Interprétation des résultats dans les essais cliniques et les analyses en sous-groupe

Emmanuel Bachy, Service d'hématologie clinique, hôpital Lyon Sud, Pierre Bénite, France

Tirés à part : E. Bachy
emmanuel.bachy@chu-lyon.fr

Liens d'intérêt : Les auteurs déclarent n'avoir aucun lien d'intérêt en rapport avec cet article.

Interpreting results in clinical trials and subgroup analyzes

Essai thérapeutique, essai clinique, critère de jugement, marqueur de substitution, quantité d'effet, analyse de sous-groupe
Therapeutic trial, clinical trial, endpoint, surrogate marker, effect amount, subgroup analysis

Résumé

Les essais cliniques sont fondamentaux pour déterminer si un nouveau médicament conduit réellement à une amélioration de la prise en charge des patients. Pourtant, beaucoup d'incompréhension demeure au sujet de la conception des essais, du choix des critères de jugement et de l'interprétation des résultats. En prenant pour exemple les essais récents dans le lymphome folliculaire, cette revue évoquera le choix du critère de jugement principal, la définition d'un marqueur de substitution, la notion de quantité d'effet et les analyses de sous-groupes. Ces concepts doivent être appréhendés par les cliniciens pour pouvoir questionner tout résultat issu d'un essai clinique.

Abstract

Clinical trials are crucial to address whether a new drug actually leads to a significant improvement in patient care. However, there is often a misunderstanding of how clinical trials are designed, why specific endpoints are chosen and how results should be interpreted. Based on recent trials in the field of follicular lymphoma, this review will deal with primary endpoint choice, surrogate endpoint definition, effect size and subgroups analysis. These statistical concepts need to be clearly understood by clinicians to challenge any result of a clinical trial.

Le choix du critère de jugement principal dans les essais cliniques est un élément essentiel à plusieurs titres. Il conditionne le nombre de sujets à inclure pour un essai de phase II ou de phase III, il doit avoir un sens médicalement pertinent pour le clinicien et doit être réaliste pour la conduite de l'essai clinique. Même préspécifiées dans les critères de jugement secondaires, les analyses de sous-groupe restent toujours sujettes à caution et ne devraient servir qu'à bâtir des hypothèses pour de nouveaux essais, jamais à poser des recommandations spécifiques de traitement. Nous allons, dans cet article, nous intéresser spécifiquement aux études de phase III conduites récemment dans le lymphome folliculaire (LF) ; néanmoins, les principes dégagés de cette analyse seront, bien entendu, extrapolables à toute autre pathologie.

Critère de jugement principal : un exercice d'équilibriste

Avant de nous concentrer sur le LF et l'interprétation des essais cliniques récents, reprenons quelques notions statistiques élémentaires afin de comprendre les problématiques en jeu dans la conception actuelle de ces essais.
On a coutume de considérer que le choix du critère de jugement principal (*primary endpoint*) est l'une des étapes cruciales dans la conception d'un essai clinique, au moins tout autant que le choix du comparateur de référence dans les

Pour citer cet article : Bachy E. Interprétation des résultats dans les essais cliniques et les analyses en sous-groupe. *Hématologie* 2019 ; 25(supplément 3) : 3-10. doi : 10.1684/hma.2019.1481

doi: 10.1684/hma2019.1481

essais de phase III. Plusieurs impératifs contradictoires s'affrontent dans le choix de tel ou tel critère. Du point de vue de la faisabilité d'un essai clinique (c'est-à-dire de son coût financier mais aussi de la capacité à recruter le nombre de patients nécessaires dans un laps de temps raisonnable), le promoteur est toujours tenté de choisir un critère pour lequel un maximum d'événements vont survenir dans un minimum de temps, la durée d'une étude conditionnant pour une large part son coût. En effet, si l'on considère la formule simplifiée suivante permettant d'estimer le nombre de patients à inclure pour une étude :

$$N = Coef \times S^2/\Delta^2 \times \left(Z_{\alpha/2} - Z_{puissance}\right)$$

Où :

