Analyse

Comprendre la multiplicité des tests

Le sujet de la multiplicité des tests ou des comparaisons multiples occupe une place centrale dans les discussions réglementaires de ces dernières années.

Par **CÉLINE KAUV,**
Directrice des Affaires Règlementaires
et **FRANÇOIS MONTESTRUC,**
Statisticien (eXYSTAT)

La Food and Drug Administration (FDA) et l'Agence Européenne du Médicament (EMA) ont largement relayé le sujet avec la publication :

- en draft en Juin 2017 de la « Guideline on multiplicity issues in clinical trials » (*https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf*)
- en version définitive en Août 2019 de la « Guideline on the investigation of subgroups in confirmatory clinical trials » (*https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf*) pour l'Europe
- de la « Guidance for Industry Multiple Endpoints in Clinical Trials » (*https://www.fda.gov/media/102657/download*) pour les Etats-Unis.

Dans une première partie, nous nous appuierons principalement sur ces trois documents réglementaires afin de discuter de la problématique de la multiplicité et décrire les méthodes les plus utilisées. Puis, dans une seconde partie, nous détaillerons l'application de ces principes pour une exploitation en publicité en proposant des recommandations simples pour éviter des refus et en illustrant par des exemples de motivation de refus de visa de publicité.

PARTIE IA : DÉFINITIONS ET TEXTES RÉGLEMENTAIRES

Pourquoi la multiplicité est-elle une vraie question ?

On comprend aisément que l'on a plus de chances de gagner à la loterie en achetant 10 tickets au lieu d'un seul. De même, plus on lance un dé, plus on a de chances de faire le 6. Dans les essais cliniques, on peut faire ce parallèle en faisant de multiples tests statistiques. La probabilité de trouver un résultat significatif simplement du fait du hasard augmente.

L'objet de la multiplicité des tests est de préserver constante cette probabilité que l'on fasse un ou plusieurs tests. Cette multiplicité peut être à gérer si on a plusieurs critères d'efficacité, des analyses en sous-groupes ou plusieurs analyses intermédiaires par exemple.

Le risque le plus contrôlé par les autorités de santé reste celui de mettre sur le marché un produit non efficace. Ce risque appelé aussi risque de première espèce ou risque alpha est le pourcentage de « chance » lors d'un essai clinique de supériorité, de conclure à une différence alors que dans la réalité cette différence n'existe pas. Pour les statisticiens « fréquentistes », on dira que c'est la probabilité de rejeter H0 (Hypothèse nulle de non différence) à tort. Dans la grande majorité des essais cliniques d'enregistrement, ce risque est fixé à 5% pour un test bilatéral[1] (ce qui revient à 2,5% pour un test unilatéral)[2].

