Université de Bordeaux Segalen

Thèse n° 2066

THÈSE

pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX SEGALEN
Mention : Sociétés, Politique, Santé Publique
Spécialité : Santé Publique - Option : Biostatistique
Arrêté ministériel : 7 août 2006

présentée par
Jérémie RIOU
Né le 18 Décembre 1985 à Brest

Thèse dirigée par **Benoit Liquet**,
co-encadrée par **Pierre Lafaye de Micheaux** et **Sébastien Marque**.

École Doctorale Sociétés, Politique et Santé Publique (ED SP2)

Multiplicité des tests, et calculs de taille d'échantillon en recherche clinique

Thèse soutenue le Mercredi 11 Décembre 2013,
devant le jury composé de :

Monsieur, Jérôme, SARACCO
Professeur des Universités, Bordeaux, Président

Monsieur, Frank, BRETZ
Professeur des Universités, Hanovre, Rapporteur

Monsieur, Philippe, BROËT
Professeur des Universités - Practicien Hospitalier, Villejuif, Rapporteur

Monsieur, Philippe, SAINT PIERRE
Maître de Conférence, Paris, Examinateur

Monsieur, Rodolphe, THIÉBAUT
Professeur des Universités - Practicien Hospitalier, Bordeaux, Examinateur

UNIVERSITÉ
BORDEAUX
SEGALEN

Thèse préparée au sein de

l'équipe Biostatistique
du centre de recherche INSERM U897
Institut de Santé Publique d'Epidémiologie et de Développement
146 rue Léo Saignat
33076 Bordeaux CEDEX

et de

l'équipe Biométrie
du Département Life Science
Danone Research
RD128, Avenue de la Vauve
91120 Palaiseau CEDEX

Remerciements

A mes directeurs de thèse

En premier lieu, à Benoit.

Merci pour m'avoir encadré pendant cette thèse. Tu as su me faire confiance pour mener à bien ce projet, me laissant faire preuve d'initiatives. J'ai profité, à travers nos nombreux échanges de toutes tes compétences en méthodologie et en statistiques. Cela a été un plaisir de travailler avec toi. Aujourd'hui une grande partie de mes connaissances, je te les dois. Encore une fois, merci.

A Sébastien.

Tu as toujours tout fait pour que cette thèse se passe au mieux. J'ai beaucoup appris de toi, tant en recherche clinique, qu'en management, et en communication. Je te remercie aussi pour tes précieux conseils.

A Pierre.

Tu as su me faire profiter de ta rigueur et de tes compétences en mathématiques et en programmation. Nos échanges m'ont toujours beaucoup apporté. Merci.

Aux membres du jury

A Jérôme Saracco.

Je vous remercie de me faire l'honneur de présider ce jury.

To Frank Bretz.

Thanks for accepting to judge my thesis, it is really a great honor for me. Thanks also for your advice and your help at the begining and at the end of this thesis. Your work and your expertise on multiple testing will allow me to have a new view of my work. Thanks a lot for the time you spent for reading the manuscript and for coming to Bordeaux.

A Philippe Broët.

Je suis très honoré que vous ayez accepté de juger ce travail. Votre grande connaissance des analyses de données génomiques, ainsi que de la gestion de la multiplicité inhérente à ce domaine, me permettront sans aucun doute d'avoir un regard nouveau sur mon travail. Je vous remercie sincèrement d'avoir accepté d'être le rapporteur de ma thèse

A Philippe Saint Pierre.

Merci d'avoir accepté de faire partie de mon jury. J'en suis très honoré.

A Rodolphe Thiébaut.

Vous me faites un grand honneur en participant à ce jury. C'est grâce à vous que je me suis lancé dans cette aventure. Votre enthousiasme ainsi que votre confiance m'ont convaincu il y a de cela quatre ans. Je vous en remercie car cette expérience a été très riche en apprentissage.

A mes enseignants

A Monsieur Denoux et Monsieur Penhoat, qui m'ont donné le goût pour la biologie et les mathématiques.

A Messieurs Petiot, Le Nouvel, Cosson, et Kermorvant ainsi que Madame Moisi qui ont été les premiers à me transmettre leurs goûts pour la statistique.

A Marthe-Aline, Valérie, Pierre, Alioum, et Karen pour votre bonne humeur, votre sens de la pédagogie et votre soutien tout au long de ces deux années de Master à l'ISPED.

Aux membres de l'ISPED

A Daniel Commenges. Je vous suis reconnaissant de m'avoir permis de réaliser ma thèse au sein de l'équipe Biostatistique. Merci aussi pour votre disponibilité. Vos conseils, votre expérience et vos connaissances m'ont toujours été très utiles pour avancer dans cette thèse.

A Polo et BoBo pour ces discussions statistiques, politiques, et philosophiques du midi. Je tiens à vous remercier pour votre soutien, et ces bons moments passés à vos côtés. Je me souviendrai en particulier de ce week-end eaux-vives qui m'a fait le plus grand bien à un moment charnière de la thèse.

A Polo, BoBo, Mathieu, Robin, Ricardo, Henri et Loïc pour ces sorties sportives qui nous ont permis de nous détendre et de nous aérer l'espace de quelques heures.

A Mélanie, Fanny, Hind, Julie et Audrey pour votre bonne humeur.

A la super équipe de stagaires de début de thèse (Pierre, petite Céline, Mélanie, Polo, Nouria, Yassin et Erwan) pour ces soirées, et toutes ces sorties sportives.

A mes collègues de bureau Mbéry, Célia, Ahmadou, Emilie pour ces trois années en salle student.

A Françoise, qui est toujours présente quand on a besoin d'elle. Mais aussi et surtout pour sa bonne humeur.

Enfin à tous les membres de l'équipe Biostat' ainsi qu'aux personnes que j'ai eu le plaisir de rencontrer pendant ces années à l'ISPED.

A la plateforme clinique de Danone Research

A Pascale, pour tes précieux conseils en gestion de projets.

A Anissa, pour tous ces échanges qui ont été si enrichissants.

A Jérôme, pour nos longues discussions cliniques et statistiques autour d'un café. Elles m'ont beaucoup aidé à prendre du recul sur mon travail.

A Céline, pour ton aide dans toute la gestion logistique de ma thèse.

A Marie, Pascaline, Romain, Rémi, Aurélien, Claire, Salomé, Laurent, Maïna pour votre bonne humeur. Et ces bons moments passés lors des évènements d'équipe.

Et de manière plus globale, à toute la plateforme clinique avec qui j'ai passé de très bonnes années.

A mes amis

A Delf', Maryline et Séverine qui m'ont permis de faire le bon choix au moment où je doutais.

A Guigui, qui malgré la distance a toujours été là. Merci pour ton amitié Bonhomme.

A Guigui, Violaine et JV pour ces soirées Parisiennes quand je montais sur Palaiseau.

A la bande de loulous : Guigui, Ben, Bidou, JV, Serge, Doud, Ju' ,Cha et Pierre pour votre soutien et votre amitié si fidèle. Merci pour ces crémaillères, soirées, vacances et week-ends qui m'ont tant apportés depuis de nombreuses années.

A ma famille et mes proches

Un merci tout spécial à mon papi, et ma mamie pour qui les études étaient si importantes, et qui m'ont tant apporté.
A mes parents qui m'ont toujours épaulé et soutenu dans mes choix. Merci pour tout le temps que vous avez passé à m'aider, et tous ces petits moments de bonheur partagés.
A ma marraine, qui a toujours été de bon conseils.
C'est grâce à vous cinq que je suis là aujourd'hui. Je vous dédie donc cette thèse.

A ma petite soeur Anaïs et mon petit frère Gaël pour leurs soutiens inconditionnels. Je vous remercie, ainsi que Sébastien et Cécile, de m'avoir permis de m'évader l'espace de vacances, de sorties ou de soirées.
Au reste de ma famille.
A la famille de Sandrine pour ces petits week-ends à Bergerac qui me faisaient le plus grand bien.

Et bien sûr le plus fort de mes remerciements va à Sandrine pour sa patience, sa bonne humeur et son réconfort quotidien. Tout au long de ces années, tu as su trouver les mots pour me réconforter et me remotiver. Avec tout mon amour, Merci.

Table des matières

Table des figures

Liste des tableaux

Préambule

L'ordre des auteurs dans le processus de publication est fortement dépendant de la discipline à laquelle nous appartenons. Si nous regardons d'un peu plus près les pratiques des disciplines connexes à la biostatistique, nous pouvons observer qu'en biologie ou en médecine, le nombre d'auteurs est souvent important et leur implication est très hétérogène. Cette pratique est assez vivement contestée dans le rapport de l'Académie des sciences dont les auteurs regrettent souvent une confusion entre auteur et collaborateur [Bach et al., 2011]. Ceci implique une surcote dans les indices de citations pour ces derniers. Les auteurs de ce rapport conseillent donc de se référer aux normes de Vancouver pour une définition précise de la notion de ≪*authorship*≫ dans le contexte biomédical [Editorship, 2010]. Dans ces disciplines, la pratique veut que le premier signataire soit la personne qui ait rédigé l'article et le dernier celle qui ait encadré le projet.

Toutefois, en mathématiques fondamentales ou mathématiques appliquées ce problème ne se pose pas car le nombre d'auteurs pour un papier est plus restreint et la contribution de chacun est donc plus homogène. Le choix concernant l'ordre des auteurs dans ces disciplines est alphabétique [Bach et al., 2011], car il est souvent difficile de différencier la contribution de chacun.

La biostatistique se situe à l'intersection entre ces deux champs disciplinaires. Un choix concernant la conduite à tenir vis à vis de l'ordre des auteurs est donc naturellement apparu. Pour cela, nous sommes partis du constat que le nombre d'auteurs était restreint sur l'ensemble des articles, et que la contribution de chacun était homogène. Nous avons donc décidé d'un ordre alphabétique des auteurs pour l'ensemble des travaux rédigés durant cette thèse.

"entia non sunt multiplicanda praeter necessitatem" [1]
Principe du Rasoir d'Ockham - Guillaume d'Ockham (1285-1349)

1. "Les multiples ne doivent pas être utilisés sans nécessité"

1 Introduction

1.1 Contexte sociétal et industriel

Si, à travers le monde, l'alimentation renvoie généralement à un sentiment de plaisir et de convivialité, elle est aussi intrinsèquement liée à la santé. Jean-Pierre Poulain et le Docteur Seignolet ont démontré, à travers leurs travaux, que l'alimentation peut être un facteur protecteur ou délétère pour de nombreuses pathologies (cancers, maladies cardiovasculaires, obésité, ostéoporose, diabète de type 2 ...) [Poulain, 2011, Seignalet, 2004]. Depuis quelques décennies, cette ambivalence autour de l'alimentation peut être illustrée d'une part par les problèmes sanitaires liés à la malnutrition infantile dans les pays à ressources limitées, mais également par les importants problèmes de Santé Publique liés à l'obésité dans les pays de l'OCDE.

L'amélioration de l'état nutritionnel de la population constitue donc, en ce début de XXI$^{\text{ième}}$ siècle, un enjeu majeur pour les politiques de Santé Publique menées en France, en Europe, et dans le Monde. Une nutrition satisfaisante et de qualité est donc un facteur de protection de la santé. C'est dans cette optique que l'état Français a lancé en 2001 le Plan National Nutrition Santé (PNNS) visant à améliorer l'état de santé de la population en agissant sur la nutrition. Pour les rapporteurs du PNNS, la nutrition s'entend comme ≪l'équilibre entre les apports liés à l'alimentation et les dépenses occasionnées par l'activité physique≫. A travers des campagnes publicitaires, il conseille par exemple une alimentation moins grasse, moins salée, moins sucrée, et plus riche en fruits et légumes.

Le principe énoncé par Hippocrate ≪Des aliments tu feras ta médecine≫ n'a donc jamais été aussi concret et pragmatique. Et c'est en se basant sur les mêmes conclusions qu'est né le groupe agroa-

limentaire Danone au sein duquel j'effectue ma thèse CIFRE. L'objectif du groupe est ≪d'offrir des produits savoureux et équilibrés apportant un bénéfice santé à un large spectre de consommateurs et adaptés spécifiquement aux problématiques de Santé Publique de chaque pays≫. Pour cela, l'entreprise s'est spécialisée dans la commercialisation des produits laitiers frais, des eaux minérales embouteillées, des produits de nutrition médicale, ainsi que des produits relevant de la nutrition infantile. Ces derniers sont d'un genre nouveau relevant à la fois du secteur de la nutrition, mais aussi de la Santé. C'est majoritairement sur cette gamme de produits que se concentrent les scientifiques des centres de Recherche et Développement.

1.2 La recherche clinique

Avec l'avènement, ces dernières décennies, des problématiques associées à la recherche d'effets santé par la nutrition, nous avons vu apparaître au sein des entreprises agroalimentaires les essais cliniques nutritionnels. Afin de comprendre la méthodologie propre à ces essais, le choix a été fait de présenter, dans un premier temps, le cadre général des essais cliniques avant de se focaliser sur les essais cliniques en nutrition.

1.2.1 Historique

De nos jours, la mise en place d'un essai clinique paraît évidente lorsqu'on veut démontrer rigoureusement l'efficacité d'une intervention d'ordre sanitaire. Cette rigueur scientifique autour des essais cliniques s'est construite avec le temps, et il est intéressant de connaître les faits importants qui ont permis la mise en place d'une telle démarche.
L'objectif de cette section va donc être de résumer l'historique des essais cliniques en se basant sur l'ouvrage rédigé par Jean-Philippe Chippaux en 2004 [Chippaux, 2004].
Comme nous le notions précédemment, la démarche scientifique qui entoure les essais cliniques s'est construite progressivement en faisant appel à de nombreuses disciplines : sémiologie clinique, pharmacologie, épidémiologie, statistiques, galénique. En parallèle de ce développement, des règles éthiques ont été instaurées, du fait des évènements historiques et sociaux qui y sont associés. On peut par exemple citer le code de Nuremberg, élaboré en 1947 dans le cadre du procès de Nuremberg intenté contre certains médecins ayant dirigé des expériences sur des détenus des camps de concentration.

Ce texte regroupe une série de dix principes et il identifie le consentement éclairé comme un préalable absolu à la conduite de recherches portant sur des sujets humains.

Historiquement, la comparaison entre deux traitements est apparue pendant l'Antiquité. A cette époque, la recherche clinique ne fait pas l'objet d'une systématisation, ou d'une formalisation. On peut par exemple citer Hippocrate, médecin et philosophe grec (425 avant J.C), considéré par beaucoup comme le ≪*père de la médecine*≫, qui fonde ses travaux sur une vision subjective. En effet, ses découvertes ne sont pas généralisables puisqu'il ne porte pas un souci particulier à rendre ses groupes comparables et représentatifs d'une potentielle population cible.
C'est seulement au Moyen-Âge qu'Avicenne (1025 après J.C), philosophe et médecin Perse, introduit et formalise le principe d'essai clinique.
La première étude qui répond à la définition actuelle d'un essai clinique a été effectuée par James Lind, médecin écossais, en 1747 [Lind, 1772, Tröhler, 2005]. Il sélectionne des marins tous atteints de scorbut, et les place dans le même environnement avant de les répartir en six groupes de deux. Il ne fera varier qu'une substance dans l'alimentation des sujets, le reste de l'alimentation restant identique. Ces substances étaient : le cidre, de l'élixir de vitriol (acide sulfurique dilué), du vinaigre, une concoction d'herbes et d'épices, de l'eau de mer et des oranges et citrons. Il découvre que le groupe qui a reçu des oranges et des citrons s'est rétabli du scorbut en 6 jours. Malheureusement il faudra attendre le début des années 1800 pour voir se généraliser la mise en place de ration journalière en citron dans la ≪Royal Navy≫ [Vale, 2008]. Il sera par la suite prouvé que la manifestation du scorbut est liée à une carence délétère en Vitamine C. La recherche clinique a beaucoup évolué dans les années qui ont suivi. Les principales étapes de son développement sont résumées dans la Figure 1.1.

1.2.1.1 Définition Actuelle

L'Organisation Mondiale de la Santé (OMS) définit l'essai clinique comme : ≪toute recherche dans laquelle les participants ou les groupes de participants sont affectés, dès le départ, à une ou des interventions d'ordre sanitaire afin d'évaluer les effets de ces dernières sur la santé. Les interventions portent, entre autres, sur les médicaments, les cellules et autres produits biologiques,

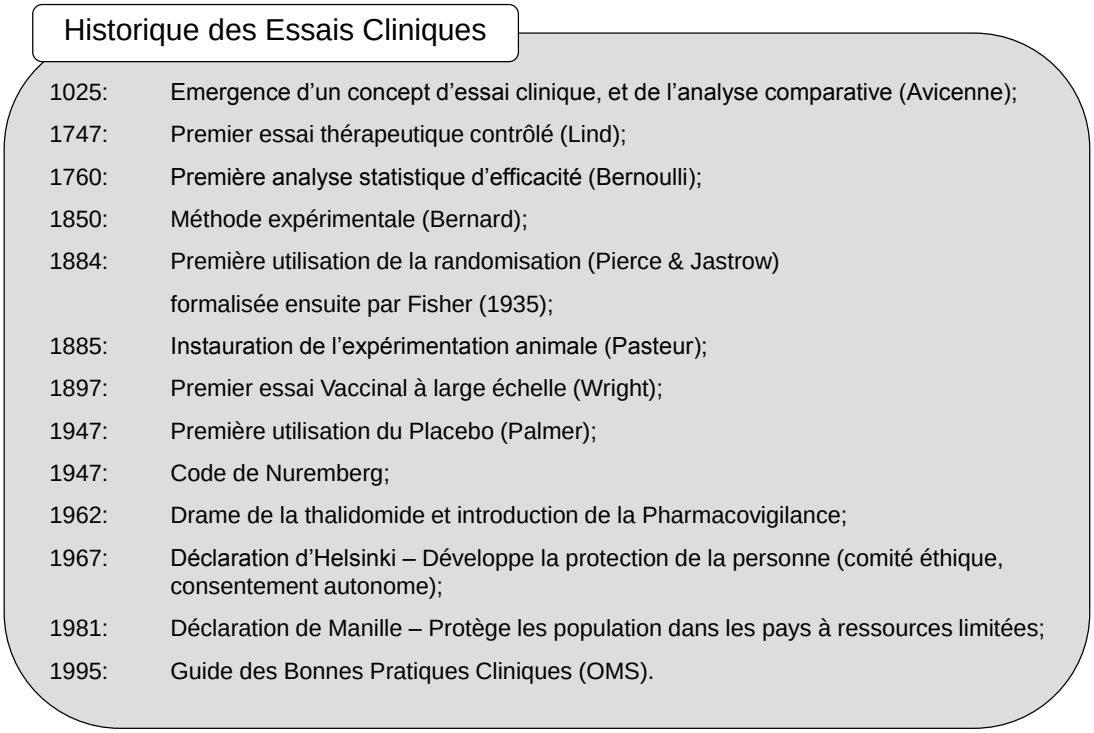


FIGURE 1.1: Les principales étapes du développement des essais cliniques

les actes chirurgicaux, les techniques radiologiques, les dispositifs, les thérapies comportementales, les changements dans les protocoles de soins, les soins préventifs, etc. Cette définition englobe les Phases d'essai de I à IV.≫ Cette définition est large et permet d'inclure les essais cliniques nutritionnels.

1.2.1.2 La demande de mise sur le marché

Lorqu'un laboratoire pharmaceutique veut commercialiser une nouvelle molécule thérapeutique, il est dans le devoir de démontrer sa qualité (fabrication), sa sécurité (développement pré-clinique) et son efficacité (développement clinique). C'est dans cette optique qu'il dépose un dossier de demande de mise sur le marché aux autorités de santé compétentes. A l'issue de l'étude de ce dossier, les autorités statuent quant à la délivrance de l'Autorisation de Mise sur le Marché (AMM).
En Europe, cette demande est faite auprès de l'EMA (≪*European Medecine Agency*≫), et aux Etats-

Unis auprès de la FDA (≪*Food and Drug Administration*≫).

1.2.1.3 Les critères de jugement

Les critères de jugement sont les variables permettant d'évaluer les objectifs d'un essai clinique. D'un point de vue clinique de nombreuses variables peuvent être définies comme telles. Cependant, elles n'ont pas toutes la même pertinence clinique, et il est nécessaire de les hiérarchiser.
On définit tout d'abord les critères de jugement principaux (≪*primary endpoints*≫) qui correspondent aux variables qui vont permettre de répondre à l'objectif principal de l'étude. En général, un seul critère de jugement principal est défini. Il arrive néanmoins qu'il soit indispensable de prendre plusieurs critères de jugement principaux. Cette situation est à éviter au maximum. Cependant, dans le cadre de critères de jugement principaux multiples, il est important de définir le vocabulaire utilisé. S'ils sont tous de même importance on parlera de ≪*co-primary endpoints*≫ en anglais. En fonction de l'objectif clinique, l'analyse de ces critères va être différente, il est donc important de le différencier. Lorsque les promoteurs souhaitent que l'ensemble des hypothèses reliées aux critères de jugement principaux soit rejeté pour conclure au succès de l'étude, alors nous parlerons de critères de jugement de type ≪*must win*≫ [Julious and McIntyre, 2012]. Julious et al. en 2012 prennent l'exemple des essais cliniques sur l'arthrose [Julious and McIntyre, 2012]. Dans le cadre de ces essais, l'efficacité du traitement ne peut être conclue que si une différence significative est observée sur l'ensemble des trois critères standards : évaluation globale du patient, Score WOMAC (index de sévérité symptomatique de l'arthrose sur les membres inférieurs) sur la douleur, Score WOMAC sur la condition physique.
Il existe aussi, dans la pratique, les critères de jugment dits ≪*at least one win*≫ [Dmitrienko et al., 2012]. Les promoteurs concluront alors au succès de l'essai si au moins un critère de jugement est rejeté parmi l'ensemble des critères principaux testés [Senn and Bretz, 2007]. Pour illustrer ces critères, prenons l'exemple d'un essai clinique en nutrition [Boge et al., 2009]. L'objectif de cet essai consiste à démontrer que la prise d'un produit laitier X augmente les réponses immunitaires dans le cadre de la vaccination contre la grippe. Il s'agit alors de comparer la réponse immunitaire pour deux stratégies prophylactiques, à savoir la vaccination seule, et la vaccination associée à la prise du produit laitier. Si la stratégie comprenant la prise de produit laitier entraîne une augmentation de la réponse immunitaire sur au moins l'une des trois souches grippales étudiées, alors le produit laitier

sera considéré comme efficace.

Pour finir, nous pouvons citer les critères de jugement ≪*at least* r *win*≫. Ils sont utilisés dans le cas où les promoteurs concluent au succès de l'essai si au moins r critères de jugement principaux parmis m sont significatifs.

Lorsque les critères principaux sont établis, il est nécessaire de définir les critères de jugement secondaires (≪*secondary endpoints*≫). Il s'agit de variables ayant un intérêt clinique mais dont l'analyse n'est pas nécessaire pour répondre à l'objectif principal de l'étude. Dans le cadre d'essais d'efficacité, ces critères pourraient être : la qualité de vie, la fréquence d'évènements indésirables, ou encore une composante d'un critère de jugement principal composite. Ces critères sont présents à titre documentaire.

Le rapport de l'EMA sur les problèmes de multiplicité conseille de limiter au maximum le nombre de critères de jugement principaux. L'une des possibilités dans le cadre d'effets multifactoriels du produit consiste à utiliser un critère de jugement composite. Celui-ci est créé artificiellement et combine plusieurs évènements cliniques. Il a le double avantage d'éviter de prendre en considération des critères de jugement très corrélés, mais aussi d'augmenter la puissance de l'effet étudié. Cependant, lorsque l'effet du produit n'est pas uniforme sur les différents évènements cliniques composant le critère composite, l'interprétation est complexe et la pertinence du critère peut alors être remise en question.

1.2.1.4 Le déroulement d'un essai clinique

Les essais cliniques sont une étape obligatoire et systématique du développement d'un produit ou d'une intervention de santé. Ils permettent de préciser l'effet de cette intervention chez l'Homme, d'en déterminer l'efficacité ainsi que les éventuels effets indésirables. Le déroulement d'un essai clinique est formalisé sous forme de phases cliniques. Celles- ci ne peuvent être réalisées chez l'Homme qu'après de multiples étapes de développement pré-clinique. Les études conduites lors du développement pré-clinique permettent de démontrer la sécurité de la nouvelle molécule. Lors de cette étape, les scientifiques récoltent et analysent des données correspondant aux études *in-vivo* et *in-vitro* du médicament : pharmacologie, toxicologie et pharmaco-cinétique principalement.

Lors du développement clinique, on peut généralement distinguer quatre phases dans les essais cliniques (phases I à IV). Cependant, il n'existe pas de limite claire entre ces phases, et beaucoup d'opinions divergent quant aux détails méthodologiques. Toutefois l'OMS a essayé de rédiger une brève description de chaque phase en se fondant sur l'objectif poursuivi dans chacune d'entre elles :

- **Phase I** : Il s'agit des premiers essais chez l'Homme d'un nouveau principe actif ou d'une nouvelle formulation. Ces essais sont généralement réalisés sur des volontaires en bonne santé, aussi appelés volontaires sains. L'objectif de cette phase consiste à permettre une évaluation préliminaire de l'innocuité ainsi que du profil pharmaco-cinétique, et si possible pharmacodynamique, du principe actif chez l'Homme.

- **Phase II** : Les essais de phase II sont effectués sur un nombre limité de sujets ; lorsqu'ils atteignent un stade plus avancé, ils sont souvent de nature comparative. Cette étude comparative se fait par rapport à un traitement de référence, ou par rapport à un Placebo si celui-ci n'est pas défini. L'objectif est ici de démontrer l'activité thérapeutique du principe actif et d'évaluer son innocuité à court terme chez des patients dont il est destiné à soulager la maladie ou à améliorer l'état. Cette phase vise aussi à déterminer l'éventail des doses ou la posologie appropriée et (si possible) à établir la relation dose/réponse, de façon à offrir une base optimale pour la conception d'essais thérapeutiques à grande échelle. De manière générale, cette phase permet donc de vérifier le concept pharmacologique chez la population cible.

- **Phase III** : Les essais de phase III portent sur des groupes de patients plus importants et éventuellement hétérogènes. Cette démarche a pour but de déterminer le rapport entre l'innocuité et l'efficacité de la ou des formulations du principe actif, et d'évaluer leur intérêt thérapeutique global et relatif. Si le produit donne lieu à des réactions indésirables fréquentes, leur profil doit être étudié, de même que certaines caractéristiques spéciales du produit. Les essais devraient de préférence être randomisés et en double aveugle. De manière générale, les conditions dans lesquelles l'essai est exécuté doivent se rapprocher au maximum des conditions normales d'utilisation. C'est à l'issue de cette phase qu'il est possible d'établir le rapport

bénéfice/risque. Le laboratoire peut demander l'Autorisation de Mise sur le Marché du produit (AMM).

– **Phase IV** : Les essais de phase IV sont effectués après la mise sur le marché du médicament. Ils sont menés en fonction des caractéristiques du produit qui ont motivé l'autorisation de mise sur le marché. Ils se présentent généralement sous la forme d'études de pharmacovigilance, d'évaluation de l'intérêt thérapeutique ou des stratégies de traitement. Bien que les méthodes utilisées puissent être différentes, ces études doivent s'appuyer sur les mêmes normes scientifiques et éthiques que les études de pré-commercialisation. Lorsqu'un produit a été mis sur le marché, les essais cliniques portant sur la recherche de nouvelles indications, de nouvelles méthodes d'administration, de nouvelles associations, sont normalement considérés comme des essais portant sur de nouveaux produits.

L'ensemble de ces informations sont résumées dans la Figure 1.2.

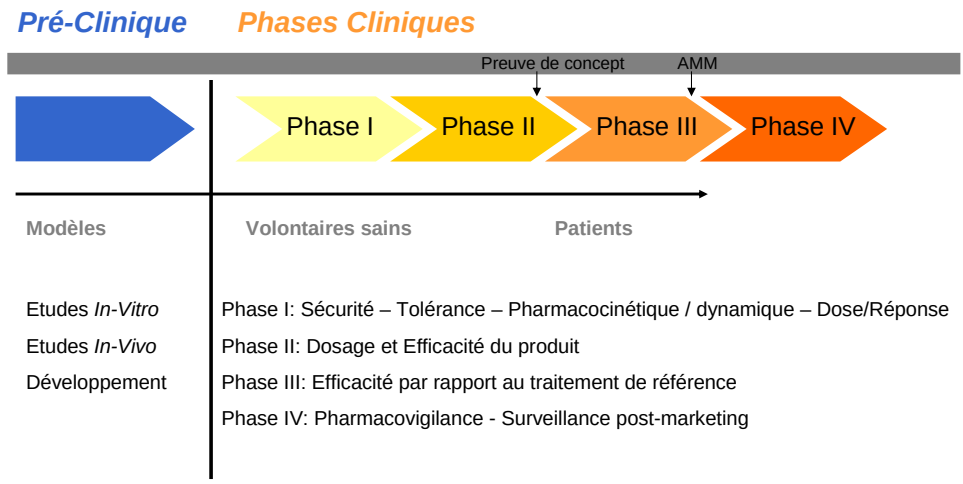


FIGURE 1.2: Le déroulement d'un essai clinique dans l'Industrie Pharmaceutique

Remarque : En France, comme partout en Europe, le lancement d'une étude clinique n'est possible qu'après autorisation des autorités de santé.

1.2.1.5 L'ICH et les Bonnes Pratiques Cliniques (BPC)

On assiste depuis plusieurs décennies à un effort d'harmonisation des différents volets du processus de réglementation pharmaceutique. Cet effort peut être illustré par la mise en place en 1990 d'une conférence internationale composée de représentants des USA, de l'Union Européenne et du Japon, ainsi que des experts de l'industrie pharmaceutique. Il s'agit de l'ICH (*Internationnal Conference on Harmonization*) et son objectif consiste à harmoniser les procédures inhérentes aux aspects scientifiques et techniques de l'enregistrement des médicaments.

En 1996, l'ICH adopte la note explicative ICH-E6 relative aux Bonnes Pratiques Cliniques (BPC). Ce document reprend les principes d'un comité d'éthique indépendant, définit les fonctions et responsabilités des investigateurs et des promoteurs d'essais cliniques, ainsi que le contenu du protocole d'un essai. En France, cette note est approuvée le 17 juillet 1996 par le Code de la Santé Publique et entre en vigueur pour les essais débutant après le 17 janvier 1997.

D'après l'OMS : ≪Les bonnes pratiques cliniques (BPC) sont des normes de qualité éthique et scientifique internationales applicables dans la conduite de la recherche biomédicale sur des sujets humains.≫ Le respect de telles normes a pour objectif de préserver les droits, le bien être, la sécurité ainsi que la confidentialité des informations concernant les sujets inclus dans l'étude conformément aux principes découlant de la Déclaration d'Helsinki. En France, pour s'assurer du respect de ces normes, les modalités des essais cliniques sont soumises avant toute action au Comité de Protection des Personnes (CPP).

Devant la prolifération de nouvelles méthodes d'analyse statistique, et l'importance de celles-ci dans les demandes d'AMM, l'ICH a conclu qu'il était nécessaire de mettre en place un document succinct. Ce dernier permettrait d'harmoniser les méthodes utilisées ainsi que de répondre aux questions statistiques inhérentes à la recherche clinique. C'est ainsi qu'est né en 1998 l'ICH-E9 : ≪*Statistical Principles for Clinical Trials*≫. Ce document fait actuellement référence dans l'analyse statistique d'essais cliniques.

1.3 La recherche clinique en nutrition

L'avènement des problématiques associées à la nutrition médicale a eu pour conséquence une augmentation croissante du nombre d'aliments faisant l'objet d'allégations nutritionnelles ou de santé. Devant ce constat, les décideurs politiques de l'Union Européenne (UE) ont adopté, en Décembre 2006, un règlement concernant l'utilisation des allégations santé et nutritionnelles pour les denrées alimentaires. Un des objectifs clés de ce règlement est de garantir que toute allégation figurant sur l'étiquette d'un aliment vendu au sein de l'UE soit claire, et justifiée par des preuves scientifiques. Avant de mettre en avant un quelconque bénéfice nutritionnel ou de santé, une entreprise agroalimentaire doit donc avoir réalisé des études scientifiques le validant.
Se basant sur cette démarche réglementaire, il est facile de comprendre pourquoi la recherche clinique est au coeur de la démarche de Recherche et Développement (R&D) du groupe Danone.
Les recommandations établies par l'ICH n'ont pas seulement une vocation pharmaceutique : ≪*The principles in this guideline may also be applied to other clinical investigations that may have an impact on the safety and well-being of human subjects.*≫
En effet, les problématiques inhérentes à la recherche clinique pharmaceutique sont très proches de celles de la nutrition. C'est la raison pour laquelle il a été choisi, au sein de la plateforme clinique de Danone Research, de prendre comme procédures de référence les BPC ainsi que les recommandations établies par l'ICH.

1.3.1 Les spécificités de la recherche clinique en nutrition

Bien que similaire sur de nombreux points à la recherche clinique pharmaceutique, la recherche clinique en nutrition possède aussi certaines spécificités qu'il est important de rappeler.
Le développement du produit se fait sur des sujets sains dans un but prophylactique. Dans ce contexte, l'effet recherché n'est pas immédiat, il est donc nécessaire de travailler au sein de l'essai sur des périodes plus longues.
Le mécanisme d'action des produits est souvent multifactoriel. Il est aussi important de noter que le but premier des aliments n'est pas de soigner. L'effet des produits est donc bien plus faible que ceux des médicaments. Toutefois, dans certains cas, l'effet du produit est cumulatif puisqu'il peut s'ajouter et compléter l'effet du médicament. Les scientifiques en nutrition se tournent principalement vers les

domaines de l'immunité, le cardiovasculaire, le digestif, le métabolisme, les allergies, les carences et la malnutrition.

1.3.2 Objectif en nutrition

Comme nous l'avons vu précédemment, bien que non concernées par le processus de l'AMM du produit de santé, les industries agro-alimentaire doivent aussi montrer l'efficacité de leurs produits. C'est la raison pour laquelle, lors du processus de développement d'un produit, des études cliniques sont conduites sur les produits pour lesquels des allégations santé sont envisagées. Cette démarche permet de démontrer et de valider leur efficacité tout en obtenant des arguments scientifiques suffisants pour soutenir devant les autorités compétentes les demandes d'allégations santé. Il s'agit par exemple de l'EFSA (*European Food Safety Administration*) en Europe et toujours de la FDA aux Etats-Unis.

1.3.3 Déroulement d'un essai chez Danone

La principale différence avec le domaine pharmaceutique réside dans le type de produit étudié. En effet, en nutrition, le produit reste un produit alimentaire pour lequel il n'est pas nécessaire d'étudier les effets secondaires. La démarche clinique se focalise donc principalement sur l'efficacité du produit. Chez Danone, la recherche clinique a commencé au début des années 90 afin d'aider les scientifiques dans la recherche et le développement de nouveaux produits. Fort de cette démarche scientifique rigoureuse, cela leur à permis par la suite de convaincre les consommateurs.
C'est en 2006, en parallèle du réglement européen légiférant l'utilisation d'allégations santé, que Danone a créé, au sein de son centre de recherche, un département spécialisé en recherche clinique. A l'heure actuelle, la conduite d'un essai clinique chez Danone peut être résumée autour de 12 étapes schématisées sur la Figure 1.3.

Le statisticien intervient quant à lui dans 6 phases développées ci dessous :

- Conception de l'étude : le nombre de sujets, le design de l'étude, le type de population à considérer, le type de produit ainsi que les marqueurs à analyser sont définis.
- Rédaction du protocole et des documents : un protocole est établi par un groupe d'experts composé de scientifiques du centre de recherche, de l'équipe en charge de l'étude clinique, ainsi

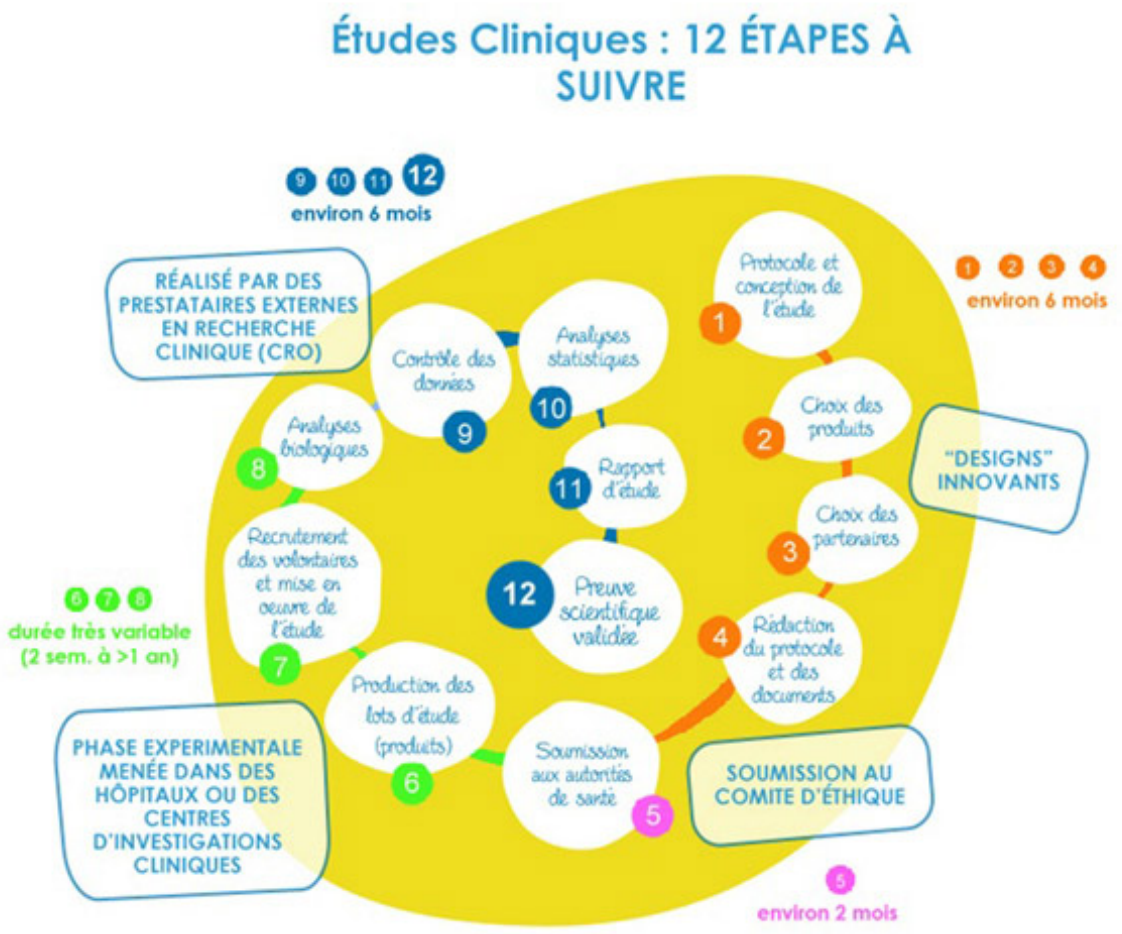


FIGURE 1.3: Le déroulement d'un essai clinique chez Danone

que d'experts extérieurs. Ce document comprend une estimation précise de l'échantillon de la population, une définition des critères à étudier et la liste de randomisation. La randomisation permet d'attribuer le produit pour chacun des sujets tout en contrôlant les facteurs de risques non mesurables ou non prévus.

