal acceptable tout en préservant l'erreur de type I nominale.

Mots-clés : *arrêt précoce séquentiel, événements indésirables graves, essais cliniques*

ABSTRACT

Several multi-stage or group sequential statistical methods have been developed for defining stopping rules in terms of efficacy in phase II and III clinical trials, but they rely on interim analyses after the inclusion of a fixed number of patients. These methods, however, are not adequate for the evaluation of serious adverse events (SAE) which can occur relatively early in the trial. A high frequency of their occurrence may require the trial to close early. The methods developed here define stopping rules after the occurrence of each SAE by comparing the total number of patients included to the number of patients satisfying maximum SAE criteria while preserving the nominal type I error.

Keywords : *early stopping rules, serious adverse events, clinical trials*

1. Introduction

Les effets indésirables graves (EIG) observés sous traitements sont devenus une préoccupation majeure en Europe et aux États-Unis. Un guide pour la centralisation des données concernant les effets secondaires est en cours de réalisation au niveau

européen (Eudravigilance). Le codage de ces effets dans le contexte d'un essai clinique est détaillé dans la Directive 2001/20/EC en vue d'une harmonisation européenne. L'objectif de cette implication est de veiller à la bonne tolérance aux médicaments pendant le déroulement d'une recherche biomédicale pour arrêter précocement l'étude ou modifier les caractéristiques des patients et/ou des traitements, si un traitement s'avère trop toxique.

Avant la mise sur le marché d'un nouveau médicament, trois phases de développement sont nécessaires : phase I, phase II et phase III. Dans les essais de phase I, la recherche de la dose maximale tolérée (DMT) est l'objectif principal, ce qui conduit à recommander une dose en phase II. Dans les essais de phase II, l'objectif principal est la recherche d'une activité thérapeutique à la dose recommandée. Dans les essais de phase III randomisés, l'objectif principal est la comparaison du nouveau médicament avec un traitement de référence. Pour chacune de ces phases, des méthodologies statistiques pour le calcul du nombre de sujets nécessaire et des règles d'arrêts sont utilisées. Ces règles d'arrêt nécessitent un contrôle de l'erreur de première espèce, car on ne peut pas multiplier de façon anarchique le nombre de tests statistiques sans augmenter le risque de se tromper dans les conclusions.

Dans les essais de phase I, des méthodes statistiques récentes permettent d'atteindre la DMT plus rapidement (O'Quigley, Shen 1996). Pour la planification des essais multi-étapes de phase II, plusieurs méthodes statistiques ont été développées pour définir des règles d'arrêt en terme de réponse au traitement en cas d'efficacité insuffisante. Néanmoins, tous ces plans reposent sur une première analyse après l'inclusion d'un nombre fixe de patients (Kramar *et al.* 1996). Ils sont construits de façon à minimiser le nombre de patients exposés à un traitement inefficace tout en préservant le niveau nominal α, le risque de déclarer un traitement efficace à tort (erreur de type I), et β, le risque de passer à côté d'un traitement efficace (erreur de type II). Ces plans multi-étapes sont construits de façon à repartir le risque global de première espèce α entre chaque analyse.

Dans les essais multi-étapes de phase II, ces méthodes ont été transposées à la situation bivariée pour prendre en compte à la fois l'efficacité et la toxicité, mais les règles d'arrêt sont seulement invoquées à la première étape de façon indépendante et symétrique pour ces deux critères (Bryant, Day 1995). Dans le cas de l'évaluation d'événements indésirables graves qui peuvent survenir tout au long de l'essai, ces méthodes ne sont pas adéquates.

Actuellement, aucune méthode statistique n'est proposée d'avance pour aider le promoteur à prendre des décisions après avoir observé un certain nombre d'EIG dans un essai de phase III.

Une approche récente a posé le problème à l'envers pour déterminer des règles d'arrêt optimales en fonction du nombre de patients inclus et du nombre d'EIG observés. Cependant, les tables publiées sont limitées à 25 patients avec une probabilité maximale pour un EIG chez un patient fixée à $5\,\%$ (Goldman, Hannan 2001). De plus, ces techniques ne contrôlent pas le risque de première espèce nominal.

L'objectif de ce travail est d'étendre cette nouvelle méthodologie à la surveillance séquentielle des événements indésirables graves par la définition de règles après chaque survenue d'un EIG en comparant le nombre total de patients inclus au nombre

maximal de patients qui vont satisfaire des critères d'EIG maximal acceptable tout en préservant l'erreur de type I nominale.