– Coef = coefficient spécifique (fixé)
– S^2 = variance (fixée)
– Δ différence minimale d'intérêt (modulable) : reliée directement au nombre d'évènements attendus et à l'effet espéré du traitement expérimental
– $z_{puissance}$ = -0,842 (si on choisit la puissance 1 - β à 80 %) (modulable dans les limites du raisonnable, rarement < 80 %)
– $z_{\alpha/2}$ = 1,96 (correspondant à α = 5 %) (fixé historiquement)

On comprend aisément que plus le nombre d'événements sera élevé, plus la différence minimale d'intérêt attendue sera grande et plus le nombre de patients à inclure sera faible à risques α et β constants. Pour mémoire, et sans entrer dans des considérations statistiques plus poussées, on rappelle que le risque α correspond au risque de conclure à tort à une différence entre deux traitements alors qu'elle n'existe pas (fixée à 5 % soit 0,05, le fameux « petit p »). Même si ce n'est pas l'objet de cette revue, rappelons au passage que ce *p* a été fixé de façon tout à fait arbitraire [1], sans signification clinique particulière si ce n'est qu'un risque égal à 5 % de mettre sur le marché, lors de chaque essai de phase III, un médicament inefficace continue à être considéré comme acceptable par l'ensemble des acteurs du monde de la santé (médecins, autorités de santé, patients, etc.), sans doute de manière mal comprise par beaucoup. Un article récent sur le sujet montre combien certains essais de phase III restent fragiles malgré un fameux « petit p » significatif [2]. Le risque β, sur lequel nous sommes souvent encore moins informés et dont les bornes sont plus fluctuantes, correspond inversement au risque de conclure à tort à une inutilité du traitement alors que celui-ci est pourtant efficace. D'où l'expression, fréquemment utilisée en statistiques, que l'absence de preuve d'efficacité d'un essai n'est souvent pas la preuve d'une inefficacité de la molécule testée : « *absence of evidence is not evidence of absence* » [3]. Elle avoisine souvent les 20 %, ce qui est considérable, expliquant souvent pourquoi plusieurs essais, posant une question semblable et d'effectifs similaires, peuvent arriver à des conclusions fort diverses. La puissance d'un essai est déterminée par ce risque β (plus exactement 1 - β).
Des contraintes s'exerçant le plus souvent sur le nombre maximal de patients qu'il est possible d'inclure dans une étude et sur la durée du suivi, pour des raisons de faisabilité financière et logistique, et le risque α étant « historiquement » fixé à 5 %, on comprend qu'il reste peu de marge de manœuvre pour la conception des essais cliniques. Sont donc très souvent employés :

– l'utilisation d'hypothèses fantaisistes sur l'inefficacité d'un traitement de référence ou l'efficacité *a priori* spectaculaire du nouveau traitement,
– le choix d'une puissance statistique faible avec un risque majeur de ne montrer aucun effet alors que celui-ci peut exister (travers de la plupart des petites séries concluant à une absence d'effet),

– le choix d'un critère de jugement permettant de recueillir un nombre élevé d'évènements.

Cette problématique est centrale dans le LF, pathologie indolente par excellence, dont la survie médiane dépasse aujourd'hui quinze ans, même pour les lymphomes dits de forte masse tumorale [4]. Il est difficile, dans ces conditions, de retenir la survie globale (SG) comme critère de jugement principal pour ce type de pathologie, car il serait trop long (et donc trop cher) pour n'importe quelle institution académique, mais aussi pour n'importe quel partenaire industriel, d'attendre un nombre de décès « suffisant ». Cette absence d'un nombre suffisant d'évènements constitue une bonne nouvelle pour nos patients, dans la mesure où ces évènements sont des décès, mais quid du *design* actuel des essais cliniques ? Peut-on considérer la survie sans progression (SSP) ou la réponse au traitement comme un marqueur de substitution ? Permettent-elles réellement d'obtenir plus tôt une information indirecte sur la survie globale ? C'est là tout l'enjeu de ce qu'on appelle communément les marqueurs de substitution (*surrogate markers*), qui font l'objet du chapitre suivant.