– Phase expérimentale : les produits actifs et de contrôle sont distribués selon la liste de randomisation pour chaque sujet. Les données sont relevées selon les critères définis dans le protocole. Une fois ce travail réalisé, le statisticien en charge de l'étude rédige le Plan d'Analyses Statistiques (SAP) en accord avec les principes de l'ICH. Les études sont généralement réalisées en

double aveugle, c'est-à-dire que ni la personne qui donne le produit, ni le sujet ne connaissent la nature du produit consommé (produit actif ou produit de contrôle).

- Contrôle des données par *Data Management* : cette étape est réalisée par le data manager et consiste à nettoyer la base de données par l'identification des valeurs atypiques (aberrantes ou extrêmes). Une fois ces valeurs identifiées, il convient de les corriger avec l'aide des attachés de recherche clinique (ARC) et des médecins. Cette étape est finalisée par la revue de données (≪*Blind review* ≫) qui consiste à valider définitivement la base de données. Cela conduit au gel de la base. C'est seulement à l'issu de ce gel de base qu'il est possible de lever l'aveugle, *i.e.* connaître le produit pris par chaque sujet.
- Analyses statistiques : les données sont analysées en s'appuyant sur le Plan d'Analyses Statistiques (SAP). Le SAP consiste à définir les objectifs, le schéma d'étude, les variables mesurées, le calcul de taille d'échantillon ainsi que les méthodes statistiques définies par l'équipe projet. Il comporte donc les analyses intermédiaires envisagées, les méthodes statistiques permettant de répondre aux objectifs cliniques, ainsi que celles qui prennent en compte tout type de problèmes statistiques prévisibles (données manquantes, multiplicité des tests, ...). Dans l'idéal, le programme informatique associé au plan d'analyse devrait pouvoir être rédigé en amont du recueil de données. Il est important que ces analyses soient reproductibles, et que la démarche effectuée soit disponible.
- Rédaction du rapport d'étude : il s'agit de la rédaction d'un rapport contenant l'ensemble des informations relevées lors des étapes précédentes. Il permet de décrire le déroulement de l'étude et de conclure quant à l'objectif de l'étude qui consiste dans la majorité des cas à affirmer ou infirmer l'efficacité du produit. Il s'attache aussi notamment à soulever les limites et forces de l'étude. Ce rapport est une partie non négligeable du rapport envoyé aux autorités pour la demande d'allégation santé.

1.4 L'émergence d'une problématique industrielle

L'une des spécificités notables de la recherche clinique en nutrition réside dans l'aspect multifactoriel de l'effet recherché et par conséquent des paramètres mesurés. Les scientifiques sont dans la majorité des cas dans l'impossibilité de définir un critère d'évaluation unique, même en passant par

un critère composite. Plusieurs critères de jugement sont donc envisagés afin de mettre en évidence l'efficacité du produit. Cependant, la multiplicité des critères va conduire à des comparaisons multiples, qui vont engendrer un problème de multiplicité [Schulz and Grimes, 2005a]. En effet, le fait de multiplier ces comparaisons au sein d'un même échantillon va augmenter le risque de fausses découvertes.

Trop souvent, ce problème est ignoré, ce qui aura pour conséquence de surestimer la significativité des résultats scientifiques. En recherche biomédicale, ce problème est bien connu par les autorités de santé qui ont décidé de mettre en place des réglementations afin d'améliorer la qualité des analyses statistiques. Par conséquent, dans le but d'améliorer les analyses statistiques, des recommandations ont été mises en place par les agences de régulation. En 1998, l'*Internationnal Conference on Harmonization* a publié une recommandation sur l'analyse statistique en recherche clinique (ICH 1998) qui intègre la problématique de la multiplicité (ICH E9). C'est aussi le cas de l'*European Medicines Agency* qui a publié un document pour l'analyse des données dans le contexte des tests multiples [EMA, 2002]. Bien que ces documents ne considèrent pas l'ajustement comme nécessaire, les autorités demandent de justifier ce choix quand celui-ci n'a pas été fait. C'est pour cela qu'il est devenu de plus en plus fréquent en recherche clinique d'incorporer dans le plan d'analyse statistique des procédures de tests multiples adéquates [Bretz et al., 2010]. Il est toutefois important de noter que l'intérêt de cette pratique n'est pas seulement de suivre les recommandations, mais il est aussi de se protéger d'une mauvaise prise de décision.

Néanmoins, devant les problématiques inhérentes à l'analyse des essais cliniques, les méthodes utilisées restent souvent les plus connues et les plus faciles d'utilisation [Hochberg, 1988, Holm, 1979]. Elles ont cependant une certaine limite puisqu'elles ne prennent pas en compte la corrélation existant entre les statistiques de tests. Cela peut se traduire par une sous-estimation de l'effet du produit, et une sur-estimation du calcul du nombre de sujets inclus dans l'étude. Ces limites ne sont pas négligeables puisqu'elles vont diminuer l'intensité des résultats scientifiques et compromettre la faisabilité interne de l'étude. Ce problème est d'autant plus important en nutrition, où l'effet recherché du produit est faible. Face à ce constat, l'équipe Biométrie de Danone Research a choisi de s'intéresser de plus près à la problématique des tests multiples en recherche clinique. L'objectif de cette thèse est de prendre en compte la corrélation dans la correction de la multiplicité sur des critères princi-

paux multiples et quantitatifs. C'est à ce moment qu'est né le projet de thèse auquel nous allons nous intéresser tout au long de ce mémoire. Pour commencer, le chapitre suivant sera l'occasion de rappeler l'état des connaissances actuelles dans le contexte des essais cliniques. Nous verrons ensuite respectivement à travers les chapitres 3 et 4, le calcul de taille d'échantillon et l'analyse de données pour un contrôle de la puissance disjonctive et de la ≪r-power≫. Nous verrons pour finir une méthode permettant de corriger le degré de signification du test dans le cadre d'une recherche de codage optimal d'une variable quantitative dans un modèle linéaire généralisé.

2

État des connaissances

Ce chapitre présente un état des lieux des recherches statistiques réalisées dans le cadre de la correction des tests multiples, et plus particulièrement pour des problématiques en recherche clinique.

A l'heure actuelle, beaucoup de projets de recherche ont pour objectif de répondre simultanément à plusieurs problématiques scientifiques ou de considérer une problématique ayant des intéractions multiples. Une analyse statistique rigoureuse des données se doit donc de prendre en compte la multiplicité induite par cette pratique, afin de ne pas biaiser l'interprétation des résultats.
Les procédures de correction des tests multiples sont fortement imprégnées des grands principes de la théorie des tests d'hypothèses. Celle-ci est formalisée au début du XX$^{\text{ème}}$ siècle par les contributions de Fisher, Neymann et Pearson, mais avait déjà été initiée, dès la fin du XIX$^{\text{ème}}$ siècle, par les travaux de Laplace, DeMoivre et Bernoulli sur la maîtrise des erreurs et de l'aléatoire. Nous avons donc choisi de commencer ce chapitre par des rappels sur cette théorie des tests.

2.1 Rappels de théorie des tests

2.1.1 Hypothèses statistiques

Un test d'hypothèse est un procédé d'inférence qui a pour but de fournir une règle de décision à une problématique scientifique. Il permet, sur la base de résultats d'échantillons aléatoires, de faire un choix entre deux hypothèses statistiques relatives à une population. Cette généralisation comporte des risques et des limites, que contrôle la démarche statistique.

L'hypothèse nulle ($\mathcal{H}_0$) et l'hypothèse alternative ($\mathcal{H}_1$) définissent les deux issues possibles du test statistique. L'hypothèse $\mathcal{H}_1$ est l'hypothèse de recherche, c'est ce que nous cherchons à démontrer. L'hypothèse complémentaire $\mathcal{H}_0$ est soumise au test, et toute la démarche s'effectue en la considérant comme vraie. La procédure de test cherche donc à réfuter l'hypothèse nulle.

2.1.2 Taux d'erreurs

L'échantillonnage apporte une part d'aléatoire qui va introduire deux types d'erreurs dans la réalisation d'un test :

- L'erreur de Type-I correspond au fait de rejeter à tort l'hypothèse nulle. La probabilité d'erreur de Type-I est notée α, et elle est définie comme :
  $\mathbb{P}[$ rejeter $\mathcal{H}_0 \mid \mathcal{H}_0$ est vraie$]$. Le niveau α d'un test correspond à la borne supérieur du risque de première espèce. Ce taux d'erreurs doit être défini par le chercheur, il est en général fixé à 0.05 ou 0.01 ;
- L'erreur de Type-II correspond quant à elle au fait de rejeter à tort l'hypothèse alternative. La probabilité d'erreur de Type-II est notée β, et elle est définie comme : $\mathbb{P}[$ rejeter $\mathcal{H}_1 \mid \mathcal{H}_1$ est vraie$]$.

La puissance est le complémentaire de la probabilité d'une erreur de Type-II. Elle est donc définie comme la probabilité de conclure à une différence sachant que la différence est vraie :

$$1 - \beta = \mathbb{P}[\text{ décider } \mathcal{H}_1 \mid \mathcal{H}_1 \text{ est vraie}].$$

Pour quantifier la puissance d'un test, il est nécessaire de connaître la loi de la statistique de test sous l'hypothèse alternative. Les différents risques associés aux tests statistiques sont résumés dans le tableau 2.1.

Nous pouvons remarquer que α et $1 - \alpha$ sont énoncés sous l'hypothèse nulle, alors que β et $1 - \beta$ sont eux énoncés sous l'hypothèse alternative. Néanmoins, pour tout test statistique, ces quatre taux coexistent puisque nous ne connaissons pas l'hypothèse qui est vraie.
La relation entre les risques d'erreurs peut quant à elle être visualisée à partir de la Figure 2.1.

TABLE 2.1: Les risques de première (α) et seconde (β) espèce

Décision	Réalité	
	$\mathcal{H}_0$ Vraie ($\mathcal{H}_1$ Fausse)	$\mathcal{H}_1$ Vraie ($\mathcal{H}_0$ Fausse)
$\mathcal{H}_0$ Non rejetée	Décision correcte (1-α)	Risque β
$\mathcal{H}_0$ Rejetée	Risque α	Décision correcte (1-β)

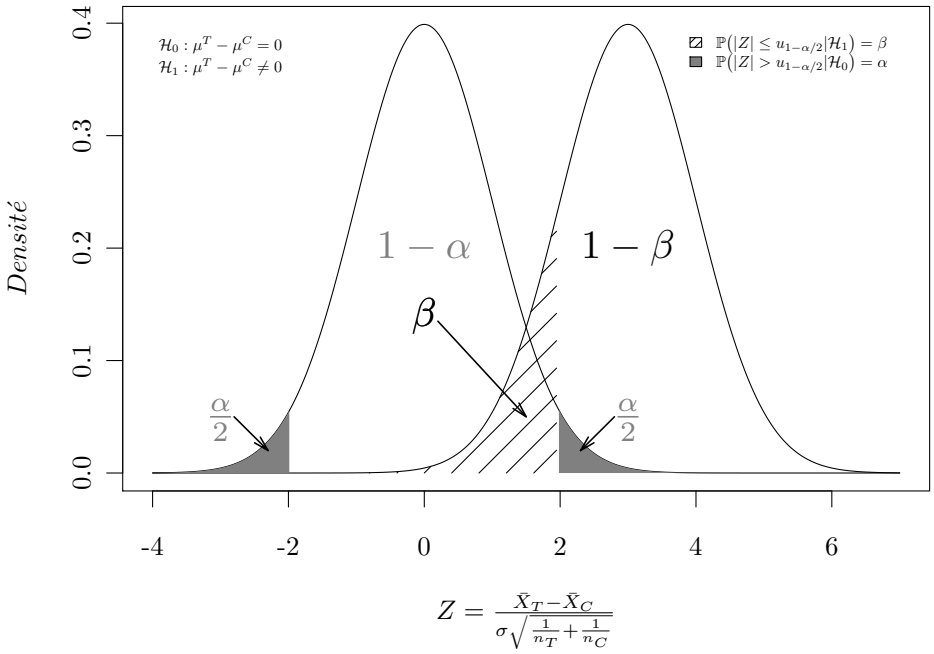


FIGURE 2.1: Relation entre le risque de première espèce (α) et le risque de seconde espèce (β) dans le cadre d'un test de comparaison de moyennes bilatéral Z

Remarques : Toutes choses étant égales par ailleurs, si on diminue le risque de première espèce (α), alors il sera plus difficile de conclure à une différence significative. De plus, on peut aussi observer sur le graphique que si l'on diminue la valeur de α alors la puissance $(1-\beta)$ diminuera aussi.

P-valeur

La p_{valeur} est définie comme la probabilité d'obtenir une statistique de test au moins aussi extrême que celle qui a été observée, en supposant que l'hypothèse nulle soit vraie. Elle peut aussi être vue comme le plus petit niveau auquel $\mathcal{H}_0$ est rejetée sachant z. Cette notion est donc très utile en théorie des tests et permet d'obtenir une mesure de significativité propre à chaque statistique observée.
Pour la calculer, il est nécessaire de connaître la distribution de la statistique (Z) sous $\mathcal{H}_0$ ainsi que la valeur de la réalisation de la statistique de test (z). Dans le cadre d'un test bilatéral de comparaison de moyennes, cette définition peut être traduite par : $\mathbb{P}(|Z| \geq z|\mathcal{H}_0)$. Cette probabilité peut aussi être vue comme l'aire sous la fonction de distribution pour lequel $|Z| \geq z$ (Figure 2.2).

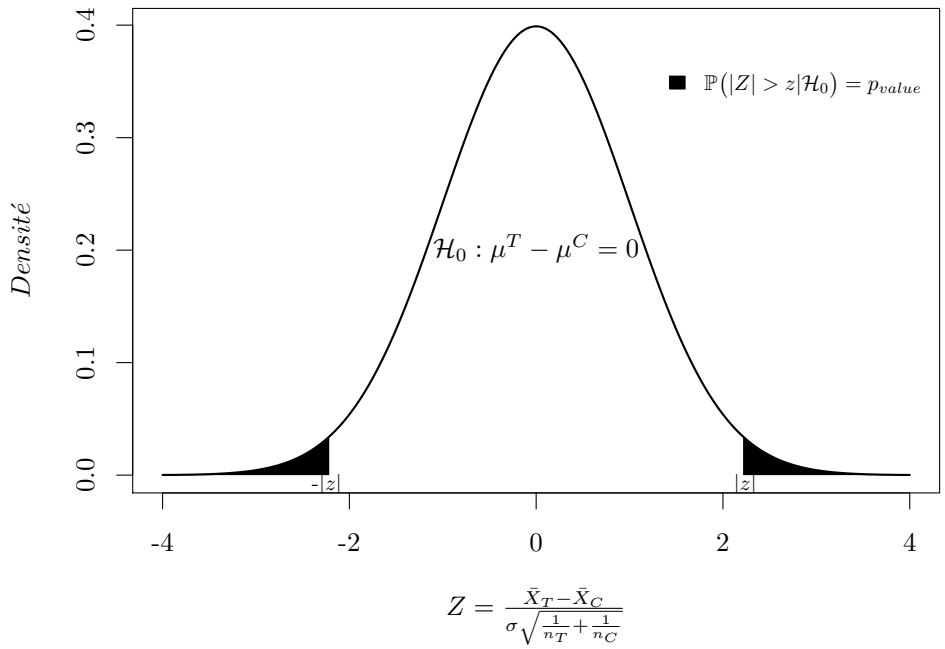


FIGURE 2.2: La p_{valeur} dans le cadre d'un test de comparaison de moyennes bilatéral Z

Le rejet de l'hypothèse nulle individuelle est réalisée si la $p_{valeur} \leq \alpha$, cette règle de décision permet un contrôle du taux d'erreurs de Type-I au seuil α.

2.1.3 Comparaison de deux moyennes

Cette section va être l'occasion de présenter les différentes approches possibles dans le cadre de tests de comparaison de deux moyennes en recherche clinique.

Pour cela, considérons un essai dont le critère principal est quantitatif, et pour lequel l'objectif est la comparaison des deux traitements. Soit X_i^k les valeurs du critère de jugement principal pour le $i^{\text{ème}}$ sujet ($1 \leq i \leq n_k$) et le traitement k ($k = T, C$). Nous posons ici l'hypothèse que les valeurs X_i^k sont indépendantes et qu'elles suivent une loi normale de moyenne μ^k et de variance σ_k^2. Nous nous intéresserons ici plus particulièrement à la vraie différence entre les moyennes des deux groupes : $\delta = \mu^T - \mu^C$. Nous définissons Δ le seuil clinique établi par les scientifiques afin de confirmer que la différence est pertinente. Nous posons ici : $\Delta \geq 0$.

En fonction de l'objectif clinique, le test de comparaison de deux moyennes peut prendre différentes formes : le test d'égalité, le test de non-infériorité, le test de supériorité ou bien le test d'équivalence. Dans la suite de cette section nous nous baserons sur des tests de comparaison de moyennes pour variance connue, souvent utilisés dans le calcul de taille d'échantillon.

2.1.3.1 Test d'égalité

Ce test est généralement utilisé dans le cas où l'objectif scientifique consiste à prouver qu'il existe une différence entre les deux groupes sur le critère de jugement étudié, ici quantitatif. Dans la majorité des cas les scientifiques choisissent $\Delta = 0$. Les hypothèses associées à ce test prennent la forme suivante :

$$\mathcal{H}_0 : \delta = \Delta \ \ versus \ \ \mathcal{H}_1 : \delta \neq \Delta.$$

Nous nous retrouvons ici dans le cadre d'un test bilatéral, où l'hypothèse nulle sera rejetée au niveau α si et seulement si :

$$\left| \frac{\bar{X}^T - \bar{X}^C - \Delta}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} \right| > u_{1-\alpha/2},$$

où $\bar{X}^k$ et σ_k^2 représentent respectivement la moyenne empirique et la variance vraie du groupe k, et $u_{1-\alpha/2}$ est le fractile d'ordre $1 - \alpha/2$ d'une loi Normale centrée réduite.

2.1.3.2 Test de non infériorité

Dans un essai clinique de non-infériorité, il existe déjà un traitement de référence sur le marché ayant fait la preuve de son efficacité contre un placebo. L'objectif scientifique consiste alors à démontrer que le nouveau produit n'est pas meilleur en efficacité, mais qu'il apporte d'autres avantages (tolérance, coût, facilité d'utilisation, ...). Dans ce cas, l'objectif statistique consiste à vérifier que la moyenne de la variable d'intérêt de la nouvelle molécule n'est pas moins bonne que celle du traitement de référence. Dans le cadre d'un essai de non-infériorité, Δ sera souvent noté Δ_{NI}.
Les hypothèses prennent donc la forme suivante :

$$\mathcal{H}_0 : \delta \leq -\Delta \quad versus \quad \mathcal{H}_1 : \delta > -\Delta.$$

Ces hypothèses sont retranscrites dans la Figure 2.3 :

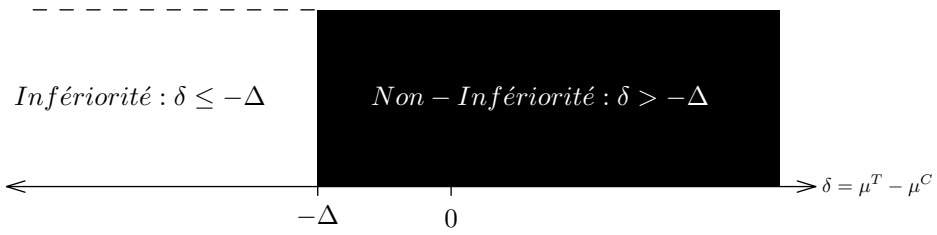


FIGURE 2.3: Le test de non-infériorité

Il s'agit ici d'un test unilatéral. L'hypothèse nulle sera rejetée au niveau α si et seulement si :

$$\frac{\bar{X}^T - \bar{X}^C + \Delta}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} > u_{1-\alpha},$$

où $u_{1-\alpha}$ est le fractile d'ordre $1-\alpha$ d'une loi Normale centrée réduite.

2.1.3.3 Test de supériorité

Dans un essai de supériorité, il n'y a pas de produit de référence et l'objectif consiste donc à montrer que ce nouveau produit est meilleur qu'un placebo ; en d'autres termes, à savoir si le nouveau produit observe un bénéfice clinique significatif, ayant pour objectif de changer la pratique actuelle. Les hypothèses statistiques sont donc de la forme suivante :

$$\mathcal{H}_0 : \delta \leq \Delta \quad versus \quad \mathcal{H}_1 : \delta > \Delta.$$

avec $\Delta \geq 0$.

Comme précédemment, il s'agit d'un test unilatéral, l'hypothèse nulle sera rejetée au niveau α si et seulement si :

$$\frac{\bar{X}^T - \bar{X}^C - \Delta}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} > u_{1-\alpha},$$

où $u_{1-\alpha}$ est le fractile d'ordre $1-\alpha$ d'une loi Normale centrée réduite.

2.1.3.4 Test d'équivalence

Dans un essai clinique d'équivalence, il existe déjà un produit de référence sur le marché ayant fait la preuve de son efficacité contre placebo. L'objectif scientifique consiste aussi à démontrer que l'innovation du nouveau produit n'est pas meilleur en efficacité, mais qu'il apporte d'autres avantages. Il s'agit ici de démontrer que la nouvelle molécule est équivalente en terme d'efficacité au traitement de référence. Ces tests sont par exemple utilisés dans le cadre du developpement des médicaments génériques, ou de médicaments dit ≪me too≫. Ces médicaments ≪me too≫ sont comparables, à certains égards, à des médicaments qui existent déjà, mais offrent tout de même une valeur ajoutée dans le traitement des patients. Il s'agit surtout pour les industriels d'une stratégie pour ne pas voir leur médicament tomber dans le domaine publique.

Dans le cadre d'essai d'équivalence, Δ sera souvent noté Δ_E.

Soit $\mathcal{R}$ une zone pour laquelle la différence entre les deux traitements peut être considérée comme équivalente du point de vue clinique : $\mathcal{R} =]-\Delta, \Delta[$.

Du point de vue statistique, les hypothèses sont :

$$\mathcal{H}_0 : |\delta| \geq \Delta \quad versus \quad \mathcal{H}_1 : |\delta| < \Delta. \tag{2.1}$$

Pour tester ces hypothèses, la méthode la plus courante consiste à réaliser deux tests unilatéraux dont les hypothèses sont :

$$\mathcal{H}_0^1 : \delta \leq -\Delta \;\; versus \;\; \mathcal{H}_1^1 : \delta > -\Delta; \; et \; \mathcal{H}_0^2 : \delta \geq \Delta \;\; versus \;\; \mathcal{H}_1^2 : \delta < \Delta. \tag{2.2}$$

L'hypothèse nulle (2.1) sera rejetée au niveau α si et seulement si les deux hypothèses nulles définies en (2.2) sont chacune rejetées au niveau $\alpha/2$. Nous remarquerons que l'hypothèse nulle $\mathcal{H}_0^1$ est en fait celle de non-infériorité , alors que $\mathcal{H}_0^2$ est celle d'un test de supériorité.
Les hypothèses d'un test d'équivalence peuvent être représentées comme dans la Figure 2.4.

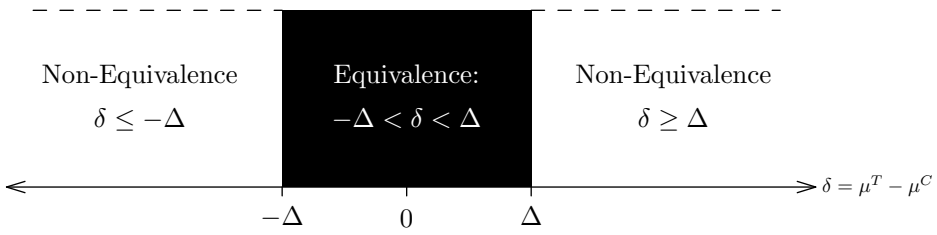


FIGURE 2.4: Le test d'équivalence

Remarques : Dans certains cas de figure, une dissymétrie dans les bornes de l'ensemble $\mathcal{R}$ peut être pertinente : $\mathcal{R} =]-\Delta_1, \Delta_2[$. Il est aussi important de noter que ces tests sont peu utilisés en recherche clinique, à l'exception de la cancérologie. Comme nous pouvons l'observer sur la Figure 2.4, si un traitement est en réalité meilleur que le traitement de référence, l'hypothèse nulle sera alors rejetée. C'est la raison pour laquelle, les tests de non-infériorité sont souvent privilégiés.

Si nous nous plaçons dans le cadre des intervalles de confiance, nous pouvons résumer les notions vues ci-dessus par la Figure 2.5.

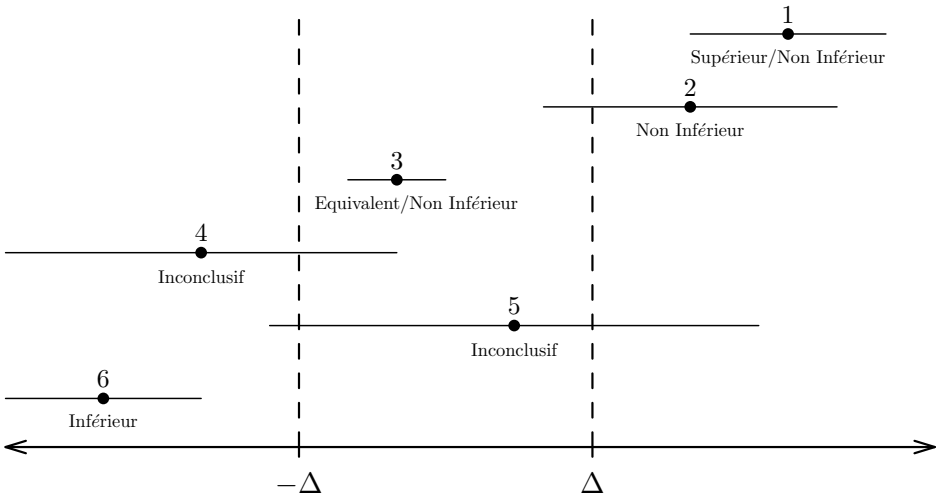


FIGURE 2.5: Différences de moyennes et leurs intervalles de confiance à 95% : notions de supériorité, d'équivalence et de non infériorité

2.1.4 Calcul de taille d'échantillon

Une fois la problématique et l'analyse définies par les scientifiques et les statisticiens, il convient de calculer le nombre de sujets qu'il est nécessaire d'inclure dans l'étude. L'objectif de ce calcul est de trouver le nombre de sujets optimal qui permettra de contrôler, avec des paramètres fixés à priori, une puissance suffisante à un seuil $1-\beta$. Pour illustrer le calcul de taille d'échantillon dans le cadre univarié, nous allons nous placer dans un contexte clinique. L'objectif de l'essai clinique consiste à déterminer s'il existe une différence sur la variable quantitative X entre le produit nouvellement développé (T : Traitement) et le produit de référence (C : Controle). La taille d'échantillon entre les deux groupes est supposée égale. Nous faisons l'hypothèse que la distribution de la variable X suit une loi normale pour chacun des deux groupes : $X_T \sim N(\mu^T, \sigma)$ et $X_C \sim N(\mu^C, \sigma)$. La variable δ est définie comme la différence de moyennes entre les deux groupes : $\delta = \mu^T - \mu^C$.

Les hypothèses individuelles sont donc de la forme :

$$\mathcal{H}_0 : \delta = 0 \; versus \; \mathcal{H}_1 : \delta \neq 0.$$

Soit la statistique de test Z définie par :

$$Z = \overline{X}^T - \overline{X}^C \sim \mathcal{N}\left(\mu^T - \mu^C, \frac{2\sigma^2}{n}\right), \; avec \begin{cases} \overline{X}^T : & \text{moyenne du groupe T;} \\ \overline{X}^C : & \text{moyenne du groupe C.} \end{cases}$$

La puissance est définie comme :

$$\mathbb{P}[\text{Rejeter } \mathcal{H}_0 | \mathcal{H}_1 \text{ est vraie}] = 1 - \beta.$$

Dans notre contexte, cette définition peut se traduire par :

$$1-\beta = \mathbb{P}\left[\left(-\infty < Z < -u_{1-\alpha/2}\frac{\sqrt{2}\sigma}{\sqrt{n}}\right) \;\middle|\; \mathcal{H}_1 \; vraie\right] + \mathbb{P}\left[\left(u_{1-\alpha/2}\frac{\sqrt{2}\sigma}{\sqrt{n}} < Z < +\infty\right) \;\middle|\; \mathcal{H}_1 \; vraie\right],$$

Comme le montre la Figure 2.1, nous pouvons considérer que l'une des deux probabilités est faible devant l'autre et elle peut donc être négligée dans la suite du calcul :

$$1 - \beta \approx \mathbb{P}\left[Z \geq u_{1-\alpha/2}\frac{\sqrt{2}\sigma}{\sqrt{n}} \;\middle|\; \mathcal{H}_1 \; est \; vraie\right],$$

Après avoir centré et réduit la statistique de test, la puissance peut être définie comme :

$$1 - \beta \approx \mathbb{P}\left[\left(\frac{Z - \delta^*}{\frac{\sqrt{2}\sigma}{\sqrt{n}}}\right) \geq \frac{u_{1-\alpha/2}(\frac{\sqrt{2}\sigma}{\sqrt{n}} - \delta^*)}{\frac{\sqrt{2}\sigma}{\sqrt{n}}} \;\middle|\; \mathcal{H}_1 \; est \; vraie\right],$$

où δ^* correspond à la valeur de δ sous $\mathcal{H}_1$. Il s'agit de l'effet que l'on cherche à détecter.
On en déduit que

$$u_{(1-\beta)} \approx u_{1-\alpha/2} - \left(\frac{\delta^*}{\frac{\sqrt{2}\sigma}{\sqrt{n}}}\right),$$

d'où

$$n \approx 2\left(\frac{\sigma}{\delta^*}\right)^2 (u_{1-\beta} + u_{1-\alpha/2})^2.$$

Pour trouver la taille d'échantillon, il reste à définir avec les scientifiques les valeurs de α, β, et de l'effet taille $\frac{\delta^*}{\sigma}$. La valeur de σ^2 (la variance vraie des groupes T et C) est généralement inconnue. Il est donc nécessaire de l'estimer par une valeur s^2 obtenue par l'avis d'experts ou à partir d'études antérieures. Du fait de l'estimation de ce paramètre, il est tout de même préférable, si cela est possible,

de prendre en compte son incertitude dans le calcul de la taille d'échantillon [Julious and Owen, 2006].

Remarque : Dans le cadre univarié, le calcul de taille d'échantillon en recherche clinique est bien expliqué dans l'ouvrage de Chow et al. [Chow et al., 2008].

2.2 Problématique des tests multiples

Dans la société actuelle, les médias font régulièrement part de découvertes sur d'éventuels facteurs de risque pour notre état de santé ou notre bien-être. Beaucoup de ces associations soulèvent néanmoins des questions quant à leur bien fondé au sein même de la communauté scientifique. On peut par exemple citer les associations entre les téléphones portables et le risque de tumeurs cérébrales [Khurana et al., 2009], ou encore l'association entre les performances cérébrales et la prise de vitamines [Peto, 1991, Whitehead, 1991]. Un grand nombre de ces associations sont fondées sur des preuves scientifiques a priori, mais certaines n'ont pas pu être confirmées par d'autres études. Ces informations contradictoires dans les médias ne confortent donc pas la confiance du grand public envers les résultats des études statistiques.

Cette contradiction quant aux résultats est souvent due au procédé d'inférence. En effet, lors d'une inférence statistique, on cherche à extrapoler les résultats obtenus au niveau de l'échantillon à la population cible. Comme nous travaillons sur un échantillon, le procédé fait intervenir une part d'aléatoire qui nécessite d'être contrôlée. C'est ce qui est fait avec le contrôle de l'erreur de Type-I dans un test statistique. Dans le cas d'un contrôle de α à 0.05, le risque de conclure à la significativité d'un test sur l'échantillon alors qu'elle n'existe pas sur la population cible est de 0.05. Le contrôle du risque α à 0.05 est communément accepté, mais il peut déjà expliquer l'existence de contradiction entre les études.

Cependant, le fait de multiplier les tests statistiques augmente la probabilité globale de se tromper quand on rejette $\mathcal{H}_0$ (risque α). Ce problème est illustré au sein de la Figure 2.6.

Afin d'illustrer ce problème de multiplicité des tests statistiques, nous pouvons prendre l'exemple bien connu de la roulette russe. Bien que l'analogie soit un peu brutale, elle a le mérite de souligner le risque encouru [Millot, 2009].

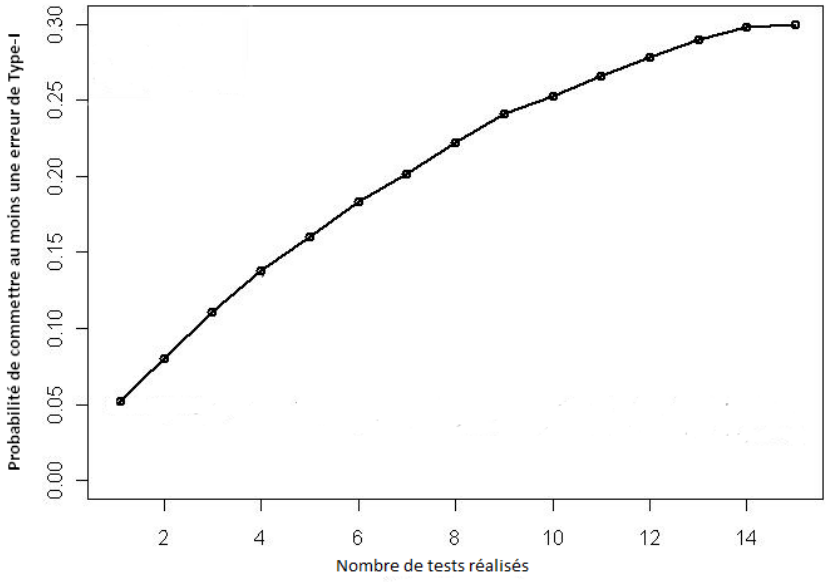


FIGURE 2.6: Inflation du risque de commettre au moins une erreur de Type-I

Posons le risque α à 0.05. Ce risque peut alors être assimilé au risque que le coup parte lors d'une partie de roulette russe avec un pistolet dont le barillet possède 20 emplacements avec une seule balle dedans. Ce risque est en effet de $\frac{1}{20} = 0.05$. Imaginons maintenant que l'expérience soit tentée plusieurs fois (à chaque tentative le barillet est tourné, et si le coup part on replace une balle dedans pour le tour suivant). Le risque encouru sur «l'ensemble du jeu» va alors augmenter, en suivant une loi binomiale $B(n, p)$, avec n correspondant au nombre d'essais et p la probabilité que le coup parte à chaque essai. La probabilité sera ici à titre d'exemple de 0.40 à 10 essais. Ce problème est similaire pour le risque α et le fait de multiplier les tests statistiques augmente la probabilité globale de rejeter une hypothèse nulle individuelle alors que celle ci est vraie. Cela montre bien que la problématique des tests multiples doit être prise en compte dans l'analyse des données afin de ne pas biaiser l'interprétation des résultats. Il est important de noter que ce problème est d'autant plus important que le nombre de tests et de comparaisons utilisés sont importants.

D'autres causes peuvent également expliquer les contradictions observées entre différentes études. Nous pouvons par exemple citer les problèmes liés aux conditions expérimentales (design non comparable, données manquantes, sélection des sujets, ...) ou encore aux limites actuelles du processus de publication (biais de publication, malédiction du vainqueur [Johnson, 2008], ...). Néanmoins l'ab-

sence de correction de la multiplicité en est l'une des raisons majeures.

En recheche clinique, le risque lié à la multiplicité peut être rencontré dans les cas présentés dans la Figure 2.7. L'objectif de l'essai va donc affecter notre choix quant à la procédure de gestion de tests multiples utilisée.

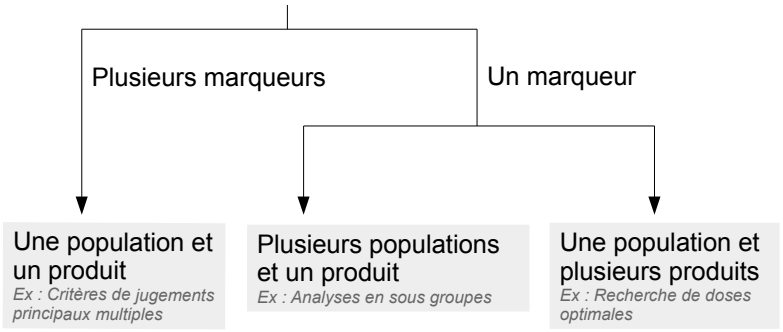


FIGURE 2.7: Multiplicité en recherche clinique

La suite de ce chapitre sera donc l'occasion de détailler les différentes notions inhérentes à la problématique des tests multiples, ainsi que les principales méthodes de corrections.

2.3 Tests multiples : taux d'erreurs

De nos jours, de nombreux projets de recherche ont pour objectif de répondre simultanément à plusieurs problématiques scientifiques. On peut cependant observer que beaucoup ignorent la multiplicité. En effet, les tests sont effectués en contrôlant l'erreur individuelle α. Ceci entraîne une augmentation non négligeable du taux de faux positifs, et donc un biais dans l'interprétation des résultats. Cette procédure sera par la suite définie comme ≪naïve≫.

Une analyse statistique rigoureuse de ces études, se doit de prendre en compte la multiplicité afin de contrôler le taux de faux positifs tout en maximisant la puissance. Afin d'avoir une interprétation plus réaliste, il est donc nécessaire, lorsque l'on réalise plusieurs tests, de contrôler le risque de commettre au moins une erreur de Type-I. Plusieurs approches existent pour contrôler ce risque, et l'objectif de cette section sera de présenter celles qui sont les plus utilisées dans la littérature. Un point sur les

erreurs de Type II et III sera également présenté.

2.3.1 Prise de décision

Comme nous l'avons vu précédemment, il existe quatre situations quant à l'issue du test d'une hypothèse nulle : on peut décider de rejeter ou non $\mathcal{H}_0$, alors qu'en réalité (mais nous ne le savons pas) $\mathcal{H}_0$ est vraie ou fausse. Il existe donc une incertitude quant à la décision à prendre, et l'enjeu sera alors de se convaincre que l'on prend la bonne décision en contrôlant le risque de se tromper.

Dans le cadre des tests multiples, sur les m hypothèses nulles testées, m_0 hypothèses nulles sont vraies et m_1 sont fausses. Les tests réalisés sur ces hypothèses, concluent à R rejets et W non rejets des hypothèses.

Parmi les m_0 vraies hypothèses nulles testées, U ne sont pas rejetées et V le sont. Nous pouvons donc considérer qu'il existe V faux positifs. De même, parmi les m_1 fausses hypothèses nulles testées, T ne sont pas rejetées et S le sont. T peut donc être considéré comme le nombre de faux négatifs. En pratique, m_0 et m_1 ne sont pas connus, seuls R et W le sont.