Ces méthodes vont permettre l'établissement d'une stratégie d'aide à la décision pour les comités de surveillance d'un essai thérapeutique afin d'arrêter précocement l'essai, en cas de toxicité inacceptable plutôt que d'être confronté à des décisions à la carte souvent trop tardives.

2. Méthodologie

Le calcul du nombre de sujets nécessaire dans un essai de phase II ou III repose sur une spécification des erreurs de type I (α) et II (β). Le niveau nominal de ces erreurs est correct seulement dans le cas où une seule analyse est effectuée à la fin de l'essai une fois que toutes les données ont été accumulées. De plus en plus d'essais cliniques prévoient des analyses intermédiaires pour pouvoir arrêter l'essai de façon précoce en faveur ou en défaveur du nouveau traitement. Si à chaque analyse j, $j = 1, \ldots, J$, on déclare une différence significative au niveau $\alpha_j = \alpha_0$, alors après J analyses, le niveau final sera $\alpha_J = 1 - \prod_{j=1}^{J}(1 - \alpha_j) = 1 - (1 - \alpha_0)^J$. Par exemple, si on fait 5 tests statistiques sur les données, chacun au niveau $\alpha_0 = 0,05$, le niveau α_J final sera égal à 0,23. On augmente ainsi le risque de conclure qu'un traitement est efficace alors qu'en réalité il ne l'est pas.

La méthodologie statistique repose sur le calcul d'un intervalle de confiance à $100(1 - \alpha)\%$ autour d'un pourcentage pour un α qui varie en fonction du moment de l'analyse. Les méthodes de calculs sont basées sur la loi binomiale et non des approximations issues de la loi gaussienne.

2.1. Intervalles de confiance à chaque analyse

Soit k, $k = 1, \ldots, n$ le nombre d'EIG pour un échantillon de taille n et p la probabilité de survenue d'un EIG. La borne inférieure, p_{inf}, de l'intervalle de confiance unilatéral à droite pour p, de niveau $1 - \alpha$, intervalle noté IC_α, est la solution de l'équation suivante :

$$\sum_{x=k}^{n} C_n^x p_{\text{inf}}^x (1 - p_{\text{inf}})^{n-x} = \alpha$$

Un intervalle de confiance à $95\,\%$ est souvent utilisé pour rapporter les résultats d'un essai thérapeutique. Cet intervalle suppose qu'une seule analyse est effectuée à la fin de l'essai. Pour illustrer la méthodologie, on considère dans un premier temps que α est fixe quelque soit le nombre d'analyses. Soit τ le niveau maximal de tolérance acceptable pour les EIG étudiés, par exemple un taux de décès toxique inférieur à $3\,\%$. On définit la règle d'arrêt si le traitement est trop toxique, c'est-à-dire si la borne inférieure de IC_α ne contient pas τ, c'est-à-dire $p_{inf} > \tau$. Puisque cette condition peut être satisfaite pour plusieurs valeurs de n, la règle de décision permet un arrêt si n_k, le nombre de patients inclus à la k^{ieme} analyse, est inférieur à N_k, où N_k correspond à la taille maximale de l'échantillon pour laquelle $p_{inf} > \tau$. Si $n_k > N_k$, la valeur de

τ sera incluse dans IC_α, et l'essai peut continuer à inclure des patients. On peut ainsi construire un schéma de règle d'arrêt qui sera invoqué à l'observation de chaque EIG. On représente $N(\tau)$=$N_1, N_2, N_3, N_4, \ldots$, avec N_k la taille de l'échantillon après l'observation de k événements pour un τ fixé.

Le tableau 1 présente les tailles maximales de l'échantillon (N_k) nécessaires pour déclarer le traitement trop toxique pour $\tau = 1\,\%, \ldots, 10\,\%$, $k = 1, \ldots, 5$ et un IC unilatéral de $95\,\%$ à chaque analyse. Pour utiliser la table, il suffit de repérer le nombre d'événements observés et le taux τ choisi initialement dans l'essai. Si le nombre de patients traités, n_k, est inférieur à la valeur dans la table, alors le taux d'EIG est plus élevé que la valeur τ choisie avec un IC de $95\,\%$. On remarque que l'on pourrait envisager d'arrêter l'essai si le premier patient présente un EIG, seulement si le taux maximal acceptable, $\tau = 3\,\%, 4\,\%$, ou $5\,\%$. Si $\tau \geqslant 6\,\%$ on ne peut pas arrêter l'essai.