Les marqueurs de substitution : attention danger

Dans le LF, comme dans la plupart des lymphomes indolents, la SSP est souvent considérée comme le reflet indirect de la SG. Le raisonnement habituel est qu'un traitement allongeant de façon significative la SSP est susceptible de prolonger la SG à long terme. Néanmoins, les données récentes que nous avons rapportées concernant le suivi de l'essai PRIMA (comparant une immunochimiothérapie ± une maintenance de deux ans pour des patients atteints de LF de forte masse tumorale) montrent que la SSP est probablement loin d'être un marqueur de substitution de la SG [4]. Ainsi, alors que la SSP est supérieure à dix ans dans le bras comportant une maintenance par rituximab, et n'est que de quatre ans dans le bras « observation seule », avec un rapport de risque (HR, pour *hazard ratio*) à 0,61 (95%CI : 0,52-0,73), aucune différence n'est observée en termes de SG, qui est de 80 % à dix ans dans les deux bras de traitement. Quelles sont les raisons susceptibles d'expliquer l'absence de bénéfice en SG malgré une telle différence en termes de SSP ? Elles sont probablement multiples avec :

– l'application plus fréquente d'une maintenance par rituximab en rechute chez les patients dans le bras « observation » en première ligne (sorte de *cross-over* spontané),
– une absence de bénéfice de la maintenance chez les patients porteurs d'une maladie agressive, qui vont malheureusement décéder assez rapidement de leur LF quel que soit le traitement utilisé,
– la présence de multiples lignes de traitement efficaces au cours de l'évolution de la maladie, diluant peu à peu l'effet potentiel de la première ligne au cours du temps [5].

Un nouveau critère est récemment venu sur le devant de la scène, dans le lymphome, communément appelé POD24, pour *progression of disease during the first 24 months* [6]. Il reprend l'idée, somme toute assez partagée en clinique, que plus une rechute est rapide après un traitement de première ligne, plus le risque de décès à court terme est important. Nous avions également montré dans le LF qu'il existe une corrélation directe entre délai à la rechute et survie ultérieure : les patients rechutant à moins de six mois du début de l'induction étaient ceux à plus haut risque, quand ceux rechutant entre douze et vingt-quatre mois étaient à risque moindre, mais toujours largement plus susceptibles de décéder dans les années suivantes que ceux qui rechutaient après vingt-quatre mois [7]. Ces résultats viennent d'être confirmés dans l'essai GALLIUM comparant un traitement par rituximab + chimiothérapie suivi d'une maintenance par rituximab à un traitement par

obinutuzumab + chimiothérapie suivi d'une maintenance par obinutuzumab [8]. Peu à peu, le *cut-off* de vingt-quatre mois s'est imposé car il permet d'identifier un nombre substantiel de patients qui restent à haut risque de décès.

Pourtant, si la POD24 est clairement corrélée à la SG, elle en constitue probablement un mauvais marqueur de substitution dans les essais cliniques. Par définition, pour qu'un paramètre serve de marqueur de substitution, il faut non seulement qu'il soit corrélé au critère qu'il tend à remplacer, mais également que l'effet d'un traitement sur ce paramètre soit le même que sur le critère auquel il doit se substituer [5, 9]. Ainsi, dans le cas de l'essai PRIMA, si la POD24 avait été un bon marqueur de substitution de la SG, l'amélioration nette de la POD24 grâce au traitement d'entretien par rituximab (82 *versus* 65 %) [10] et la corrélation forte entre la POD24 et la SG [11] se serait traduite par une amélioration significative de la SG, ce qui n'est pas le cas [4]. Dès lors, il n'est pas possible de prétendre qu'un traitement permettant d'augmenter la POD24 est susceptible d'améliorer la SG à long terme.

Il est important de saisir également que ce qui est valable dans un sous-type histologique ne l'est pas forcément pour un autre sous-type. Un article récent démontre par exemple que la SSP et la POD24 sont des marqueurs de substitution probablement fiables de la SG dans le lymphome B diffus à grandes cellules (LBDGC) [12]. Cela n'est pas forcément étonnant, d'un point de vue conceptuel, une rechute dans le LBDGC n'ayant pas la même signification en termes de gravité qu'une rechute dans le LF.

À l'heure actuelle, dans le LF, l'une des seules démonstrations relativement solides qui existe en matière de marqueur de substitution concerne l'utilisation de la réponse complète à trente mois comme marqueur de la SSP [13].