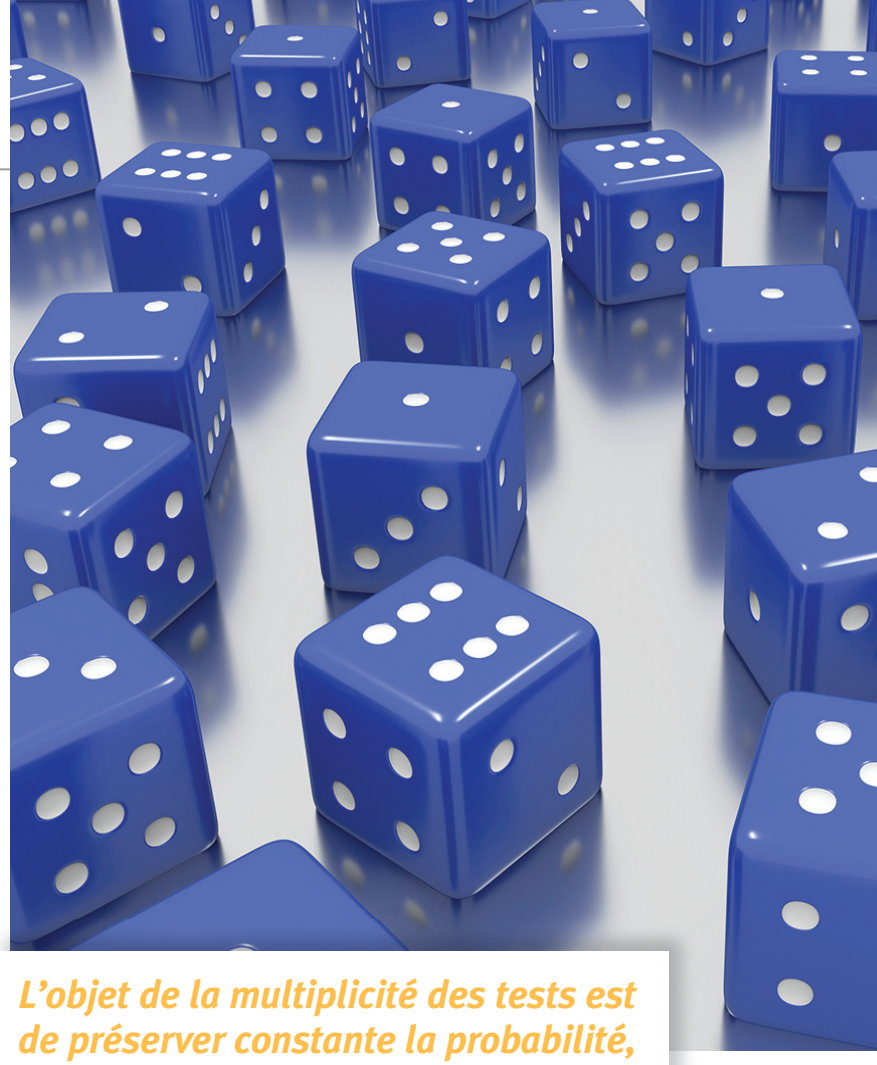
Pour les essais de non-infériorité, de la même manière, le risque de première espèce est de conclure à la non-infériorité à tort.

Si l'essai clinique comporte 2 groupes, un critère principal, une population unique sur laquelle porte la seule comparaison pré-spécifiée, aucune analyse intermédiaire ni d'analyse en sous-groupes, alors aucun problème de multiplicité n'est à l'ordre du jour et le risque fixé à 5% n'a aucunement besoin d'un ajustement puisqu'on ne jette le dé qu'une fois (!). **MAIS** cela n'arrive évidemment que très rarement car on imagine mal la mise en place d'un essai clinique d'enregistrement qui miserait tout sur un seul test. La problématique d'ajustement du risque est donc à prévoir dans la grande majorité de ces études de phase III ; évidemment en amont de l'étude dès la conception du protocole. Le mot ajustement du risque est donc utilisé dans le cadre de la multiplicité car ce risque de 5% doit rester constant que l'on effectue un ou plusieurs tests. Si l'on effectue un test à 5%, il n'est pas utile d'ajuster. Si l'on fait 2 tests à un niveau de 5% sans ajuster le risque, le risque global de conclure au moins une fois à tort passe de 5% à 9.75% ainsi le niveau de 5% n'est pas respecté. Chacun des 2 tests devra être fait à un niveau de risque inférieur à 5% ; c'est ce que l'on appelle l'ajustement.

Il est à noter une autre situation où l'ajustement est inutile à savoir lorsque pour conclure à une efficacité, la positivité de l'essai est déterminée sur **tous** les critères d'analyse au seuil alpha (situation du ET). Dans cette situation, nul besoin d'ajuster le risque alpha car le même niveau d'exigence s'applique à tous les critères qui doivent tous être significatifs. Un seul critère ne concluant pas à l'efficacité rendra l'étude négative quels que soient les résultats des autres critères.

1 Test bilatéral : on envisage que le produit A peut être supérieur au produit B et que le produit B peut être supérieur au produit A. La comparaison pourra se faire dans les deux sens.

*2 Test unilatéral : on envisage **uniquement** que le produit A peut être supérieur au produit. La comparaison devra se faire dans un seul sens.*



On comprendra que dans cette situation c'est la puissance qui est affectée (c'est-à-dire la capacité de l'étude à conclure avec raison à l'efficacité du produit) et que le nombre de patients doit ainsi être augmenté.