TABLEAU 1

N_k pour déclarer le traitement trop toxique en fonction du nombre d'EIG et le taux maximal acceptable de toxicité pour $\alpha = 0,05$ à chaque analyse

EIG	Taux maximal acceptable de toxicité τ									
k	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
1	5	2	1	1	1	-	-	-	-	-
2	35	18	12	9	7	6	5	4	4	3
3	82	41	27	21	16	14	12	10	9	8
4	137	69	46	34	28	23	20	17	16	14
5	198	99	66	50	40	33	29	25	22	20

Par exemple, supposons qu'un taux d'EIG de $3\,\%$ est le taux maximal acceptable dans notre population de patients. Si l'on observe un premier EIG, alors on rejettera notre traitement en étant trop toxique si cet EIG est observé chez le premier patient. La décision d'arrêter l'essai dès le premier patient, n'est pas obligatoire, mais elle nécessitera un regard sérieux quant à l'évaluation du bon déroulement du protocole thérapeutique, surtout quand c'est le premier patient, ou un traitement compliqué à appliquer. Dans ce cas, le taux estimé est de 100% et la borne inférieure de l'$IC_{0,05}$ = 5,0003 %. Puisque $\tau = 3\,\%$, le taux maximal acceptable d'EIG, est inférieur à cette borne, on rejette le traitement. Par contre, si cet événement est seulement observé chez le deuxième patient, alors le taux estimé est de $50\,\%$ et la borne inférieure de l'$IC_{0,05}$ = $2,532\,\%$. Puisque $\tau = 3\,\%$ est supérieur à cette borne, on ne rejette pas le traitement.

Au moment où un deuxième EIG est observé, on rejettera alors le traitement si les deux événements ont été observés parmi les 12 premiers patients. Dans ce cas, le taux estimé est $16,7\,\%$ et la borne inférieure de l'$IC_{0,05} = 3,04616\,\%$. Puisque $\tau = 3\,\%$ est inférieur à cette borne, on rejette le traitement et on conclut que le traitement est significativement plus toxique que la limite maximale acceptable. Si on

avait observé deux EIG parmi les 13 premiers patients, alors on n'aurait pas rejeté le traitement, puisque la borne inférieure de l'$IC_{0,05} = 2,805\,\%$ est inférieur à $\tau = 3\,\%$, le taux maximal acceptable de toxicité.

Les intervalles de confiance dans le tableau 1, ont été calculés avec un niveau α égal à 0,05 à l'observation de chaque EIG. Ceci n'est pas satisfaisant, car le risque global n'est plus égal à $0,05$, et on a besoin d'adapter le niveau α au fur et à mesure des analyses, comme pour les tests statistiques sur des critères multiples pour préserver le niveau α nominal.

2.2. Préservation du niveau α nominal

Puisque plusieurs analyses intermédiaires vont être réalisées, il convient d'utiliser des méthodes adéquates, par exemple, celles de Pocock (1977) ou O'Brien et Fleming (1979), pour préserver le niveau nominal. Ces méthodes ont été élaborées dans le contexte d'un essai clinique qui prévoit des analyses séquentielles groupées sur le critère principal, après l'inclusion d'un nombre fixe de patients. Pour diminuer la probabilité d'une conclusion précoce erronée, différentes valeurs de α ont été proposées à chaque analyse intermédiaire. Selon la méthode de Pocock (1977), chaque test statistique est associé à un niveau de signification constant, déterminé de façon à ce que le niveau global de α soit égal à 0,05. Selon la méthode de O'Brien et Fleming (1979), chaque test statistique est associé à un niveau de signification calculé préalablement et répertorié dans une table. Par exemple, pour 3 analyses intermédiaires prévues d'avance et espacées de façon égale, les tests statistiques sont associés au niveau de signification 0,0002, 0,012, et 0,038, alors que ceux de Pocock sont associés au niveau de signification 0,017 à chaque analyse. Ce sont des valeurs seuils pour le calcul de la significativité des tests. Si le résultat du test statistique à la k^{ieme} analyse donne une valeur de signification p_k inférieure au niveau de significativité α_k, le niveau adopté dans cette analyse, alors on peut arrêter l'essai. On voit bien que dans la situation de O'Brien et Fleming, il devient difficile d'arrêter un essai de façon précoce, car le niveau de α est très faible au début, mais il augmente de façon *convexe* pour mieux protéger contre des faux positifs (déclarer à tort un résultat statistiquement significatif alors qu'en réalité il n'y a pas de différence). La décision d'arrêter ou non l'essai de façon précoce, dépendra également sur d'autres critères que le seul critère principal.