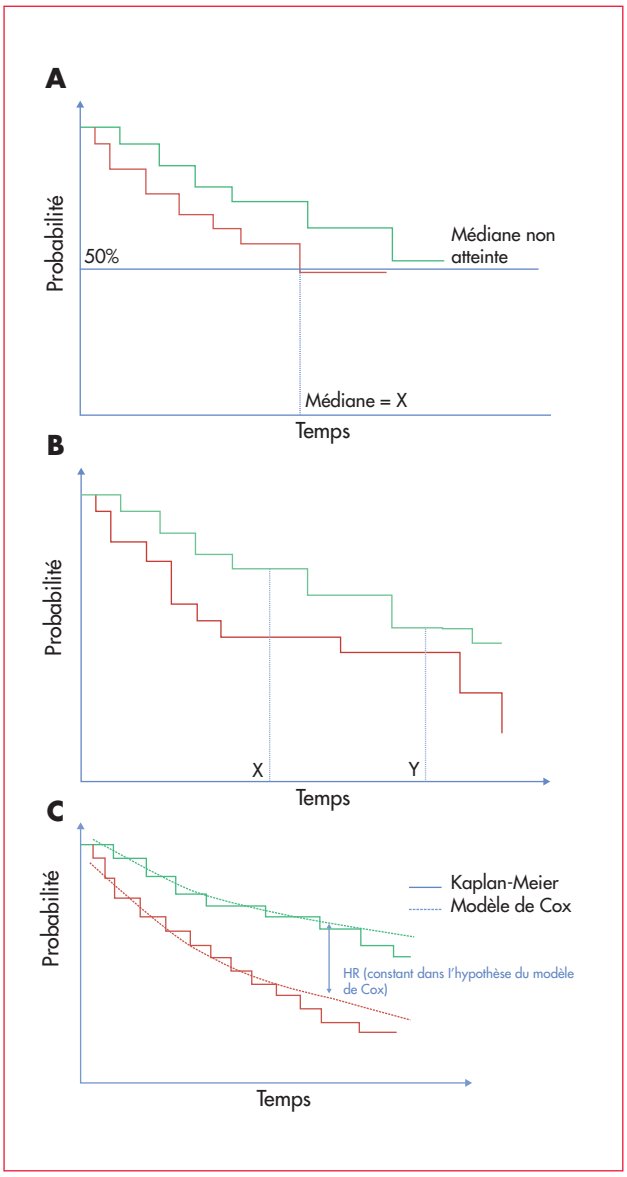
Pour tenter de résumer ce concept, assez complexe, de critère de substitution, il est tout à fait licite de dire qu'un patient rechutant dans les deux ans de son LF est un patient grave et plus à risque de décès dans les années à venir, mais il est erroné de dire qu'un traitement améliorant la POD24 améliore probablement la SG à long terme.

Significativité et importance d'effet : ne pas confondre

L'un des paramètres essentiels dans l'interprétation des essais cliniques est l'évaluation de la quantité d'effet d'un traitement par rapport à une thérapie de référence, que les données concernant cette dernière soient issues de la littérature (comme souvent dans la conception d'un essai de phase II non randomisé) ou observée dans un bras réel de référence (comme dans un essai de phase III randomisé) [14]. Plusieurs données peuvent servir à évaluer cette importance d'effet pour des données de survie : la différence de survie à un temps donné (SSP à deux, à trois ou à cinq ans entre les deux bras par exemple), la médiane de survie (*i.e.* temps avant que 50 % des patients aient présenté l'événement), ou le *hazard ratio* (rapport de risque dit instantané, qui correspond plus exactement au rapport de risque dans un laps de temps infinitésimal, en faisant l'hypothèse que celui-ci est constant au cours du temps entre les deux bras de traitement pour le modèle de Cox). De tous ces paramètres, le *hazard ratio* constitue sans doute l'un des meilleurs, en considérant la survie dans son ensemble et pas seulement à un point de temps donné. La *figure 1* montre de quelle façon l'évaluation des médianes de survie et de la survie à un point de temps donné peuvent induire en erreur sur l'effet réel d'un traitement dans son ensemble. Selon le point de temps considéré, on voit comment les résultats peuvent être présentés de façon tout à fait avantageuse pour un traitement, de même que la mention d'une médiane de survie non atteinte peut refléter des situations tout à fait distinctes. C'est finalement le *hazard ratio* qui reflète sans doute le mieux la différence globale d'effet sur l'ensemble de la durée de l'étude.