En revanche, lorsque l'on veut conclure à une étude positive lorsqu'au moins un des critères est significatif (Situation du OU), la nécessité de l'ajustement s'impose.

Dans la première situation du ET, on lance plusieurs dés et l'étude est positive si l'on n'a que des 6. Dans la deuxième situation du OU, l'étude est positive si on a au moins un 6.

Cette deuxième situation est une situation de multiplicité et nécessite d'aborder la gestion du risque alpha (appelé aussi family-wise error rate ou FWER). La situation du OU se retrouve dans les situations suivantes :

- Analyses de sensibilité sur le critère principal ou plusieurs critères secondaires
- Analyses en sous-groupes ou dans différentes populations ou avec plusieurs doses
- Critères composites
- Analyses intermédiaires (Futilité, efficacité précoce, …)
- *Designs* adaptatifs *(stepwise designed studies, Enrichment designs,* …)

Nous nous concentrerons pour la suite sur les deux premiers points. Les autres situations seront abordées dans un prochain article.

Un exemple souvent cité dans la littérature afin d'expliquer la nécessité d'ajuster le risque est le suivant : Si pour chaque critère/comparaison pris individuellement, le risque est de 5% et que les critères sont indépendants, le risque monte à 1-(0.95)² pour 2 critères/comparaisons, 1-(0.95)3 soit 14% pour 3 critères/comparaisons et à 1-(0.95)10 soit 40% pour 10 critères/comparaisons.

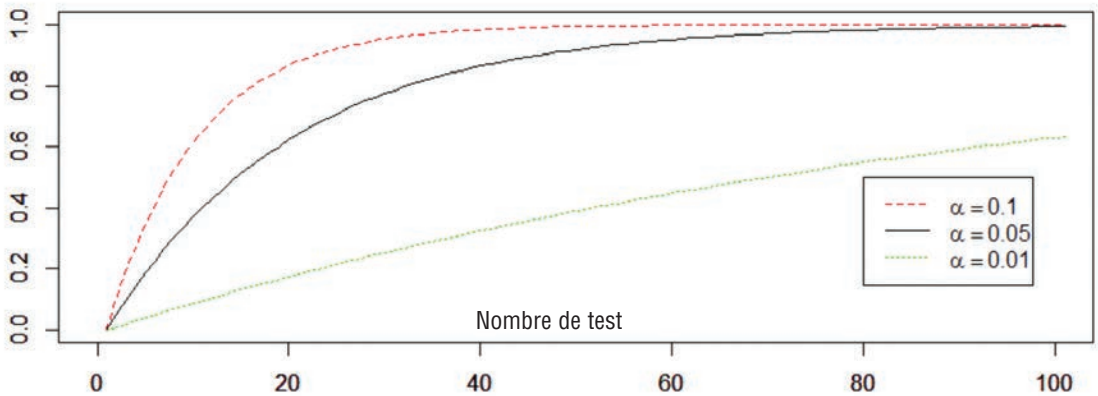
Le graphique ci-dessous résume la situation :

L'objet de la multiplicité des tests est de préserver constante la probabilité, que l'on fasse un ou plusieurs tests.

On comprend donc aisément la nécessité de contrôler ce risque. De manière imagée, si l'on me dit que la probabilité d'avoir un accident mortel en traversant la rue est de 5%, cette probabilité augmente à 40% si je traverse 10 fois la rue. Il sera donc légitime d'ajuster le risque initial de sorte que le risque global sur toute une vie soit de 5% par exemple.

Les analyses en sous-groupes ou sur plus d'un critère ou sur plusieurs populations ou plusieurs doses utilisent globalement les mêmes méthodes d'ajustement du risque alpha qui peuvent se diviser en 2 familles : les méthodes à une étape ou à plusieurs étapes.

Probabilité de commettre au moins une erreur de type 1



PARTIE IB : LES PRINCIPALES MÉTHODES D'AJUSTEMENT (OU DE CONTRÔLE) DU RISQUE ALPHA

Nous décrirons ici les 3 méthodes les plus utilisées pour la problématique de la gestion de la multiplicité des tests. Il convient cependant de dire que de nombreuses méthodes existent et que la réflexion doit être conduite en amont lors de la rédaction du protocole de l'étude et du plan d'analyse statistique. Le bon choix peut être déterminant lors de l'analyse finale et « l'excès de confiance » peut parfois se révéler fatal !