L'ensemble $\mathcal{M}$ définit l'ensemble des hypothèses nulles testées ($\mathcal{M} = \{1, \ldots, m\}$), $\mathcal{M}_0$ est défini comme l'ensemble des vraies hypothèses nulles ($\mathcal{M}_0 = \{1, \ldots, m_0\}$), et $\mathcal{M}_1$ quant à lui est défini comme l'ensemble des fausses hypothèses nulles ($\mathcal{M}_1 = \{1, \ldots, m_1\}$).

Le tableau 2.2 aide à la visualisation des différents scénarios possibles dans le cadre du test des m hypothèses.

TABLE 2.2: Scénarios possibles en présence de m hypothèses

Hypothèses nulles	Non Rejetées	Rejetées	Total
Vraies	U	V	m_0
Fausses	T	S	m_1
Total	W	R	m

2.3.2 Erreur de Type-I

Dans le cas d'un test unique, l'erreur de Type-I est défini par α. Lors de tests multiples, il est nécessaire de généraliser cette définition.
Cette généralisation de l'erreur de Type-I à la problématique des tests multiples a donné lieu à de nombreuses approches, dont les plus connues et les plus couramment utilisées seront présentées dans cette section.

2.3.2.1 Taux d'erreurs par comparaison : PCER

Le PCER pour ≪*Per Comparison Error Rate*≫ en anglais, peut être traduit comme le taux d'erreurs par comparaison. Il est défini comme l'espérance du taux d'erreurs de Type-I attendues parmi les m hypothèses :

$$PCER = \frac{\mathbb{E}(V)}{m}.$$

Si chacune des m hypothèses est testée séparément à un niveau pré-spécifié α, le PCER est alors défini comme : $PCER = \frac{\alpha m_0}{m} \leq \alpha$. Ce taux ne prend pas compte de la multiplicité des tests. C'est la raison pour laquelle dans la plupart des applications le contrôle du PCER n'est pas considéré comme adéquat.

2.3.2.2 Taux d'erreurs par famille : PFER

Une famille d'hypothèses est un ensemble d'hypothèses considérées conjointement. Le taux d'erreurs par famille (PFER : *Per Family Error Rate*), n'est pas une probabilité comme la majorité des taux d'erreurs, mais il représente le nombre attendu d'erreur par famille d'hypothèses. En d'autres termes, il correspond à l'espérance du nombre d'erreurs de Type-I fait dans une famille d'hypothèses. Sur un ensemble de m hypothèses testées conjointement on peut donc définir le PFER comme suit :

$$PFER = \mathbb{E}(V).$$

Nous voyons assez facilement la relation entre le PFER et PCER : $PFER = m \times PCER$.

2.3.2.3 FamilyWise Error Rate : FWER

Historiquement, le FWER est le taux d'erreurs le plus utilisé dans les tests multiples. Il est d'autant plus utilisé quand le nombre de comparaisons est modéré, et/ou lorsque des preuves évidentes sont nécessaires.

Pour comprendre le *FamilyWise Error Rate* (FWER), il faut aussi considérer l'ensemble des m hypothèses testées conjointement comme une famille. Le FWER sera alors défini comme la probabilité de commettre au moins une erreur de Type-I :

$$FWER = \mathbb{P}(V > 0).$$

Les procédures qui contrôlent le PFER sont plus conservatrices que celles qui contrôlent le FWER, qui sont elles mêmes plus conservatrices que celles qui contrôlent le PCER [Dudoit and Van der Laan, 2008, p. 24].

Démonstration 1.

Soit $\mathcal{H}_0$ l'hypothèse nulle globale ($\mathcal{H}_0 = \bigcap_{j \in \mathcal{M}} \mathcal{H}_0^j$). Supposons dans ce contexte, que toute les hypothèses nulles $\mathcal{H}_0^j$ soient testées au risque exact de première espèce individuel α_j. La décision de rejeter cette hypothèse ne dépendra donc que de la statistique de test correspondante T_j.

Dans ces conditions, le PCER correspond à la moyenne des risques de première espèce individuels α_j, et le PFER à la somme des risques de première espèce individuels. Le FWER quant à lui ne peut pas être exprimé seulement avec les α_j, en effet il dépend aussi de la distribution conjointe des statistiques de test individuelles T_j.

Nous obtenons donc [Dudoit and Van der Laan, 2008, p. 24] :

$$\begin{aligned} PCER &= \frac{1}{m}(\alpha_1 + \ldots + \alpha_m) \\ &\leq \texttt{max}\{\alpha_1, \ldots, \alpha_m\} \\ &\leq FWER \\ &\leq \alpha_1 + \ldots + \alpha_m = PFER \end{aligned}$$

2.3.2.4 Generalized FamilyWise Error Rate : gFWER

Quand le nombre d'hypothèses testées est trop important, le contrôle du FWER peut s'avérer trop strict. Dans ce cas, une extension consiste à contrôler la probabilité de commettre au moins q erreurs de Type-I. Si le nombre d'hypothèses testées m est suffisamment important, alors le contrôle du taux d'erreurs avec une faible valeur q est acceptable [Bretz et al., 2010]. Le gFWER est donc défini [Victor, 1982, Hommel and Hoffmann, 1988, Lehmann and Romano, 2005] comme suit :

$$gFWER = \mathbb{P}(V \geq q), q \in \mathcal{M},$$

Ce taux n'est donc utilisé que lorsqu'il est acceptable de commettre plus d'une erreur de Type-I. En pratique, on lui préfère le FWER car on cherche le plus souvent à minimiser les erreurs commises.

2.3.2.5 False Discovery Rate : FDR

Le *False Discovery Rate* (FDR) a gagné en popularité ces dernières années avec l'avènement des analyses en grande dimension (*e.g.* analyses de données -omiques). Il est défini comme l'espérance de la proportion d'erreurs de Type-I parmi les hypothèses rejetées :

$$\begin{aligned} FDR &= \mathbb{E}\left(\frac{V}{R}\right) \\ &= \mathbb{E}\left(\frac{V}{R}\middle| R>0\right)\mathbb{P}(R>0) + 0\cdot\mathbb{P}(R=0) \\ &= \mathbb{E}\left(\frac{V}{R}\middle| R>0\right)\mathbb{P}(R>0). \end{aligned}$$

Cette approche du contrôle de l'erreur de Type-I est introduite par Benjamini & Hochberg en 1995 [Benjamini and Hochberg, 1995]. Les plus anciennes références au FDR sont présentes dans les travaux précurseurs de Seeger en 1968 [Seeger, 1968] et de Sorić en 1989 [Soric, 1989]. La proportion d'erreurs de Type-I parmi les hypothèses rejetées est défini dans la littérature par l'anglicisme : *False Discovery Proportion (FDP)* Ce taux est souvent utilisé en phase exploratoire (*e.g.* analyse des effets indésirables, analyse de données -omiques), quand le nombre d'hypothèses testées est trop important, et que le contrôle du FWER semble trop strict. Quand l'ensemble des hypothèses nulles sont vraies, le fait de contrôler le FDR revient à contrôler le FWER. Cette démonstration, bien connue, est redéfinie ci-dessous [Foulkes, 2009] :

Démonstration 2.

Supposons que toutes les hypothèses nulles soient vraies. Dans ce cas, on a $V = R$ et :

$$V/R = \begin{cases} 0 \; si \; V = 0 \\ 1 \; si \; V \geq 1 \end{cases}$$

$$\begin{aligned} \mathbb{E}(V/R) &= 0\cdot\mathbb{P}(V=0) + 1\cdot\mathbb{P}(V\geq 1) \\ &= \mathbb{P}(V\geq 1) \\ &= FWER \end{aligned}$$

De manière plus générale, toute procédure qui contrôle le FWER, contrôle aussi le FDR.

2.3.2.6 Positive False Discovery Rate : pFDR

Storey en 2003 [Storey, 2003] propose la notion de pFDR (*positive False Discovery Rate*). Il s'agit en fait du FDR conditionné au fait qu'au moins une découverte positive est réalisée. On le définit comme suit :

$$pFDR = \mathbb{E}(V/R|R > 0).$$

Le pFDR est utilisé lorsque le nombre d'hypothèses nulles m est très grand. Tout comme le FDR, on le retrouve le plus souvent lorsqu'on analyse des données de grandes dimensions. Il est cependant important de noter que le pFDR n'a pas été retenu par Benjamini et Hochberg dans leur article princeps en 1995 [Benjamini and Hochberg, 1995], car celui-ci n'est pas contrôlable. En effet, sous l'hypothèse nulle globale, tous les positifs sont par définition de faux positifs, et le pFDR est donc égal à 1.

2.3.2.7 Tail Probability for the Proportion of False Positives : TPPFP

Nous pouvons aussi citer comme alternative au FDR, pour l'analyse de grande dimensions le TPPFP (*Tail Probability for the Proportion of False Positives* [Dudoit and Van der Laan, 2008, p. 20]. Cette approche du contrôle de l'erreur de Type-I est définie comme la probabilité que le taux de faux positifs parmi les hypothèses nulles rejetées soit supérieur à une constante $q \in (0, 1)$ définie par l'utilisateur :

$$TPPFP = \mathbb{P}\left(\frac{V}{R} > q\right).$$

2.3.2.8 Contrôle fort et faible du taux d'erreurs de Type-I

Les définitions des taux d'erreurs sont conditionnelles au nombre d'hypothèses vraies. On retrouve donc dans la littérature deux contrôles de ces risques ; le contrôle faible (*weak*), et le contrôle fort (*strong*) [Bretz et al., 2010, Westfall et al., 2011, Hochberg and Tamhane, 1987]. On parle de contrôle faible de ces taux, si l'erreur de Type-I est contrôlée seulement sous l'hypothèse nulle globale :

$$\mathcal{H}_0 = \bigcap_{j \in \mathcal{M}_0} \mathcal{H}_0^j, \ \mathcal{M}_0 = \mathcal{M},$$

où l'on considère que l'ensemble des hypothèses nulles, $\{\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m\}$ sont vraies. On obtient donc dans le cadre du contrôle faible du FWER :

$$\mathbb{P}(V > 0 | \mathcal{H}_0) \leq \alpha.$$

Considérons une étude dont l'objectif principal est de prouver l'efficacité d'un traitement sur un ensemble de critères de jugement. Le contrôle faible du FWER dans cette étude impliquera donc de contrôler la probabilité de déclarer un effet sur au moins un des critères de jugement, alors qu'il n'y a en fait aucun effet sur aucune variable. Nous avons aussi pu voir à travers la démonstration 2 que le contrôle du FDR, entraîne un contrôle faible du FWER.

Cependant, en pratique il est très rare que toutes les hypothèses nulles soient vraies, et l'hypothèse globale $\mathcal{H}_0$ est rarement vérifiée. C'est pourquoi il est souvent nécessaire d'utiliser un contrôle fort du taux d'erreurs, qui demande moins de conditions à vérifier. On définit alors le contrôle fort du taux d'erreurs de Type-I (*strong*). Cela correspond au fait que le taux d'erreurs de Type-I est contrôlé pour n'importe quelle configuration de vraies ou fausses hypothèses nulles. Dans le cas du FWER, le contrôle fort est défini comme :

$$\max_{\mathcal{J} \subseteq \mathcal{M}} \mathbb{P}\left(V > 0 \middle| \bigcap_{j \in \mathcal{J}} \mathcal{H}_0^j \right) \leq \alpha,$$

où le maximum est atteint sous toutes les configurations possibles d'hypothèses nulles, $\varnothing \neq \mathcal{J} \subseteq \mathcal{M}$. Si l'on reprend l'exemple précédent, le contrôle fort du FWER, implique de contrôler la probabilité de déclarer un effet sur au moins un critère de jugement, indépendamment de l'effet de chacune d'entre elles. Il est donc souhaitable d'utiliser une procédure qui a un contrôle fort du taux d'erreurs quand on ne connaît pas l'ensemble des vraies hypothèses nulles.

Remarque : Il est important de noter qu'en pratique la véracité ou non des hypothèses testées n'est pas vérifiable.

2.3.2.9 Choix du taux d'erreurs de Type-I

La relation entre les différents taux d'erreurs, peut être définie comme suit [Dudoit and Van der Laan, 2008, p. 24] :

$$PCER \leq FDR \leq FWER \leq PFER.$$

Une procédure qui contrôle le PFER contrôle donc le FWER, le FDR et le PCER.

Le PCER et le PFER sont des critères controversés [Hochberg and Tamhane, 1987] et peu utilisés, nous ne discuterons donc par la suite que du FWER et du FDR.

Historiquement, les premières procédures de tests multiples qui visent à contrôler l'erreur de Type-I se sont intéressées au contrôle du FWER [Simes, 1986, Dunn, 1958, Dunn, 1961].

De nos jours, les études dont l'objectif principal est décisionnel auront tendance à choisir une procédure qui contrôle le FWER. En effet, dans ce genre d'étude le moindre faux positif peut être lourd de conséquence. Il est donc important de choisir le FWER, dont l'intérêt réside dans un contrôle strict du nombre de faux positifs. Par exemple, contrôler le FWER au seuil de 5% permet d'être confiant à 95% de n'avoir aucun faux positif. Le FWER est à privilégier dans les études à but décisionnel, ou lorsque le nombre de tests n'est pas trop important.

Cependant, dans des analyses exploratoires pour lesquelles il est permis d'avoir quelques faux positifs, le contrôle du FWER peut paraître trop contraignant. Il est donc préféré, dans ces analyses, un contrôle du FDR (ou du pFDR).

Ces dernières années, avec l'avènement des biotechnologies et des analyses génétiques en grandes dimensions, ce taux d'erreurs est de plus en plus utilisé et par conséquent de plus en plus présent dans la littérature. De nombreux travaux portent d'ailleurs sur l'amélioration et la généralisation du FDR [Benjamini and Yekutieli, 2001, Genovese and Wasserman, 2002, Storey, 2003, Sun and Cai, 2007, Tang and Zhang, 2007, Efron, 2008, Sarkar et al., 2008, Roquain and van de Wiel, 2009, Chen et al., 2009].

Néanmoins, ces deux approches ne sont pas forcément à opposer et peuvent même être complémentaires. Ainsi, lorsque l'on va choisir après une étude exploratoire (contrôle du FDR), de lancer une étude confirmatoire, on va privilégier dans la seconde étude un contrôle du FWER permettant un contrôle plus strict du nombre de faux positifs.

2.3.3 Erreur de Type-II

Il est nécessaire, dans toutes procédures de tests, de maximiser la puissance, et par la même occasion de minimiser l'erreur de Type-II, pour une erreur de Type-I prédéfinie. Le contrôle de la puissance est une partie non négligeable de la mise en place d'une étude scientifique, et en particulier dans le calcul du nombre de sujets. C'est en effet le nombre d'individus inclus dans une étude qui va permettre de contrôler la puissance et donc l'erreur de Type-II. Comme pour le taux d'erreurs de Type-I, la généralisation de la puissance aux tests multiples peut être vue sous différents angles. L'objectif de cette section est donc de présenter les définitions de la puissance les plus rencontrées dans la littérature [Maurer and Mellein, 1988, Westfall et al., 2011, Senn and Bretz, 2007].

2.3.3.1 Puissance individuelle

La puissance individuelle (*individual power*), comme son nom l'indique, correspond au concept de puissance lorsque l'on teste une seule hypothèse. Elle est définie comme suit :

$$\pi_{ind}^{j} = \mathbb{P}(rejeter\ \mathcal{H}_0^j |\ \mathcal{H}_1^j\ est\ vraie),$$

Cette puissance est aussi appelée puissance par paire (*per-pair power*).

2.3.3.2 Puissance moyenne

La puissance moyenne (*average power*), est directement liée au concept de puissance individuelle, puisqu'il s'agit de la moyenne du nombre attendu d'hypothèses nulles rejetées avec raison.

$$\pi_{ave} = \frac{\mathbb{E}(S)}{m_1} = \frac{1}{m_1} \sum_{j \in \mathcal{M}_1} \pi_{ind}^{j}.$$

où $m_1 = m - m_0$ correspond au nombre de fausses hypothèses nulles.

2.3.3.3 Puissance disjonctive

La puissance disjonctive (*disjunctive power*) correspond quant à elle, à la probabilité de rejeter au moins une fausse hypothèse nulle :

$$\pi_{dis} = \mathbb{P}(S \geq 1).$$

Cette puissance peut aussi être appelée puissance minimale (*minimal power*) ou puissance de n'importe quelle paire (*any pair power*).

2.3.3.4 Puissance conjonctive

A contrario, il existe la puissance conjonctive (*conjunctive power*) qui est définie comme la probabilité de rejeter toutes les fausses hypothèses nulles :

$$\pi_{con} = \mathbb{P}(S = m_1).$$

Elle est aussi appelée puissance complète (*complete power*), totale (*totale power*), ou encore puissance de toutes les paires (*all-pair power*).

2.3.3.5 True Discovery Rate : TDR

Le taux de vraies découvertes (*TDR : True Discovery Rate*) peut être défini comme l'espérance des vrais négatifs parmi les hypothèses nulles testées [Dudoit and Van der Laan, 2008, p. 23] :

$$TDR = \mathbb{E}\left(\frac{S}{R}\right) = \mathbb{E}\left(\frac{R-V}{R}\right).$$

De plus, si on pose que quand $R = 0$, $(R - V)/R \equiv 0$, alors le TDR peut être réécrit comme suit :

$$TDR = \mathbb{E}\left(\frac{R-V}{R}\middle| R > 0\right)\mathbb{P}(R > 0) = \mathbb{P}(R > 0) - FDR.$$

Ce taux de contrôle de la puissance est souvent associé aux procédures qui contrôlent le taux d'erreurs de Type-I par le FDR. Si toutes les hypothèses nulles sont fausses (*i.e.*, $m_0 = 0$), alors le TDR est égal à la puissance disjonctive.

Remarque : De façon similaire le taux d'erreurs de Type-II peut être mesuré par le NDR (*Non-Discovery Rate*) qui est défini comme l'espérance du taux de faux négatifs parmi les hypothèses alternatives [Craiu and Sun, 2008] : $NDR = \mathbb{E}\left(\frac{T}{m_1}\right)$. Mais aussi par le FNR (*False Non-discovery Rate*) qui est défini comme l'espérance de la proportion d'erreur de Type-II parmi l'ensemble des hypothèses nulles non rejetées [Genovese and Wasserman, 2002] : $FNR = \mathbb{E}\left(\frac{T}{W}\right)$.

2.3.3.6 r power

La définition générale de la puissance est donnée par Dunnett & Tamhane [Dunnett and Tamhane, 1992] :

$$\pi_r = \mathbb{P}(S \geq r).$$

Elle peut être vue comme étant la probabilité de rejeter au moins r fausses hypothèses nulles.

2.3.3.7 Choix d'un taux d'erreurs de Type-II

Une question pertinente consiste à savoir quelle est la définition de la puissance la plus appropriée à une étude donnée. Si l'objectif principal de l'étude est d'observer tous les effets existants, alors il faut utiliser la puissance conjonctive. Si l'objectif de l'étude est plus ouvert, et que l'on cherche à trouver au moins un effet sur l'ensemble des effets réels, alors notre choix se portera plus sur la puissance disjonctive. La ≪r-power≫ quant à elle, nous permet une solution alternative. En effet, l'objectif de l'étude peut être de prouver un effet sur r tests parmi les m_1 fausses hypothèses nulles. La puissance moyenne est moins utilisée, mais peut servir pour comparer des procédures de tests multiples par exemple. Enfin, l'utilisation du TDR est quant à elle souvent associée à l'utilisation du FDR dans le contrôle de l'erreur de Type-I, *i.e.* principalement dans des analyses à but exploratoire.

2.3.4 Erreur de Type-III

Dans un test bilatéral, quand on rejette l'hypothèse nulle, on recherche souvent à avoir une conclusion sur le signe de l'effet qui vient d'être trouvé. Cependant, une telle conclusion nécessite d'une part un contrôle de l'erreur de Type-I, mais aussi un contrôle du signe de l'effet non nul. L'erreur de Type-III (aussi connu sous le nom d'erreur directionnelle (*directional error*)) est donc définie comme : $DE = \mathbb{P}(\mathcal{A}_2)$, où $\mathcal{A}_2$ est l'évènement défini comme le fait de réaliser au moins une erreur de signe parmi les véritables effets non nuls.

Mais ce qui nous interesse vraiment en pratique c'est le principe d'erreur combinée notée CER pour ≪Combined Error Rate≫. Il est défini comme suit : $CER = \mathbb{P}(\mathcal{A}_1 \cup \mathcal{A}_2)$, où $\mathcal{A}_1$ est défini comme l'évènement qu'au moins une erreur de Type-I existe.

2.4 Rappels et procédures de gestion des tests multiples

Dans les années 1930, les premières procédures de gestion des tests multiples sont proposées par Fisher [Hochberg and Tamhane, 1987, Rafter et al., 2002] pour les tests simultanés de plusieurs contrastes dans le cadre du modèle linéaire d'analyse de la variance.
Depuis, de nombreux concepts et de nombreuses procédures ont été proposés. Toutes correspondent à une problématique donnée, et leur objectif est de minimiser le nombre de faux-négatifs (erreurs de Type-II), tout en contrôlant le nombre de faux-positifs (erreurs de Type-I).
Nous nous intéresserons dans cette section aux différents concepts, et aux différentes procédures permettant la gestion des tests multiples. Les exemples d'application seront principalement orientés vers le secteur biomédical, où les applications sont très nombreuses. C'est par exemple le cas en épidémiologie (*e.g.* recherche d'associations entre une pathologie et plusieurs facteurs de risques), en génétique (*e.g.* comparaison des profils génétiques), ou encore dans l'analyse des essais cliniques (*e.g.* comparaison de plusieurs groupes de traitements à un traitement de référence).

Pvaleur ajustée

Le principe de p_{valeur} ajustée consiste à généraliser la p_{valeur} marginale au champ des tests multiples, tout en gardant le fait qu'elle puisse être comparable au même seuil α. Cette mesure est appelée p_{valeur} ajustée, et elle est définie comme le plus petit taux d'erreurs de Type-I pour lequel on rejette l'hypothèse nulle $\mathcal{H}_0^j,\ j \in \mathcal{M}$.
Dans le cadre du contrôle du FWER, on obtient [Westfall and Young, 1993, Wright, 1992, Bretz et al., 2010] :

$$q_j = \texttt{inf}\left\{\alpha \in (0,1)|\ \mathcal{H}_0^j \text{ est rejetée } \text{ au niveau } \alpha\right\}, \text{ si un tel } \alpha \text{ existe, sinon } q_j = 1.$$

Dans le cas, où $q_j \leq \alpha$ l'hypothèse nulle individuelle est rejetée, tout en contrôlant le FWER au niveau α. La p_{valeur} marginale (p_j) sera par la suite appelée p_{valeur} non ajustée.

2.4.1 Méthode de construction des procédures de tests multiples

Cette section a pour objectif d'expliquer les principales méthodes qui permettent la construction des procédures de tests multiples.

2.4.1.1 Test d'Union Intersection

Historiquement, il s'agit de la première méthode de construction de procédures de comparaisons multiples qui a été mise en place [Roy, 1953, Roy and Bose, 1953].

Elle permet de gérer les tests multiples dans une étude dont l'objectif est d'observer au moins un effet sur une des variables testées. Cette méthode contrôle la puissance disjonctive.

Nous définissons une famille d'hypothèses nulles $\mathcal{H}_0^j$, $j \in \mathcal{M}$. Celles-ci sont associées aux hypothèses alternatives $\mathcal{H}_1^j$. L'objectif de la méthode sera alors de tester l'hypothèse nulle globale $\mathcal{H}_0 = \bigcap_{j\in\mathcal{M}} \mathcal{H}_0^j$. L'approche consiste à utiliser les statistiques de test T_j, et à rejeter $\mathcal{H}_0$ si au moins un T_j est supérieur à la valeur critique associée s_j. La région de rejet sera alors une union des régions de rejet, $\bigcup_{j\in\mathcal{M}}\{T_j > s_j\}$, donnant lieu au terme «union» de la procédure [Bretz et al., 2010, p.21]. Les tests d'Union Intersection testent l'intersection des hypothèses nulles versus l'union des hypothèses alternatives, cela peut donc se traduire par :

$$\mathcal{H}_0 = \bigcap_{j\in\mathcal{M}} \mathcal{H}_0^j \; versus \; \mathcal{H}_1 = \bigcup_{j\in\mathcal{M}} \mathcal{H}_1^j.$$

Cette méthode teste globalement l'ensemble des hypothèses nulles, sans formellement permettre une évaluation des hypothèses nulles individuelles. Cette limite peut toutefois être détournée en appliquant un principe de fermeture (*closure principle*), ou par la construction d'intervalles de confiance simultanés [Lehmann, 1986, p.90] [Dmitrienko et al., 2009, p.47-48].

Une notion très importante dans les tests d'Union Intersection est celle du *max-t test*, puisqu'il s'agit d'une classe de tests d'Union Intersection [Bretz et al., 2010, p.21]. Considérons un ensemble de statistiques de test individuel $\{T_1, \ldots, T_m\}$ associé à l'ensemble des hypothèses nulles $\{\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m\}$. Une approche intuitive consiste alors à étudier le maximum des statistiques individuelles. Cette notion est définie dans le cadre de tests unilatéraux comme :

$$T_{max} = \texttt{max}\{T_1, \ldots, T_m\}.$$

L'hypothèse nulle globale $\mathcal{H}_0$ est rejetée si et seulement si $T_{max} \geq s$. La valeur critique s est déterminée pour un contrôle du risque de première espèce α : $\mathbb{P}(T_{max} \geq s|\mathcal{H}_0) = \alpha$. La constante s est donc calculée à partir de la distribution conjointe des variables aléatoires $\{T_1, \ldots, T_m\}$. Dans de nombreux cas, la fonction de distribution conjointe des statistiques de test est impossible à obtenir.

Le choix de la procédure de gestion des tests multiples se porte alors sur des méthodes plus conservatrices. Dans le cas des tests bilatéraux, le maximum des tests est défini en utilisant les valeurs absolues des valeurs des statistiques de test : $T_{max} = \max\{|T_1|, \ldots, |T_m|\}$. Il est aussi important de noter que cette notion peut aussi être vue sous l'angle des $p_{valeurs}$. On utilisera dans ce cas le minimum des $p_{valeurs}$ (*min* p_{value}) qui est surtout intéressant quand toutes les statistiques de test ne suivent pas la même loi [Westfall, 2005].

Remarques : En 1969, Gabriel a montré que le fait d'appliquer le principe du *max-t test*, amène à utiliser des procédures cohérentes et consonantes (se référer au principe de ≪closed testing≫ pour une définition de ces notions).
Beaucoup de procédures très connues utilisent ce principe dans leur construction, c'est par exemple le cas de la procédure de Bonferroni, de Dunnett et de Tukey.

2.4.1.2 Test d'Intersection Union

A l'inverse du principe d'Union Intersection, ce principe a pour objectif de gérer la multiplicité sur des études dont le but est de déterminer un effet sur l'ensemble des variables testées. Nous pouvons par exemple citer des essais de Phase III sur l'arthrose, où pour conclure à l'efficacité d'un produit il est nécessaire de prouver une différence significative sur l'ensemble des trois critères standards : évaluation globale du patient, Score WOMAC (index de sévérité symptomatique de l'arthrose sur les membres inférieurs) sur la douleur, Score WOMAC sur la condition physique [Julious and McIntyre, 2012]. L'efficacité du produit ne sera conclue que si toutes les hypothèses nulles individuelles sont rejetées. Cette méthode contrôle la puissance conjonctive. Dans le cadre général, les hypothèses globales sont définies comme suit :

$$\mathcal{H}_0 = \bigcup_{j \in \mathcal{M}} \mathcal{H}_0^j \; versus \; \mathcal{H}_1 = \bigcap_{j \in \mathcal{M}} \mathcal{H}_1^j.$$

On pourra rejeter l'hypothèse nulle globale à un niveau α, si toutes les hypothèses nulles individuelles sont rejetées à un niveau individuel α' [Berger, 1982]. Comme pour la procédure individuelle, la notion de *min-t test* peut être envisagée [Laska and Meisner, 1989]. En effet, si toutes les statistiques de test individuel $T_j,\ j \in \mathcal{M}$, ont la même distribution marginale, alors le principe des tests d'Intersection

Union rejettera l'hypothèse nulle globale si et seulement si : $T_{min} \geq s$, où $T_{min} = \texttt{min}\{T_1, \ldots, T_m\}$, et s est le quantile d'ordre $(1-\alpha)$ de la distribution marginale.

2.4.1.3 Principe de ≪closed testing≫

Lorsque nous utilisons un test d'Union Intersection, l'objectif est de tester l'hypothèse globale $\mathcal{H}_0$, mais il ne permet en aucun cas de conclure quant à une hypothèse individuelle. Le principe de ≪closed testing≫ aussi défini par l'anglicisme *closure principle* est introduit par Marcus et al. en 1976 [Marcus et al., 1976]. Cette méthode a pour objectif de pallier ce problème. Les auteurs ont basé ce principe sur une procédure séquentielle, dont le but est d'obtenir une conclusion individuelle pour chaque hypothèse nulle $\mathcal{H}_0^j$.

Soit un ensemble de m hypothèses nulles individuelles $\mathcal{H}_0^j$ $j \in \mathcal{M}$. Le principe de ≪closed testing≫ consiste à prendre en compte l'ensemble des intersections construites à partir des hypothèses nulles individuelles. Chaque niveau d'intersection est testé à un niveau local α. Pour l'inférence finale, l'hypothèse nulle individuelle ne peut être rejetée que si toutes les hypothèses faisant intervenir une intersection comprenant l'hypothèse individuelle sont aussi rejetées. Il a été montré que cette procédure contrôle fortement le FWER [Marcus et al., 1976].

Afin d'illustrer le principe d'intersection des hypothèses, reprenons les exemples présents dans l'ouvrage de Bretz et al. [Bretz et al., 2010]. Intéressons nous, dans un premier temps, au test de trois hypothèses nulles. La Figure 2.8 représente les hypothèses individuelles, $\mathcal{H}_0^j, j = 1, 2, 3$ et leurs intersections à l'aide d'un diagramme de Venn. Dans cet exemple, les hypothèses individuelles sont jointes. Il est donc important de tenir compte, dans la procédure de tests, de l'intersection entre les hypothèses.

La Figure 2.9 illustre quant à elle le principe de ≪closed testing≫ dans ce contexte.

A partir de ce diagramme, il est assez simple de construire et d'appliquer cette procédure. Il permet de visualiser les niveaux de tests (de familles), ainsi que les dépendances entre les hypothèses. Dans cet exemple on peut observer trois familles d'hypothèses qui sont chacune testées au niveau local α. La procédure commence par tester, au niveau local α, l'hypothèse globale $\mathcal{H}_0^{123} = \mathcal{H}_0^1 \cap \mathcal{H}_0^2 \cap \mathcal{H}_0^3$. Si celle-ci est rejetée alors on peut passer à l'étape suivante, sinon on arrête la procédure.

L'étape 2 consiste alors à tester individuellement les hypothèses de la deuxième famille : $\mathcal{H}_0^{12}$, $\mathcal{H}_0^{13}$

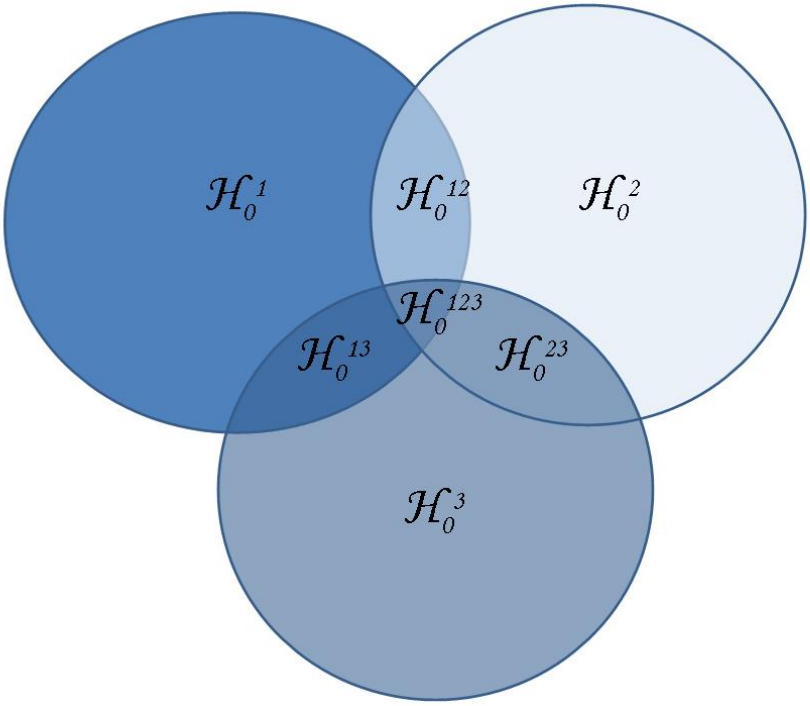


FIGURE 2.8: Représentation des trois hypothèses nulles $\mathcal{H}_0^1$, $\mathcal{H}_0^2$ et $\mathcal{H}_0^3$, ainsi que de leurs intersections en utilisant un diagramme de Venn

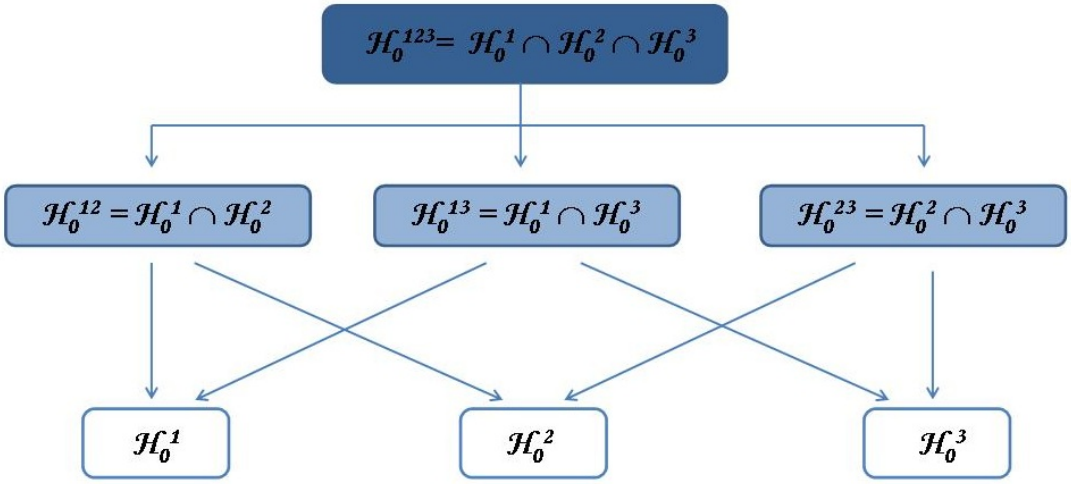


FIGURE 2.9: Représentation par un diagramme du principe de «closed testing» pour trois hypothèses nulles $\mathcal{H}_0^1$, $\mathcal{H}_0^2$ et $\mathcal{H}_0^3$, ainsi que leurs intersections

et $\mathcal{H}_0^{23}$ au niveau local α. Si les hypothèses ne sont pas rejetées alors il n'est pas nécessaire de tester les hypothèses de niveau inférieur qui lui sont reliées. En effet, l'hypothèse individuelle $\mathcal{H}_0^1$ ne pourra être rejetée que si $\mathcal{H}_0^{123}$, $\mathcal{H}_0^{12}$, $\mathcal{H}_0^{13}$, et $\mathcal{H}_0^1$ sont toutes rejetées. Il est important de noter que les hypothèses individuelles $\mathcal{H}_0^1$, $\mathcal{H}_0^2$, et $\mathcal{H}_0^3$ sont aussi considérées comme une famille, et qu'elles sont testées à un niveau local α.

De nombreuses procédures de gestion de tests multiples utilisent ce principe, comme par exemple la procédure séquentielle descendante de Holm [Holm, 1979] ou les procédures dites de *Gatekeeping* [Bauer et al., 1998, Westfall and Krishen, 2001, Dmitrienko et al., 2003].

2.4.1.4 Cohérence et Consonance :

Une des propriétés essentielles d'une procédure de gestion des tests multiples est la cohérence (≪*coherence*≫). Les procédures de gestion des tests multiples sont définies comme cohérentes si elles répondent à la propriété suivante : si $\mathcal{H}_0^i \subseteq \mathcal{H}_0^j$ et $\mathcal{H}_0^j$ est rejetée, alors $\mathcal{H}_0^i$ sera aussi rejetée [Gabriel, 1969]. Les procédures qui se basent sur le principe de ≪closed testing≫ sont par construction cohérentes.

Une autre propriété souhaitée pour une procédure de gestion des tests multiples est la consonance [Bretz et al., 2010]. Soit $\mathcal{H}_0^I = \cap_{i \in I} \mathcal{H}_0^i$ qui représente l'intersection des hypothèses nulles pour un ensemble d'indices tel que $I \subseteq \mathcal{M}$. De plus, on considère une hypothèse $\mathcal{H}_0^I$ comme non-maximale s'il existe au moins un J tel que $J \subseteq \mathcal{M}$ avec $\mathcal{H}_0^J \supsetneq \mathcal{H}_0^I$. Dans le cas contraire on considère l'hypothèse nulle $\mathcal{H}_0^I$ comme maximale. Le principe de consonance implique donc que si une hypothèse non-maximale est rejetée, alors on peut rejeter au moins une hypothèse maximale [Gabriel, 1969]. Dans la majorité des applications, les hypothèses individuelles sont maximales. Le principe de consonance signifie que si une hypothèse nulle basée sur l'intersection des hypothèses individuelles $\mathcal{H}_0^I$ est rejetée, alors au moins une hypothèse individuelle $\mathcal{H}_0^i$ le sera, avec $i \in I$.

2.4.1.5 Principe de partitionnement

Le principe de construction des procédures de tests multiples basé sur le partitionnement est assez récent. C'est en 2002 que Finner & Strassburger [Finner and Straßburger, 2002] ont introduit cette méthode, en se basant sur les travaux de Stefansson et al. [Stefansson et al., 1988], et Hayter et Hsu

[Hayter and Hsu, 1994]. Comme nous l'expliquent Bretz et al.[Bretz et al., 2010], l'idée principale réside dans la création de sous ensembles disjoints dans l'espace des paramètres étudiés Θ. Ceci peut être illustré par la Figure 2.10, que les auteurs utilisent dans leur ouvrage pour le test de deux hypothèses individuelles.