En ce qui concerne l'interprétation des résultats d'essais cliniques dans le LF, il est intéressant de noter qu'une quantité d'effet relativement similaire a pu être perçue

Figure 1



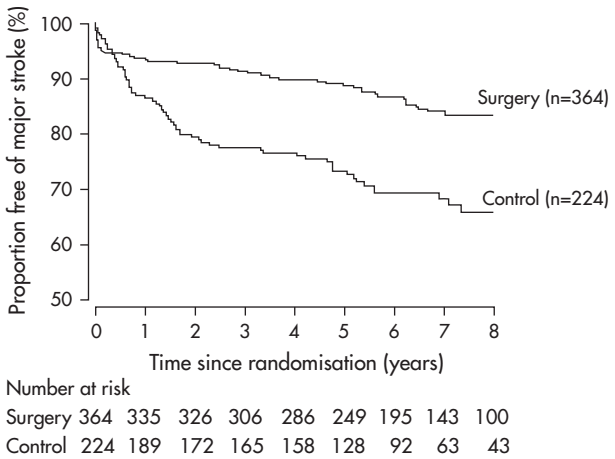
Estimation de la quantité d'effet selon plusieurs métriques permettant d'apprécier les biais éventuels. **A)** La survie médiane : malgré une différence d'effet minime, la médiane de survie est « non atteinte » pour le groupe en vert versus de X années pour le groupe en rouge, témoignant de la précarité de cet indicateur. **B)** La survie à X années : il existe clairement un biais à choisir spécifiquement ce point de temps pour exprimer la différence de survie entre les deux groupes, on voit que choisir le point de temps Y aurait donné une toute autre valeur à la comparaison. **C)** Le hazard ratio : dans le modèle de Cox, qui fait l'hypothèse d'un risque proportionnel, celui-ci permet d'estimer l'effet global du traitement expérimental sur l'ensemble de la courbe (schématisé de façon simplifiée comme une forme de régression de la courbe de Kaplan-Meier, ce qui n'est pas le cas en réalité).

par les hématologues de façon différente entre les essais GALLIUM et PRIMA. Le *hazard ratio* en faveur de la maintenance par rituximab est de 0,61 (95%CI : 0,52-0,73) [4] tandis que celui en faveur du remplacement du rituximab par l'obinutuzumab dans l'essai GALLIUM est de 0,66 (95%CI : 0,51-0,85) [15], soit une réduction du risque de progression sensiblement comparable dans les deux études, d'autant plus difficile à mettre en évidence que le bras de référence

constituait déjà un excellent traitement contrôle dans l'essai GALLIUM. Pourtant, l'adoption de la maintenance dans le LF de façon large a été extrêmement rapide, alors que plusieurs centres questionnent encore l'intérêt de l'obinutuzumab en première ligne thérapeutique. Certains des freins expliquant cette attitude peuvent être attribués aux études de sous-groupes rapportées dans la présentation de l'essai, notamment l'absence de significativité dans le groupe de patients à faible risque (*follicular lymphoma international prognostic index* [FLIPI] 0 ou 1) [15]. Néanmoins, et comme nous allons le voir dans la partie suivante, les analyses de sous-groupes d'un essai de phase III ne devraient jamais guider la décision thérapeutique et servent juste à questionner l'intérêt de conduire une nouvelle étude dans un sous-groupe particulier ou à vérifier l'absence d'interaction entre un sous-groupe et l'un des bras de traitement.

FIGURE 2

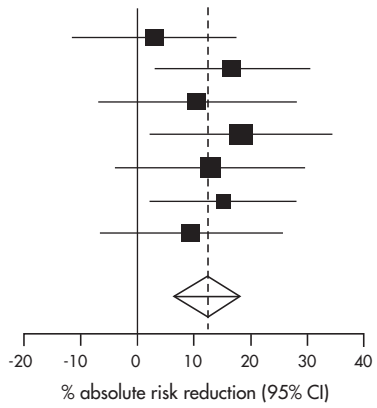
A.



B.