1. La méthode la plus simple mais aussi la moins puissante (c'est-à-dire celle risquant de ne pas mettre en évidence une différence qui existe) est la **Méthode d'ajustement de Bonferroni.** Cette méthode est une procédure à une étape, qui compare les m p-valeurs à un seuil commun α/m (revient à diviser la p-valeur par le nombre de tests m). On rejettera l'hypothèse H0 si $p < \alpha/m$. Par exemple, si 5 tests sont réalisés avec les p-valeurs associées suivantes {0.001,0.03,0.0004,0.048, 0.0007} on ne pourra rejeter les tests N°2 et N°4 au seuil global de 5% car les valeurs 0.03 et 0.05 sont supérieures au seuil de 0.01 = 0.05/5.

2. La méthode sans doute la plus utilisée est la méthode d'ajustement séquentielle (« closed tests » ou « hierarchical sequential testing » en anglais). Cette méthode à plusieurs étapes hiérarchise les critères ou les comparaisons dans un ordre prédéfini. Lors de l'analyse de l'essai, les critères sont considérés les uns après les autres, séquentiellement, en suivant la hiérarchie pré-établie. On pourra conclure sur toutes les comparaisons pour tous les premiers critères statistiquement significatifs au seuil habituel (5%) jusqu'au premier non significatif (>5%). Aucun ajustement du seuil de signification n'est effectué. L'intérêt de cette approche est de pouvoir éventuellement (en fonction des résultats obtenus) conclure à la démonstration de l'effet simultanément sur plusieurs critères à partir d'un seul essai et cela sans inflation du risque alpha. De ce fait, cette méthode est très utilisée d'autant qu'elle est référencée depuis 2002 dans le guideline européen sur la multiplicité. Le choix de la « séquence » est évidemment primordial car si l'on veut par exemple effectuer 5 comparaisons et que la 2ème n'est pas significative, les comparaisons 3 à 5 ne pourront être utilisées même si elles sont statistiquement significatives. Par exemple, si 5 tests sont réalisés avec les p-valeurs associées suivantes {0.001,0.06,0.0004,0.048,0.0007}, le test N°1 est significatif mais les tests N°2 à N°5 ne peuvent être considérés comme l'étant. On voit dans cet exemple que la Méthode d'ajustement de Bonferroni aurait conclu positivement (rejet de H0) pour les tests 1, 3 et 5, illustrant ainsi l'importance du choix de la méthode d'ajustement. Dans cet exemple, les résultats des tests N°2 à N°5 ne seront sans doute pas considérés dans l'évaluation de l'Autorisation de Mise sur le Marché (AMM) du médicament ni dans l'avis de transparence et par conséquent ne pourront être utilisés en publicité.

3. La troisième méthode d'ajustement est une synthèse des deux précédentes et est sans doute la méthode à recommander car la plus puissante tout en étant la plus complexe. Cette méthode à plusieurs étapes appelée « **Fallback procedure** » a été développée en 2007 par Dmitrienko (*doi.org/10.1080/10543400500265660*). Elle permet de tester de manière séquentielle comme précédemment mais de ne pas s'arrêter si un test est négatif. On est alors autorisé à aller à des niveaux inférieurs mais avec un seuil de significativité beaucoup plus bas et défini à l'avance. Le α « complet » de 5% est divisé par le nombre de tests et pondéré en fonction de la pertinence clinique de chaque critère, chaque séquence aura un risque α différent. La guidance FDA décrit de manière précise cette procédure. Par exemple, si 5 tests sont réalisés avec les p-valeurs associées suivantes {0.001, 0.06,0.0004,0.048,0.0007}, les tests N°1, N°3 et N°5 seront significatifs malgré le fait que les tests N°2 et N°4 ne le soient pas.