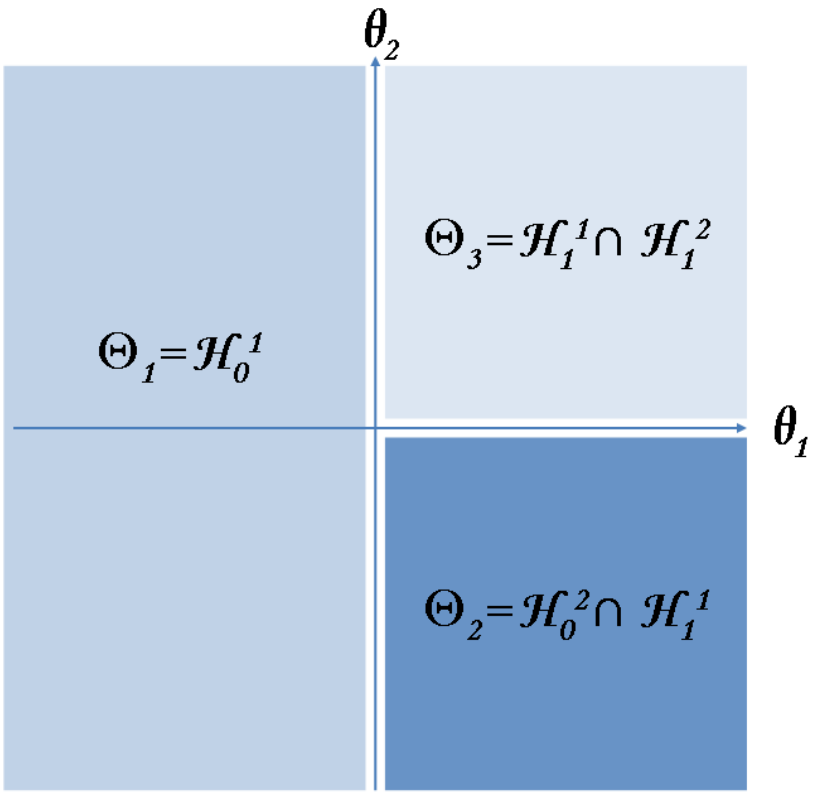


FIGURE 2.10: Principe de partition pour deux hypothèses nulles $\mathcal{H}_0^1$ et $\mathcal{H}_0^2$ dans l'espace $\mathbb{R}^2$

En se basant sur l'hypothèse de sous-ensembles disjoints, nous pouvons affirmer que le vrai vecteur des paramètres θ n'appartient qu'à un seul sous ensemble. Chacun d'entre eux peut donc être testé au niveau α.

Les procédures basées sur ce principe ont aussi l'avantage de contrôler fortement le FWER au niveau α. Dans le cas général de m hypothèses, la procédure est définie ci-dessous :

i. Choisir une partition appropriée $\{\Theta_l : l \in L\}$ de l'espace des paramètres Θ pour un certain ensemble d'indices L ;

ii. Tester chaque Θ_l au niveau α ;

iii. Rejeter l'hypothèse nulle $\mathcal{H}_0^j$ si tous les Θ_l tels que $\Theta_l \cap \mathcal{H}_0^j \neq \emptyset$ sont rejetés ;

iv. L'union de tous les Θ_l qui ne sont pas rejetés constituent une zone de confiance de θ au niveau $1 - \alpha$.

Remarques : Des extensions à cette procédure ont aussi été développées dans le papier de Finner & Strassburger en 2002 [Finner and Straßburger, 2002]. Cette méthode peut être très utile pour des applications dans le domaine de la recherche de doses optimales [Bretz et al., 2003, Liu et al., 2007, Strassburger et al., 2007], des tests d'équivalence [Finner et al., 2006], de l'estimation d'intervalles de confiance simultanés pour des procédures séquentielles [Finner and Straßburger, 2006, Strassburger and Bretz, 2008], mais aussi pour des tests d'Intersection Union [Strassburger et al., 2004].

2.4.2 Procédures en une étape

Une procédure *single-step* est caractérisée par le fait que le rejet ou non d'une hypothèse nulle est indépendante du rejet des autres hypothèses [Dmitrienko et al., 2009]. Par conséquent, l'ordre dans lequel les hypothèses sont testées n'est pas important. Nous pouvons alors considérer qu'elles sont évaluées simultanément, d'où l'utilisation du terme *single-step* (une seule étape). Les méthodes les plus connues sont celles de Bonferroni, Simes [Simes, 1986] et Šidák [Dunn, 1958] que nous exposons dans cette section.

2.4.2.1 Procédure de Bonferroni

Certains affirment que cette méthode est due à Fisher [Hochberg and Tamhane, 1987, Rafter et al., 2002], alors que d'autres prétendent que c'est plutôt Dunn [Dunn, 1961] qui en est le père [Toothaker, 1991]. Tous s'entendent cependant pour dire qu'elle se base sur l'inégalité de Boole, laquelle stipule que la probabilité d'un ou de plusieurs événements $(\mathcal{E}_j)$ est inférieure ou égale à la

somme des probabilités de ces événements séparés [Rom, 1990] :

$$\mathbb{P}\left(\bigcup_{j=1}^{m}\mathcal{E}_j\right) \leq \sum_{j=1}^{m}\mathbb{P}\left(\mathcal{E}_j\right).$$

Dans le cadre de la procédure de Bonferroni, on a $\mathcal{E}_j = \{p_j \leq \alpha/m\}$. Il s'agit donc d'une procédure en une étape (*single step procedure*), qui compare les p_{valeur} non ajustées $p_1, \ldots, p_m$ à un seuil commun α/m, où m est le nombre d'hypothèses testées. De manière équivalente, une hypothèse nulle $\mathcal{H}_0^j,\ j \in\ \mathcal{M}$ est rejetée, si la p_{valeur} ajustée $q_j = \texttt{min}\{1, mp_j\} \leq \alpha$.

La démonstration 3 nous montre que cette procédure contrôle fortement le FWER.

Démonstration 3.

Soit $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$ la famille d'hypothèses nulles , et $p_1, \ldots, p_m$ les $p_{valeurs}$ correspondantes.

Considérons $\mathcal{M}_0$ l'ensemble des vraies hypothèses nulles, qui comporte m_0 membres.

Sous l'hypothèse nulle $\mathcal{H}_0^j$, la p_{valeur} p_j suit une loi uniforme sur l'intervalle $[0, 1]$.

Le FWER est défini comme le fait de rejeter au moins l'un des membres de $\mathcal{M}_0$, *i.e.* de commettre au moins une erreur de Type-I. En nous basant sur l'inégalité de Boole, nous obtenons la preuve suivante :

$$\begin{aligned} FWER &= \mathbb{P}(V > 0) \\ &= \mathbb{P}\left(\bigcup_{j\in\mathcal{M}_0}\{p_j \leq \frac{\alpha}{m}\}\right) \\ &\leq \sum_{j\in\mathcal{M}_0}\mathbb{P}(p_j \leq \frac{\alpha}{m}) \ \leq\ m_0\frac{\alpha}{m} \ \leq \alpha, \text{ car } \frac{m_0}{m} \leq 1. \end{aligned}$$

2.4.2.2 Procédure de Simes

Simes [Simes, 1986] propose une modification de la procédure de Bonferroni, qui consiste à tester l'hypothèse nulle globale. Cette hypothèse est définie comme l'intersection de l'ensemble des hypothèses nulles individuelles. Ceci consiste à rechercher l'absence d'effet, *i.e.* qu'aucune hypothèse n'est significative :

$$\mathcal{H}_0 = \bigcap_{j\in\mathcal{M}}\mathcal{H}_0^j.$$

Posons $p_{(1)} \leq \ldots \leq p_{(m)}$, l'ensemble des $p_{valeurs}$ ordonnées, non ajustées, associées aux hypothèses $\mathcal{H}_0^{(1)} \ldots \mathcal{H}_0^{(m)}$. Ainsi, le test de Simes rejette $\mathcal{H}_0$ dès lors qu'une des m hypothèses ordonnées est rejetée au niveau $(j\alpha/m)$.

Remarques : Il est important de noter que cette procédure ne permet pas de conclure quant aux hypothèses individuelles, et qu'elle ne contrôle pas le FWER [Hommel, 1988].

2.4.2.3 Procédure de Šidák

Le développement de la procédure de Šidák se base sur les travaux de Dunn en 1958 [Dunn, 1958] et sur l'inégalité multiplicative de Šidák de 1967 [Šidák, 1967]. Cette procédure est construite sur l'hypothèse d'indépendance des statistiques de test. Dans ce contexte, le FWER devient :

$$FWER = \mathbb{P}(V \geq 0) = 1 - \mathbb{P}(V = 0) = 1 - (1 - \alpha)^m.$$

La procédure de Šidák consiste à rejeter une hypothèse nulle $\mathcal{H}_0^j$, si $p_j \leq 1-(1-\alpha)^{1/m}$, ou de manière équivalente si la p_{valeur} ajustée $q_j = 1-(1-p_j)^m \leq \alpha$. Cette procédure contrôle fortement le FWER.

Remarques : La procédure de Šidák est plus puissante que celle de Bonferroni, même si le gain de puissance est négligeable en pratique. Cependant, comme son champ d'application est beaucoup plus restreint (indépendance) elle est moins utilisée.

2.4.3 Comparaisons multiples de moyennes

Dans ce contexte, nous utilisons l'ANOVA (Analyse de la Variance). Cette méthode ne permet d'obtenir qu'un résultat global quant à la problématique initiale, *i.e* nous pouvons conclure qu'il existe au moins une différence de moyennes significative mais nous ne savons ni combien, ni lesquelles. Cependant en recherche clinique, l'intérêt des scientifiques se porte plus souvent sur une conclusion individuelle. Il est donc nécessaire de réaliser plusieurs tests et donc de corriger le degré de signification dû à la multiplicité des tests. Considérons une ANOVA à 1 facteur et à p niveaux. Nous pouvons alors distinguer plusieurs situations possibles [Games, 1971] :

1. Si le plan d'analyses statistiques prévoit de tester au plus $p-1$ comparaisons, il convient alors d'évaluer les ≪contrastes≫ à l'aide de techniques de comparaisons a priori orthogonales, où l'usage veut que le risque de première espèce soit fixé au seuil α.
2. Si au contraire, le plan d'analyses statistiques prévoit plus de $p-1$ comparaisons, alors les tests associés aux contrastes sont souvent évalués en utilisant la méthode de Bonferroni ou de Šidák ;
3. Si le nombre de comparaisons n'est pas prévu, il est nécessaire d'utiliser des techniques de comparaisons a posteriori. Parmi celles-ci, le choix de la méthode va dépendre de l'objectif de l'analyse :
   - Si le choix se porte sur la comparaison de toutes les moyennes entre elles, alors la méthode la plus utilisée est celle de Tukey ;
   - Si le choix se porte sur la comparaison des moyennes avec la moyenne d'un groupe témoin, il convient alors d'utiliser la méthode de Dunnett [Dunnett, 1955].

   Ces deux dernières procédures sont développées dans la suite de la section. Il faut toutefois s'assurer que l'effet global de l'ANOVA est significatif avant de les utiliser.

2.4.3.1 Procédure de Dunnett

La procédure de Dunnett [Dunnett, 1955], a pour objectif de comparer plusieurs traitements à un traitement de référence. Cette méthode est basée sur la distribution conjointe de tous les tests. De cette façon, elle tient ainsi compte de la corrélation entre les statistiques de test. Pour ce faire, Dunnett a développé la distribution du maximum de plusieurs variables aléatoires T qui suivent individuellement une loi de Student. Cette distribution s'apparente à une distribution de Student multivariée. La fonction $F(x|m,v)$ définie ci-dessous est la fonction de distribution cumulative d'une distribution unilatérale de *Dunnett* [Dmitrienko et al., 2009, p.74] :

$$F(x|m,v) = \mathbb{P}(\texttt{max}\{T_1,\ldots,T_m\} \leq x),$$

où la probabilité est évaluée sous l'hypothèse nulle globale : $\theta_1 = \ldots = \theta_m = 0$, $\theta_j = \mu_j - \mu_0,\ j \in \mathcal{M}$ (0 correspond au traitement de référence). La procédure rejette les hypothèses $\mathcal{H}_0^j$ lorsque $T_j \geq u_\alpha(m,v)$, où $u_\alpha(m,v)$ est le fractile d'ordre $(1-\alpha)$ de la distribution de Dunnett, où

$v = (m + 1)(n - 1)$.

Remarque : Dans le but d'améliorer la puissance de cette méthode, des procédures séquentielles ont été développées. Elles seront présentées dans la section suivante. On peut aussi noter que Hasler & Hothorn [Hasler and Hothorn, 2011] ont généralisé la procédure de Dunnett à l'évaluation de critères de jugement multiples.

2.4.3.2 Test de Tukey

Le test de Tukey est une procédure standard qui permet de répondre à un objectif portant sur la comparaison deux à deux de tous les groupes entre eux. Les hypothèses statistiques pour un test de comparaison de deux groupes T et C sont donc de la forme :

$$\mathcal{H}_0 : \mu^T - \mu^C = 0 \; versus \; \mathcal{H}_1 : \mu^T - \mu^C \neq 0$$

La statistique de test est définie comme suit :

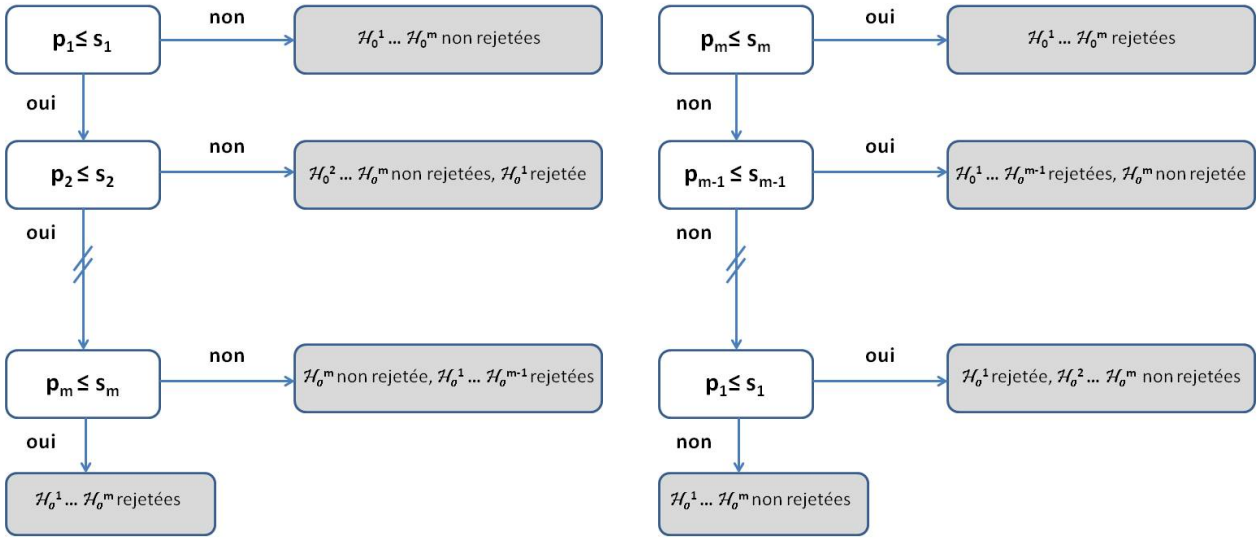
$$Q = \frac{|\bar{X}^T - \bar{X}^C|}{\sqrt{\frac{CM_e}{N}}}$$

où CM_e représente les carrés moyens résiduels, et $n = n_T = n_C$ représente l'effectif par groupe. Pour une analyse à un facteur, $N = n$, et pour une analyse à deux facteurs, $N = n \times q$ puisque l'on fusionne les q niveaux du deuxième facteur. La statistique suit une distribution dite *Studentized Range Distribution* avec (V, dl_e) degrés de liberté. Le premier degré de liberté, V, représente toujours le nombre de groupes qui seront comparés, *i.e.* le nombre de niveaux de la variable. Le deuxième degré de liberté, dl_e, représente les degrés de liberté résiduels.

Remarque : Le test de Tukey est très conservateur et nécessite que les échantillons soient de même taille. On peut noter qu'il existe une extension séquentielle de cette procédure, lui permettant d'être un peu plus puissante [Finner, 1988]. Le test de Scheffé [Scheffé, 1999] permet d'analyser des échantillons de tailles différentes, mais il souffre aussi d'un manque de puissance. Un bon compromis pourrait être l'utilisation de la procédure de test de Newman-Keuls [Newman, 1939, Keuls, 1952, Begun and Gabriel, 1981] qui est plus puissante et permet de prendre en compte des effectifs inégaux.

2.4.4 Procédures séquentielles

Les procédures séquentielles sont divisées entre les procédures descendantes et ascendantes. Chaque type de procédure considère un ensemble d'hypothèses $\{\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m\}$. Pour les procédures descendantes, les hypothèses statistiques sont considérées dans l'ordre croissant de leur p_{valeur} respectives. Dès que l'on ne rejette pas une hypothèse nulle $\mathcal{H}_0^j$, la procédure séquentielle descendante s'arrête, et aucune autre hypothèse n'est rejetée - voir la Figure 2.11(a). A contrario, si l'on décide de rejeter une hypothèse nulle $\mathcal{H}_0^j$, la procédure séquentielle ascendante s'arrête et toutes les autres sont rejetées - voir la Figure 2.11(b).



(a) Procédure séquentielle descendante (b) Procédure séquentielle ascendante

FIGURE 2.11: Principe des procédures séquentielles de gestion de tests multiples. Les $p_{valeurs}$ sont ordonnées par ordre croissant, p_j représente la $j^{\text{ème}}$ p_{valeur} ordonnée, et s_j le seuil de comparaison associé à p_j.

L'objectif de cette section est de présenter les principales procédures ascendantes et descendantes observées dans la littérature.

2.4.4.1 Procédures séquentielles ascendantes

Dans une procédure ascendante, les hypothèses sont testées une-à-une de la moins significative à la plus significative. Le fait de commencer le test par l'hypothèse la moins significative permet de répondre à la question : « Est-ce que toutes les hypothèses nulles peuvent être rejetées ? ». Lorsque la réponse est négative, le reste de la procédure permet d'identifier toutes les hypothèses pouvant être « retenues » [Tamhane and Dunnett, 1999]. L'objectif de cette sous-section sera donc de présenter les principales procédures ascendantes.

Procédure de Hochberg : La procédure ascendante de Hochberg [Hochberg and Tamhane, 1987] se base sur un test global de Simes [Simes, 1986]. Il est important de noter que cette méthode utilise des $p_{valeurs}$ issues de tests univariés. Le seuil s_j dans cette procédure est fixé à α/j.

L'algorithme est le suivant :

(1) Ordonner les $p_{valeurs}$ par ordre croissant, $p_1 \leq p_2 \leq \ldots \leq p_m$ correspondant aux hypothèses ordonnées $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m$.

(2) Ne pas rejeter $\mathcal{H}_0^m$ si $p_m > \alpha$, et continuer la procédure. Sinon, rejeter l'ensemble des hypothèses nulles $\{\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m\}$ et arrêter la procédure.

(3) Ne pas rejeter $\mathcal{H}_0^j$ si $p_j > \alpha/(m-j+1)$, et continuer la procédure. Sinon, rejeter l'ensemble des hypothèses nulles $\{\mathcal{H}_0^1, \ldots, \mathcal{H}_0^j\}$ et arrêter la procédure.

(4) Ne pas rejeter $\mathcal{H}_0^1$ si $p_1 > \alpha/m$. Sinon, rejeter l'hypothèse nulle $\mathcal{H}_0^1$.

Remarques : Cette procédure ne garantit pas le contrôle du FWER pour tout type de dépendance. On peut de plus observer que cette méthode est conservatrice lorsque le nombre de tests est important et lorsqu'il existe une forte dépendance positive entre les $p_{valeurs}$.
La p_{valeur} ajustée correspondant à cette procédure est donnée par $q_j = \texttt{min}\{1, \texttt{min}[(m-j+1)p_j, q_{j+1}]\},\ j \in \mathcal{M}$.

Procédure de Rom : En 1990, Rom propose une amélioration de la procédure de Hochberg, permet-

tant de pallier certaines de ces limites [Rom, 1990]. Sous l'hypothèse d'indépendance des statistiques de test, le FWER est contrôlé à un niveau légèrement inférieur à α pour la procédure de Hochberg, et exactement au niveau α pour celle de Rom. Cette procédure est donc légèrement plus puissante que celle de Hochberg. Pour améliorer cette procédure, Rom propose simplement de remplacer les seuils de comparaison $s_j = \alpha/j$ par un seuil s_j, où s_j est défini par la formule de récurrence ci-dessous [Dunnett and Tamhane, 1993] [Bernhard et al., 2004].

$$s_1 = \alpha,$$

et

$$\sum_{j=1}^{k-1} s_1^j = \sum_{j=1}^{k-1} \left(\begin{array}{c} k \\ j \end{array} \right) s_{j+1}^{k-j}, \text{ pour } 2 \leq k \leq m.$$

Remarques : La procédure de Rom est moins connue que celles d'Hochberg et d'Hommel, ce qui implique que l'accessibilité sur les principaux logiciels statistiques est moins importante. Par exemple sous le logiciel SAS, cette procédure n'est accessible qu'à partir d'une macro [Westfall et al., 2011]. Néanmoins, mis à part ce problème d'accessibilité, la procédure est assez simple à mettre en oeuvre et elle est plus puissante que celle d'Hochberg dans le cas de tests indépendants. Malheureusement, lorsque les statistiques de test sont dépendantes les propriétés de controle du FWER ne sont pas connues.

Procédure de Hommel : Tout comme la procédure d'Hochberg, cette procédure [Hommel, 1988] se base sur le test global de Simes [Simes, 1986]. Néanmoins, sa mise en oeuvre est plus complexe. Cette procédure consiste à appliquer un test de Simes à chaque hypothèse nulle d'intersection de la ≪*closed testing procedure*≫.

L'algorithme peut être défini comme suit :

(1) Ordonner les $p_{valeurs}$ par ordre croissant, $p_1 \leq p_2 \leq \ldots \leq p_m$ correspondant aux hypothèses ordonnées $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m$.

(2) Ne pas rejeter $\mathcal{H}_0^m$ si $p_m > \alpha$, et continuer la procédure. Sinon, rejeter l'ensemble des hypothèses nulles $\{\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m\}$ et arrêter la procédure.

(3) Ne pas rejeter $\mathcal{H}_0^{m-j+1}$ si $p_{m-l+1} > (j-l+1)\cdot(\alpha/j)$, pour $2 \leq j \leq m-1$ et $1 \leq l \leq j$, et continuer la procédure. Sinon, rejeter l'ensemble des hypothèses nulles $\{\mathcal{H}_0^1, \ldots, \mathcal{H}_0^{m-j+1}\}$ et arrêter la procédure.

(4) Ne pas rejeter $\mathcal{H}_0^1$ si $p_{m-l+1} > (m-l+1)\cdot(\alpha/m)$, pour $1 \leq l \leq m$. Sinon, rejeter l'hypothèse nulle $\mathcal{H}_0^1$.

Cette procédure est cohérente, mais pas consonante, cela signifie qu'il est possible de rejeter l'hypothèse nulle globale sans pour autant rejeter une hypothèse individuelle.

Remarques : Bien que cette procédure soit uniformément plus puissante que celle de Hochberg [Hommel, 1989], Dunnett & Tamhane [Dunnett and Tamhane, 1993] conseillent l'utilisation de la procédure de Rom ou celle de Hochberg, car le gain en puissance n'est pas suffisant pour utiliser cette procédure qui est plus complexe à mettre en oeuvre. Cependant, cette procédure est implémentée sous les logiciels statistiques et doit donc être privilégiée. Cette procédure, comme l'ensemble des procédures qui se basent sur la procédure de Simes (*e.g.* la procédure de Hochberg) ne garantissent pas le contrôle du FWER pour tout type de dépendance. Sarkar et al. établissent les conditions nécessaires pour que les procédures de Hochberg et Hommel contrôlent fortement le FWER [Sarkar and Chang, 1997, Sarkar, 1998]. Plus récemment, Sarkar a prouvé que les procédures basées sur la procédure de Simes préservent le FWER lorsque la distribution conjointe des statistiques de test suit une loi normale multivariée pour laquelle l'ensemble des coefficients de corrélation sont non-négatifs [Sarkar, 2008].

2.4.4.2 Procédures séquentielles descendantes

Dans une procédure descendante, les hypothèses sont testées de la plus significative à la moins significative. Ainsi, le fait de commencer le test par l'hypothèse la plus significative permet de répondre à la question : « Est-ce qu'au moins une hypothèse nulle peut être rejetée ? ». Lorsque la réponse est positive, le reste de la procédure permet d'identifier toutes les hypothèses pouvant être rejetées [Tamhane and Dunnett, 1999]. L'objectif de cette sous-section sera donc de présenter les principales

procédures descendantes.

Procédure de Holm : En 1979, Holm [Holm, 1979] introduit une procédure de gestion des tests multiples qui améliore celle de Bonferroni. A l'heure actuelle, il s'agit de la plus connue et la plus utilisée des procédures séquentielles descendantes. Cette procédure contrôle fortement le FWER. Son principe consiste à appliquer l'inégalité de Bonferroni à chaque niveau de la procédure, tout en testant les hypothèses dans un ordre dépendant des données.
Son algorithme peut être écrit de la façon suivante :

(1) Ordonner les $p_{valeurs}$ par ordre croissant, $p_1 \leq p_2 \leq \ldots \leq p_m$ correspondant aux hypothèses ordonnées $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m$.

(2) Rejeter $\mathcal{H}_0^1$ si $p_1 \leq \alpha/m$, et continuer la procédure. Sinon, ne rejeter aucune hypothèse nulle, et arrêter la procédure.

(3) Rejeter $\mathcal{H}_0^j$ si $p_j \leq \alpha/(m-j+1)$, et continuer la procédure. Sinon, ne rejeter aucune des hypothèses nulles $\{\mathcal{H}_0^j, \ldots, \mathcal{H}_0^m\}$ et arrêter la procédure.

(4) Rejeter $\mathcal{H}_0^m$ si $p_m \leq \alpha$. Sinon, seule l'hypothèse nulle $\mathcal{H}_0^m$ n'est pas rejetée.

Remarques : La p_{valeur} ajustée correspondant à la procédure de Holm est donnée par $q_j = \texttt{min}\{1, \texttt{max}[(m-j+1)p_j, q_{j-1}]\},\ j \in \mathcal{M}$.
Shaffer [Shaffer, 1986] a étendu la procédure de Holm dans le cadre des dépendances logiques entre les hypothèses. Le terme de dépendance logique est utilisé lorsque la véracité d'un sous-groupe d'hypothèses implique nécessairement la véracité d'une ou plusieurs autres hypothèses. Quand on est dans ce cas, la procédure de Shaffer va permettre d'améliorer la puissance des tests fermés (*closed tests*). Cette procédure sera donc très utile en recherche biomédicale, pour la comparaison de plusieurs groupes avec un témoin ou lors d'études de recherche de dose maximale (*dose-finding studies*).
Une extension paramétrique de cette procédure a été développée par Westfall & Tobias [Westfall and Tobias, 2007]. Leur idée est d'utiliser la dépendance des données au lieu de l'inégalité de Bonferroni, ce qui a pour conséquence d'augmenter la puissance de cette procédure.
Il est aussi intéressant de citer les travaux de Tamhane et al. [Tamhane et al., 1998a] qui ont développé une procédure séquentielle ascendante-descendante (*step up-down procedure*). Celle-ci s'intéresse à

la comparaison de plusieurs traitements à un témoin, et l'objectif est de pouvoir répondre à la question : ≪Est-ce qu'au moins r hypothèses nulles peuvent êtres rejetées ?≫ Cette procédure contrôle la r-power.

2.4.5 Procédures globales

2.4.5.1 Test d'Hotelling

Dans de nombreux essais cliniques, l'objectif principal consiste à comparer un groupe traitement à un groupe témoin sur un ensemble de critères de jugement. Pour cela considérons l'ensemble des variables aléatoires d'intérêts X_{ijk}, où j représente le critère de jugement $(1 \leq j \leq m)$, k le groupe $(k = T, C)$, et i l'individu $(1 \leq i \leq n_k)$. Nous nous interesserons donc plus particulièrement ici à la différence de moyenne entre les deux groupes pour l'ensemble des critères de jugement : $\delta_j = \mu_j^T - \mu_j^C$.

La procédure de test d'Hotelling [Hotelling, 1931] fait l'hypothèse d'indépendance et d'homoscédasticité des échantillons.

Les hypothèses statistiques du test d'Hotelling sont définies comme suit :

$$\mathcal{H}_0 : \boldsymbol{\delta} = \boldsymbol{\mu}^T - \boldsymbol{\mu}^C = \mathbf{0}_m, \ versus \ \mathcal{H}_1 : \boldsymbol{\delta} = \boldsymbol{\mu}^T - \boldsymbol{\mu}^C \neq \mathbf{0}_m,$$

où $\boldsymbol{\mu}^k$ est le vecteur de moyennes du groupe k, $\boldsymbol{\delta}$ le vecteur de différence de moyennes entre le groupe traitement et le groupe témoin, et $\mathbf{0}_m$ est un vecteur nul. Tous ces vecteurs sont de longueur m. Les hypothèses peuvent donc aussi s'écrire comme suit :

$$\mathcal{H}_0 : \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ versus \ \mathcal{H}_1 : \exists j, \ \delta_j \neq 0.$$

Les paramètres $\bar{X}^T, \bar{X}^C, S_C, S_T$ représentent les estimateurs des moyennes empiriques et des variances pour les groupes T et C. Sous l'hypothèse d'homoscédasticité, les matrices de variance covariance des deux groupes sont supposées identiques, une estimation de cette matrice est donc nécessaire :

$$S_p = \frac{(n_T - 1)S_T + (n_C - 1)S_C}{n_T + n_C - 2}.$$

La statistique de test du T^2 de Hotelling, qui est une généralisation multivariée du test de Student s'écrit :

$$T^2 = (\bar{X}^T - \bar{X}^C)' \left[S_p \left(\frac{1}{n_T} + \frac{1}{n_C} \right) \right]^{-1} (\bar{X}^T - \bar{X}^C).$$

Cette statistique peut également s'écrire sous la forme :

$$T^2 = \frac{n_T n_C}{n_T + n_C} (\bar{X}^T - \bar{X}^C)' \left[S_p \right]^{-1} (\bar{X}^T - \bar{X}^C).$$

Sous l'hypothèse nulle, cette statistique suit une loi de Hotelling, mais dans la pratique une transformation de cette statistique est réalisée afin qu'elle suive une loi de Fisher :

$$F = \frac{n_T + n_C - m - 1}{m(n_T + n_C - 2)} \cdot T^2.$$

Ici, F suit une loi de Fisher à $(m, n_T + n_C - m - 1)$ degrés de liberté.

Remarques : Ce test multivarié est assez robuste à l'hypothèse sur la distribution multinormale des X, tout comme le test de Student dans le cadre unidimensionnel. Cependant, il est moins robuste à l'hypothèse d'homoscédasticité, et plus spécifiquement pour des effectifs déséquilibrés entre les groupes.

Il est aussi important de noter que la principale limite des tests globaux réside dans le fait qu'ils ne fournissent qu'une conclusion globale quant à la question de recherche, *i.e.* les résultats individuels ne sont pas disponibles.

2.4.5.2 Test des moindres carrés d'O'Brien

Comme pour le test d'Hotelling, on s'intéresse ici à la comparaison d'un groupe traitement à un groupe témoin sur un ensemble de critères de jugement. Les notations restent donc les mêmes, et la seule différence réside dans le fait que nous réalisons des tests unilatéraux.

L'idée de O'Brien est de simplifier la problématique, et de poser l'hypothèse que le vecteur des différences de moyennes entre ces deux groupes $\delta = (\delta_1, \ldots, \delta_m)'$ soit égale à $\lambda(\sigma_1^2, \ldots, \sigma_m^2)'$, avec $\lambda \geq 0$, *i.e.* si $\delta_j / \sigma_j = \lambda_j$ correspond à l'effet traitement standardisé pour le $j^{\text{ème}}$ critère de jugement, alors O'Brien considère que $\lambda_j = \lambda \geq 0$ pour tout j. Dans ce cas, l'hypothèse nulle globale se simplifie et peut s'écrire :

$$\mathcal{H}_0^* : \lambda = 0 \; versus \; \mathcal{H}_1^* : \lambda > 0$$

L'hypothèse nulle globale $\mathcal{H}_0^*$ est basée sur la réponse standardisée, $Y_{ijk} = X_{ijk}/\sigma_j$. Sous l'hypothèse que les différences de moyennes standardisées soient les mêmes ($\frac{\delta_j}{\sigma_j} = \lambda$), alors O'Brien propose de considérer une régression linéaire dans le but de modéliser la variable réponse standardisée :

$$Y_{ijk} = \frac{X_{ijk}}{\sqrt{\sigma_j^2}} = \frac{\mu_j}{\sqrt{\sigma_j^2}} + \frac{\lambda}{2} I_{ijk} + \epsilon_{ijk}, \ (k = T, C; 1 \leq i \leq n_k; 1 \leq j \leq m),$$

où $\mu_j = (\mu_j^T - \mu_j^C)/2$, $I_{ijk} = 1$ si $k = T$ et $I_{ijk} = -1$ si $k = C$, et $\epsilon_{ijk} \sim \mathcal{N}(0, 1)$. Les corrélations sont alors définies par : $Corr(\epsilon_{ijk}, \epsilon_{i'j'k'}) = \rho_{jj'} = \frac{\sigma_{jj'}}{\sqrt{\sigma_j^2 \sigma^2 j'}}$ si $i = i'$ et $j = j'$, autrement $Corr(\epsilon_{ijk}, \epsilon_{i'j'k'}) = 0$.

Il est important de noter que les vecteurs $\mathbf{Y_{ik}} = (Y_{ik1}, \ldots, Y_{ikm})'$ sont indépendants avec une matrice de corrélation commune $R = \{\rho_{jj'}\}$.

C'est en se basant sur cette idée qu'il propose en 1984 [O'Brien, 1984] deux tests globaux qui ont pour but de démontrer un effet global positif du traitement sur un ensemble de critères quantitatifs. Il s'agit des tests des moindres carrés ordinaux (*ordinary least square* (OLS)), et des moindres carrés généralisés (*generalized least square* (GLS)).

Test des moindres carrés ordinaux (O'Brien OLS test) : La première procédure développée par O'Brien est basée sur une estimation des moindres carrés ordinaux de l'effet λ. On considère respectivement $\hat{\lambda}_{OLS}$ l'estimation des moindres carrés ordinaux de l'effet λ, et $SD_{\hat{\lambda}_{OLS}}$ son écart type. Il peut alors être démontré que la statistique de test des moindres carrés, pour le test de l'hypothèse nulle globale $\mathcal{H}_0^*$, peut être donnée par :

$$t_{OLS} = \frac{\hat{\lambda}_{OLS}}{SD_{\hat{\lambda}_{OLS}}} = \frac{J't}{\sqrt{J'\hat{R}^{-1}J}},$$

où J est un vecteur unitaire de longueur m, et T est le vecteur des statistiques de test. La statistique de test pour le $j^{\text{ème}}$ critère de jugement est définie comme :

$$t_j = \frac{n_C n_T}{n_T + n_C} \frac{\bar{x}_j^T - \bar{x}_j^C}{s_j},$$

où $\bar{x}_j^k$ représente la moyenne observée de la variable réponse pour le $k^{\text{ème}}$ groupe et le $j^{\text{ème}}$ critère de jugement. On pose aussi S la matrice de variance covariance poolée. Les éléments diagonaux de cette matrice sont notés : $s_1^2, \ldots, s_m^2$.

Test des moindres carrés généralisés (O'Brien GLS test) : Si les termes d'erreurs du modèle de régression sont corrélés, alors il est préférable d'utiliser une estimation de λ par la méthode des moindres carrés généralisés, qui implique la définition de la statistique suivante :

$$t_{GLS} = \frac{\hat{\lambda}_{GLS}}{SD_{\hat{\lambda}_{GLS}}} = \frac{J'\hat{R}^{-1}t}{\sqrt{J'\hat{R}^{-1}J}}.$$

Remarques : En 2004, Logan et Tamhane [Logan and Tamhane, 2004] ont généralisé les tests de moindres carrés ordinaux (OLS), et généralisés (GLS), au cas hétéroscédastique. Il est recommandé d'utiliser en recherche clinique la procédure basée sur les moindres carrés ordinaux [Dmitrienko et al., 2009, p.152]. En effet, la procédure OLS est généralement plus puissante que la procédure GLS, mais la procédure GLS est aussi plus difficile à mettre en oeuvre à cause des pondérations négatives.

2.4.6 Resampling based methods

Dans de nombreuses situations, les distributions conjointes et/ou marginales des statistiques de test sont inconnues. Face à ce problème, des méthodes basées sur le rééchantillonnage permettent d'estimer les $p_{valeurs}$ ajustées, sans aucune connaissance de la distribution conjointe. Considérons la problématique du test simultané de m hypothèses nulles $(\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m)$ et leurs $p_{valeurs}$ respectives $p_1, \ldots, p_m$. Supposons un instant que nous connaissons la distribution exacte des $p_{valeurs}$ individuelles. Westfall & Young [Westfall and Young, 1989] ont introduit la notion de méthode par rééchantillonnage et propose dans ce contexte de définir les $p_{valeurs}$ ajustées de la manière suivante :

$$q_j = \mathbb{P}\{\texttt{min}(P_1, \ldots, P_m) \leq p_j\},\ j = 1, \ldots, m,$$

où $P_1, \ldots, P_m$ sont des variables aléatoires qui suivent la même distribution que $p_1, \ldots, p_m$ lorsque les m hypothèses nulles sont simultanément vraies. Une fois que les $p_{valeurs}$ ajustées ont été calculées, les hypothèses nulles peuvent être testées directement. En effet, l'hypothèse nulle $\mathcal{H}_0^j$ est rejetée si $q_j \leq \alpha$.

Ces méthodes développées par Westfall & Young [Westfall and Young, 1993] utilisent donc des méthodes de rééchantillonnage par Bootstrap, ou permutation, afin d'estimer une distribution de référence des $p_{valeurs}$ individuelles. Elles contrôlent fortement le FWER dans certaines conditions (*e.g.*

échangeabilité pour la permutation), et permettent de prendre en compte les corrélations existant entre les tests. Ces méthodes prennent en compte la structure de corrélation empirique des $p_{valeurs}$ individuelles. Dans le cadre des procédures en une étape, elles sont par exemple plus puissantes que les méthodes de Bonferroni et de Šidák.

L'ouvrage de référence concernant la gestion des tests multiples par rééchantillonnage est celui de Westfall & Young [Westfall and Young, 1993]. On peut aussi citer les travaux de Dudoit & Van der Laan [Dudoit and Van der Laan, 2008, p. 289-321] et de Yu & Liang [Yu et al., 2011] sur la gestion des tests multiples par rééchantillonnage dans l'analyse de données génomiques. Westfall & Troendle [Westfall and Troendle, 2008] se sont quant à eux concentrés sur l'analyse comparative de plusieurs échantillons sur des critères de jugement multiples par rééchantillonnage.

2.4.7 Gatekeeping Procedures

Afin d'amortir le coût d'un essai clinique, les scientifiques se fixent des objectifs multiples qu'il est nécessaire de hiérarchiser. Dans ce cas, ils doivent définir des critères primaires et secondaires (pour une classification détaillée des critères primaires et secondaires, il convient de se référer à l'article d'Agostino [D'Agostino Sr, 2000]). L'objectif primaire va souvent être d'observer l'effet principal du traitement. L'objectif secondaire quant à lui regroupe d'une part l'étude des critères de jugement secondaires, mais aussi les analyses en sous-groupes. L'analyse secondaire a donc un rôle de soutien à l'objectif principal. Les procédures de Gatekeeping développées par Westfall et al. [Westfall and Krishen, 2001, Dmitrienko et al., 2003] consistent à prendre en compte la hiérarchie des critères de jugement, tout en contrôlant le FWER. Cette méthodologie a le mérite d'être en adéquation avec les recommandations des agences de sécurité sanitaire. Cette procédure permet de conclure à un effet du critère secondaire, uniquement si un effet a été observé sur le critère principal [O'Neill, 1997]. Elle a aussi le mérite de contrôler le FWER.