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	7/56	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

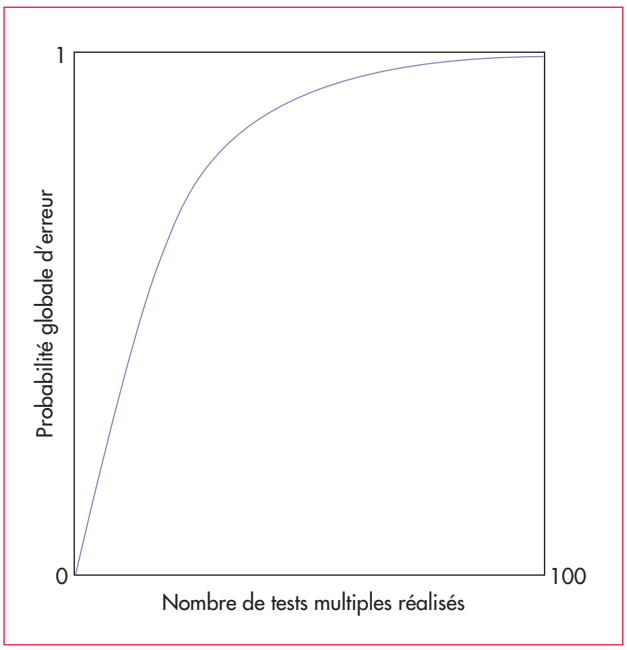


A) Comparaison de l'effet de l'endartériectomie et du traitement conventionnel sur le risque d'accident vasculaire cérébral. **B)** Analyse de l'effet par sous-groupe selon le jour de la semaine où est né le patient.

Les analyses de sous-groupe : rarement une bonne idée

Plusieurs excellentes démonstrations par l'absurde ont été publiées, visant à montrer l'importance de ne pas surestimer l'interprétation des analyses de sous-groupes dans les essais cliniques. S'il n'en existe pas d'exemple évident dans le champ de l'hématologie, une étude publiée en 2005 dans le *Lancet* (citée plus de 770 fois) l'illustre très bien. Cet article reprenait l'analyse d'un essai comparant traitement médical seul et endartériectomie chez les patients atteints de sténose carotidienne, et montrant un intérêt global net en faveur de la prise en charge chirurgicale (*figure 2A*) [16]. L'auteur de l'article du *Lancet* avait judicieusement réalisé un graphique (appelé *forrest plot*) des analyses de sous-groupes en fonction du jour de la semaine où étaient nés les patients (*figure 2B*). De cette analyse, on pourrait malencontreusement conclure que les patients nés un dimanche ne bénéficient pas du traitement chirurgical, tandis qu'un traitement médical pur est clairement inadapté pour ceux nés un lundi. Ce type de conclusion invite à sourire, et, pourtant, les mêmes travers statistiques sous-tendent les conclusions faites dans le LF quand on suppose que les patients présentant un FLIPI faible ne bénéficient sans doute pas du remplacement du rituximab par l'obinutuzumab. Attention, cependant : si un rationnel biologique ou clinique fort sous-tend les résultats d'une analyse de sous-groupe (*e.g.,* avantage potentiel de l'ajout d'ibrutinib chez les patients jeunes de type ABC dans l'essai PHOENIX) ou si plusieurs analyses de sous-groupes issues d'essais différents vont dans le même sens (*e.g.,* avantage de l'ibrutinib en première ligne dans les leucémies lymphoïdes chroniques à parties variables des chaînes lourdes des immunoglobulines non mutées), bien évidemment, le poids de ce type d'analyse se trouve renforcé. L'explication sous-jacente permettant de comprendre pourquoi les analyses de sous-groupe conduisent le plus souvent à des interprétations erronées réside dans la notion de tests statistiques multiples. Si vous considérez lors de chaque test qu'il existe un risque inférieur à 5 % (risque α usuel du « petit p ») d'obtenir un résultat faux, le fait de multiplier le nombre de tests augmente de façon drastique le risque

FIGURE 3



Augmentation du risque de conclure à tort à un effet significatif lors de tests multiples.

de trouver par hasard un résultat significatif par erreur, selon la formule $X = 1 - (1 - \alpha)^N$ où N est le nombre de tests réalisés (*figure 3*) [17]. Pour trente analyses de sous-groupe, on estime à près de 80 % le risque de trouver un résultat significatif par erreur et donc de conclure à tort qu'un traitement pourrait être efficace pour un sous-groupe particulier de patients, alors qu'il ne l'est pas. D'où l'intérêt de remettre ce type d'analyse dans son contexte, c'est-à-dire essentiellement dans un but exploratoire méritant éventuellement confirmation par un autre essai.