PARTIE II : IMPACT EN PUBLICITÉ

L'exploitation en publicité de ces analyses en sous-groupes nécessite de nombreuses précautions. L'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM) régit le contrôle de la publicité à destination des professionnels de santé et du grand public. Elle édicte de nombreuses recommandations permettant aux industriels de clarifier et de préciser les obligations réglementaires en matière de publicité, notamment celles du Code de la Santé Publique.

Il est intéressant de constater que l'ANSM définit les sources de données utilisables et non exploitables en publicité, à travers la recommandation du même nom **« Sources des données »**, et le type d'études utilisables au sens méthodologiquement acceptable, avec quelques rappels de base tels que : *« Les études de type explicatif sont les plus aptes à sous-tendre une efficacité ou une sécurité d'emploi, notamment en publicité comparative. Elles doivent être prospectives, contrôlées, randomisées et, si possible (et en fonction des cas) conduites en l'aveugle, avec des effectifs justifiés permettant d'avoir une puissance suffisante ».*
Il est évident que ces études doivent être conformes au libellé de l'AMM ainsi qu'aux conclusions de la Commission de la Transparence, avant toute analyse méthodologique.

La présentation des résultats de ces études est détaillée au travers de nombreuses autres recommandations. La recommandation ***« Publicité comparative »*** aborde les critères de comparaison avec la notion statistique de significativité intégrée : *« La comparaison doit être la plus exhaustive possible sans privilégier exclusivement les éléments favorables. Afin d'être objective, la comparaison doit porter sur des caractéristiques essentielles, significatives, pertinentes et vérifiables »*

Pour les données d'efficacité, nous disposons :
- **de la recommandation « Critères principaux et secondaires d'une étude clinique »** qui explique que la présentation du(es) critère(s) principal(aux) doit être privilégiée, sans en expliquer le fondement statistique ni apporter de précisions quant à la possibilité ou plutôt les restrictions de présentation des critères secondaires ;
- **de la recommandation « Ciblage d'une sous-population »** qui précise que la communication sur une sous-population *« n'est acceptable que si elle est fondée sur des résultats d'études cliniques* ***méthodologiquement correctes*** *permettant de l'affirmer »*

Là encore, une méthodologie correcte est évoquée et semble indispensable, mais aucun détail ni explication quant à sa vérification pour l'utilisation des résultats n'est précisé.

Aussi, quels sont les écueils à éviter lors de la présentation en publicité de sous-populations ou de sous-groupes d'une étude clinique ? La présentation des critères secondaires est-elle toujours possible si la méthodologie de l'étude est correcte ? Quels sont les motifs de refus les plus fréquents, notifiés par l'ANSM sur ces motifs méthodologiques ?

PARTIE IIA : LES SOUS-GROUPES DÉFINIS *A POSTERIORI*

Nous avons vu précédemment les définitions et les risques associés à l'utilisation de ces sous-groupes, il nous faut désormais en **préciser concrètement l'exploitation en publicité.**

Très souvent, dans une publication, l'analyse sur le critère principal est faite sur la population globale, sans se concentrer sur une caractéristique particulière de la population inclue. La puissance de l'étude a été définie sur un nombre de patients déterminé initialement, pour mettre en évidence une différence clinique pertinente définie quantitativement et qualitativement. L'étude est conclue positive si le résultat sur ce critère clinique (critère principal) est significatif sur la totalité des patients inclus. Il arrive cependant, que ce résultat puisse être quantitativement amélioré si une partie de la population est ciblée, sélectionnée. Par exemple, prendre uniquement la sous-population des sujets âgés, des patients en insuffisance rénale, ou encore des patients présentant tel symptôme ou telle mutation à l'inclusion. Or, si cette sous-population n'a pas été définie initialement dans le protocole, c'est-à-dire que l'impact de la diminution du nombre de patients sur la puissance de l'étude n'a pas été pris en compte, alors la conclusion sur cette sous-population définie *a posteriori* sera entachée d'un risque alpha supérieur de conclure à tort, invalidant ainsi la robustesse méthodologique des résultats. Et ce même si le critère principal est significatif : l'exploitation de sous-groupes pour identifier les « facteurs » de succès ayant permis la significativité du critère principal ne saurait être méthodologiquement acceptable si ces critères n'ont pas été identifiés lors de la rédaction du protocole de l'étude (définis *a priori*) et l'ajustement du risque alpha pour leurs analyses prévu au protocole.