Dans le cas d'événements indésirables graves, il est plus éthique d'utiliser une procédure qui va donner plus de chance d'arrêter un essai de façon précoce si on observe trop d'EIG. Pour ceci il faut construire une procédure qui va permettre de dépenser α plus vite au début de l'essai, car si le traitement engendre des EIG trop souvent il vaut mieux le savoir tôt dans l'essai plutôt qu'à la fin, tout en préservant le niveau α nominal. C'est pour cette raison, qu'il vaut mieux utiliser une fonction croissante en α qui a une forme *concave*. La fonction de la famille γ (Hwang *et al.* 1990) est suffisamment flexible pour cette situation car elle permet de généraliser ces deux situations. Elle se calcule de la manière suivante pour $\gamma \neq 0$:

$$\alpha(\gamma, t_k) = \alpha_0 \frac{1 - \exp^{-\gamma t_k}}{1 - \exp^{-\gamma}}$$

où t_k, $(0 < t_k \leqslant 1)$ correspond à la fraction d'information au moment de la k^{ieme} analyse, c'est-à-dire le nombre de sujets au moment de l'analyse par rapport au nombre total de sujets prévus dans l'étude : $t_k = \frac{n_k}{N_{max}}$, et $\alpha_0 = \alpha(\gamma, 1)$.

Pour $\gamma = 1$, on retrouve les frontières séquentielles de Pocock (1977). Pour $\gamma = -4$, on retrouve les frontières séquentielles de O'Brien et Fleming (1979). Pour $\gamma = 4$, cette fonction est bien adaptée pour l'évaluation des effets indésirable graves, car elle dépense α plus rapidement au début de l'étude, ce qui permet un arrêt précoce en cas d'événements indésirables graves trop fréquents.

Pour préserver un niveau α nominal $= 0,05$ (ou $0,10$), il est nécessaire de recalculer chaque $\alpha(\gamma, t_k)$ qu'on notera aussi $\alpha(t_k)$ en fonction de la fraction d'information pour une valeur fixe de γ. Ces nouvelles valeurs de $\alpha(t_k)$ peuvent ensuite être utilisées pour le calcul des valeurs c_k qui leurs sont associées d'après l'algorithme itératif ci-après.

À la première analyse, on cherche la valeur de c_k qui satisfait la relation suivante :

$$P_0(W(t_1) \geqslant c_1) = \alpha(t_1)$$

où $W(t_j) = \sqrt{t_j} Z(t_j)$ est un processus stochastique, avec $Z(t_j)$ la statistique utilisée pour exprimer les résultats des données qui ont une forme binomiale, normale, ou données de survie (Scharfstein *et al.* 1997), le cas binomial étant le cas qui nous intéresse ici.

À la deuxième analyse, on cherche la valeur de c_2 qui satisfait la relation suivante :

$$\alpha(t_1) + P_0(W(t_1) < c_1, W(t_2) \geqslant c_2) = \alpha(t_2)$$

Pour les autres valeurs on procède de façon itérative par l'algorithme récursif AMR (Armitage *et al.* 1969).

$$\alpha(t_{k-1}) + P_0(W(k_1) < c_1, \ldots, W(t_{k-1}) < c_{k-1}, W(t_k) \geqslant c_k) = \alpha(t_k)$$

Le tableau 2 donne quelques valeurs de $\alpha(t_k)$ pour $\gamma = 1$, 4, et 7, $\alpha = 0,05$ et $0,10$ et 5 étapes, dont les 4 premières sont espacées de façon régulière au début de l'étude. Pour d'autres valeurs de la fraction d'information, il est nécessaire de calculer numériquement les nouvelles valeurs de c_k à partir de la fonction $\alpha(t_k)$. La règle d'arrêt sera alors fonction de $\alpha(t_k)$ et non plus de α. Cette règle permet un arrêt si n_k est inférieur à N_k^*, où N_k^* correspond à la taille maximale de l'échantillon pour laquelle la borne inférieure de l'intervalle de confiance à $100(\Phi(c_k))\%$, qui correspond à $IC_{1-\Phi(c_k)}$, est supérieure à τ. $\Phi(c_k)$ correspond à la fonction de répartition d'une loi normale centrée réduite évaluée à la valeur c_k.