Afin d'illustrer cette procédure, nous allons introduire l'exemple présenté dans la Figure 2.12. On considère un essai clinique dans lequel les scientifiques ont hiérarchisé leurs objectifs en trois niveaux. L'objectif principal consiste à observer l'effet du traitement sur au moins un des critères de jugement principaux étudiés. Si l'objectif principal est atteint, alors on peut tester l'effet du traitement sur le critère de jugement 3. Les critères secondaire de niveau 2 ne seront alors étudiés que si le traitement

entraîne un effet significatif sur le critère de jugement 3. Le principe de cette procédure consiste donc tout d'abord à hiérarchiser les objectifs, en créant des familles d'hypothèses à tester. Chaque famille d'hypothèses sera alors testée au niveau global α désiré par les scientifiques, ici fixé à 0.05. Dans cet exemple, on pourra conclure à un effet du traitement sur le critère de jugement 6 que si on a observé un effet sur le critère de jugement 3, et sur au moins un des critères principaux 1 ou 2.

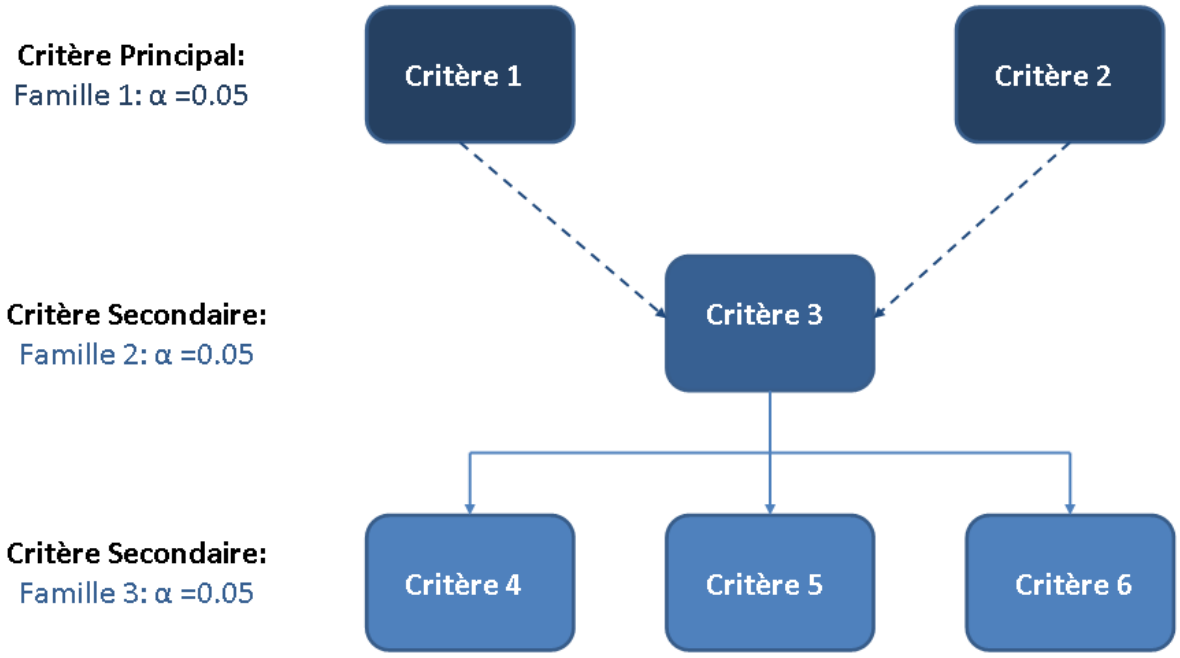


FIGURE 2.12: Principe des procédures de Gatekeeping. Le trait plein est utilisé pour indiquer que tous les tests de la famille de niveau supérieure doivent être significatifs pour passer à l'étude de la famille de niveau inférieure. Le trait en pointillé indique qu'il est nécessaire d'avoir au moins un test significatif dans la famille de niveau supérieur pour passer à la famille de niveau inférieur.

3

Analyse et calcul de taille d'échantillon, avec un contrôle de la puissance disjonctive, dans le contexte de critères de jugement co-principaux.

Résumé

Les différentes agences de sécurité sanitaire (FDA, EMA) ainsi que l'ICH [EMA, 2002, ICH, 1999] conseillent de diminuer au maximum le nombre de critères de jugement principaux dans la mise en place d'un essai. Cependant, même en passant par un critère composite, il arrive qu'il soit difficile de se restreindre à un seul critère.

L'analyse de ces essais implique donc de tester l'efficacité du nouveau traitement sur l'ensemble des critères de jugement principaux. Cette pratique engendre un problème de multiplicité, qu'il est possible auquel il est possible de répondre en utilisant des méthodes de correction de tests multiples [Hommel, 1988, Hochberg, 1988, Holm, 1979]. Cependant la majorité d'entre elles ne prennent pas en compte la corrélation qui existe entre les tests, et sont donc conservatrices.

Dans le cadre de critères de jugement principaux multiples, quantitatifs et corrélés, notre objectif a donc été de développer des méthodes moins conservatrices permettant de calculer le nombre de sujets nécessaires, ainsi que d'analyser l'effet du produit tout en contrôlant le FWER ainsi que la puissance disjonctive. Ce chapitre sera donc l'occasion de vous présenter le contexte inhérent à cette problématique, puis les différentes méthodes proposées à travers un article publié dans le ≪*Journal of Biopharmaceutical Statistics*≫.

Contexte

L'utilisation de critères de jugement principaux multiples dans les essais de Phase III est devenue de plus en plus courante. Cependant l'utilisation de tels critères est source de débat [Sankoh et al., 1997], et les problématiques inhérentes sont d'ailleurs un point d'ancrage à de nombreux articles dans les revues scientifiques. L'Internationnal Conference on Harmonization (ICH E9 Expert Working Group, 1999), qui fait référence dans le domaine des essais cliniques, conseille de limiter au maximum le nombre de critères de jugement principaux. Cette recommandation a un triple avantage puisqu'elle permettra de définir de manière précise l'objectif de l'essai, de limiter le problème lié à la multiplicité, et enfin de prouver plus facilement l'efficacité du produit.

Néanmoins, cette stratégie possède de réelles limites. A l'heure actuelle, dans de nombreux essais de recherche d'efficacité (Phase II et III), il est impossible de résumer l'effet recherché à un seul critère principal, même en passant par un critère composite. Cela peut s'expliquer par le fait que les agences réglementaires (FDA, EMA, ...) demandent souvent de prouver l'efficacité du produit sur un ensemble de paramètres. Plusieurs cas sont alors possibles, cependant nous nous focaliserons ici sur celui pour lequel l'efficacité est conclue si au moins un des critères de jugement principaux est significatif.

Dans ce contexte, la mise en place d'un essai possèdant plusieurs critères de jugement principaux devient donc un réel défi pour les statisticiens [O'Brien, 1984, Cook and Farewell, 1996, Pocock et al., 1987]. C'est une des raisons pour lesquelles les scientifiques considèrent souvent les critères de jugement principaux comme ayant la même importance. Dans ce cas de figure, ils sont alors définis par l'anglicisme ≪*co-primary endpoints*≫ [Julious and McIntyre, 2012]. Nous nous placerons dans ce cas de figure. Cependant dans la pratique courante, il pourrait aussi être envisageable de définir un poids différent pour chaque critère de jugement [Bretz et al., 2009, Bretz et al., 2011, Burman et al., 2009].

La littérature disponible est abondante concernant l'analyse et le calcul de taille d'échantillons de tels essais [Senn and Bretz, 2007, Dunnett and Tamhane, 1992, Chuang-Stein et al., 2007]. Toutefois, nous pouvons distinguer deux stratégies bien distinctes.

Certaines stratégies sont développées dans la même philisophie que les recommandations de l'ICH et consistent à réduire au maximum le nombre de critères de jugement [Neuhäuser, 2006]. Néanmoins,

ces stratégies ont tendance à être trop restrictives et vont donc perdre une partie de l'information pertinente. Ces stratégies ne permettront pas non plus une réponse optimale à la problématique scientifique.

L'autre stratégie d'analyse consiste à analyser l'ensemble des critères de jugement principaux. Cette stratégie sera moins puissante mais permettra une réponse plus appropriée aux attentes scientifiques. L'analyse de tels critères nécessitera la prise en compte de la multiplicité. Pour cela, il existe une pléthore de méthodes de correction permettant, d'une part le calcul de la taille d'échantillon, mais aussi l'analyse des critères. Les méthodes les plus couramment utilisées [Hommel, 1988, Hochberg, 1988, Holm, 1979] ont l'avantage d'être assez simple d'utilisation, mais elles ont aussi l'inconvénient de ne pas prendre en compte la corrélation existant entre les statistiques de test calculées. Il s'agit ici d'une importante limite car ces méthodes seront, dans le cadre de critères de jugement corrélés, conservatrices. Ceci aura pour conséquence de diminuer la puissance de chaque test et donc de ne pas mettre en évidence de faibles effets. On comprend bien que cette limite est d'autant plus importante en nutrition, domaine pour lequel l'effet recherché est beaucoup plus faible et la variabilité plus importante que dans le domaine pharmaceutique.

Les méthodes de Gatekeeping [Dmitrienko et al., 2003] peuvent aussi être utilisées quand les scientifiques sont capables de hiérarchiser les critères de jugement, mais elles utilisent généralement les méthodes précédentes et présentent donc les mêmes limites.

Une méthode alternative consiste alors à utiliser les méthodes se basant sur les tests d'Union Intersection développés par Roy en 1953 [Roy, 1953]. Ces méthodes, qui se basent sur la distribution conjointe des statistiques de tests, contrôlent le FWER et prennent en compte la corrélation entre les tests. Il est cependant nécessaire de connaître la distribution conjointe des statistiques réalisées pour pouvoir l'appliquer.

Une dernière possibilité existant dans la littérature concerne les méthodes globales, telles que le T^2 d'Hotelling [Hotelling, 1931] ou les tests des moindres carrés d'O'Brien [O'Brien, 1984]. Ces méthodes prennent en compte la corrélation, mais elles ont l'inconvénient de ne donner qu'un résultat global, et la procédure d'Hotelling est non directionnelle [Sankoh et al., 1999]. De plus, il a été démontré qu'en présence de données manquantes complètement aléatoires (MCAR), ces méthodes sont peu puissantes [Yoon et al., 2011].

Dans le cadre de critères de jugement principaux multiples, quantitatifs et corrélés, notre objectif a donc été de développer des méthodes moins conservatrices permettant de calculer le nombre de sujets nécessaires, ainsi que d'analyser l'effet du produit tout en contrôlant le FWER ainsi que la puissance disjonctive. Pour cela, nous nous sommes basés sur une méthode d'Union Intersection, ainsi qu'une généralisation du T^2 d'Hotelling.

Les méthodes développées au sein de ce chapitre ont été implémentées sous le logiciel ≪open-source≫ de statistiques R au sein du package Sample.

Ce travail a fait l'objet d'une publication dans le ≪*Journal of Biopharmaceutical Statistics*≫.

Journal of Biopharmaceutical Statistics
Volume 00, Number 0, Pages 000–000
S ????-????(XX)0000-0

POWER AND SAMPLE SIZE DETERMINATION IN CLINICAL TRIALS WITH MULTIPLE PRIMARY CONTINUOUS CORRELATED END POINTS

PIERRE LAFAYE DE MICHEAUX, BENOIT LIQUET, SÉBASTIEN MARQUE, AND JÉRÉMIE RIOU

Abstract. The use of two or more primary correlated end points is becoming increasingly common. A mandatory approach when analyzing data from such clinical trials is to control the Familywise Error Rate (FWER). In this context, we provide formulas for computation of sample size, and for data analysis. Two approaches are discussed: an individual method based on a union-intersection procedure and a global procedure based on a multivariate model which can take into account adjustment variables. These methods are illustrated with simulation studies and applications. An `R` package known as `Sample` is also available.

Keywords: Sample Size Determination, Correlated Endpoints, Family Wise Error Rate, Multivariate Normal Distribution, Multiple Continuous Endpoints, Power.

1. Introduction

The use of multiple end points to characterize product efficacy and safety measures is an increasingly common feature in recent clinical trials. Efficacy is often defined not by a unique end point but by a combination of several parameters. Regulatory agencies commonly require more than one end point to measure different aspects of product efficacy in confirmatory clinical trials. However, the use of multiple end points is a source of debate (see Sankoh et al. (1997)) and a lot of statistical literature on the subject has been published. In general, national health authorities recommend, on the basis of the biostatistics guideline developed by the International Conference on Harmonization (ICH) (ICH E9 Expert Working Group, 1999), the selection of one primary end point to provide strong scientific evidence of the efficacy of a test treatment. However this strategy has clear limitations, notably when it leads to arbitrary classification of different end points. While it reduces the dimension of the problem, if the classification is not sufficiently robust real effects may be ignored or left undetected. Consequently, many clinical trials incorporate multiple primary end points to demonstrate the efficacy of the product.
Consideration of multiple end points nonetheless brings with it several challenges to the design and analysis of trial data (O'Brien, 1984; Pocock et al., 1987; Cook and Farewell, 1996). Several authors have discussed power calculations in clinical trials when two or more primary end points are given as continuous variables (Senn and Bretz, 2007; Chuang-Stein et al., 2007; Dunnett and Tamhane, 1992). On the one hand, one strategy, following the same philosophy as that of the guidelines, is

to reduce as far as possible the number of end points (Neuhäuser, 2006). However, this strategy may result in a loss of information concerning end points and does not address the scientific problem of the selection of parameters. On the other hand, an alternative strategy is to consider all primary end points. We have focussed on the situation of treating all endpoints equally, thus referred to as co-primary endpoints. However, in clinical practice, it can also be interesting to consider a different weighting for each endpoint (Bretz et al., 2009, 2011; Burman et al., 2009). Depending on the scientific question raised, statisticians may be interested in "'or"' comparisons (detecting at least one significant primary end point) or multiple must-win comparisons (detecting at least r among m comparisons); see Julious and McIntyre (2012). Several authors developed multiple testing procedures in the context of a "'win"' on all co-primary end points; e.g. Berger (1982) and Sozu et al. (2006, 2010, 2011). We have limited this report to the detection of "at least one primary end point" for the "two treatments" case. In this context, the most common strategy is to use either single-step (Simes, 1986; Sidak, 1967) or stepwise (Holm, 1979; Hochberg, 1988; Hommel, 1988) procedures. For single-step methods, the rejection or non-rejection of a single hypothesis does not account for the outcome of any other hypotheses. A well known example of single-step procedures is the Bonferroni test. In contrast, for stepwise methods, the rejection or non-rejection of a particular hypothesis may take into account the outcome of other hypotheses. Stepwise methods are more powerful than single-step procedures. The equally well known Holm procedure is a stepwise extension of the Bonferroni test using a closure principle. Both types of procedures are conservative (lead to wrongly "accepting" the null hypothesis) and might lead to biased test decisions, as information about correlations of the end points is not exploited. This implies a strong control of the Type-I error probability and consequently, a decrease in the power of each test. An extensive work has been done by Sankoh et al. (1997) in order to characterize the advantages and limit of adjusted methods. Gatekeeping procedures (Dmitrienko et al., 2003), which consist of scheduling the hypotheses and analyzing the data with multiple families of null hypotheses, suffer from similar problems and need an order of priority among the end points. Another alternative is to use the union-intersection test procedure (Roy, 1953). This method can control the Family Wise Error Rate (FWER) and correlations among end points can be taken into account. Finally, global methods such as the T^2 test of Hotelling (1953) which take correlations into account, can be used "where the endpoints are alternative measures of the same fundamental quantity" (see Sankoh et al. (1997)). One limitation of this procedure is that it gives a global and non-directional result. This problem is mentioned by Sankoh et al. (1999). Furthermore, when data are missing completely at random (MCAR), Yoon et al. (2011) recently showed that this is a less powerful method.

The aim of this article is to provide sample size calculation methods, as well as corrections for Type-I errors probabilities based on a global method with a multivariate linear model or on an individual method involving a union-intersection procedure. The approach of the global method is to generalize the T^2 test of Hotelling to deal with adjustment variables. Finally, we compare power and FWER control of both methods with common methods for different scenarios of correlation and adjustment. In section 2 we present the statistical methods related to simultaneous testing as well as power and sample size calculations. In Section 3, we present the

results of a simulation study, and an application in two nutritional clinical studies. Lastly, the results are discussed and a conclusion including limitations and perspectives is provided in Section 4.

2. Methods

Two different approaches are presented in this section. First, we present an individual testing procedure with an exact control of the FWER. In this context, the power and the sample size determination are defined under different assumptions. Secondly, we propose a global procedure based on a multivariate model involving adjustment variables.

2.1. **Overview.** We consider the context where a vector $\mathbf{X} = (X_1, \ldots, X_m)^\intercal$ of m quantitative variables (end points) is measured in a group of $2n$ subjects taken at random in two sub-populations: a control group (C) and a test group (T). Let $\mathbf{X}_1^j, \ldots, \mathbf{X}_n^j$ be n independent and identically distributed (*i.i.d.*) (conditional on group j) random vectors, with expectation $\boldsymbol{\mu}^j$ and same covariance matrix Σ, where $j = C$ stands for the control group and $j = T$ stands for the test product. The k^{th} component $X_{i,k}^j$ of vector $\mathbf{X}_i^j$ denotes the i^{th} observation ($1 \leq i \leq n$), on the k^{th} end point ($1 \leq k \leq m$) for product j. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^\intercal = \boldsymbol{\mu}^T - \boldsymbol{\mu}^C$, with $\delta_k = \mu_k^T - \mu_k^C$ be the vector of true mean differences between the test and control products respectively, where $^\intercal$ denotes vector or matrix transposition. The test product will be considered to be different from the control product on the k^{th} end point, if $\delta_k \neq 0$. The clinical aim is to be able to detect a mean difference between the test and the control product for at least one end point among m. This can be stated under a statistical hypothesis formalism as:

$$\mathcal{H}^0 : \boldsymbol{\delta} = \boldsymbol{0}_m \quad \text{versus} \quad \mathcal{H}^1 : \boldsymbol{\delta} \neq \boldsymbol{0}_m \tag{1}$$

where $\mathbf{0}_m = (0, \ldots, 0)^\intercal$ is the null vector of length m. In section 2.3, we use a global test of $\mathcal{H}^0$ to address this problem. Another avenue is to consider a so-called individual testing procedure based on the m following single hypotheses:

$$\mathcal{H}_k^0 : \delta_k = 0 \quad \text{versus} \quad \mathcal{H}_k^1 : \delta_k \neq 0, \tag{2}$$

noting that

$$\mathcal{H}^0 = \bigcap_{k=1}^{m} \mathcal{H}_k^0 \quad \text{and} \quad \mathcal{H}^1 = \bigcup_{k=1}^{m} \mathcal{H}_k^1. \tag{3}$$

This latter approach, based on the family hypothesis $\{\mathcal{H}_1^0, \ldots, \mathcal{H}_m^0\}$, is considered first.

2.2. **Individual Testing Procedure.** In the context of individual testing procedures, we need to define all the test statistics used. When the variances $\sigma_k^2 = \mathbb{V}\text{ar}(X_{1,k}^j)$, $1 \leq k \leq m$ are known, the standardized test statistic that will be used to test (2) is:

$$Z_k^{(n)} = \frac{\bar{X}_k^T - \bar{X}_k^C}{\sqrt{\frac{2}{n}}\sigma_k}, \tag{4}$$

where $\bar{X}_k^j = \frac{1}{n}\sum_{i=1}^{n} X_{i,k}^j$ is the sample mean for group j.

When the σ_k^2's are unknown, they will be estimated by the pooled variance

$$\widehat{\sigma}_k^2 = \frac{1}{2n-2}\sum_{i=1}^{n}\left[\left(X_{i,k}^C - \bar{X}_k^C\right)^2 + \left(X_{i,k}^T - \bar{X}_k^T\right)^2\right]$$

and the "studentized" test will be used instead

$$T_k^{(n)} = \frac{\bar{X}_k^T - \bar{X}_k^C}{\sqrt{\frac{2}{n}}\widehat{\sigma}_k}. \tag{5}$$

In the sequel, $Z_k^{(n)}$ will be replaced with $T_k^{(n)}$ when the σ_k's are unknown.

2.2.1. *A direct approach to control the FWER.* We reject the individual null hypothesis $\mathcal{H}_k^0$ if $|Z_k^{(n)}|$ is larger than a suitable multiplicity adjusted critical point c_α. Since $\mathcal{H}^1 = \cup_{k=1}^m \mathcal{H}_k^1$, it seems natural to decide $\mathcal{H}^1$ if at least one member of the family $\{\mathcal{H}_1^0, \ldots, \mathcal{H}_m^0\}$ is rejected using an individual procedure. The type-I error probability of the global procedure is then exactly equal to the FWER of the multiple procedure defined as:

$$\begin{aligned} FWER &= \mathsf{P}(\text{reject at least one } \mathcal{H}_k^0,\ 1 \le k \le m \mid \mathcal{H}^0 \text{ is true}) \\ &= 1 - \mathsf{P}\left\{\left(|Z_1^{(n)}| \le c_\alpha\right) \cap \ldots \cap \left(|Z_m^{(n)}| \le c_\alpha\right) \mid \mathcal{H}^0 \text{ is true}\right\}. \end{aligned} \tag{6}$$

The adjusted critical value c_α is chosen to satisfy $FWER = \alpha$, for a fixed significance level α. Obviously, the joint distribution of $\mathbf{Z}_n = \left(Z_1^{(n)}, \ldots, Z_m^{(n)}\right)^\mathsf{T}$, or of $\mathbf{T}_n = \left(T_1^{(n)}, \ldots, T_m^{(n)}\right)^\mathsf{T}$ when the σ_k's are estimated, has to be known or at least approximated to some degree (see 2.2.3). Note that this procedure allows us to explicitly specify the value of the FWER, which is better than controlling its value using an upper limit, as is usually the case.

2.2.2. *Power and sample size determination.* An important task in the design phase of clinical trials, is to determine the sample size n which guarantees a pre-specified power, noted hereafter as $1-\beta$. In single testing situations, power is defined as the probability of rejecting the null hypothesis under investigation, when it is false. For multiple testing and multiple comparisons, Westfall et al. (1999) propose other definitions of power. The clinical interest here is to detect at least one significant end point among m with a given power, so we use the the so-called minimal power (referred to as disjunctive power by Senn and Bretz (2007)), which is given by

$$\begin{aligned} 1-\beta &= \mathsf{P}\left(\text{reject at least one } \mathcal{H}_k^0,\ 1 \le k \le m \mid \mathcal{H}^1 \text{ is true}\right) \\ &= 1 - \mathsf{P}\left\{\left(|Z_1^{(n)}| \le c_\alpha\right) \cap \ldots \cap \left(|Z_m^{(n)}| \le c_\alpha\right) \mid \mathcal{H}^1 \text{ is true}\right\}. \end{aligned} \tag{7}$$

In this paper, we aim to determine the common adjusted critical value c_α, as well as the sample size n, in order to control the FWER at a fixed significance level α and to guarantee a pre-specified minimal power $1-\beta$. We use an iterative procedure based on equations (6) and (7) with two unknown parameters (c_α and n). Clearly the joint distribution of the test statistics used in equations (6) and (7) has to be known under $\mathcal{H}^0$, as well as under $\mathcal{H}^1$. In the latter case, this distribution will depend on the value of the vector of mean differences between the test and control products (reported hereafter as $\boldsymbol{\delta}^* \neq 0$), and also on Σ or an estimate of it.

This is investigated thereafter.

Remark: The equation (6) can also be used alone for determining c_α in order to control the FWER when the aim is to analyse a data set.

2.2.3. *Distribution of* $\mathbf{Z}_n$ *and* $\mathbf{T}_n$.

Normality assumption and known covariance matrix. We assume that the *i.i.d.* random vectors $\mathbf{X}_1^j, \ldots, \mathbf{X}_n^j$ follow a $\mathcal{N}_m(\boldsymbol{\mu}^j, \Sigma)$ distribution with Σ known. In this context, it is easy to show that

$$\mathbf{Z}_n \overset{\mathcal{H}^0}{\sim} \mathcal{N}_m\left(\boldsymbol{0}_m, R\right) \qquad \text{and} \qquad \mathbf{Z}_n \overset{\mathcal{H}^1}{\sim} \mathcal{N}_m\left(\sqrt{\frac{n}{2}} P\boldsymbol{\delta}^*, R\right),$$

where $\boldsymbol{\delta}^* \neq \boldsymbol{0}_m$ is the value of $\boldsymbol{\delta}$ under $\mathcal{H}^1$ and where $R = P\Sigma P$ is the $m \times m$ correlation matrix associated with Σ, with P the diagonal matrix whose k^{th} element is $1/\sigma_k$.

Remark: Senn and Bretz (2007) proposed an alternative method based on a common latent variable in the case where you have a single unvarying pairwise correlation and if the components of $P\boldsymbol{\delta}^*$ (non-centrality parameters) are all the same.

Normality assumption and unknown covariance matrix. In this context, allowing $Y_k = \frac{\bar{X}_k^T - \bar{X}_k^C}{\sqrt{\frac{2}{n}}\sigma_k}$ and $U_k = \nu\frac{\widehat{\sigma}_k^2}{\sigma_k^2}$, we use the vector

$$\mathbf{T}_n = \left(T_1^{(n)}, \ldots, T_m^{(n)}\right)^\intercal = \left(\frac{Y_1}{\sqrt{U_1/\nu}}, \ldots, \frac{Y_m}{\sqrt{U_m/\nu}}\right)^\intercal,$$

where, under the global null hypothesis $\mathcal{H}^0$, the vector $\mathbf{Y} = (Y_1, \ldots, Y_m)^\intercal$ follows a m-dimensional normal distribution with correlation matrix R and where $U_1, \ldots, U_m$ are dependent χ^2 random variables with $\nu = 2n-2$ degrees of freedom. The distribution of the vector $\mathbf{T}_n$ is a type-II multivariate Student distribution with ν degrees of freedom generalization of a bivariate t-distribution considered by Siddiqui (1967), representing the situation of two end points. It has not been possible, as far as we know, to obtain an expression of the density or distribution function of this law in a closed form. Hasler and Hothorn (2011) propose, without justification, to approximate this distribution by an m-variate type-I t-distribution with $\nu = 2n-2$ degrees of freedom and with correlation matrix $\widehat{R}$, an estimate of R. Using the same approximation as these authors, the distribution of the vector $\mathbf{T}_n$ under the alternative hypothesis is approximated by a m-variate type-I t-distribution with the non-centrality parameter $\sqrt{\frac{n}{2}}P\boldsymbol{\delta}^*$ and with $\nu = 2n-2$ degrees of freedom.

Asymptotic context. In order to be more general, we can consider that the covariance matrices differ between the control and test group respectively, namely that we have $\mathbb{V}\text{ar}\left(\mathbf{X}_1^j\right) = \Sigma^j$, $j = C, T$. Then, the usual individual test statistic $T_k^{(n)}$ is defined by

$$T_k^{(n)} = \frac{\bar{X}_k^T - \bar{X}_k^C}{\sqrt{\frac{\widehat{\sigma}_{k,C}^2}{n} + \frac{\widehat{\sigma}_{k,T}^2}{n}}}, \tag{8}$$

where $\widehat{\sigma}^2_{k,j} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X^j_{i,k} - \bar{X}^j_k\right)^2$ for $j = C, T$. The multivariate central limit theorem enables us to state

$$\sqrt{n}\left[(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C) - (\boldsymbol{\mu}^T - \boldsymbol{\mu}^C)\right] \xrightarrow{L} \mathcal{N}_m(\boldsymbol{0}_m, \Sigma),$$

where $\bar{\mathbf{X}}^j = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^j_i$ and where $\Sigma = \Sigma^C + \Sigma^T$ since the two groups are independent. We thus have

$$R^{-1/2}\left[\sqrt{n}V(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C) - \sqrt{n}V\boldsymbol{\delta}^*\right] \xrightarrow{L} \mathcal{N}_m(\boldsymbol{0}_m, I_m), \tag{9}$$

where here $R = V\Sigma V^{\mathsf{T}}$ with $V = \text{diag}\left(1/\sqrt{\sigma^2_{k,C} + \sigma^2_{k,T}}\right)$. In this context, under very general conditions (Cox and Hinkley, 1994, p. 258-266), we can estimate Σ^j by :

$$\widehat{\Sigma}^j = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}^j_i - \bar{\mathbf{X}}^j\right)\left(\mathbf{X}^j_i - \bar{\mathbf{X}}^j\right)^{\mathsf{T}}.$$

Then, $\widehat{R} = \widehat{V}\ \widehat{\Sigma}\ \widehat{V}$ is a consistent estimator of R, the correlation matrix of $\mathbf{T}_n = \sqrt{n}\widehat{V}(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C)$, where $\widehat{V} = \text{diag}\left(1/\sqrt{\widehat{\sigma}^2_{k,C} + \widehat{\sigma}^2_{k,T}}\right)$ and $\widehat{\Sigma} = \widehat{\Sigma}^C + \widehat{\Sigma}^T$. Now, using the Slutsky's theorem, we obtain

$$\widehat{R}^{-1/2}\mathbf{T}_n \xrightarrow{L} \mathcal{N}_m\left(\boldsymbol{0}_m, I_m\right), \text{ under } \mathcal{H}^0,$$

and

$$\widehat{R}^{-1/2}\left(\mathbf{T}_n - \sqrt{n}\widehat{V}\boldsymbol{\delta}^*\right) \xrightarrow{L} \mathcal{N}_m\left(\boldsymbol{0}_m, I_m\right), \text{ under } \mathcal{H}^1 : \boldsymbol{\delta} = \boldsymbol{\delta}^* \neq \boldsymbol{0}_m.$$

2.2.4. *Practical implementation.* Computation of the adjusted critical value c_α and determination of the sample size n are done in **R**, an open source statistical software (**R** Development Core Team, 2011). We used the `pmvnorm()` and `pmvt()` functions from the `mvtnorm` package (Genz and Bretz, 2009; Genz et al., 2012) for the computation of the multivariate normal and of the multivariate type-I t-distribution probabilities. The sample size computation involves an effect size parameter. We recall that the effect size for the k^{th} end point is defined as $\Delta_k = \frac{\mu^T_k - \mu^C_k}{\sigma^*_k}$, where σ^*_k is the population standard deviation of variable X_k. Note that σ^*_k can be expressed in terms of the standard deviations of the variables X^C_k and X^T_k. In our framework of normality assumption with known or unknown covariance matrix, the standard deviation σ^*_k equals σ_k and the vector of effect size for the m end points corresponds to the term $P\boldsymbol{\delta}^*$. In the asymptotic context, as we consider a standard deviation in the control group which is different from the test group ($\sigma_{k,T} \neq \sigma_{k,C}$), the vector of effect size corresponds to $\sqrt{2}V\boldsymbol{\delta}^*$. In this latter definition, we consider that $\sigma^*_k = \sqrt{\frac{\sigma^2_{k,C} + \sigma^2_{k,T}}{2}}$.

Two iterative procedures have been defined according to the assumptions made.

Normality assumption and known covariance matrix. Briefly, the procedure based on the `pmvnorm()` function consists of performing the following steps:

(i) Specifying the effect size $P\boldsymbol{\delta}^*$ for all end points, the correlation matrix R, the significance level α and the desired power $1 - \beta$;
(ii) determining c_α as a solution of $FWER = \alpha$ in equation (6);

(iii) for a starting value n_0 of sample size, computing the minimal power $1-\beta$ using equation (7) with n_0 and c_α from step (ii);
(iv) going back to (iii) with an incremented or decremented sample size n_0 until the desired power.

Normality assumption and unknown covariance matrix. The procedure (based on the `pmvt()` function) is slightly different because the distribution of $\mathbf{T}_n$ under the null hypothesis depends on the sample size:

(i) Specifying the effect size for all end points defined by $P\boldsymbol{\delta}^*$, the correlation matrix $\widehat{R}$ (which can be given by a pilot study), the significance level α and the desired power $1-\beta$;
(ii) for a starting value n_0 of sample size, determining c_α as a solution of $FWER = \alpha$ in equation (6);
(iii) computing the minimal power $1-\beta$ using equation (7) with n_0 and c_α from step (ii);
(iv) going back to (ii) with an incremented or decremented sample size n_0 until the desired power.

Asymptotic context. The principle of the procedure is the same as the procedure for the *Normality assumption and known covariance matrix* case. Only the specification of the effect size for all end points defined by $\sqrt{2}V\boldsymbol{\delta}^*$ and of the correlation matrix $\widehat{R}$ change and can be defined by a pilot study. This definition of the effect size parameter permits a homogeneous notation with both previous cases.

2.3. **Global procedure.**

2.3.1. *Model.* We propose the following multivariate linear regression model to represent the data generation process:

$$\mathbf{Y} = \Gamma B + \mathbf{E}, \tag{10}$$

where $\mathbf{Y}^\intercal = \left[\mathbf{X}_1^C; \ldots; \mathbf{X}_n^C; \mathbf{X}_1^T; \ldots; \mathbf{X}_n^T\right]$ is a $m \times 2n$ matrix, $\Gamma = [\mathbf{1}_{2n}; \mathbf{G}; \mathbf{A}]$ is the $2n \times (p+2)$ design matrix, with $\mathbf{1}_{2n} = (1, \ldots, 1)^\intercal$, $\mathbf{G} = (\mathbf{1}_n^\intercal, \mathbf{0}_n^\intercal)^\intercal$, being an indicator variable of each group (1 for the control group and 0 for the treatment group), $\mathbf{A}$ is a $2n \times p$ matrix whose l^{th} column $\boldsymbol{a}_l = (a_{l1}^C, \ldots, a_{ln}^C, a_{l1}^T, \ldots, a_{ln}^T)^\intercal$ contains the measurements of the l^{th} adjustment variable ($1 \leq l \leq p$) on the $2n$ subjects, B is a $(p+2) \times m$ matrix of unknown coefficients associated with the design matrix and $\mathbf{E}$ is a $2n \times m$ random matrix of errors such that $\text{vec}(\mathbf{E})$ follows a $\mathcal{N}_{2n\times m}(\boldsymbol{0}_m; I_{2n} \otimes \Sigma)$ distribution, where $\text{vec}(\cdot)$ denotes the column-stacking operator and $\otimes$ denotes the Kronecker symbol. We let $\boldsymbol{\delta}$ be the second row of the matrix B, which represents the adjusted group effect for the m end points. It is worthwhile mentioning that $\boldsymbol{\delta} = \mathbb{E}[\mathbf{X}_i^T - \mathbf{X}_i^C | \mathbf{a}_i]$. In the sequel, we will note $\bar{a}_l^j = (1/n)\sum_{i=1}^n a_{li}^j$ and $\overline{a_l a_{l'}}^j = (1/n)\sum_{i=1}^n a_{li}^j a_{l'i}^j$, $j = C, T$.

To overcome the multiple testing problem, we propose using a global test of the hypothesis

$$\mathcal{H}^0 : \boldsymbol{\delta} = \boldsymbol{0}_m \quad \text{versus} \quad \mathcal{H}^1 : \boldsymbol{\delta} \neq \boldsymbol{0}_m.$$

In the framework of model (10), this can be restated as

$$\mathcal{H}^0 : \boldsymbol{C}B = \boldsymbol{0}_m^\intercal \text{ versus } \mathcal{H}^1 : \boldsymbol{C}B \neq \boldsymbol{0}_m^\intercal,$$

using the contrast row vector $\boldsymbol{C} = (0, 1, \mathbf{0}_p^\intercal)$ of size $1 \times (p+2)$.

Under the multinormality assumption of the disturbances, the least square estimator of matrix B is given by:

$$\hat{\mathbf{B}} = \left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y} \sim \mathcal{N}_{(p+2)\times m}\left(B, \left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1} \otimes \Sigma\right).$$

We can thus state

$$\boldsymbol{C}\hat{\mathbf{B}} = \boldsymbol{C}\left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y} \sim \mathcal{N}_m\left(\boldsymbol{C}B, W\right),$$

where

$$W = \left(\boldsymbol{C} \otimes I_m\right)\left[\left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1} \otimes \Sigma\right]\left(\boldsymbol{C}^{\mathsf{T}} \otimes I_m\right) : m \times m. \tag{11}$$

This leads to

$$W^{-1/2}\left[\boldsymbol{C}\hat{\mathbf{B}} - \boldsymbol{C}B\right] = W^{-1/2}\left[\boldsymbol{C}\left(\Gamma^{\mathsf{T}}\Gamma\right)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y} - \boldsymbol{C}B\right] \sim \mathcal{N}_m(\boldsymbol{0}_m, I_m).$$

2.3.2. *Statistical procedure and distribution.*

Known covariance matrix Σ. In this context, the test statistic considered is:

$$Z_n^2 = \left(\boldsymbol{C}\hat{\mathbf{B}}\right) W^{-1} \left(\boldsymbol{C}\hat{\mathbf{B}}\right)^{\mathsf{T}}.$$

After some relatively easy but tedious computations, we were able to show, using a formula for matrix inversion in block form, that W from (11) can be written as

$$W = (\Gamma^{\mathsf{T}}\Gamma)^{-1}_{2,2}\Sigma = \frac{1}{n}\left(2 + \boldsymbol{v}^{\mathsf{T}}M^{-1}\boldsymbol{v}\right)\Sigma,$$

where $\boldsymbol{v}$ is a $p \times 1$ vector whose l^{th} component is $v_l = \bar{a}_l^T - \bar{a}_l^C$, and where M is a $p \times p$ matrix with general term $M_{l,l'} = \left(\overline{a_l a_{l'}}^C - \bar{a}_l^C\bar{a}_{l'}^C\right) + \left(\overline{a_l a_{l'}}^T - \bar{a}_l^T\bar{a}_{l'}^T\right)$.