Les motifs de refus de l'ANSM les plus souvent retrouvés sur ces analyses en sous-groupe définis *a posteriori* (ou en *post hoc*) sont ainsi rédigés :

- Les résultats relatifs au sous-groupe de patients X sont issus d'une analyse en sous-groupe de patients constitué *a posteriori,* ce qui n'est pas méthodologiquement pertinent et qui confère aux résultats un caractère exploratoire. La présentation des résultats du sous-groupe n'est donc pas objective ;
- La publication ne prend pas en compte la multiplicité des analyses sur les différents critères par une méthode d'ajustement du risque α, permettant de réduire le risque de conclure à tort. Les résultats référencés par les publications sont issus d'analyses *post-hoc* non prévues initialement au protocole et revêtent un caractère purement exploratoire, rendant ainsi la présentation non objective ;
- La présentation des résultats du sous-groupe de patients âgés de plus de 75 ans d'une étude clinique, en l'absence de stratification du tirage au sort, a été jugée comme étant purement exploratoire.

Ainsi, à travers ces exemples, nous comprenons un peu mieux ce que veut dire l'ANSM quand elle écrit que la communication sur une sous-population *« n'est acceptable que si elle est fondée sur des résultats d'études cliniques* ***méthodologiquement correctes*** *permettant de l'affirmer ».*

PARTIE IIB : L'ANALYSE SÉQUENTIELLE

Comme défini précédemment avec la situation du ET, l'analyse séquentielle utilise une analyse statistique hiérarchisée des critères secondaires de l'étude afin de réduire le risque de conclure à tort; c'est-à-dire que l'analyse est effectuée jusqu'au 1er critère secondaire non significatif. Aussi, le(s) critère(s) secondaire(s) ayant atteint le seuil de significativité, au risque alpha initial identique à celui du critère principal, peuvent être présentés, mais le premier critère secondaire non significatif entrainera sa non-présentation ainsi que celle des critères secondaires suivants, en l'absence d'analyse statistique avec un contrôle du risque alpha pour que leur analyse ne revête pas un caractère exploratoire. Car en l'état, aucune significativité ne peut être établie sur ces critères secondaires, donc leur présentation en publicité sera jugée non objective car méthodologiquement incorrecte.

Les motifs de refus de l'ANSM sur cette analyse séquentielle de critères hiérarchisés sont par exemple :

- Un document présente les résultats sur le critère principal significatifs, sur les critères secondaires clés, et sur d'autres critères secondaires. Or, un résultat non significatif est obtenu sur les deux premiers critères secondaires clés. Aussi, compte tenu de l'analyse hiérarchisée et de l'absence de contrôle du risque alpha, les autres critères secondaires revêtent un caractère exploratoire qui rend leur présentation non objective.
- Un document présente les résultats des deux critères secondaires d'une étude, en plus de celle prioritaire du critère principal significatif. La méthodologie de l'étude prévoit une analyse des critères secondaires selon un ordre hiérarchique. Or, aucune différence significative n'a été mise en évidence sur le premier critère secondaire donc aucune significativité ne peut être établie sur le deuxième critère secondaire. Aussi, la présentation des deux critères secondaires n'est pas objective.

Ainsi ces exemples de motifs de refus viennent illustrer les notions abordées dans les recommandations ANSM, et préciser les termes « significativité » ou encore « méthodologiquement correcte » utilisés dans les recommandations précitées.

CONCLUSION

Le sujet de la multiplicité des tests ou des comparaisons multiples concerne de multiples situations dans des essais cliniques d'enregistrement : analyses intermédiaires, plusieurs critères d'analyses, analyses en sous-groupes ou à des temps répétés. **Les méthodes choisies pour garantir aux analyses un risque constant malgré la répétition des tests doivent être clairement définis dans le protocole et le plan d'analyse statistique de l'étude.** Ces choix s'avèrent cruciaux au moment de l'analyse en vue des étapes réglementaires allant de la demande d'AMM à l'exploitation en publicité.