TABLEAU 2
$\alpha(t_k)$ en fonction de la fraction d'information, γ et α

		Fraction d'information				
α	γ	0,05	0,10	0,15	0,20	1,00
0,05	1	0,0038	0,0075	0,0110	0,0143	0,0500
	4	0,0092	0,0168	0,0230	0,0280	0,0500
	7	0,0147	0,0252	0,0325	0,0377	0,0500
0,10	1	0,0077	0,0151	0,0220	0,0287	0,1000
	4	0,0184	0,0336	0,0459	0,0561	0,1000
	7	0,0294	0,0503	0,0650	0,0754	0,1000

3. Exemple

L'exemple présenté ici concerne un essai randomisé comparant deux traitements de chimiothérapie chez des patients présentant une tumeur germinale en rechute (Rosti *et al.* 2002). Dans le bras standard on s'attend à observer un taux de décès toxique de 3 %. Dans le bras d'intensification, on s'attend à observer plus de décès toxiques et on est prêt à accepter un taux $\tau = 5\,\%$ et un α unilatéral égal à $0,10$, pour donner plus de chance à certains patients de guérir de leur maladie, ou au moins prolonger le plus possible la survie sans rechute. On prévoit 5 analyses et on décide de prendre $\gamma = 4$, car on préfère dépenser plus de α au début de l'essai pour ce critère.

La tableau 3 présente les valeurs de c_k nécessaires pour déterminer les valeurs de N_k^* pour les 5 premiers décès toxiques, survenus après l'inclusion de 24, 35, 43 et 52 patients sur l'ensemble des 140 patients. Les valeurs de c_k sont obtenues en supposant que la première analyse est réalisée dès le deuxième décès toxique, car les intervalles de confiance recouvriront les taux de toxicités acceptables pour tout $\tau > 2,5\,\%$, quelque soit le nombre de patients. Les valeurs de N_k en bas du tableau correspondent aux valeurs obtenues pour $\alpha = 0,05$ à chaque analyse (*cf.* tableau 1).

Sur ce modèle, si on observe le deuxième décès toxique parmi les N_2^*=7 premiers patients ou le troisième parmi les N_3^*=14 premiers etc…, on peut conclure que le taux de décès toxique est significativement supérieur à 5%, et recommander un arrêt de l'essai pour un taux de décès toxique inacceptable. Dans cet exemple, le taux de décès toxique a toujours été considéré «acceptable», car $n_k > N_k^*$ pour $k = 2, \ldots 5$.

Une autre façon de présenter les informations et d'aider les décideurs est de construire des valeurs de N_k^* pour différentes valeurs de τ (tableau 4). Cette représentation permet d'obtenir une estimation approximative du taux réel des EIG en situant le dénominateur n_k par rapport à deux valeurs voisines de N_k^*. Par exemple, le deuxième EIG a été observé sur 24 patients. Puisque 24 se situe entre 18 et 36, le taux de toxicité se situe entre 1% et 2%. Le cinquième EIG a été observé sur 52 patients, qui se situe entre 43 et 57, donc le taux de toxicité se situe entre 3% et 4%, etc.

TABLEAU 3
N_k^* en fonction de la fraction d'information, $\gamma = 4$, α unilateral $= 0,10$ et $\tau = 5\,\%$
(N_k est associé à un α unilateral=0,05)

Paramètre	Analyse intermédiare			
k	2	3	4	5
n_k	24	35	43	52
$\hat{\tau} = \frac{k}{n_k}$	0,083	0,086	0,093	0,096
t_k	0,171	0,25	0,307	0,371
$\alpha(t_k)$	0,050	0,064	0,072	0,079
c_k	1,640	1,807	1,880	1,906
$\Phi(c_k)$	0,949	0,965	0,970	0,972
N_k^*	7	14	24	34
N_k	7	16	28	40

TABLEAU 4
Valeurs de N_k^* en fonction de τ pour les données de l'exemple avec $\gamma = 4$ et α unilateral $= 0,10$