It can be shown (Bilodeau and Brenner, 1999) that, under the null hypothesis, Z_n^2 follows a χ^2 distribution with m degrees of freedom, reported as χ^2_m. We then reject the null hypothesis $\mathcal{H}^0$ if the observed value of the test statistic Z_n^2 is greater than $q^m_{1-\alpha}$, the quantile of the order $1-\alpha$ of the χ^2_m. Under the alternative hypothesis $\mathcal{H}^1 : \boldsymbol{C}B = \boldsymbol{\delta}^{*\mathsf{T}}$ (with $\boldsymbol{\delta}^* \neq \boldsymbol{0}_m$), the test statistic Z_n^2 follows a non-central χ^2 distribution with m degrees of freedom and decentrality parameter

$$\lambda_n = \boldsymbol{\delta}^{*\mathsf{T}}W^{-1}\boldsymbol{\delta}^*. \tag{12}$$

Unknown covariance matrix Σ. In this context, the test statistic considered is

$$T_n^2 = \left(\boldsymbol{C}\hat{\mathbf{B}}\right)\widehat{W}_n^{-1}\left(\boldsymbol{C}\hat{\mathbf{B}}\right)^{\mathsf{T}},$$

where $\widehat{W}_n = \left(\boldsymbol{C} \otimes I_m\right)\left[(\Gamma^{\mathsf{T}}\Gamma)^{-1} \otimes \widehat{\Sigma}\right]\left(\boldsymbol{C}^{\mathsf{T}} \otimes I_m\right) = \frac{1}{n}\left(2 + \boldsymbol{v}^{\mathsf{T}}M^{-1}\boldsymbol{v}\right)\widehat{\Sigma}$, with $\widehat{\Sigma}$ an unbiased estimator of Σ (see Appendix 2 for an explicit definition). In this case, under the null hypothesis, T_n^2 converges in distribution to a χ^2 distribution with m degrees of freedom. We also prove that, under the alternative hypothesis $\mathcal{H}^1 : \boldsymbol{C}B = \boldsymbol{\delta}^{*\mathsf{T}} \neq \boldsymbol{0}_m^{\mathsf{T}}$, the test statistic T_n^2 converges in distribution to a non-central χ^2 distribution with m degrees of freedom and decentrality parameter

$$\lambda_n = \boldsymbol{\delta}^{*\mathsf{T}}\widehat{W}_n^{-1}\boldsymbol{\delta}^*. \tag{13}$$

We note that, without adjustment variables in model (10), the test statistic T_n^2 reduces to the classical Hotelling's test statistic (see Appendix 2 for a proof of this result).

2.3.3. *Power and sample size determination.* In the context of this global test, the power function for the statistic Z_n^2 (or T_n^2) is

$$1-\beta = \mathsf{P}\left(Z_n^2 > q_{1-\alpha}^m | \mathcal{H}^1\right) = 1 - F_{\chi^2_m(\lambda_n)}(q_{1-\alpha}^m), \tag{14}$$

where $F_{\chi^2_m(\lambda_n)}(\cdot)$ is the cumulative distribution function of the non-central χ^2_m with decentrality parameter λ_n. The sample size required to achieve the desired power $1-\beta$ is given as the smallest integer satisfying equation (14), using the decentrality parameter λ_n as given in (12) (or (13)).

2.3.4. *Practical implementation.* To achieve the sample size computation, the user specifies the vector $\boldsymbol{\delta}^*$ of mean differences between the test and the control products, the covariance matrix Σ between the outcomes, the desired significance level α, and the desired power $1-\beta$. The R program we developed enables computation of the decentrality parameter λ_n using equation (12), and the sample size using equation (14) (or (13)).

In the presence of a single adjustment variable, the decentrality parameter λ_n from (13) reduces to

$$\lambda_n = \boldsymbol{\delta}^{*\mathsf{T}} \left\{ \frac{1}{n} \left(2 + \frac{(\bar{a}_1^T - \bar{a}_1^C)^2}{\overline{a_1^2}^C - (\bar{a}_1^C)^2 + \overline{a_1^2}^T - (\bar{a}_1^T)^2} \right) \hat{\Sigma} \right\}^{-1} \boldsymbol{\delta}^*,$$

where $\bar{a}_1^j = (1/n)\sum_{i=1}^n a_{1i}^j$, $\overline{a_1^2}^j = (1/n)\sum_{i=1}^n a_{1i}^{2\ j}$ and where a_{1i}^j is the value of the adjustment variable for the i-th subject for treatment j ($j = C$: control; $j = T$: treatment). From a practical point of view, the user has to specify this parameter, which can be evaluated after a pilot study has been conducted. Note that for a binary adjustment variable, for example gender (1=women and 0=men), the user only has to specify the frequency of women in each group (since in this case $\overline{a_1^2}^C = \bar{a}_1^C$). Moreover, if the number of women in each group is the same, the computation of the decentrality parameter λ_n reduces to the case without an adjustment variable.

3. Results

3.1. **Simulations.** A simulation study was performed to evaluate the performance of the two proposed approaches. We first studied how the individual testing procedure was able to control the FWER. We investigated three different assumptions: normality and known covariance matrix (termed "NormKnown"), normality and unknown covariance matrix (termed "NormUnKnown"), and the asymptotic context (termed "Asympt"). We also compared the results obtained with the Bonferroni method and with the "naive method", which consists of choosing the most significant test without any correction of the FWER. We then investigated the power of our proposed approaches and compared them to standard approaches such as Holm (1979), Hochberg (1988), Bonferroni and Hotelling (1953); methods which are implemented using the `multtest` R package (Pollard et al., 2005).

All data for these simulations came from the model defined in (10) with one adjustment variable ($p = 1$), which follows a Bernoulli distribution with a probability of success $\pi = 0.6$. Each simulation was carried out on 200 subjects in each group. To simplify the interpretation and shorten the simulation study, we considered a compound symmetric covariance matrix Σ with $\mathrm{diag}(\Sigma) = \mathbf{1}_m$. Note that we thus

have a constant pairwise correlation ρ for all pairs of outcomes. Moreover, as multiple end points are often correlated in the same direction, we only investigated positive correlations ($\rho > 0$). We used 5 000 replications for all simulations. In the sequel, we define $B^{\mathsf{T}} = [\boldsymbol{b_0}, \boldsymbol{\delta}, \boldsymbol{b_1}]$ where the vector $\boldsymbol{b_0}$ $(m \times 1)$ represents the intercept of the model and where $\boldsymbol{b_1}$ represents the coefficient vector associated with the adjustment variable.

3.1.1. *FWER.* We first investigated, for different numbers of end points ($m \in \{1, \ldots, 15\}$), the control of the FWER for the individual testing procedure (under the three assumptions), for the Bonferroni method and for the naive approach. Under the null hypothesis ($\boldsymbol{\delta} = \boldsymbol{0}_m$), the FWER was estimated by the proportion of the Monte-Carlo experiments that lead to a rejection of the null hypothesis ($p_{value} < 0.05$). In this simulation, we considered a model without an adjustment variable ($\boldsymbol{b}_1 = \boldsymbol{0}_m$), the intercept vector was fixed at $\boldsymbol{b}_0 = \boldsymbol{1}_m$ and finally ρ was set to 0.6. Figure 1 shows the evolution of the FWER for different numbers of end points.

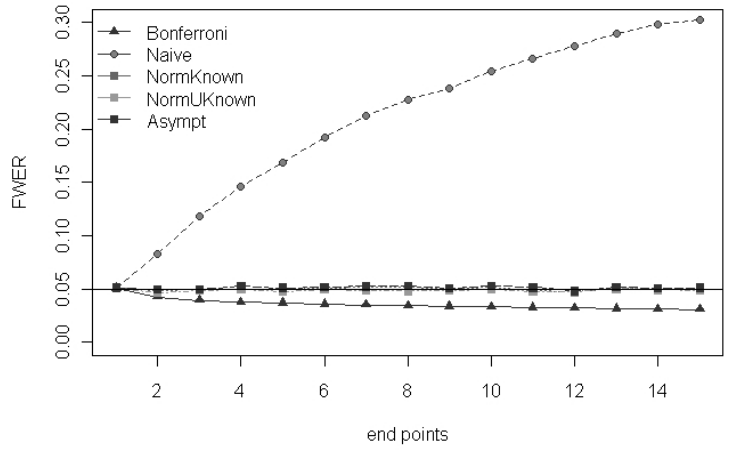


FIGURE 1. FWER as a function of the number of end points for the Naive, Bonferroni, and individual testing procedure

The *naive* method without correction for the multiple testing problem, increased with the number of end points. The error rate calculated by the Bonferroni method decreased with the number of end points. This correction was therefore too conservative whereas the individual testing procedure gave a Type-I error rate close to the nominal 0.05 value, under the three assumptions.

3.1.2. *Power.* We investigated the performance of the proposed approaches with varying correlations among the outcomes. We considered $m = 3$ correlated end points. First, we investigated the case where the adjustment variable has no effect ($\boldsymbol{b_1} = \boldsymbol{0}_m$) on the outcomes. We then used $\boldsymbol{b_1} = (1.0, 0.8, 0.6)^{\mathsf{T}}$ in order to determine the effect of an adjustment variable. In the two simulations cases, we fixed

the treatment effect at $\boldsymbol{\delta} = (0.31, 0.13, 0.19)^{\mathsf{T}}$ and the intercept at $\boldsymbol{b_0} = (2, 3, 1)^{\mathsf{T}}$.

No effect of the adjustment variable.

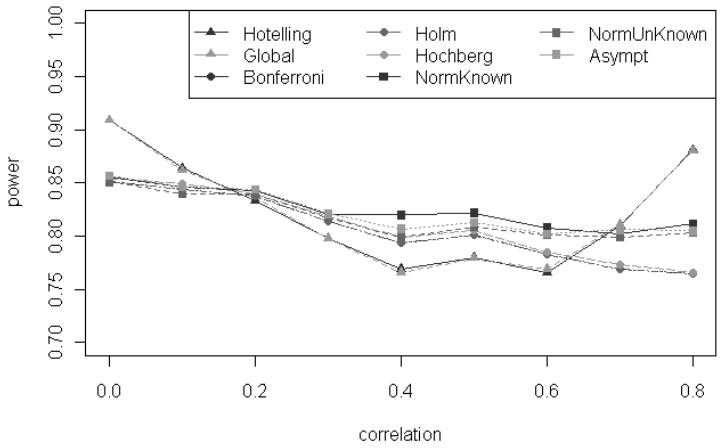


FIGURE 2. Minimal power study of different methods as a function of the correlation coefficient ρ from data which came from a model with no adjustment variable effect.

The results of the minimal power for the different methods are presented in Figure 2. We can see that for low correlation ($\rho < 0.2$) and high correlation ($\rho > 0.7$), the global method seems to be more powerful than the individual testing procedure which was more powerful than the other methods. For medium correlation, the individual testing method was the most powerful and the global procedures were less so. Finally, for high correlation ($\rho > 0.6$), as expected, the methods which do not take into account the correlation between the end points were the least powerful.

Effect of the adjustment variable. The results of the minimal power for the different methods are presented in Figure 3. In this simulation, we can see that the global procedure and the individual testing procedure with known covariance matrix assumption were more powerful than the others. Global method was more powerful for low and high correlation whereas the individual testing procedure for known covariance matrix assumption gave better results for medium correlation. In this situation, the global method outperformed the Hotelling test. Taking into account an adjustment variable improves the estimation of the covariance matrix, which results in an increase of the power with the global method. The other methods are less powerful in this situation since they do not take into account the adjustment variable. However, the individual testing procedure was still more powerful than methods which do not take into account the correlation between end points.

3.2. **Sample Size Computation.** We present some results about sample size calculation in the context of three end points with the following parameters: the

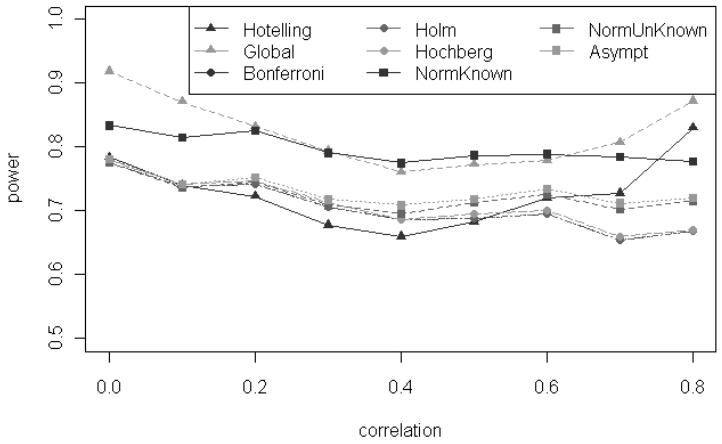


FIGURE 3. Minimal power study of different methods as a function of the correlation coefficient ρ for data generated from a model involving an adjustment variable.

vector $\boldsymbol{\delta}^* = (0.2, 0.3, 0.4)^\intercal$ of mean differences between the test and the control products, a compound symmetric covariance matrix Σ between the outcomes, with $\text{diag}(\Sigma) = (1.1^2, 1.2^2, 2.3^2)^\intercal$. The desired significance level was chosen at $\alpha = 0.05$. For the global method with a binary adjustment variable, the frequency of this variable in each group is fixed for samples of any size to $\bar{a}^C = 0.4$ and $\bar{a}^T = 0.6$. The results in Table 1 are coherent with the previous power simulation study. The sample size of each group required to reach a desired minimal power increases with the correlation coefficient for the Bonferroni procedure. We can also observe that, for low ($\rho < 0.2$) and high ($\rho > 0.6$) correlations, the procedure based on the model requires fewer subjects than the other methods. For medium correlations, it is the individual testing procedure which requires the smallest sample size. We note that the global method with an adjustment variable (MA) requires more subjects than a model without an adjustment variable (M). While this may appear strange in regards to the previous power simulation study, in fact, the sample size calculation assumes that Σ is known. Therefore adding an adjustment variable to improve the estimation of Σ is not useful. However, in practice Σ is unknown (even if it has been estimated in a previous study with a very large sample size). We recommend that the sample size be determined using an adjustment variable if further data analysis is to be performed.

3.3. **Application in clinical studies in nutrition.** The purpose of this section is to present the results obtained using our methods, in terms of sample size determination and the statistical data analysis. The two following applications deal with clinical studies performed in nutrition. Both studies were double-blind Randomized Controlled Trials (DB-RCT) performed according to Good Clinical Practices (ICH-GCP).

TABLE 1. Sample size n of each group required to achieve the desired level of minimal power: for Bonferroni (B), for our individual testing procedure for known (K) or unknown (U) covariance matrix, and in the asymptotic context (A), for our global method based on a multivariate model without an adjustment variable (M), and with a binary adjustment variable (MA), for various correlations ρ and with a FWER=0.05.

ρ	Power (0.80)					Power (0.90)				
	B	K/A	U	M	MA	B	K/A	U	M	MA
0	221	219	222	174	181	287	285	288	226	235
0.1	233	231	233	207	215	304	303	305	268	280
0.2	246	243	245	238	248	322	319	321	309	322
0.3	258	255	256	268	279	340	336	337	349	363
0.4	272	265	267	296	308	358	350	352	385	401
0.5	285	276	277	320	334	376	365	366	416	434
0.6	299	285	286	339	354	393	376	378	441	459
0.7	312	292	293	349	363	409	386	387	453	472
0.8	325	295	297	338	352	423	390	391	440	458
0.9	333	291	292	278	289	431	383	385	361	376

3.3.1. *Example 1: Sample size calculation.* The first application is the sample size determination of a new DB-RCT with the objective of demonstrating the efficacy of the consumption of a dairy product on seric antibody titres for the three strains of Influenza virus. So, for k^{th} strain, the individual null hypothesis is $\mathcal{H}_k^0 : \delta_k = \mu_k^1 - \mu_k^0 = 0$. The product will be considered as effective if at least one out of the three strains is statistically significant. According to the usual standard, the Type-II error probability is fixed at 20% in order to obtain a power of 80% and the Family Wise Error rate must be controlled at 5%.

Two pilot studies were planned to define the product effects and variability. Both were DB-RCT multicentric studies conducted in France among elderly volunteers during the two vaccination seasons 2005 & 2006. Details are reported in Boge et al. (2009). Based on the results, the product's effects defined by means and correlations were calculated. The mean differences between both groups is $\hat{\boldsymbol{\delta}} = (0.35, 0.28, 0.46)^{\mathsf{T}}$ and the covariance matrix was $\widehat{\Sigma} = \begin{pmatrix} 5.58 & 2.00 & 1.24 \\ 2.00 & 4.29 & 1.59 \\ 1.24 & 1.59 & 4.09 \end{pmatrix}$. Following experts consensus the mean differences obtained could be considered as clinically relevant. Based on these assumptions, we compared the most powerful methods, namely the global and the individual procedure for known covariance matrix. Table 2 shows that the sample size may be reduced significantly depending on the method used. Indeed, with an individual procedure for known covariance matrix and an adjusted Type-I error probability at 0.0178, the sample size falls to 336 subjects required to see a significant difference for at least one outcome, versus 359 for the global method.

3.3.2. *Example 2: Analysis of Clinical Study Data.* In order to demonstrate the effect of a fermented dairy product on the immune system, a monocentric, DB-RCT, parallel study with two groups was performed in 1 000 healthy subjects.

TABLE 2. Sample size computation with Global method and Individual Procedure

Method	Type-I error	Sample size (n)
Global	0.05	359
Indiv	0.0178	336

Global: Global method based on multivariate model;

Indiv: Individual procedure for known covariance matrix.

Results are reported in Guillemard et al. (2009). As an exploratory analysis, the immune function of interest was characterized by a set of 11 biomarkers. According to the exploratory concept of this analysis, the product efficacy was assumed if at least one out of the 11 markers was statistically significant. During this analysis, the covariance matrix between parameters and the means vector was estimated on the actual data.
The statistical analysis was performed using a test of comparison of means with common multiple testing procedures and with the proposed asymptotic individual testing procedure defined in section 2. A global procedure involving a model without an adjustment variable was also used. The functions `indiv.1m.analysis()` and `global.1m.analysis()` from **R** package `Sample` were used to perform the analysis. The results for the individual method (three assumptions) in terms of "adjusted p_{value}" estimation are summarized in Table 3.

TABLE 3. Adjusted p_{value} estimation on eleven immunological markers for various multiple testing procedures, in an efficacy study of fermented dairy product.

Endpoint	Naive	Bonferroni	Holm	Hochberg	Asympt
1	0.13	1.00	0.74	0.60	0.62
2	0.15	1.00	0.74	0.60	0.67
3	0.45	1.00	0.90	0.67	0.98
4	0.67	1.00	0.90	0.67	1.00
5	0.10	1.00	0.70	0.60	0.52
6	0.02*	0.21	0.19	0.18	0.14
7	0.00*	0.01*	0.01*	0.01*	0.01*
8	0.30	1.00	0.90	0.67	0.91
9	0.12	1.00	0.74	0.60	0.59
10	0.07	0.76	0.55	0.55	0.40
11	0.02*	0.22	0.19	0.18	0.14

*: significant association.

The importance of using a Type-I error correction can be seen in this Table. Without any correction ("naive method"), we could conclude that there are three significant end points (markers 6, 7 & 11). But, when a correction method is used, only one significant endpoint is found (marker 7). This conducts to conclude to the efficacy of the product. The global method which gives us only a single result, is also significant with a p_{value} less than 0.01 and confirms that we have at least one marker which is significant.

4. Concluding remarks

In this paper, we considered two approaches (individual and global) for sample size determination and for the analysis of data with multiple continuous end points. The global method we have developed leads to a generalization of the well known Hotelling (1953)'s statistic, involving adjustment variables. The methods developed allow consideration of cases when the covariance matrix is known or estimated. When designing clinical studies, assumptions for the sample size calculation may come from different sources which are more or less biased. The best situation is to be able to gather data from a well-designed pilot study, among defined populations on relevant and well-measured endpoints. In this case, the estimator of the mean differences could be considered as non-or slightly biased. An interesting approach would be to consider the lower and upper boundaries of the confidence interval of the estimator and to calculate the two sample sizes associated to them. Regarding the covariance matrix, the bias may be more sensitive but several approaches might be used in order to correct this bias as described, for instance, by Julious and Owen (2006). In any case, it is important not to consider the results from literature or previous studies as "known values" but always as estimations with their variability.

Simulation studies showed that the method based on the global model with adjustment variables is powerful method when the true covariance matrix is unknown. Consequently, better sample size computations are possible. When adjustment variables are not available, the individual method seems to be more powerful, except for low and high correlation cases. Furthermore, in this context, the methods developed perform favorably compared to common procedures. Therefore, the choice of the method depends mainly on the aim of the study; for example the global method gives only a global result and not a directional one. However the choice also depends on the value of the correlation coefficient between the end points. Work on the generalization of the individual method in the context of detecting r significant end points among m is ongoing. In conclusion, we have implemented various methods in an `R` package called `Sample`. In this report, we have focused our analysis on bilateral tests, however this package also takes into account the unilateral case for the individual procedure.

Acknowledgement

We would like to thank the Danone Research Clinical Study Platform for making the data available. The work of the first author was supported by a grant from the Natural Science and Engineering Research Council of Canada.

References

Berger, R. (1982). Multiparameter hypothesis testing and acceptance sampling. Technometrics , 295–300.

Bilodeau, M., Brenner, D. (1999). Theory of multivariate statistics. Springer.

Boge, T., Rémigy, M., Vaudaine, S., Tanguy, J., Bourdet-Sicard, R., Van der Werf, S. (2009). A probiotic fermented dairy drink improves antibody response to influenza vaccination in the elderly in two randomised controlled trials. Vaccine 27, 5677–5684.

Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in medicine 28, 586–604.

Bretz, F., Maurer, W., Hommel, G. (2011). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. Statistics in medicine 30, 1489–1501.

Burman, C., Sonesson, C., Guilbaud, O. (2009). A recycling framework for the construction of bonferroni-based multiple tests. Statistics in Medicine 28, 739–761.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., Offen, W. (2007). Challenge of multiple co-primary endpoints: a new approach. Statistics in medicine 26, 1181–1192.

Cook, R., Farewell, V. (1996). Multiplicity considerations in the design and analysis of clinical trials. Journal of the Royal Statistical Society. Series A (Statistics in Society) , 93–110.

Cox, D., Hinkley, D. (1994). Theoretical statistics. Chapman & Hall.

R Development Core Team (2011). **R**: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Dmitrienko, A., Offen, W., Westfall, P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statistics In Medicine 22, 2387–2400.

Dunnett, C., Tamhane, A. (1992). A step-up multiple test procedure. Journal of the American Statistical Association , 162–170.

Genz, A., Bretz, F. (2009). Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T. (2012). mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9992.

Guillemard, E., Tondu, F., Lacoin, F., Schrezenmeir, J. (2009). Consumption of a fermented dairy product containing the probiotic lactobacillus casei dn-114001 reduces the duration of respiratory infections in the elderly in a randomised controlled trial. British Journal of Nutrition 103, 58–68.

Hasler, M., Hothorn, L.a. (2011). A Dunnett-Type Procedure for Multiple Endpoints. The International Journal of Biostatistics , 74–81.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika 75, 383–386.

Hotelling, H. (1953). A generalised t test and measure of multivariate dispersion. Proceedings of the second Berkeley Symposium on Mathematical Statistics and Probability , 23–41.

Julious, S., McIntyre, N. (2012). Sample sizes for trials involving multiple correlated must-win comparisons. Pharmaceutical Statistics .

Julious, S., Owen, R. (2006). Sample size calculations for clinical studies allowing for uncertainty about the variance. Pharmaceutical statistics 5, 29–37.

Neuhäuser, M. (2006). How to deal with multiple endpoints in clinical trials. Fundamental & Clinical Pharmacology 20, 515–23.

O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. Biometrics , 1079–1087.

Pocock, S., Geller, N., Tsiatis, A. (1987). The analysis of multiple endpoints in clinical trials. Biometrics , 487–498.

Pollard, K.S., Gilbert, H.N., Ge, Y., Taylor, S., Dudoit, S. (2005). multtest: Resampling-based multiple hypothesis testing. R package version 2.6.0.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics 24, 2387–2400.

Sankoh, A., Huque, M., Dubey, S. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Statistics in medicine 16, 2529–2542.

Sankoh, A., Huque, M., Russel, H., D'Agostino, R. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. Drug Information Journal 33, 119–140.

Senn, S., Bretz, F. (2007). Power and sample size when multiple endpoints are considered. Pharmaceutical statistics 6, 161–170.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62, 626–633.

Siddiqui, M, M. (1967). A bivariate t-distribution. Annals of Mathematical Statistics 38, 162–166.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

Sozu, T., Kanou, T., Hamada, C., Yoshimura, I. (2006). Power and sample size calculations in clinical trials with multiple primary variables. Japanese Journal of Biometrics , 83–96.

Sozu, T., Sugimoto, T., Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. Statistics in Medicine 29, 2169–79.

Sozu, T., Sugimoto, T., Hamasaki, T. (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. Journal of Biopharmaceutical Statistics, 21, 650–68.

Westfall, P., Tobias, R., Rom, D., Wolfinger, R., Hochberg, Y. (1999). Multiple comparisons and multiple tests. SAS Press.

Williams, J.D., Woodall, W.H., Birch, J.B., Sullivan, J.H. (2004). Distributional Properties of the Multivariate T2 Statistic Based on the Successive Differences Covariance Matrix Estimator. Technical Report. Department of Statistics, Virginia Polytechnic Institute and State University.

Yoon, F.B., Fitzmaurice, G.M., Lipsitz, S.R., Horton, N.J., Laird, N.M., Normand, S.L.T. (2011). Alternative methods for testing treatment effects on the basis of multiple outcomes: Simulation and case study. Statistics in Medicine 30, 1917–32.

Appendix 1

Sample Package

The `Sample` package was developed in **R**, an open source statistical software available at `http://www.r-project.org`. It contains, for the moment, five functions: three for the sample size computation (one for the individual procedure, one for the global method, and one for the Bonferroni procedure) and two for the analysis of real data in order to solve the multiple testing problem (one for the individual procedure, one for the global method). We present an illustration of the main `Sample` functions below.

Briefly, concerning the sample size determination, the user needs to specify in the `indiv.1m.ssc()` function the alternative hypothesis (bilateral or unilateral), the effect size and the correlation between the end points. However, for the global method, the user also needs to specify in the `global.1m.ssc()` function, the difference of means between the two groups ($\boldsymbol{\delta}^*$), the vector of the standard deviations of each end point, instead of the effect size. With an adjustment variable, the user also needs to enter the mean of the adjustment variable for each group. The functions for sample size computation provide the adjusted significance level and the required sample size. The `bonferroni.1m.ssc()` function is not displayed below.

```
# Sample size computation for the individual method:
> indiv.1m.ssc(method="Known", ES=c(0.1,0.2,0.3),cor=diag(1,3))

Sample size:  183
Adjusted significance level:  0.0170

# Sample size computation for the global method:
> global.1m.ssc(method="Adj.Model",mean.diff=c(0.1,0.2,0.3),
sd=c(1,1,1),cor=diag(1,3),v=-0.2,M=0.46)

Sample size:  163
```

Concerning the analysis of the data, the user needs to specify in the `indiv.1m.analysis()` function, the alternative hypothesis (bilateral or unilateral) and the assumption used: asymptotic test or normality assumption. The function provides the adjusted p_{value} associated to each end point. For the analysis based on the multivariate model, the user specifies the adjustment covariable in the `global.1m.analysis()` function that returns the p_{value} of the global test.

```
> data(data.sim)
> n <-  nrow(data)/2
> XC <- data[1:n,1:3]
> XT <- data[(n+1):(2*n),1:3]

# Data analysis for the individual method:
> indiv.1m.analysis(method="UnKnown",XC=XC,XT=XT,n=n)
```

```
Endpoint           1        2        3
Adjusted p-value   0.4164   0.1419   0.0076

# Data analysis for the global method:
> global.1m.analysis(XC=XC,XT=XT,A=data[,5],n=n)

p-value:  0.0023
```

Appendix 2

Statistic T_n^2 reduces to Hotelling's test statistic. Without adjustment covariates in the multivariate model, the matrix $\widehat{W}_n = (\boldsymbol{C}\otimes I_m)\left[(\Gamma^{\mathsf{T}}\Gamma)^{-1}\otimes\widehat{\Sigma}\right](\boldsymbol{C}^{\mathsf{T}}\otimes I_m)$ reduces to $\widehat{W}_n = \frac{2}{n}\hat{\Sigma}$. Then, the statistic T_n^2 is defined as

$$T_n^2 = \frac{n}{2}(\boldsymbol{C}\hat{\mathbf{B}})\hat{\Sigma}^{-1}(\boldsymbol{C}\hat{\mathbf{B}})^{\mathsf{T}},$$

where

$$\hat{\mathbf{B}} = (\Gamma^{\mathsf{T}}\Gamma)^{-1}\Gamma^{\mathsf{T}}\mathbf{Y}$$

and

$$\widehat{\Sigma} = \frac{1}{2n-(2+p)}(\mathbf{Y}-\Gamma\hat{\mathbf{B}})^{\mathsf{T}}(\mathbf{Y}-\Gamma\hat{\mathbf{B}}),$$

with p the number of adjustment variables. In the context of no adjustment variable, we obtain

$$\hat{\mathbf{B}} = \left[\bar{\mathbf{X}}^C, \bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C\right]^{\mathsf{T}}$$

and thus

$$\widehat{\Sigma} = \frac{1}{2n-2}\sum_{i=1}^{n}\left[\left(\mathbf{X}_i^C - \bar{\mathbf{X}}^C\right)\left(\mathbf{X}_i^C - \bar{\mathbf{X}}^C\right)^{\mathsf{T}} + \left(\mathbf{X}_i^T - \bar{\mathbf{X}}^T\right)\left(\mathbf{X}_i^T - \bar{\mathbf{X}}^T\right)^{\mathsf{T}}\right].$$

Finally, the statistic T_n^2 can be written as

$$T_n^2 = \frac{n}{2}(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C)^{\mathsf{T}}\widehat{\Sigma}^{-1}(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C).$$

This statistic is also known as the Hotelling's two-sample T-squared statistic T^2 when the two groups have the same sample size ($n_T = n_C$) and follows a Hotelling's T-squared distribution:

$$T^2 = \frac{n_T n_C}{n_T + n_C}\left(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C\right)^{\mathsf{T}}\widehat{\Sigma}^{-1}\left(\bar{\mathbf{X}}^T - \bar{\mathbf{X}}^C\right) \sim T^2(m, n_T + n_C - 2).$$

Note that Williams et al. (2004) showed that

$$T^2 \xrightarrow{L} \chi_m^2.$$

Thus, using a quantile based on the chi-squared distribution or based on the Hotelling's T-squared distribution will logically give the same results.

4

Calcul de taille d'échantillon pour un contrôle de la «r-Power»

Résumé

A l'heure actuelle en recherche clinique, un nombre croissant de plans expérimentaux utilisent des critères de jugement principaux multiples. Dans ce contexte, l'étude sera perçue comme un succès par les promoteurs s'il est possible de rejeter au moins r hypothèses nulles parmi l'ensemble des m hypothèses nulles testées. Dans ce contexte, le statisticien se doit de prendre en compte la multiplicité induite par cette pratique. Pour cela, il peut s'appuyer sur une littérature abondante pour les méthodes d'analyse ainsi que pour le calcul de taille d'échantillon quand $r = 1$ ou $r = m$. L'objectif de notre travail consiste ici à développer une méthodologie permettant le calcul de taille d'échantillon pour les procédures les plus couramment utilisées en recherche clinique, à savoir les procédures *single-step* et *step-wise*, et ce quelle que soit la valeur de r ($1 \leq r \leq m$).

Ce chapitre sera donc l'occasion de présenter dans un premier temps le contexte inhérent à cette problématique, puis de présenter la méthodologie utilisée via un article en cours de soumission.

Contexte

Les différentes agences de sécurité sanitaire (FDA, EMA) se basent sur les recommandations de l'ICH et conseillent de diminuer au maximum le nombre de critères de jugement principaux dans la mise en place d'un essai. Cette recommandation a un double avantage. Le premier consiste à définir le plus précisément possible l'objectif de l'essai, et le second à limiter les problèmes liés à la multiplicité.

Cependant, cette directive n'est pas toujours applicable puisque l'objectif de nombreux essais consiste à étudier un effet multifactoriel du produit. Cela implique donc des problématiques complexes pour lesquels l'objectif principal ne peut être résumé à un seul critère de jugement, même en utilisant un critère composite [Sankoh et al., 2003]. Ces essais sont surtout effectués pour des analyses de tolérance en Phase I, ou en Phases II et III pour étudier l'efficacité du produit.

Dans ce contexte, les scientifiques font face à des critères de jugement multiples, qu'ils considèrent dans la majorité des cas comme ayant la même importance. Ces critères sont alors définis par l'anglicisme ≪co-primary endpoints≫. En fonction de l'objectif global de l'étude, les attentes diffèrent et la définition du critère principal aussi. En effet, les promoteurs concluent au succès de l'essai si au moins r critères de jugement principaux parmis m sont significatifs. Cette définition quant au succès de l'étude reste très générale, et dans la pratique il convient de distinguer trois scénarios :

1. At least one win ($r = 1$) [Dmitrienko et al., 2012, Sankoh et al., 2003, Senn and Bretz, 2007]. Pour illustrer ce cas de figure, nous pouvons prendre l'exemple utilisé dans le chapitre précédent concernant l'effet du produit sur la réponse vaccinale [Boge et al., 2009]. Il s'agissait de tester la réponse vaccinale du produit versus placebo sur trois souches grippales. L'efficacité était conclue si un effet du produit sur au moins l'une d'entre elles était observé ;
2. All win ($r = m$) [Dmitrienko et al., 2012, Sankoh et al., 2003, Senn and Bretz, 2007]. Ce type de critère est défini par l'anglicisme ≪must win≫ par Julious et al. en 2012 [Julious and McIntyre, 2012]. Ils utilisent l'exemple des essais cliniques sur l'arthrose, où pour montrer l'efficacité du traitement il est nécessaire de trouver une différence significative sur l'ensemble des trois critères standards : évaluation globale du patient, Score WOMAC (index de sévérité symptomatique de l'arthrose sur les membres inférieurs) sur la douleur, Score WOMAC sur la

condition physique ;

3. At least r win ($r \in \{1, \ldots, m\}$) [Hung and Wang, 2009, Sankoh et al., 2003]. Afin d'illustrer ce scénario, nous pouvons nous appuyer sur l'article de Pedrono et al. [Pédrono et al., 2009]. Ils proposent dans ce papier une nouvelle définition des critères de jugement dans le cadre d'un essai vaccinal contre le pneumocoque, pour des patients infectés par le VIH (Virus de l'Immunodéficience Humaine). Dans ce contexte, la définition d'un critère composite ne semble pas adaptée. L'une des solutions qui semble ressortir de cette étude serait alors de conclure à l'efficacité du vaccin si une différence significative est observée sur au moins 5 des 7 critères de jugement principaux.

Le scénario envisagé dépend donc directement de l'objectif scientifique de l'étude. Celui-ci va donc avoir un impact sur la méthodologie choisie par le statisticien qui à l'heure actuelle peut s'appuyer sur une littérature abondante pour les méthodes d'analyse et le calcul de taille d'échantillon dans le cadre des scénarios 1 et 2 [Sankoh et al., 2003, Senn and Bretz, 2007, Sozu et al., 2010, Sozu et al., 2011].

Le scénario 1 représente le cas où l'efficacité du produit ne peut être conclue que s'il existe une différence significative entre le produit et son contrôle sur au moins un critère de jugement principal. Dans ce cas, les méthodes courantes d'analyse, telles que Bonferonni, Holm, Hochberg et Hommel [Hommel, 1988, Hochberg, 1988, Holm, 1979] sont les plus couramment utilisées [Bretz et al., 2010]. Bien que conservatrices dans le cadre de critères de jugement principaux corrélés, elles ont le mérite d'être simples d'utilisation et de contrôler le FWER (probabilité de commettre au moins une erreur de Type-I). Dans le cadre de la mise en place de l'essai, il est aussi nécessaire de calculer la taille d'échantillon. Ce calcul se fait en contrôlant la puissance disjonctive [Senn and Bretz, 2007].

A l'inverse, le scénario 2 correspond à l'analyse de critères de jugement principaux $\ll$ must-win $\gg$. Lorsque les statisticiens se retrouvent face à cette problématique, il n'est pas nécessaire de corriger l'erreur de Type-I pour l'analyse [Sankoh et al., 2003, EMA, 2002, Dmitrienko et al., 2012] mais il est cependant important de contrôler la puissance conjonctive dans le calcul de taille d'échantillon [Senn and Bretz, 2007, Sozu et al., 2011].

Bien qu'il existe beaucoup de travaux s'intéressant aux deux premiers scénarios, très peu s'intéressent au troisième qui a pourtant l'avantage d'être plus général et de regrouper les deux premiers.
C'est donc dans une optique de généralisation, que nous avons choisi de nous intéresser au scenario 3. L'efficacité du produit est alors conclue si un effet est démontré sur au moins $\ll r \gg$ critères de jugement parmi $\ll m \gg$.
Notre démarche consiste à généraliser le calcul de taille d'échantillon aux procédures ≪single-step≫ et ≪step wise≫, dont les plus couramment utilisées sont celles de Bonferroni, Holm et Hochberg. L'utilisation de ces méthodes nous permettra dans une majorité des cas un contrôle fort du FWER, ce qui permet de minimiser l'erreur de Type-I. Ce contrôle est exigé par les agences de sécurité sanitaire pour toute étude à but confirmatoire. Ceci s'explique par le fait que dans le cadre de ce contrôle la probabilité de fausses découvertes pour tout type de configuration des hypothèses nulles n'excédera jamais le risque de première espèce marginal α déterminé par le statisticien.
Malgré les avantages évidents du contrôle du FWER en confirmatoire, celui ci possède des limites dans les études à but exploratoire. Plusieurs statisticiens ont pointé le fait que ce contrôle est souvent trop strict quand le nombre d'hypothèses testées est grand. En effet, le contrôle du FWER impliquera une correction stricte du risque de première espèce réajusté, ce qui empêchera de mettre en évidence certaines associations qui pourraient être intéressantes dans un cadre exploratoire.
Lehmann et Romano [Lehmann and Romano, 2005] ont donc proposé le ≪*generalized Family-Wise Error Rate*≫ (gFWER) permettant de pallier cette limite :

$$gFWER = \mathbb{P}(\text{commettre au moins } q \text{ erreurs de Type-I}).$$

Il consiste à généraliser le FWER. Lorsque $q = 1$, la définition du gFWER est la même que celle du FWER, et lorsque $q \geq 1$ cela permet d'être plus permissif.
Dans ce travail, nous nous sommes interessés au calcul de taille d'échantillon pour les procédures développées par Lehmann & Romano [Lehmann and Romano, 2005] ainsi que Romano & Shaikh [Romano and Shaikh, 2006] permettant la généralisation des procédures de Bonferroni, Holm et Hochberg pour un contrôle du gFWER. Le choix d'appliquer notre méthode à ces procédures nous permettra donc d'intégrer parfaitement notre travail au sein des pratiques cliniques actuelles, notamment grâce au développement d'un package R.

Ici pour le calcul de taille d'échantillon nous contrôlons la ≪r-Power≫ développée par Dunnett et Tamhane en 1992 [Dunnett and Tamhane, 1992]. Cette puissance est définie comme la probabilité de rejeter au moins r fausses hypothèse nulles et nous permet de répondre totalement à notre problématique scientifique. Toutefois, en l'état, cette définition ne nous permet pas le calcul de taille d'échantillon. Notre objectif consiste donc à partir de cette définition et à l'appliquer aux différentes procédures définies ci-dessus, jusqu'à obtenir une expression qui dépend de la loi conjointe des statistiques de tests. Chen et al. en 2011 [Chen et al., 2011] ont tenté de répondre à ce problème, non sans mal puisqu'il existe certaines faiblesses dans leur définition de la puissance et donc dans leur calcul de taille d'échantillon.

Cependant, notre travail ne se cantonne pas à une simple généralisation des méthodes existantes, puisqu'il permet aussi de répondre à un réel besoin en recherche clinique. Ceci peut être illustré d'une part par l'exemple cité précédemment mais il sera aussi très utile dans des études de type ≪*proof of concept*≫ qui consistent à vérifier les hypothèses d'efficacité du produit durant la Phase II du développement.