Conclusion

Beaucoup d'embûches émaillent le parcours d'un essai clinique, depuis sa conception jusqu'à son interprétation. Nous avons ici essayé de montrer dans le cas du LF :

– comment le choix du critère de jugement principal conditionne dès le départ l'échec potentiel ou le succès d'un essai clinique,
– comment la recherche de critères de substitution est actuellement un enjeu majeur pour les hémopathies lymphoïdes indolentes,
– comment un « petit p » significatif n'est qu'une partie de la réussite d'un essai clinique à mettre en perspective avec la quantité d'effet obtenue,
– comment les analyses de sous-groupes sont à regarder avec circonspection, la seule conclusion valable restant celle portant sur le critère de jugement principal défini *a priori*.

Références

[1] Fisher RA. Statistical methods for research workers. Edinburg : Oliver and Boyd, 1925.

[2] Del Paggi JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicines: a retrospective analysis. *Lancet Oncol* 2019 ; 20 (8) : 1065-9.

[3] Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995 ; 311 (7003) : 485.

[4] Bachy E, Seymour JF, Feugier P, *et al.* Sustained progression-free survival benefit of rituximab maintenance in patients with follicular lymphoma: long-term results of the PRIMA study; J Clin Oncol 2019 ; 37 : 2815-24.

[5] Ciani O, Davis S, Tappenden P, *et al.* Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results and implications for policy makers. *Int J Technol Assess Health Care* 2014 ; 30 (3) : 312-24.

[6] Casulo C, Byrtek M, Dawson KL, *et al.* Early relapse of follicular lymphoma after rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone defines patients at high risk for death: an analysis from the National LymphoCare study. *J Clin Oncol* 2015 ; 33 (23) : 2516-22.

[7] Maurer MJ, Bachy E, Ghesquieres H, *et al.* Early event status informs subsequent outcome in newly diagnosed follicular lymphoma. *Am J Hematol* 2016 ; 91 (11) : 1096-101.

[8] Seymour JF, Marcus R, Davies A, *et al.* Association of early disease progression and very poor survival in the GALLIUM study in follicular lymphoma: benefit of obinutuzumab in reducing the rate of early progression. *Haematologica* 2019 ; 104 (6) : 1202-8.

[9] Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Control Clin Trials* 2002 ; 23 (6) : 607-25.

[10] Salles G, Seymour JF, Offner F, *et al.* Rituximab maintenance for 2 years in patients with high tumour burden follicular lymphoma responding to rituximab plus chemotherapy (PRIMA): a phase 3, randomised controlled trial. *Lancet* 2011 ; 377 (9759) : 42-51.

[11] Bachy E, Maurer MJ, Habermann TM, *et al.* A simplified scoring system in *de novo* follicular lymphoma treated initially with immunochemotherapy. *Blood* 2018 ; 132 (1) : 49-58.

[12] Shi Q, Schmitz N, Ou FS, *et al.* Progression-free survival as a surrogate end point for overall survival in first-line diffuse large B-cell lymphoma: an individual patient-level analysis of multiple randomized trials (SEAL). *J Clin Oncol* 2018 ; 36 (25) : 2593-602.

[13] Shi Q, Flowers CR, Hiddemann W, *et al.* Thirty-month complete response as a surrogate end point in first-line follicular lymphoma therapy: an individual patient-level analysis of multiple randomized trials. *J Clin Oncol* 2017 ; 35 (5) : 552-60.

[14] Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. *J Grad Med Educ* 2012 ; 4 (3) : 279-82.

[15] Marcus R, Davies A, Ando K, *et al.* Obinutuzumab for the first-line treatment of follicular lymphoma. *N Engl J Med* 2017 ; 377 (14) : 1331-44.

[16] Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet* 2005 ; 365 (9455) : 256-65.

[17] Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications and interpretation. *Lancet* 2005 ; 365 (9454) : 176-86.