EIG	Taux maximal acceptable de toxicité τ									
k/n_k	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
1/2	—	—	—	—	—	—	—	—	—	—
2/24	36	18	12	9	7	6	5	4	4	3
3/35	71	36	24	18	14	12	10	9	8	7
4/43	116	58	39	29	24	20	17	15	13	12
5/52	>140	85	57	43	34	29	25	22	19	18
6/72	>140	115	77	58	47	39	34	30	26	24
7/95	>140	>140	94	70	57	47	41	36	32	29
8/96	>140	>140	110	83	66	56	48	42	38	34
9/115	>140	>140	128	97	78	65	56	49	44	40

4. Conclusions

L'utilisation des méthodes présentées ici pourrait conduire à l'amélioration de l'écriture des protocoles d'essais cliniques de phase II et III, par la définition a priori des règles d'arrêt pour la tolérance au traitement. Actuellement, la partie du protocole concernant les règles d'arrêt pour une toxicité excessive est souvent décrite de façon sommaire sans règle statistique comme pour l'efficacité. Pour les essais de phase II et en l'absence d'autres techniques, des méthodes développées pour le critère d'efficacité ont été transposées au critère de toxicité de façon symétrique (Bryant et Day 1995), mais elles ne sont pas adéquates dans le cas de l'évaluation d'événements comme des EIG qui peuvent être rares ou qui peuvent survenir de façon précoce. Cette technique prévoit un arrêt précoce de l'essai après l'inclusion d'un nombre fixe de patients, alors que le seuil de tolérance a peut-être été dépassé plus tôt.

Dans une situation où on est prêt à accepter un taux de décès toxique de 15 % pour permettre à une partie de la population de patients de bénéficier d'une greffe allogénique, on prévoit un arrêt soit pour une toxicité excessive, si on observe au moins 3 décès toxiques, soit pour une efficacité insuffisante si on observe moins de 5 succès thérapeutiques parmi les 20 premiers patients. Par la méthodologie exposée ici, en première approximation, l'essai pourrait être arrêté après l'observation du $2^{\text{ème}}$, $3^{\text{ème}}$, $4^{\text{ème}}$ et $5^{\text{ème}}$ premiers décès toxiques s'ils surviennent parmi 3, 7, 12 et 17 patients respectivement. Les règles d'arrêt par la méthode de Bryant et Day nous conseillent d'arrêter l'essai uniquement après 3 décès toxiques sur 20 patients, ce qui implicitement se traduira par l'arrêt dès que l'on observe 3 décès toxiques sans attendre l'inclusion de 20 patients.

La méthodologie développée ici pourra permettre une prise de décision plus rapide du promoteur en cas de toxicité excessive au fur et à mesure de leur observation. Le promoteur aura alors des arguments statistiques en plus des arguments médicaux pour décider s'il faut arrêter les inclusions dans l'essai, même de façon temporaire, de modifier les critères d'éligibilité ou de conduire une investigation plus approfondie des causes des EIG survenus dans un centre particulier, par exemple.

Avec l'avancée des nouvelles technologies liées aux recueils et traitement des données via internet, l'arrivée rapide des données validées permettrait l'utilisation des analyses séquentielles proprement dites, surtout lorsqu'il s'agit d'événements graves, tels que des EIG qui se produisent pendant la période de traitement. Néanmoins, certaines données peuvent être en attente de validation malgré la rapidité de transmission. L'application des règles strictes sera difficile si le rythme des inclusions est trop rapide. Les patients qui viennent de commencer le traitement et pour lesquels l'évaluation des EIG est en cours peuvent poser un problème supplémentaire pour la prise de décision. Nous recommandons l'utilisation de ces méthodes sur la population de patients ayant terminé leur traitement.

Cependant, la mise en place de ces méthodes nécessite un choix des paramètres γ et α. Dans notre expérience, une valeur de $\gamma = 4$ et α unilatéral $= 0,10$ semble satisfaire les exigences statistiques avec des contraintes raisonnables, car un arrêt précoce permettrait néanmoins de garder le niveau de l'intervalle de confiance proche de 95 % en dépensant plus de α au début car γ est positif.

Ceci se voit dans le tableau 2, car à la quatrième analyse intermédiaire, 56 % (0,056/0,10) de α aurait été dépensé au moment où 20 % des inclusions ont

été réalisées. Dans cette situation on prend un risque plus grand que d'habitude de rejeter un traitement que l'on considère trop toxique alors qu'en réalité ce taux était «acceptable».