Afin d'illustrer cette problématique, nous nous baserons sur deux études cliniques récentes.

La première application présentée dans ce chapitre porte sur un essai randomisé de Phase II dont l'objectif est de comparer l'immunogénécité de deux stratégies vaccinales pour des patients atteints par le Virus de l'Immunodéficience Humaine (VIH) [Pédrono et al., 2009]. Dans le cadre du vaccin anti-pneumococcique, l'immunogénécité est mesurée sur 7 sérotypes, cependant aucune définition claire de l'immunogénécité n'est donnée dans la littérature. L'objectif de cet exemple sera donc de calculer le nombre de sujets nécessaires pour différentes définitions possibles de l'immunogénécité et différentes procédures de contrôle de la multiplicité. Cette application sera aussi l'occasion de réaliser une étude de sensibilité.

Le second essai est détaillé dans l'article de Davison et al. [Davison et al., 2011]. Il s'agit d'une étude de type ≪*proof of concept*≫, dont l'objectif est de valider l'efficacité du traitement (Relaxin©) contre les insuffisances cardiaques aigües. L'efficacité du traitement est alors conclue si et seulement si un effet significatif du traitement est observé sur la majorité des critères de jugement principaux. Dans ce contexte, il est souhaitable d'utiliser la méthode développée pour le calcul du nombre de sujets nécessaires. Celle-ci permettra d'être en adéquation avec la définition du critère d'intérêt (efficacité

du produit), et donc de ne pas surestimer ou sous-estimer de manière importante le nombre de sujets nécessaires à l'étude. Si l'investigateur juge trop important le nombre de sujets pour une Phase-II, notre méthode lui permettra de connaître la puissance de son analyse pour un nombre de sujets maximum. Une fois cette puissance calculée, l'investigateur aura en sa possession l'ensemble des informations nécessaires lui permettant d'évaluer la viabilité de l'essai.

Ce travail est soumis dans la revue Biometrika, et sera présenté dans la suite du chapitre.

Biometrika (2013), **xx**, x, *pp*. 1–10

Printed in Great Britain

Type-II Generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination

BY PHILIPPE DELORME, PIERRE LAFAYE DE MICHEAUX

Département de mathématiques et de statistique, Université de Montréal, 2920, chemin de la Tour, Montréal, Québec H3T 1J4, Canada

delorme@dms.umontreal.ca lafaye@dms.umontreal.ca

BENOIT LIQUET

The University of Queensland, School of Mathematics and Physics, Brisbane St Lucia, QLD 4072, Australia

b.liquet@uq.edu.au

JÉRÉMIE RIOU

Danone Research, Equipe Biométrie, Avenue de la Vauve, 91700 Palaiseau Cedex, France

jeremie.riou@isped.u-bordeaux2.fr

SUMMARY

Co-primary endpoints are increasingly used in clinical trials. The significance of some of these clinical trials is established if at least r null hypotheses are rejected among m that are simultaneously tested. The usual approach in multiple hypothesis testing is to control the family-wise error rate, which is defined as the probability that at least one type-I error is made. More recently, the q-generalized family-wise error rate has been introduced to control the probability of making at least q false rejections. For procedures controlling this global type-I error rate, we define a type-II r-generalized family-wise error rate which is directly related to the r-power defined as the probability of rejecting at least r false null hypotheses. We obtain very general power formulas which can be used to compute the sample size for single-step and step-wise procedures. These are implemented in our `R` package `Sample` and complexities of the formulas are presented to gain insight into computation time issues. We then compute sample sizes for two clinical trials involving co-primary endpoints; one designed to investigate the effectiveness of a drug against acute heart failure, the other for the immunogenicity of a vaccine against pneumococcus.

Some key words: Clinical research; Co-primary endpoint; Multiple testing; r-power; Sample size determination.

1. INTRODUCTION

In order to capture a multi-factorial effect of some product, it is now increasingly common in clinical research to define multiple co-primary endpoints and then test simultaneously a finite number of null hypotheses. In this situation where many hypotheses are tested and each individual test has a specified individual type-I error probability, the probability that at least some type-I errors (false rejections or false positives) are made increases with the number of hypotheses. Multiple hypothesis testing methods have been proposed for dealing with this problem. The most common ones in clinical trials are the single-step Bonferroni procedures or the stepwise

(Holm, 1979)'s and (Hochberg, 1988)'s procedures. Their popularity could be explained by their ease of use and also for their control of the family-wise error rate, which is defined as the probability of one or more false rejections among the family of the m null hypotheses considered. However, some statisticians have noted the overly stringent control of the family-wise error rate, which tends to wipe out some interesting effects when many tests are performed. To overcome this limit, Benjamini & Hochberg (1995) have proposed the false discovery rate. More recently, Lehmann & Romano (2005) have introduced the q-generalized family-wise error rate, defined as the probability to make at least q false rejections ($q = 1, \ldots, m$), which corresponds to the family-wise error rate when $q = 1$. They have also provided modifications of Bonferroni and Holm's methods to control this new global error rate; see also (Wang & Xu, 2012; Romano & Shaikh, 2006; Sarkar, 2007) for step-up procedures controlling it.

Now, in most clinical trials, statistical significance needs to be demonstrated for all primary endpoints (Offen et al., 2007) or for at least one primary endpoint (Sankoh et al., 2003). A new treatment may also be preferred to a placebo if it shows efficacy on at least r out of m endpoints (Tamhane et al., 1998), where r is any specified integer between 1 and m. In the same spirit as in (Julious & McIntyre, 2012; Dmitrienko et al., 2012), this can be termed an at-least-r must-win problem. For example, in some proof-of-concept studies (Davison et al., 2011), a majority ($r = \lceil m/2 \rceil$) of primary endpoints are required to be significant; see also (Dmitrienko et al., 2010, p. 6). Extensive literature exists on sample size computation when $r = 1$ or $r = m$ (Lafaye de Micheaux et al., 2013; Sozu et al., 2011; Senn & Bretz, 2007). Chen et al. (2011) have proposed an extension to the multiple must-win context, but limited to the $m = 3$ case. To the best of our knowledge, this seems to be the only article dealing with this problem, but some of their formulas are erroneous. We address the sample size computation problem for the general multiple must-win context. To this purpose, we extend the overall type-II family-wise error rate, defined in (Gabbay et al., 2011, p.504) as the probability of at least one acceptance of a false null hypothesis (false negative or individual type-II error), in the same way as Lehmann & Romano (2005) generalize the (type-I) family-wise error rate. More formally, suppose we plan to collect some data from a true model (e.g., given by an entirely specified distribution). Let us suppose a model P to be the true model for which p null hypotheses are false and $m - p$ are true. Then, for some $r \leq p$, our global type-II r-generalized family-wise error rate is defined by:

$$\beta_{r,m}(P) = \mathrm{pr}\left(\text{make at least } p - r + 1 \text{ individual type-II errors among the } p \text{ false hypotheses}\right),$$

which is one minus the r-power of Dunnett & Tamhane (1992), defined by

$$1 - \beta_{r,m}(P) = \mathrm{pr}\left(\text{reject at least } r \text{ of the } p \text{ false null hypotheses}\right)$$

and also called generalized disjunctive power by Dmitrienko & D'Agostino Sr (2013).

Let $\mathcal{I}_p = \{1, \ldots, p\}$ for all $p = 1, \ldots, m$. In what follows, we consider the simultaneous testing of the m null hypotheses $\mathcal{H}_0^k$, $k \in \mathcal{I}_m$, using m possibly dependent test statistics $T_1, \ldots, T_m$, whose order statistics are denoted by $T_{1:m}, \ldots, T_{m:m}$. We suppose that rejection regions are in the form $\{T_k > c_{k,m}\}$, where the critical values $c_{k,m}$ are determined to control either the (generalized) family-wise error rate as in Lehmann & Romano (2005), or the false discovery rate (Benjamini & Hochberg, 1995), or any such method that aims at controlling the global error rate. The aim of this paper is to show how to compute the necessary sample size in order to control weakly or strongly the type-II r-generalized family-wise error rate, for procedures that already control strongly the type-I q-generalized family-wise error rate. We say that weak control at level β (e.g., 20%) of the type-II r-generalized family-wise error rate is reached when $\beta_{r,m}(P) \leq \beta$ for a potential choice P of the true model under which all null hypotheses tested are false (i.e.,

$p = m$). Strong control at level β occurs when $\beta_{r,m}(P) \leq \beta$ for all potential choices P of the true model such that $p \geq r$ null hypotheses are false.

In Section 2, we derive general power formulas for single-step and stepwise procedures, in the non-exchangeable and exchangeable cases, for statistics whose distribution can be of any type. To ease the numerical computations of sample sizes, we express these formulas in terms of probabilities of non-ordered test statistics. In Section 3, we show how to specialize these formulas for the simultaneous testing of means differences of continuous endpoints. We consider various scenarios of covariance structures among the endpoints. Finally, we compute sample sizes for two clinical trials using our `R` package `Sample`, which is available upon request. Proofs are given in a supplementary section.

2. POWER FORMULAS

2·1. *Single-step methods*

For any given model P, we suppose from now on and without loss of generality that the false null hypotheses are associated with the p first (unordered) test statistics $T_1, \ldots, T_p$. The r-power for any single-step procedure is given below.

THEOREM 1. *Let $\cap_\emptyset = \Omega$. We have*

$$1 - \beta^s_{r,m}(P) = 1 + \sum_{k=p-r+1}^{p} (-1)^{k-p+r} \binom{k-1}{p-r} \sum_{J \subset \mathcal{I}_p; |J|=k} \text{pr}\left\{\cap_{j\in J}(T_j \leq c_{j,m})\right\}$$
$$= \sum_{k=r}^{p} (-1)^{k-r} \binom{k-1}{k-r} \sum_{J \subset \mathcal{I}_p; |J|=k} \text{pr}\left\{\cap_{j\in J}(T_j > c_{j,m})\right\}.$$

When $p - r + 1 > r$, obtaining the r-power is faster using the first expression. Otherwise, one should use the second one, involving $C(p, r)\,{}_2F_1(1, r - p; r + 1; -1)$ probability terms, where $C(p, r)$ is another notation for binomial coefficients and ${}_2F_1(a, b; c; x)$ is the Gauss's hypergeometric function. This being said, in the context of sample size computation in clinical trials, it is often assumed that test statistics are exchangeable, namely that their joint distribution is defined up to a permutation of their indices (Olshen, 1974). This permits a substantial speed up in computations.

COROLLARY 1. *For exchangeable statistics and when rejection regions are under the form $\{T_k > c_m\}$ for some common critical value c_m, the r-power is simplified to*

$$1 - \beta^{s,e}_{r,m}(P) = 1 + \sum_{k=p-r+1}^{p} (-1)^{k-p+r} \binom{p}{k} \binom{k-1}{p-r} \text{pr}\left\{\cap_{j=1}^{k}(T_j \leq c_m)\right\}$$
$$= \sum_{k=r}^{p} (-1)^{k-r} \binom{k-1}{k-r} \binom{p}{k} \text{pr}\left\{\cap_{j=1}^{k}(T_j > c_m)\right\}.$$

2·2. *Step-up methods*

Let $\underline{a} = (a_1, \ldots, a_q)^{\mathrm{T}} \in \mathbb{N}^q$ and note $\underline{a}^* = (a_2, \ldots, a_{q+1})^{\mathrm{T}}$ with $a_{q+1} = p$. Let also $a_0 = 0$, $\underline{a}_+ = \sum_{i=1}^{q} a_i$ and $\Delta a_i = a_i - a_{i-1}$ for $i \in \mathcal{I}_{q+1}$. We introduce the set

$$\mathcal{J}(\underline{a}, p) = \left\{\underline{j} \in \mathcal{I}_p^{a_q}; j_r < j_{r+1} \text{ for } r \in \{a_{h-1} + 1, \ldots, a_h - 1\}, h \in \mathcal{I}_q \text{ and } j_r \neq j_s, 1 \leq r < s \leq a_q\right\}.$$

It has a cardinality equal to the multinomial coefficient $C(p, \Delta^*_{\underline{a}})$ where $\Delta^*_{\underline{a}} = (\Delta a_1, \ldots, \Delta a_{q+1})^{\mathrm{T}}$. For any given $\underline{\ell} = (\ell_1, \ldots, \ell_q)^{\mathrm{T}} \in \mathbb{R}^q$, we define the following summation which involves $\nu(\underline{\ell}; p)$ terms:

$$\sum_{\underline{a}=\underline{\ell}}^{\underline{a}^*} f(\underline{a}) = \sum_{a_q=l_q}^{p} \sum_{a_{q-1}=l_{q-1}}^{a_q} \cdots \sum_{a_1=l_1}^{a_2} f(a_1, \ldots, a_q).$$

When $\underline{\ell} = (1, \ldots, q)$, we will denote $\nu(\underline{\ell})$ by $\nu(q; p) = C(p+q, q) - C(p+q, q-1)$, which are called the Catalan's triangle numbers (Glueck et al., 2008).

THEOREM 2. *Given a step-up method, we let $u_1 \leq \cdots \leq u_m$ be the critical values associated with $T_{1:m}, \ldots, T_{m:m}$ and we define $v_i = u_{m-p+i}$, $i = 1, \ldots, p$. The r-power is*

$$1 - \beta^u_{r,m}(P) \geq 1 - (-1)^{(p-r+1)(p-r+2)/2} \sum_{\underline{a}=\underline{w}}^{\underline{a}^*} (-1)^{\underline{a}+} \mathbb{P}_{\underline{a}} \prod_{h=1}^{p-r+1} \binom{(\Delta a_h) - 1}{a_h - h}, \qquad (1)$$

where $\underline{w} = (1, \ldots, p-r+1)$ and $\mathbb{P}_{\underline{a}} = \sum_{\underline{j} \in \mathcal{J}(\underline{a}, p)} \mathrm{pr}\left[\cap_{i=0}^{p-r} \left\{\cap_{k=a_i+1}^{a_{i+1}} (T_{j_k} \leq v_{i+1})\right\}\right]$. When $p = m$, namely for a weak control of the type-II r-generalized family-wise error rate, inequality (1) *becomes an equality.*

The complexity of (1) is $O((p-r+2)^p)$. This could become intractable for $p \geq 10$, but note that standard clinical trials usually involve a small number of co-primary endpoints. For further discussion, see Section 4.

COROLLARY 2. *For exchangeable statistics, we can simplify $\mathbb{P}_{\underline{a}}$ in* (1) *to*

$$\mathbb{P}_{\underline{a}} = \binom{p}{\Delta^*_{\underline{a}}} \mathrm{pr}\left[\cap_{i=0}^{p-r} \left\{\cap_{k=a_i+1}^{a_{i+1}} (T_k \leq v_{i+1})\right\}\right].$$

In this case, complexity reduces to $\nu(p-r+1; p)$. This is tractable if $p \leq 15$ and if $p-r+1$ is not too large.

Remark 1. In an all must-win ($r = m$) context, Eaton & Muirhead (2007) showed that there is no inflation of the type-I error rate. Thus, when one wants to prove an effect of the product on all endpoints, one just needs to show that $\min(T_1, \ldots, T_m) \geq c_{1-\alpha}$, where $c_{1-\alpha}$ is the α-level critical value for an individual test. This decision rule corresponds to Hochberg's correction, which is thus recommended for sample size computations when $r = m$.

2·3. *Step-down methods*

The following results are valid for any step-down method.

THEOREM 3. *Let $h_1 \leq \cdots \leq h_m$ be the critical values associated with $T_{1:m}, \ldots, T_{m:m}$ and define $d_i = h_{m-p+i}$, $i = 1, \ldots, p$. The r-power is*

$$1 - \beta^d_{r,m}(P) \geq (-1)^{r(r+1)/2} \sum_{\underline{a}=\underline{t}}^{\underline{a}^*} (-1)^{\underline{a}+} \tilde{\mathbb{P}}_{\underline{a}} \prod_{h=1}^{r} \binom{(\Delta a_h) - 1}{a_h - h}, \qquad (2)$$

where $\underline{t} = (1, \ldots, r)$ and $\tilde{\mathbb{P}}_{\underline{a}} = \sum_{\underline{j} \in \mathcal{J}(\underline{a}, p)} \mathrm{pr}\left[\cap_{i=0}^{r-1} \left\{\cap_{k=a_i+1}^{a_{i+1}} (T_{j_k} > d_{p-i})\right\}\right]$. When $p = m$, inequality (2) *becomes an equality.*

The complexity of the above formula is $O((r+1)^p)$ while the one below is $\nu(r; p)$.

COROLLARY 3. *For exchangeable statistics, we can simplify* $\tilde{\mathbb{P}}_{\underline{a}}$ *in* (2) *to*

$$\tilde{\mathbb{P}}_{\underline{a}} = \begin{pmatrix} p \\ \Delta^* \underline{a} \end{pmatrix} \text{pr}\left[\cap_{i=0}^{r-1}\left\{\cap_{k=a_i+1}^{a_{i+1}}\left(T_k > d_{p-i}\right)\right\}\right].$$

Remark 2. Based on the complexities given previously, we recommend using a step-down method when $p - r + 1 > r$ and a step-up method otherwise.

3. JOINT DISTRIBUTION OF TEST STATISTICS FOR CONTINUOUS CO-PRIMARY ENDPOINTS

3·1. *Overview and notations*

Let us consider two independent groups, each one denoted by an index $g \in \{E, C\}$, E for experimental and C for control, for which we plan to observe n_g independent and identically distributed continuous random vectors, consisting of m co-primary endpoints, $\boldsymbol{X}_i^g = (X_{i,1}^g, \ldots, X_{i,m}^g)^{\mathrm{T}}$, $i = 1, \ldots, n_g$, with expectation vector $\boldsymbol{\mu}^g = (\mu_1^g, \ldots, \mu_m^g)^{\mathrm{T}}$ and variance-covariance matrix Σ^g whose diagonal elements are $\sigma_k^{2,g}$. We want to test the m following hypotheses :

$$\mathcal{H}_0^k : \mu_k^E - \mu_k^C \leq d_k \text{ versus } \mathcal{H}_1^k : \mu_k^E - \mu_k^C > d_k$$

for some d_k's. This is done using the following test statistics:

$$T_k = \left(\widehat{\text{Var}}\left(\bar{X}_k^E - \bar{X}_k^C - d_k\right)\right)^{-1/2}(\bar{X}_k^E - \bar{X}_k^C - d_k),$$

where $\bar{X}_k^g = n_g^{-1}\sum_{i=1}^{n_g} X_{i,k}^g$. When the joint distribution of the vector of test statistics $\boldsymbol{T} = (T_1, \ldots, T_m)^{\mathrm{T}}$ is known, it becomes possible to compute the required sample sizes n_g for reaching some pre-specified r-power $1 - \beta_{r,m}(P)$, for a given model P. To obtain tractable joint distributions, we investigate the case of a multivariate Gaussian distribution

$$\left(\boldsymbol{X}_1^g, \ldots, \boldsymbol{X}_{n_g}^g\right)^{\mathrm{T}} \sim \mathcal{N}_m^{n_g}\left((\boldsymbol{\mu}^g, \ldots, \boldsymbol{\mu}^g)^{\mathrm{T}}, \mathcal{I}_{n_g} \otimes \Sigma^g\right), \tag{3}$$

which is related to the multivariate Behrens-Fisher problem (Belloni & Didier, 2008), for various classical scenarios on the structure of the covariance matrices Σ_j.

We introduce some notation before presenting the distribution of $\boldsymbol{T}$ for an unstructured covariance matrix and a multiple compound symmetry covariance matrix. Let $\boldsymbol{\delta} = \boldsymbol{\mu}^E - \boldsymbol{\mu}^C - \boldsymbol{d}$ where $\boldsymbol{d} = (d_1, \ldots, d_m)^{\mathrm{T}}$. We will write $\mathrm{Diag}(\boldsymbol{v})$ for the diagonal matrix whose diagonal consists of the elements of the vector $\boldsymbol{v}$, $\mathrm{diag}(M)$ for the vector constituted from the diagonal elements of matrix M. We also let $\mathrm{tr}(\cdot)$ be the trace matrix operator.

3·2. *Unstructured Covariance Matrix*

Let us recall the definition of the multivariate type-II Student distribution (Jensen, 2005).

DEFINITION 1. *If* $\boldsymbol{Z} \sim \mathcal{N}_m(\boldsymbol{\mu}, \Sigma)$ *and if the* $m \times m$ *matrix* $\nu S = \nu(S_{i,j})$ *follows, independently of* $\boldsymbol{Z}$*, a* $W_m(\nu, \Xi)$ *Wishart distribution such that the diagonal elements of* Σ *and* Ξ *are the same, then*

$$\boldsymbol{T} = \left(S_{1,1}^{-1/2}Z_1, \ldots, S_{m,m}^{-1/2}Z_m\right)^{\mathrm{T}} \sim t_m^{II}\left(\boldsymbol{\mu}, \Sigma, \Xi, \nu\right),$$

a multivariate type-II Student distribution.

PROPOSITION 1. *Let* $\Sigma_{EC} = \Sigma_E + \Sigma_C$ *with* $\Sigma_E = n_E^{-1}\Sigma^E$ *and* $\Sigma_C = n_C^{-1}\Sigma^C$*. Let* $V = \mathrm{Diag}(\mathrm{diag}(\Sigma_{EC}))$ *and* $R = V^{-1/2}\Sigma_{EC}V^{-1/2}$*. Under the multivariate Gaussian model* (3)*, we*

obtain a non-asymptotic approximation of the joint distribution of $\boldsymbol{T}$, using the method of moments to find the best matching Wishart distribution for $\widehat{\Sigma}_{EC}$, a consistent estimator of Σ_{EC}:

$$\boldsymbol{T} \overset{approx}{\sim} t_m^{II}\left(V^{-1/2}\boldsymbol{\delta}, R, R, f\right),$$

where

$$f = \frac{\operatorname{tr}\left(\Sigma_{EC}^2\right) + \operatorname{tr}^2\left(\Sigma_{EC}\right)}{(n_E-1)^{-1}\left(\operatorname{tr}\left(\Sigma_E^2\right) + \operatorname{tr}^2\left(\Sigma_E\right)\right) + (n_C-1)^{-1}\left(\operatorname{tr}\left(\Sigma_C^2\right) + \operatorname{tr}^2\left(\Sigma_C\right)\right)}.$$

When $\Sigma^E = \Sigma^C$, we have

$$\boldsymbol{T} \sim t_m^{II}\left(V^{-1/2}\boldsymbol{\delta}, R, R, n_E + n_C - 2\right).$$

To the best of our knowledge, no simple procedure is available to compute accurately probabilities for the t_m^{II} distribution. To overcome this issue, one can rely on an asymptotic distribution for $\boldsymbol{T}$ proposed by Lafaye de Micheaux et al. (2013) when $n_E = n_C = n$:

$$\widehat{R}^{-1/2}\left(\boldsymbol{T} - n^{1/2}\widehat{V}^{-1/2}\boldsymbol{\delta}\right) \xrightarrow{L} \mathcal{N}_m\left(\boldsymbol{0}_m, I_m\right),$$

where $\widehat{V}$ and $\widehat{R}$ are natural plugin estimators of V and R using $\widehat{\Sigma}_{EC}$. When Σ_{EC} is assumed known, the joint distribution of $\boldsymbol{T}$ is a multivariate Gaussian distribution.

Remark 3. Another option is to follow Hasler & Hothorn (2011)'s advice to approximate the t_m^{II} distribution with a Kshirsagar m-variate t-distribution.

3·3. *Multisample Compound Symmetry Covariance Matrix*

The definition of the multivariate type-I Student distribution is given in a 1994 Harvard University PhD thesis by C. Liu.

DEFINITION 2. *Let $\boldsymbol{Z} \sim \mathcal{N}_m(\boldsymbol{\mu}_Z, \Sigma)$ and $\nu\tau \sim \chi^2_\nu$, independent of $\boldsymbol{Z}$. Then*

$$\boldsymbol{T} = \tau^{-1/2}\boldsymbol{Z} + \boldsymbol{\mu} \sim t_m(\boldsymbol{\mu}, \boldsymbol{\mu}_Z, \Sigma, \nu),$$

has a noncentral multivariate t distribution. When $\boldsymbol{\mu} = \boldsymbol{0}$, we obtain the $t_m(\boldsymbol{0}, \boldsymbol{\mu}_Z, \Sigma, \nu)$ distribution, also called the $t_m^K(\boldsymbol{\mu}_Z, \Sigma, \nu)$ Kshirsagar (1961) noncentral m-variate t-distribution.

PROPOSITION 2. *Suppose that the covariance matrix has a multisample compound symmetry structure $\Sigma^g = \sigma^{2,g}K_\rho$ with $K_\rho = (1-\rho)I_m + \rho\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$, where ρ is given, I_m is the identity matrix, and $\mathbf{1}_m$ is the m-length vector of ones. We get*

$$\boldsymbol{T} \overset{approx}{\sim} t_m^K\left(\sigma_{EC}^{-1}\boldsymbol{\delta}, K_\rho, f\right),$$

where $\sigma_{EC}^2 = \sigma_E^2 + \sigma_C^2$ with $\sigma_E^2 = n_E^{-1}\sigma^{2,E}$ and $\sigma_C^2 = n_C^{-1}\sigma^{2,C}$ and where

$$f = m\{(n_E-1)^{-1}\sigma_E^4 + (n_C-1)^{-1}\sigma_C^4\}^{-1}\sigma_{EC}^2$$

is an approximation of the degrees of freedom obtained by applying the Cochran-Satterthwaite's method. For $\Sigma^E = \Sigma^C = \sigma^2 K_\rho$, we have

$$\boldsymbol{T} \sim t_m^K\left((n_E^{-1} + n_C^1)^{-1/2}\sigma^{-1}\boldsymbol{\delta}, K_\rho, m(n_E + n_C - 2)\right).$$

Remark 4. We have supposed that ρ is known because its estimation involves solving a polynomial equation of order 4, which would lead to an untractable distribution for $\boldsymbol{T}$. When $\rho = 0$, multisample compound symmetry is called sphericity (Pillai & Nagarsenker, 1979).

4. APPLICATION: COMPUTING SAMPLE SIZE FOR m CO-PRIMARY ENDPOINTS

4·1. *Considerations on computation time*

Considering a multivariate Gaussian model P, and using the distribution of $\boldsymbol{T}$ under this model, one can compute the r-power for any specified values of n_E and n_C. But first, it is worthwhile to make a few comments about computation time. In standard clinical trials when dealing with one endpoint, one only needs to specify the effect size $\Delta = \delta/\sigma$, which is a measure of the improvement δ of a new treatment E compared to a control group C over the population standard deviation σ of the studied endpoint. When m continuous primary endpoints are considered, one needs to specify $\boldsymbol{\delta}$ and Σ_{EC}, which can be estimated from pilot studies. The test statistics are exchangeable when both the effect sizes and the correlations are the same for all the endpoints. A particular case is to assume a multisample compound symmetry covariance matrix for model P and the same measure of improvement for all endpoints. This enables us to use our simplified r-power formulas which greatly decreases the combinatorial complexity. Moreover, in clinical trials, it is customary to assume an effect for all endpoints, corresponding to a weak control (specified by a unique model P) of the r-power. Given the huge number of models involved in a strong control of r-power, $\sum_{j=r}^{m} C(m,j)$, this presents an added advantage in terms of complexity. Currently, the main bottleneck in terms of computation speed of our `R` implementation resides in the use of the `pmvt()` function from package `mvtnorm` (Genz et al., 2013) to compute probabilities for t_m^K distributions. Thus, any improvement in the computation speed of this function would have a great impact on our own computations. Other avenues of future research are to compute bounds on the r-power (Lee, 1992) or to avoid computing negligible terms in our power formulas, as advocated by M. Schonlau, in Section 6.3.1 of his 1997 University of Waterloo PhD thesis, who underlines that $\mathrm{pr}(A \cap B)$ is always smaller than $\mathrm{pr}(A)$. This is easy to see in Corollary 1. See also (Glueck et al., 2008; Kwong & Chan, 2008) for algorithms to compute the distribution of order statistics for independent but non identically distributed observations.

Remark 5. We have compared the results given by our `R` package `Sample` to those obtained recently by Dmitrienko & D'Agostino Sr (2013) who used a Monte Carlo approach (see supplementary section). Our methodology is around 1,000 times faster in their context.

4·2. *The Pneumovac Trial*

The Pneumovac phase-II randomized trial (Lesprit et al., 2007) has been conducted to compare the immunogenicity of two vaccine strategies against pneumococcal infection among adult patients infected by the Human Immunodeficiency Virus (HIV). This comparison has been realized on log-transformed antibody concentrations for seven serotypes of interest, and t-tests where used. Pédrono et al. (2009) suggest that one vaccine strategy might be considered as superior to the other when at least 3, 5 or 7 serotypes are found significant. We compute sample sizes necessary for a weak control of the r-power for $r = 3, 5, 7$ using estimations of effect sizes and correlations found in that paper. As in Pédrono et al. (2009), we assume a common unstructured covariance matrix for both vaccinal strategies. The multivariate normal distribution(asymptotic case) and the Kshirsagar 7-variates t-distribution (approximation of the t_m^{II} distribution suggested previously) are considered. Our results are presented in Table 1 for single-step (Bonferroni), step-up (Hochberg) and step-down (Holm) methods controlling the family-wise error rate ($q = 1$).

Table 1. *Sample size computation for the Pneumovac trial assuming a common unstructured covariance matrix,* $\beta = 0.2$*,* $\alpha = 0.05$*,* $m = 7$ *and* $n = n_E = n_C$*.*

	Normal			Kshirsagar		
	$r = 3$	$r = 5$	$r = 7$	$r = 3$	$r = 5$	$r = 7$
Bonferroni	22	51	*	22	51	*
Holm	21	42	*	21	42	*
Hochberg	21	41	116	21	41	116

(*) values not computed (see Remark 1).

As expected, sample size is lower for stepwise procedures, greater for increasing values of r, with the difference between both types of procedures increasing with r. Results are similar for the Normal and Kshirsagar cases, validating the approximation chosen.

To facilitate a decision concerning the final choice of sample size, a sensibility analysis has been conducted (see Figure 1) for the Kshirsagar case with a common value of correlation ρ, standard deviation σ and difference of means δ, for varying values of each one of these parameters, the two others being fixed to their observed mean obtained from the pilot study. We observe a slight impact of ρ and, as expected, a decrease of the sample size with the effect size.

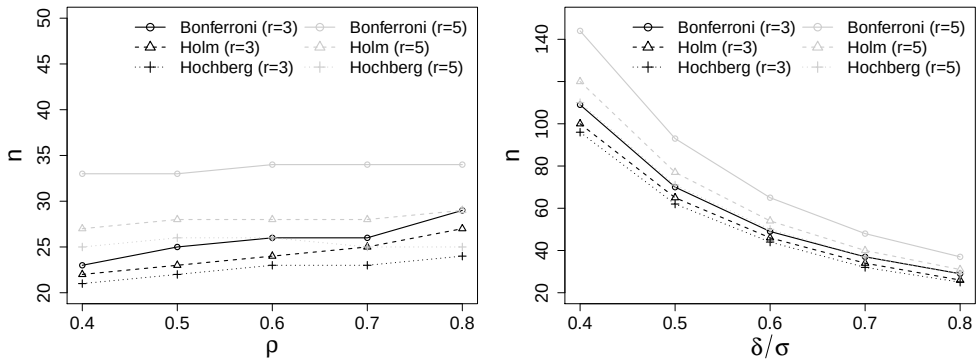


Fig. 1. Sensibility analysis for sample size computation: $\beta = 0.2, \alpha = 0.05, m = 7, n = n_E = n_C$. Left: Effect of ρ ($\sigma = 0.59, \delta = 0.5$) Right: Effect of δ/σ ($\rho = 0.6$).

4·3. *The Pre-RELAX-AHF Trial*

The Pre-RELAX-AHF proof of study trial (Teerlink et al., 2009) is a multicentre, randomized, placebo-controlled, parallel-group, dose-finding Phase-IIb study, where three doses of a new treatment (Relaxin) were tested on nine biological parameters to help people suffering from acute heart failure. A current acceptance criterion for proof of concept studies is to show an effect on the majority of co-primary endpoints (Davison et al., 2011). These authors showed an efficacy on 6 out of 9 endpoints for the 30 $\mu g/kg/day$ dose of Relaxin treatment versus placebo. For confirmatory purpose, we compute the required sample size for a weak control of the $r = 6$-power, assuming a common compound symmetric covariance matrix as suggested by Genz &

Bretz (2009, Section 3.2). This gives rise to a Kshirsagar type-I 9-variates t-distribution, whose values are provided by Davison et al. (2011). The results for various correlations ρ between the endpoints, and three multiple testing procedures are presented in Table 2.

Table 2. *Sample size computation in the Pre-RELAX-AHF trial: multisample compound symmetry covariance matrix,* $\beta = 0.2$, $\alpha = 0.1$, $r = 6$, $m = 9$, $n = n_E = n_C$.

Type-I Kshirsagar distribution										
ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Bonferroni	367	373	379	383	387	389	390	389	385	379
Holm	286	293	299	305	308	311	314	314	313	307
Hochberg	285	292	298	304	308	311	313	313	311	305

ACKNOWLEDGEMENT

The second author was supported by a grant from the Natural Science and Engineering Research Council of Canada.

Supplementary material available at *Biometrika* online includes proofs of all results and a comparison with Monte-Carlo strategy.

REFERENCES

BELLONI, A. & DIDIER, G. (2008). On the behrens-fisher problem: A globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. *Ann. Stat* **36**, 2377–2408.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B. Met* **57**, 289–300.

CHEN, J., LUO, J., LIU, K. & MEHROTRA, D. (2011). On power and sample size computation for multiple testing procedures. *Comput. Stat. Data. An* **55**, 110–122.

DAVISON, B. A., COTTER, G., SUN, H., CHEN, L., TEERLINK, J. R., METRA, M., FELKER, G. M., VOORS, A. A., PONIKOWSKI, P., FILIPPATOS, G. et al. (2011). Permutation criteria to evaluate multiple clinical endpoints in a proof-of-concept study: lessons from Pre-RELAX-AHF. *Clin. Res. Cardiol* **100**, 745–753.

DMITRIENKO, A., D'AGOSTINO, R. B. & HUQUE, M. F. (2012). Key multiplicity issues in clinical drug development. *Stat. Med* **32**.

DMITRIENKO, A. & D'AGOSTINO SR, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Stat. Med* , n/a–n/a.

DMITRIENKO, A., TAMHANE, A. C. & BRETZ, F. (2010). *Multiple testing problems in pharmaceutical statistics*. Chapman & Hall/ CRC Biostatistics Series. Taylor & Francis.

DUNNETT, C. & TAMHANE, A. (1992). A step-up multiple test procedure. *J. Am. Stat. Assoc* **87**, 162–170.

EATON, M. L. & MUIRHEAD, R. J. (2007). On a multiple endpoints testing problem. *J. Stat. Plan. Infer* **137**, 3416–3429.

GABBAY, D., THAGARD, P., WOODS, J., BANDYOPADHYAY, P. & FORSTER, M. (2011). *Philosophy of statistics*, vol. 7. North Holland.

GENZ, A. & BRETZ, F. (2009). *Computation of multivariate normal and t probabilities*, vol. 195 of *Lecture Notes in Statistics*. Dordrecht: Springer.

GENZ, A., BRETZ, F., MIWA, T., MI, X., LEISCH, F., SCHEIPL, F. & HOTHORN, T. (2013). *mvtnorm: Multivariate Normal and t Distributions*. R package version 0.9-9995.

GLUECK, D. H., KARIMPOUR-FARD, A., MANDEL, J., HUNTER, L. & MULLER, K. E. (2008). Fast computation by block permanents of cumulative distribution functions of order statistics from several populations. *Comm. Statist. Theory Methods* **37**, 2815–2824.

HASLER, M. & HOTHORN, L. A. (2011). A dunnett-type procedure for multiple endpoints. *Int. J. Biostat* **7**, 1–15.

HOCHBERG, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat* **6**, 65–70.

JENSEN, D. (2005). Multivariate t distribution. In *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds., vol. 2. J. Wiley, pp. 3870–3872.

JULIOUS, S. A. & MCINTYRE, N. E. (2012). Sample sizes for trials involving multiple correlated must-win comparisons. *Pharm. Stat.* **11**, 177–185.

KSHIRSAGAR, A. M. (1961). Some extensions of the multivariate generalization t-distribution and the multivariate generalization of the distribution of the regression coefficient. *Math. Proc. Cambridge Philos. Soc.* **57**, 80–85.

KWONG, K.-S. & CHAN, Y.-M. (2008). On the evaluation of the joint distribution of order statistics. *Comput. Statist. Data Anal.* **52**, 5091–5099.

LAFAYE DE MICHEAUX, P., LIQUET, B., MARQUE, S. & RIOU, J. (2013). Power and sample size determination in clinical trials whith multiple primary correlated endpoints. *J. Biopharm. Stat* , IN PRESS.

LEE, M.-Y. (1992). Bivariate Bonferroni inequalities. *Aequationes Math.* **44**, 220–225.

LEHMANN, E. & ROMANO, J. (2005). Generalizations of the familywise error rate. *Ann. Stat* **33**, 1138–1154.

LESPRIT, P., PÉDRONO, G., MOLINA, J.-M., GOUJARD, C., GIRARD, P.-M., SARRAZIN, N., KATLAMA, C., YÉNI, P., MORINEAU, P., DELFRAISSY, J.-F. et al. (2007). Immunological efficacy of a prime-boost pneumococcal vaccination in HIV-infected adults. *Aids* **21**, 2425–2434.

OFFEN, W., CHUANG-STEIN, C., DMITRIENKO, A., LITTMAN, G., MACA, J., MEYERSON, L., MUIRHEAD, R., STRYSZAK, P., BADDY, A., CHEN, K. et al. (2007). Multiple co-primary endpoints: Medical and statistical solutions: A report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of america. *Drug. Inf. J* **41**, 31–46.

OLSHEN, R. (1974). A note on exchangeable sequences. *Probab. Theory. Rel* **28**, 317–321.

PÉDRONO, G., THIÉBAUT, R., ALIOUM, A., LESPRIT, P., FRITZELL, B., LÉVY, Y. & CHÊNE, G. (2009). A new endpoint definition improved clinical relevance and statistical power in a vaccine trial. *J. Clin. Epidemiol* **62**, 1054–1061.

PILLAI, A. K. C. S. & NAGARSENKER, B. N. (1979). Distribution of the Likelihood Ratio Statistic for Testing Sphericity Structure for a Normal Covariance Matrix and Its Percentage Points. *Sankhyā. Ser. B* **41**, 154–169.

ROMANO, J. & SHAIKH, A. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Stat* **34**, 1850–1873.

SANKOH, A. J., D'AGOSTINO, R. B. & HUQUE, M. F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Stat. Med* **22**, 3133–3150.

SARKAR, S. K. (2007). Step-up procedures controlling generalized FWER and generalized FDR. *Ann. Stat* **35**, 2405–2420.

SENN, S. & BRETZ, F. (2007). Power and sample size when multiple endpoints are considered. *Pharm. Stat.* **6**, 161–170.

SOZU, T., SUGIMOTO, T. & HAMASAKI, T. (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *J. Biopharm. Stat* **21**, 650–68.