Le choix de la valeur du paramètre τ est également important et doit être fixé au début de l'essai. Sa valeur dépend de l'essai et du critère qui définit les événements considérés comme inacceptables. Il prend une valeur généralement inférieure à 15 %. Dans l'exemple présenté, les estimations de τ au fur et à mesure de l'essai étaient toutes plus proches de 9 % que du niveau acceptable pressenti au départ pour $\tau = 5\,\%$. Ceci est dû au fait que la valeur de $\tau = 5\,\%$ est comparée à la borne inférieure de l'intervalle de confiance p_{inf}, et non à $\hat{\tau}$.

Une autre remarque concerne l'influence de N_{max}, la taille de l'échantillon prévue au départ, sur le calcul des N_k^*. Puisque les valeurs N_k^* dépendent de $IC_{1-\Phi(c_k)}$, qui dépend de c_k, qui dépend de $\alpha(t_k)$, qui dépend de t_k, la fraction d'information, les seuils pour l'arrêt ou non de l'essai seront différents. Dans l'exemple présenté, si la taille de l'échantillon était deux fois plus grande, c'est-à-dire 280 au lieu de 140, les valeurs de N_k^* seront égales à $\{-, 5, 12, 19, 32\}$ au lieu de $\{-, 7, 14, 24, 34\}$ (*cf.* tableau 3). Ces différences pourront conduire à des décisions différentes. Par exemple, si on observe 3 décès toxiques sur 13 patients, dans l'essai où on a prévu 140 patients au total, on pourrait arrêter l'essai, car la frontière pour l'arrêt est sur 14 patients, alors que dans l'essai où on a prévu 280 patients au total, l'essai pourra continuer, car la frontière est de 12 patients.

Ces méthodes peuvent également s'appliquer à tout autre événement rare, tel que la surveillance des rechutes dans une maladie où le taux de guérison est élevé, comme par exemple dans les tumeurs germinales non séminomateuses.

Références

[1] ARMITAGE P., MCPHERSON C.K., ROWE B.C. (1969), Repeated signifiance test on accumulating data, *Journal of the Royal Statistical Society, Series A*,132 :232-244.

[2] BRYANT R., DAY R. (1995), Incorporating toxicity considerations into the design of two-stage phase II trials, *Biometrics*,51 :656-664.

[3] EUDRAVIGILANCE-CLINICAL TRIALS MODULE (JUILLET 2002), Detailed guidance on the European data base of suspected unexpected serious adverse reactions, draft 2.8.

[4] GOLDMAN A.I., HANNAN P.J. (2001), Optimal continuous sequential boundaries for monitoring toxicity in clinical trials : a restricted search algorithm, *Statistics in Medicine*, 20 :1575-1589.

[5] HWANG I.K., SHIH W.J., DECANI J.S. (1990), Group sequential designs using a family of type I error probability spending functions, *Statistics in Medicine*, 9 :1439-1445.

[6] KRAMAR A., POTVIN D., HILL C. (1996), Plans expérimentaux pour l'inclusion de patients dans les essais de phase II, *Revue d'Epidémiologie et Santé Publique*, 44 :364-371.

[7] LEE Y.J., WESLEY R.A. (1981), Statistical contributions to phase II trials in cancer : interpretation, analysis and design, *Seminars in Oncology*, 8 :403-416.

[8] O'BRIEN P.C., FLEMING T.R. (1979), A multiple testing procedure for clinical trials, *Biometrics*, 35 :549-556.

[9] O'0QUIGLEY J., SHEN L.Z. (1996), Continual reassessment method : a likelihood approach, *Biometrics*, 52 :673-684.

[10] POCOCK S.J. (1977), Group sequential methods in the design and analysis of clinical trials, *Biometrika*, 64 :191-199.

[11] ROSTI G., PICO J.-L., WANDT H., KOZA V., SALVIONI R., THEODORE C., LELLI G., SEIGERT W., HORWICH A., MARANGOLO M., SCHMOLL H.J., LINKESCH W., PIZZOCARO G., BOUZY J., KRAMAR A., DROZ J.-P. (2002), High dose chemotherapy in the salvage treatment of patients failing first-line platinum chemotherapy for advanced germ cell tumour; first results of a prospective randomised trial of the European Group for Blood and Marrow Transplantation (EBMT) : IT94 *Proceeding of ASCO*,132 :A716.

[12] SCHARFSTEIN D.O., TSIATIS A.A., ROBINS J.-M. (1997), Semiparametric efficiency and its implication on the design and analysis of group sequential studies, *Journal of the American Statistical Association*, 92 :1342-1350.