TAMHANE, A., LIU, W. & DUNNETT, C. (1998). A generalized step-up-down multiple test procedure. *Can. J. Stat* **26**, 353–363.

TEERLINK, J. R., METRA, M., FELKER, G. M., PONIKOWSKI, P., VOORS, A. A., WEATHERLEY, B. D., MARMOR, A., KATZ, A., GRZYBOWSKI, J., UNEMORI, E. et al. (2009). Relaxin for the treatment of patients with acute heart failure (pre-relax-ahf): a multicentre, randomised, placebo-controlled, parallel-group, dose-finding phase iib study. *Lancet* **373**, 1429–1439.

WANG, L. & XU, X. (2012). Step-up procedure controlling generalized family-wise error rate. *Stat. Probabil. Lett* **82**, 775–782.

[*Received XXX XXXX. Revised XXX XXXX*]

Supplementary material for Type-II Generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination

Proof of Theorem 1 and its corollary (Single step)

An application of Waring's theorem (Macdonald, 2006) gives the probability that at least q events among p occur.

THEOREM 1. *Let $A_1, \ldots, A_p$ be some events and note $S_k = \sum_{J\subset\{1,\ldots,p\};|J|=k} P(\cap_{j\in J} A_j)$ with $\cap_{j\in\emptyset} A_j = \Omega$. The probability that at least q events occur is*

$$Q_{r:p} = \sum_{k=q}^{p} (-1)^{k-q} \binom{k-1}{k-q} S_k.$$

In the exchangeable case, we have

$$S_k = \binom{p}{k} P(\cap_{j=1}^{k} A_j).$$

Now, we have m null hypoteses and p among them are false (let us say the first ones). The first equation of Theorem 1 is obtained using Theorem 1 on the events $\{T_k \leq c_{k,m}\}$ in order to compute $\beta_{r,m}(P)$. The second equation is a direct application of Theorem 1 on the events $\{T_k > c_{k,m}\}$.

Proof of Theorem 2 (Step-up)

We focus on the p false null hypotheses $H_0^1, \ldots, H_0^p$ associated to the test statistics $T_1, \ldots, T_p$ and define the p corresponding ordered test statistics $T_{1:p}, \ldots, T_{p:p}$. Since in step-up procedures we start by considering $T_{1:p}$, we have

$$\begin{aligned} 1 - \beta_{r,m}^u(P) &= 1 - \mathrm{pr}(\text{Accept at least } p-(r-1) \text{ false null hypotheses}) \\ &\geq 1 - \mathrm{pr}\left\{ \bigcap_{k=1}^{p-r+1} (T_{k:p} \leq v_k) \right\}. \end{aligned}$$

To conclude this proof, we replace the ordered statistics with unordered ones by using the following result, which has been adapted to suit our notation.

THEOREM 2. *(Maurer & Margolin, 1976, Theorem 3.1) Let $T_1, \ldots, T_m$ be some statistics. Let $\underline{l} = (l_1, \ldots, l_q)^{\mathrm{T}}$ be a vector of integers such that $1 \leq l_1 \leq \cdots \leq l_q \leq m$ (and similarly for $\underline{t} = (t_1, \ldots, t_q)^{\mathrm{T}}$) and $c_1 \leq \cdots \leq c_q$ be real numbers. Then*

$$\mathrm{pr}\left\{ \bigcap_{h=1}^{q} (T_{l_h:m} > c_h) \right\} = (-1)^{q(m+1)-\underline{l}_+} \sum_{\underline{a}=\underline{t}}^{\underline{a}^*} (-1)^{\underline{a}_+} \tilde{\mathbb{P}}_{\underline{a}} \prod_{i=1}^{q} \binom{(\Delta a_i) - 1}{a_i - t_i}$$

where $t_j = m + 1 - l_{q+1-j}$, $j = 1, \ldots, q$ and

$$\mathrm{pr}\left\{ \bigcap_{h=1}^{q} (T_{l_h:m} \leq c_h) \right\} = (-1)^{\underline{l}_+} \sum_{\underline{a}=\underline{l}}^{\underline{a}^*} (-1)^{\underline{a}_+} \mathbb{P}_{\underline{a}} \prod_{i=1}^{q} \binom{(\Delta a_i) - 1}{a_i - l_i},$$

where $\underline{l}_+ = \sum_{i=1}^{q} l_i$, $\sum_{\underline{a}=\underline{l}}^{\underline{a}^*} f(\underline{a}) = \sum_{a_a=l_q}^{m} \sum_{a_{q-1}=l_{q-1}}^{a_q} \cdots \sum_{a_1=l_1}^{a_2} f(a_1, \ldots, a_q)$, $\Delta a_i = a_i - a_{i-1}$, $a_0 = 0$, $a_{q+1} = m$ *and where*

$$\tilde{\mathbb{P}}_{\underline{a}} = \sum_{\underline{j} \in \mathcal{J}(\underline{a},m)} \mathrm{pr}\left[\bigcap_{i=0}^{q-1}\left\{\bigcap_{k=a_i+1}^{a_{i+1}} (T_{j_k} > c_{q-i})\right\}\right], \qquad \mathbb{P}_{\underline{a}} = \sum_{\underline{j} \in \mathcal{J}(\underline{a},m)} \mathrm{pr}\left[\bigcap_{i=0}^{q-1}\left\{\bigcap_{k=a_i+1}^{a_{i+1}} (T_{j_k} \leq c_{i+1})\right\}\right],$$

with

$$\mathcal{J}(\underline{a}, m) \equiv \left\{(j_1, \ldots, j_{a_q}) \in \{1, \ldots, m\}^{a_q}; j_r < j_{r+1} \textit{ for } r \in \mathcal{R}(\underline{a}, h) \equiv \{a_h + 1, \ldots, a_{h+1} - 1\}, \right.$$
$$\left. h = 0, \ldots, q-1, \textit{ and } j_r \neq j_s \textit{ when } 1 \leq r < s \leq a_q\right\}.$$

Remark 1. Note that another equivalent formula involving events with superiority signs can be obtained by noting that

$$1 - \mathrm{pr}\left\{\bigcap_{k=1}^{p-r+1} (T_{k:p} \leq v_k)\right\} = \mathrm{pr}\left\{\bigcup_{k=1}^{p-r+1} (T_{k:p} > v_k)\right\},$$

followed by an application of Theorem 1 (with $q = 1$) and Theorem 2. Computation time being greater for this formula, it has not been included in the paper.

PROOF OF THEOREM 3 (STEP-DOWN)

Proof is similar as the one above, noting that for step-down procedures we start with $T_{p:p}$. We thus have

$$1 - \beta^d_{r,m}(P) = \mathrm{pr}(\text{Reject at least } r \text{ false null hypotheses})$$
$$\geq \mathrm{pr}\left\{\bigcap_{k=p-r+1}^{p} (T_{k:p} > d_k)\right\} = \mathrm{pr}\left\{\bigcap_{k=1}^{r} (T_{(p-r+k):p} > d_{p-r+k})\right\}.$$

An application of Theorem 2 concludes the proof. It is straightforward to obtain another equivalent formula involving inferiority signs as explained in Remark 1.

PROOF OF PROPOSITION 1 (UNSTRUCTURED COVARIANCE MATRIX)

Let $\bar{\boldsymbol{X}}^g = n_g^{-1} \sum_{i=1}^{n_g} \boldsymbol{X}_i^g$. We have

$$\boldsymbol{T} = \frac{\bar{\boldsymbol{X}}^E - \bar{\boldsymbol{X}}^C - \boldsymbol{d}}{\sqrt{\mathrm{diag}(\hat{\Sigma}_{EC})}}$$

where $\frac{\boldsymbol{u}}{\boldsymbol{v}}$ denotes the vector constituted by term-by-term division of the two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and $\sqrt{\boldsymbol{u}}$ denotes the vector consisting of square roots of the elements of $\boldsymbol{u}$. We have

$$\bar{\boldsymbol{X}}^E - \bar{\boldsymbol{X}}^C - \boldsymbol{d} \sim \mathcal{N}_m\left(\boldsymbol{\delta}, \Sigma_{EC}\right).$$

Let us now study the distribution of $\hat{\Sigma}_{EC}$. An unbiased estimator of Σ^g based on a modification of its maximum likelihood estimator (Bilodeau & Brenner (1999, p.81); Lütkepohl (2005, p. 660-661)) is

$$\widehat{\Sigma}^g = (n_g - 1)^{-1}(Y^g - \widehat{M}^g)^{\mathrm{T}}(Y^g - \widehat{M}^g),$$

and we have:

$$(n_g - 1)\widehat{\Sigma}^g \sim W_m(n_g - 1, \Sigma^g).$$

Using results from Nel & Van Der Merwe (1986) we get

$$\widehat{\Sigma}_{EC} = n_E^{-1}\widehat{\Sigma}^E + n_C^{-1}\widehat{\Sigma}^C \sim SoW_m\left(n_E - 1, n_C - 1, (n_E(n_E-1))^{-1}\Sigma^E, (n_C(n_C-1))^{-1}\Sigma^C\right),$$

a sum of Wishart distributions which they approximate by a $W_m(f, \Xi)$ where $\Xi = f^{-1}\Sigma_{EC}$ and f is given in the Theorem, so that $f\widehat{\Sigma}_{EC} \overset{approx}{\sim} W_m(f, \Sigma_{EC})$; see also (Krishnamoorthy & Yu, 2012, Lemma A.1) and (Zhang & Xu, 2009, Equation (3.18) p.1294) for other approximations.

The proof is concluded by noting that $t_m^{II}(\boldsymbol{\mu}, \Sigma, \Xi, \nu) = t_m^{II}(A\boldsymbol{\mu}, A\Sigma A_{\mathrm{T}}, A\Xi A_{\mathrm{T}}, \nu)$ for any $m \times m$ diagonal matrix A.

When $\Sigma^E = \Sigma^C \equiv \Sigma$, we have $\Sigma_{EC} = (n_E n_C)^{-1}(n_E + n_C)\Sigma$. We take the following unbiased estimator derived from a simple modification of the maximum likelihood estimator of Σ:

$$\widehat{\Sigma} = (n_E + n_C - 2)^{-1} \sum_{g=E,C} \sum_{i=1}^{n_g} (\boldsymbol{X}_i^g - \bar{\boldsymbol{X}}^g)(\boldsymbol{X}_i^g - \bar{\boldsymbol{X}}^g)^{\mathrm{T}}$$

and we have $(n_E + n_C - 2)\widehat{\Sigma}_{EC} \sim W_m(n_E + n_C - 2, \Sigma_{EC})$.

Proof of Proposition 2 (multisample compound symmetry)

We have

$$\boldsymbol{T} = (\hat{\sigma}_{EC}/\sigma_{EC})^{-1}(\sigma_{EC}^{-1}(\bar{\boldsymbol{X}}^E - \bar{\boldsymbol{X}}^C - \boldsymbol{d}),$$

where $\sigma_{EC}^2 = n_E^{-1}\sigma^{2,E} + n_c^{-1}\sigma^{2,C}$ with $\hat{\sigma}_{EC}^2$ an unbiased estimator of σ_{EC}^2 obtained by pluging in the following simple modification of the maximum likelihood estimator of $\sigma^{2,g}$:

$$\hat{\sigma}^{2,g} = (m(n_g - 1))^{-1}\mathrm{tr}\left\{\sum_{i=1}^{n_g}(\boldsymbol{X}_i^g - \bar{\boldsymbol{X}}^g)^{\mathrm{T}}K_\rho^{-1}(\boldsymbol{X}_i^g - \bar{\boldsymbol{X}}^g)\right\}$$

if $\sigma^{2,E} \neq \sigma^{2,C}$ and

$$\hat{\sigma}_{EC}^2 = (n_E + n_C - 2)^{-1}(n_E^{-1} + n_C^{-1})\left\{(n_E - 1)\hat{\sigma}^{2,E} + (n_C - 1)\hat{\sigma}^{2,C}\right\}$$

otherwise. Now

$$\sigma_{EC}^{-1}(\bar{\boldsymbol{X}}^E - \bar{\boldsymbol{X}}^C - \boldsymbol{d}) \sim \mathcal{N}_m(\sigma_{EC}\boldsymbol{\delta}, K_\rho)$$

and

$$f\sigma_{EC}^{-1}\hat{\sigma}_{EC} \sim \chi_f^2,$$

where f is approximated by a Cochran-Satterthwaite method and given in the Theorem when $\sigma^{2,E} \neq \sigma^{2,C}$, and where $f = m(n_E + n_C - 2)$ otherwise. The independence between the numerator and the denominator is proved using Theorem 3.1.2 in (Muirhead, 2008, p.80).

Comparison with Monte-Carlo strategy

Recently, Dmitrienko & D'Agostino Sr (2013) have used a Monte-Carlo simulation in order to compute the generalized disjunctive power of a procedure in a clinical trial. The aim of this

section is to compare it with our approach, both in terms of power and computation time. Thus, we applied the two strategies on the example developed by Dmitrienko & D'Agostino Sr (2013) in their section 10.3. In this example, the endpoints are continuous, and the true mean changes are expected to be given by vector $\boldsymbol{\delta} = (5, 5, 3.5)^{\mathrm{T}}$. We considered $\alpha = 0.025$, $n = 260$, the same standard deviation for each endpoint ($\sigma_k = 18$) and each group, and the same correlation between all tests ($\rho = 0.5$) for each group. The results are presented in Figure : power (top), computation times for the Monte Carlo strategy (bottom left) and using our `Sample` package (bottom right).

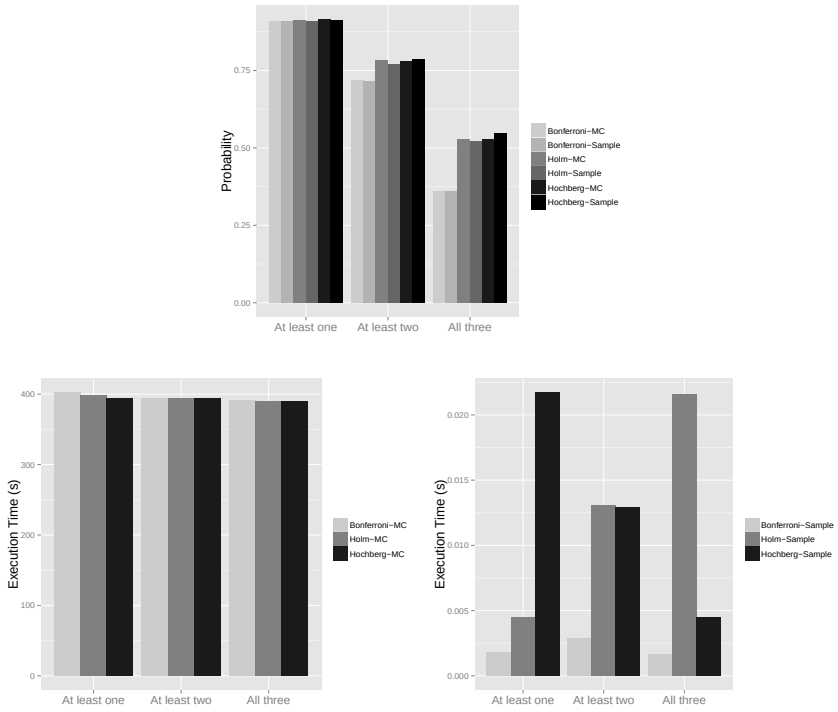


Fig. 1. Comparison of Monte Carlo and `Sample` Package. Top: Generalized Disjunctive Power Comparison. Bottom left: computation time for the Monte Carlo strategy. Bottom right: computation time for package `Sample`.

As expected, results are similar in terms of power for both approaches. However, our strategy is much faster ($< 0.05s$) than the Monte Carlo one ($\approx 400s$).

REFERENCES

BILODEAU, M. & BRENNER, D. (1999). *Theory of multivariate statistics*. Springer.

DMITRIENKO, A. & D'AGOSTINO SR, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Stat. Med* , n/a–n/a.

KRISHNAMOORTHY, K. & YU, J. (2012). Multivariate Behrens-Fisher problem with missing data. *J. Multivariate Anal.* **105**, 141–150.

LÜTKEPOHL, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer-Verlag.

MACDONALD, A. S. (2006). Waring's theorem. In *Encyclopedia of Actuarial Science*, J. L. Teugels & B. Sundt, eds., vol. 3. John Wiley & Sons, Ltd, pp. 1749–1750.

MAURER, W. & MARGOLIN, B. (1976). The multivariate inclusion-exclusion formula and order statistics from dependent variates. *Ann. Stat* **4**, 1190–1199.

MUIRHEAD, R. J. (2008). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc.

NEL, D. & VAN DER MERWE, C. (1986). A solution to the multivariate Behrens-Fisher problem. *Comm. Statist. Theory Methods* **15**, 3719–3735.

ZHANG, J. & XU, J. (2009). On the k-sample Behrens-Fisher problem for high-dimensional data. *Sci. China Ser. A* **52**, 1285–1304.

Package Sample

Nous avons implémenté une fonction sous le logiciel «open-source» de statistiques R, permettant l'accessibilité de ce travail à l'ensemble de la communauté scientifique. La fonction $\mathrm{indiv.rm.ssc()}$ fait partie intégrante du package Sample présenté dans le chapitre précédent. Nous allons voir dans cette section les différents arguments qui la composent ainsi qu'un exemple d'application.
Afin d'utiliser cette fonction, l'utilisateur doit rentrer la procédure de gestion de tests multiples utilisée à travers l'argument method. Il a le choix entre les procédures de Bonferroni (Bonf), Holm (Holm) et Hochberg (Hoch). Il est nécessaire de spécifier le contrôle du gFWER utilisé en précisant la valeur de g :

$$gFWER = \mathbb{P}(\text{commettre au moins } g \text{ erreurs de Type-I}).$$

L'utilisateur doit préciser la puissance (power), le risque de première espèce attendu (alpha), ainsi que la valeur de r et de m. Le calcul de taille d'échantillon nécessite, bien évidemment, les vecteurs des moyennes attendues ainsi que les matrices de variance covariance pour chaque groupe à travers les arguments muC, muT, sigmaC, sigmaT. La définition des hypothèses statistiques fait intervenir le vecteur $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^T$ définit par l'argument delta, où :

$$\mathcal{H}_0^k : \ \mu_k^T - \mu_k^C \leq \delta_k \text{ versus } \mathcal{H}_1^k : \ \mu_k^T - \mu_k^C > \delta_k$$

L'utilisateur doit préciser si il souhaite utiliser une estimation asymptotique de la distribution conjointe (multivariée normale) à travers l'argument $\mathrm{asympt{=}TRUE}$, si au contraire il souhaite utiliser une distribution de student multivariée de type Kshirsagar pour la distribution conjointe ($\mathrm{asympt{=}FALSE}$). Cet argument vaut par défault FALSE.
Enfin, bien qu'il existe une valeur par défault, l'utilisateur peut spécifier l'intervalle de valeurs attendu pour la taille d'échantillon par groupe à travers l'argument $\mathrm{interval}$. Au final, l'utilisateur obtient en sortie la valeur du nombre de sujets calculée par groupe.

Afin d'illustrer cette fonction, reprenons l'exemple de l'essai clinique PNEUMOVAC. Comme nous l'avons spécifié précédemment, l'étude préliminaire de Pedrono suppose la même matrice de variance covariance entre les groupes (sigmaC=sigmaT=cov), et ne nous donne que la différence d'immunogénécité entre les deux stratégies vaccinales. Nous avons donc rentré pour le vecteur muT la

différence d'immunogénicité entre les groupes, puis un vecteur nul pour l'argument muC. Dans cet exemple, nous utilisons la procédure de Bonferroni avec laquelle nous souhaitons contrôler le FWER $(g = 1)$. Les scientifiques souhaitent obtenir un effet de la nouvelle stratégie vaccinale sur au moins 3 sérotypes parmis les 7 sérotypes testés. Nous fixons donc $r = 3$ et $m = 7$. Enfin, nous souhaitons contrôler une puissance à 0.8 et un FWER à 0.05, pour des tests de supériorité où delta est un vecteur nul. La fonction $\mathrm{Indiv.rm.ssc}()$ est donc utilisée comme suit :

```
######################### Chargement du package ###################################

library(Sample)

############### Définition des paramètres estimés ############################

# différences attendues entre les deux strategies vaccinales
muT <- c(0.55,0.34,0.38,0.20,0.70,0.38,0.86)

# variances attendues
var <- c(0.352^2,0.622^2,0.543^2,0.608^2,0.628^2,0.553^2,0.807^2)

# matrice de variance covariance attendue
cov <- matrix(1,ncol=7,nrow=7)
cov[1,2:7] <- cov[2:7,1] <- c(0.134,0.137,0.075,0.140,0.128,0.161)
cov[2,3:7] <- cov[3:7,2] <- c(0.287,0.185,0.316,0.295,0.396)
cov[3,4:7] <- cov[4:7,3] <- c(0.199,0.274,0.237,0.342)
cov[4,5:7] <- cov[5:7,4] <- c(0.192,0.156,0.238)
cov[5,6:7] <- cov[6:7,5] <- c(0.264,0.397)
cov[6,7] <- cov[7,6] <- c(0.335)

diag(cov) <- var
```

```
################ Appel de la fonction ##############################

n3ss <- indiv.rm.ssc(method="Bonf", asympt=FALSE, r=3,m=7,muT=muT,muC=rep(0,7),
sigmaC=cov,sigmaT=cov,delta=rep(0,7),power=0.8,alpha=0.05,interval=c(10,2000),g=1)

################ Sortie de la fonction ################

Sample size: 20
```

5

Correction du degré de signification engendré par la recherche du codage «optimal» d'une variable explicative continue dans un modèle linéaire généralisé.

Résumé

L'objectif de ce chapitre consiste à présenter le travail réalisé dans le cadre de la recherche d'un codage optimal dans un Modèle Linéaire Généralisé (GLM). Cette pratique, couramment utilisée en statistique appliquée, n'est pas recommandée, notamment à cause de problèmes liés à la multiplicité. L'objectif de cette démarche est donc corrective et consiste à corriger la multiplicité induite. Liquet et Commenges ont commencé à travailler sur cette problématique dans le cadre d'une transformation binaire ou de Box-Cox pour une régression logistique en 2001 [Liquet and Commenges, 2001], puis l'ont généralisé au modèle GLM en 2005 [Liquet and Commenges, 2005].

Notre idée à été ici de généraliser leur travail à un codage en m classes. Ce chapitre s'orientera tout d'abord vers la présentation du contexte, puis nous nous intéresserons à la méthodologie utilisée.

Contexte

En statistique appliquée, une pratique courante consiste à mesurer l'association existant entre une variable explicative d'intérêt et une variable à expliquer. Il est par exemple courant en épidémiologie de focaliser son analyse sur un facteur de risque particulier. Le problème scientifique sous-jacent consistera alors à connaître l'impact de ce facteur de risque sur une maladie, un traitement, ou n'importe quel autre type d'≪outcome≫.

Dans l'optique de répondre à cette problématique, les statisticiens utilisent généralement un modèle de régression où le facteur de risque est ici représenté par une variable continue X ajustée sur $p-1$ facteurs de risque déjà connus. Toutefois, quand la variable explicative d'intérêt est quantitative, comme ici, le choix scientifique se porte souvent sur une transformation de cette variable [Bennette and Vickers, 2012]. Le codage d'une variable continue en variable catégorielle est très utilisé en épidémiologie afin de rendre l'interprétation plus facile, ou parce que les scientifiques suspectent l'existence d'un effet seuil de la variable continue. D'autres transformations peuvent aussi être utilisées, telles que les transformations continues de type Box-Cox.

Néanmoins, quand aucun ≪Gold Standard≫ n'existe pour le codage de la variable, le choix du codage est souvent subjectif. Du fait de cette subjectivité, la pratique la plus couramment utilisée consiste à tester plus d'un ensemble de points de coupure dans le but de trouver celui qui paraît le plus adapté à l'analyse effectuée. Pour analyser les données, le statisticien va donc vérifier quelle est la transformation ≪optimale≫ dans ce contexte. Pour cela, il va réaliser un test pour lequel l'hypothèse nulle $(\mathcal{H}_0)$ est définie par $\ll\beta_{\mathbf{k}} = \mathbf{0}\gg$ (où $\beta_{\mathbf{k}}$ représente le vecteur des coefficients associés à la transformation k de notre variable continue X). Au final, le codage retenu sera celui pour lequel la p_{valeur} est la plus faible. Toutefois, cette démarche peut entraîner certaines limites [Royston et al., 2006, Bennette and Vickers, 2012], notamment en termes de multiplicité. De manière générale, les chercheurs omettent de corriger leur p_{valeur}, ce qui aura pour conséquence d'augmenter l'erreur de Type-I et de surestimer l'effet recherché. Il est donc nécessaire de corriger ce problème de multiplicité en ajustant la p_{valeur}.

Pour cela plusieurs méthodes existent, mais les méthodes classiquement utilisées [Simes, 1986, Šidák, 1967, Holm, 1979, Hommel, 1988, Hochberg, 1988] ne prennent pas compte de la corrélation existante entre les tests, et sont donc, dans notre contexte, conservatrices. Ceci aura pour conséquence

de diminuer la puissance des tests.

C'est dans le but de pallier ce problème que Liquet et Commenges [Liquet and Commenges, 2001, Liquet and Commenges, 2005] ont proposé une correction exacte de la multiplicité dans le cadre de codages binaires ou de transformations de type Box-Cox pour des modèles linéaires généralisés. Hashemi et Commenges [Hashemi and Commenges, 2002] ont ensuite repris ce travail, qu'ils ont généralisé aux modèles à risques proportionnels.

Cependant, le codage binaire n'est pas le seul codage à être utilisé. Notre objectif a donc été de développer une méthode alternative peu conservatrice permettant un codage en plus de deux classes. Des transformations en plus de deux classes auront l'avantage d'être plus flexibles quant à la distribution de l'effet qui reste inconnue. Au niveau statistique, il n'existe pas dans ce contexte une forme analytique de la distribution conjointe des statistiques réalisées, ce qui nous empêche d'obtenir une correction exacte des $p_{valeurs}$. C'est pour cette raison que nous avons choisi d'orienter notre travail vers des méthodes basées sur le rééchantillonage. Ces méthodes, développées par Westfall [Westfall and Young, 1993], permettent d'estimer la distribution de référence par rééchantillonage et donc de prendre en compte la corrélation existant entre les statistiques calculées. Ces méthodes ne sont donc pas conservatrices.

Pour finir, dans l'optique de rendre ces méthodes accessibles au plus grand nombre, nous avons choisi de réaliser le Package CPMCGLM disponible sous le logiciel R.

Nous allons vous présenter ce travail qui a fait l'objet d'une publication dans ≪*BMC Medical Research Methodology*≫.

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models

Benoit Liquet[1,2,3] and Jérémie Riou[1,2,4*]

Abstract

Background: In statistical modeling, finding the most favorable coding for an exploratory quantitative variable involves many tests. This process involves multiple testing problems and requires the correction of the significance level.

Methods: For each coding, a test on the nullity of the coefficient associated with the new coded variable is computed. The selected coding corresponds to that associated with the largest statistical test (or equivalently the smallest p_{value}). In the context of the Generalized Linear Model, Liquet and Commenges (Stat Probability Lett,71:33–38,2005) proposed an asymptotic correction of the significance level. This procedure, based on the score test, has been developed for dichotomous and Box-Cox transformations. In this paper, we suggest the use of resampling methods to estimate the significance level for categorical transformations with more than two levels and, by definition those that involve more than one parameter in the model. The categorical transformation is a more flexible way to explore the unknown shape of the effect between an explanatory and a dependent variable.

Results: The simulations we ran in this study showed good performances of the proposed methods. These methods were illustrated using the data from a study of the relationship between cholesterol and dementia.

Conclusion: The algorithms were implemented using R, and the associated CPMCGLM R package is available on the CRAN.

Keywords: Bonferroni procedure, Generalized linear model, Multiple coding, Parametric bootstrap, Permutation, p_{value}, Resampling procedure

Background

In applied studies, the relationship between an explanatory and a dependent variable is routinely measured using a statistical model. For instance, in epidemiology it is quite common that a study focuses on one particular risk factor. The scientific problem is to analyze whether this risk factor has an influence on the risk of occurrence of a disease, a biological trait , or another outcome. To answer to this question, a regression model is often used in which the risk factor will be represented by a continuous X, allowing adjustment on $p-1$ known risk factors of the studied trait. However, the form of the effect (or the dose-effect relationship) is not known in advance, and as such, the continuous variable X is often transformed, typically into categorical variables, by grouping values into two or more categories. An example of this is seen in an *The American Journal of Epidemiology* (October 2009, volume 170, number 8), where four of six papers with continuous exposure used categorization, and only two kept the variable as continuous [1].

*Correspondence: jeremie.riou@isped.u-bordeaux2.fr
[1] University Bordeaux, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France
[2] INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France
Full list of author information is available at the end of the article

Binary coding is often used in epidemiology, either to make interpretation easier, or because a threshold effect is suspected. In a regression model with multiple explanatory variables, the interpretation of the regression coefficient for a binary variable may be easier to understand than a change in one unit of the continuous variable. Dichotomous transformations of a variable X are defined as:

$$X(k) = \begin{cases} 1 & \text{if } X \geq c_k \\ 0 & \text{if } X < c_k \end{cases}$$

Other transformations are also used, in particular Box-Cox transformations which have been defined as:

$$X(k) = \begin{cases} \lambda_k^{-1}(X^{\lambda_k} - 1) & \text{if } \lambda_k > 0 \\ \log X & \text{if } \lambda_k = 0, \end{cases}$$

but the choice of the transformation is often subjective. The arbitrariness of the choice of cutpoints may lead to the idea of trying more than one set of values. Hence to analyze data, the statistician may have to use several transformations, and for each the statistician applies a test for "$\beta = 0$" (where β is the coefficient representing the effect of the risk factor of interest). The most favorable transformation is then chosen. The cutpoint giving the minimum p_{value} is often termed "optimal" [2,3]. When testing several codings of a variable, there is a problem with the multiplicity of tests performed, leading to an incorrect p_{value} and possible overestimation of effects [4]. Generally, researchers fail to consider this problem and do not correct the significance level in relation to the number of tests performed [3], which can lead to an increase in the Type-I error [5]. The p_{value} should thus be corrected to take into account the multiplicity of tests.

In many cases, it is now widely recognized that categorization of a continuous variable could introduce major problems to an analysis and interpretation of the associated model [1,3]. It is important to note that the aim of this paper is not to defend this practice, but to improve a practice commonly used by epidemiologists in terms of multiple testing. Furthermore, despite known loss of power following dichotomization in the univariate case, Westfall [6] shown that dichotomizing continuous data can greatly improve the power when multiple comparisons are performed.

Many methods of correction exist, the most simple and well known being the Bonferroni rule. Several authors have improved this method to make it more powerful, however most do not take into account the correlation between the tests [7-11]. If the tests are independent, or moderately dependent, then they provide an upper bound which may be satisfactory. Efron [12] proposed a correction that account for the correlation between two consecutive tests if there is a natural order between the tests, with high correlation between adjacent tests. Liquet and Commenges [13,14] and Hashemi and Commenges [15] proposed a more exact correction, accounting for the whole correlation matrix, for score tests obtained in logistic regression, generalized linear model and proportional hazards models.

Here, we proprose extending these studies to a categorical transformation (with $m > 2$ categories) of the continuous variable by involving more than one parameter in the model; $m-1$ dummy variables are introduced in the model. The categorical transformation is a more flexible way to explore the unknown shape of the effect. In this context, we propose a method and an R program based on resampling approaches to determine the significance level for a series of several transformations (including dichotomous, Box-Cox and categorical transformations) of an explanatory variable in a Generalized Linear Model. The problem of correcting the estimation of the effect will not be examined here.

First, we revisit the example proposed by Liquet and Commenges [14] on the relationship between cholesterol and dementia [16] to provide a framework for our discussion. In section 'Methods: Statistical context', we present the statistical contexts relating to multiple testing; the model, the maximum test and the minimum p_{value} procedure and finally the score tests are exposed. Section 'Methods: Significance level correction' presents the different methods of correction of the Type-I error. A simulation study for the different strategies of coding, and application of the model to the initial example are presented in the section 'Results'. Concluding remarks are given in the two last sections.

Example: revisiting the PAQUID cohort example

We revisited the example presented in the article of Liquet and Commenges [13] for a coding of a binary variable in a logistic regression. This example is based on the work of Bonarek et al. [16], who studied the relationship between serum cholesterol levels and dementia. The data came from a nested case-control study of 334 elderly French subjects aged 73 and over who participated in the PAQUID cohort (37 subjects with dementia and 297 controls). The variables age, sex, level of education and wine consumption were considered as adjustment variables. The analysis focused on the influence of HDL-cholesterol(high-density lipoprotein) on the risk of dementia. Bonarek et al. [16] first considered HDL-cholesterol as a continuous variable; then, to ease clinical interpretation, they chose to transform the HDL-cholesterol into a categorical variable with four classes. Finally, as there was no significant difference between the first three quartiles, HDL-cholesterol was split into two categories with a cutpoint at the last quartile. The best p_{value}, 0.007, was obtained in the latter analysis

and was selected for interpretation. However, this p_{value} did not take into account the numerous transformations performed to determine the best representation of the variable of interest. Legitimate questions arising from this include the following: What is the real association between dementia and HDL-cholesterol, with a correction of the Type-I error? Is it really significant? Liquet and Commenges [14] proposed correcting the p_{value} associated with multiple transformation including dichotomous and Box-Cox transformation, however, their method cannot be used with categorical transformation.

Methods

Statistical context

Model

Let us consider a Generalized Linear Model with p explanatory variables [17], where Y_i ($1 \leq i \leq n$) are independently distributed with probability density function in the exponential family defined as follows:

$$f_{Y_i}(y_i, \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}; \quad (1)$$

with $\mathbb{E}[\,Y_i] = \mu_i = b'(\theta_i)$,$\mathbb{V}ar[\,Y_i] = b''(\theta_i)a(\phi)$ and where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known and differentiable functions. $b(\cdot)$ is three times differentiable, and its first derivative $b'(\cdot)$ is invertible. Parameters (θ_i, ϕ) belong to $\Omega \subset \mathbb{R}^2$, where θ_i is the canonical parameter and ϕ is the dispersion parameter.

In this context, we wished to test the association between the outcome Y_i and explanatory variable of interest X_i, adjusted on a vector of explanatory variables $\mathbf{Z_i}$. The form of the effect of X_i is unknown, so we may consider K transformations of this variable $\mathbf{X_i(k)} = g_k(X_i)$ with $k = 1, \ldots, K$.

For example, if we transform the continuous variable in m_k classes, $m_k - 1$ dummy variables are defined from the function $g_k(\cdot)$: $\mathbf{X_i(k)} = g_k(X_i) = (X_i^1(k), \ldots, X_i^{m_k-1}(k))$. Different numbers of level m_k of the categorical transformation are possible.

The model for a transformation k can then be obtained by modeling the canonical parameter θ_i as:

$$\theta_i(k) = \boldsymbol{\gamma}\mathbf{Z_i} + \boldsymbol{\beta_k}\mathbf{X_i(k)}, \;\; i = 1, \ldots, n;$$

where $\mathbf{Z_i} = (1, Z_i^1, \ldots, Z_i^{p-1})$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{p-1})^T$ is a $p-1$ vector of coefficients, and $\boldsymbol{\beta_k}$ is the $m_k - 1$ vector of coefficients associated with a categorical transformation k of the variable X_i. For dichotomous or Box-Cox transformations $\boldsymbol{\beta_k}$ reduce to a scalar ($\boldsymbol{\beta_k} \in \mathbb{R}$).

The hypothesis of the test for the transformation k is defined as follows:

$$\mathcal{H}_0(k) : \boldsymbol{\beta_k} = \mathbf{0}_{m_k-1} \;\text{ versus }\; \mathcal{H}_1(k) : \boldsymbol{\beta_k} \neq \mathbf{0}_{m_k-1},$$

where $\mathbf{0}_{m_k-1}$ is a null vector of dimension $m_k - 1$. Under the null hypothesis $\mathcal{H}_0(k)$ we have $\theta_i(k) = \boldsymbol{\gamma}\mathbf{Z_i}$, which do not depend on k. Thus all the null hypotheses are the same, and denote it by $\mathcal{H}_0$.

Maximum test and minimum P-value procedures

For each coding, k, of the variable X_i, a test statistic T_k is performed on the nullity of the vector $\boldsymbol{\beta_k}$. We then have a vector of test statistics $\mathbf{T} = (T_1, \ldots, T_K)$ for the same null hypothesis (no effect of the risk factor of interest). In the context of dichotomous and Box-Cox transformations, each test statisitic, T_k, has asymptotically, a standard normal distribution. Thus rejecting the null hypothesis if one of the absolute values of the test T_k is larger than a critical value c_α, is equivalent to rejecting the null hypothesis if $T_{max} > c_\alpha$ where $T_{max} = max(|T_1|, \ldots, |T_K|)$. To cope with the multiplicity problem, Liquet and Commenges [13,14] proposed that the probablity of Type-I error for the statistic T_{max} under the null hypothesis be computed as:

$$p_{value} = P(T_{max} \geq t_{max}) = 1 - P(|T_1| < t_{max}, \ldots, |T_{max}| < t_{max},) \quad (2)$$

where t_{max} is the realization of T_{max}.

An equivalent approach is to use a procedure based on the individual p_{value} of each test T_k noted $P_k = P(|T_k| > |t_k|)$ (where t_k is the realization of T_k). The minimum of the K realized p_{value} corresponds to the test k which obtains the highest realization (in absolute values; $k/\ t_{max} = |t_k|$). Then, we have:

$$p_{value} = P(P_{min} \leq p_{min}) \quad (3)$$

where $P_{min} = min(P_1, \ldots, P_K)$ and p_{min} is the realization of P_{min}. The interest of using a procedure based on the p_{value} is the possibility of combining statistical tests which do not follow the same distribution. In the current context, we will combine dichotomous, Box-Cox and categorical transformations with more than two levels.

Score test

We briefly present the score test used for all of the K transformations where the same null hypothesis is tested (*i.e.* $\mathcal{H}_0$: "$\boldsymbol{\beta_k} = \mathbf{0}_{m_k-1}$" given by $\theta_i(k) = \boldsymbol{\gamma}\mathbf{Z_i}$ (with different alternatives)). We present the main results obtained by Liquet and Commenges [14] for the Generalized Linear Model in the context of dichotomous and Box-Cox transformations, and then consider the score test for categorical transformations.

Dichotomous and Box-cox Transformations In the context of dichotomous and Box-Cox transformations, the score test used for testing the effect of the transformed

variable ($\boldsymbol{\beta_k} = 0$ with $\boldsymbol{\beta_k} \in \mathbb{R}$) follows asymptotically a standard normal distribution:

$$T_k = \frac{X(k)^T \hat{R}}{\sqrt{X(k)^T (I - H) V X(k)}}$$

where $\hat{R}$ is the vector of residuals $\hat{R}_i = Y_i - \hat{\mu}_i$ computed under the null hypothesis, V is a diagonal matrix such that $v_{ii} = \hat{Var}(Y_i)$, $H = VZ(Z^T VZ)Z^T$, and Z the $n \times p$ matrix with rows $\mathbf{Z_i}$, $i = 1, \ldots, n$.

The correlation between the different tests has been defined by Liquet and Commenges [14]. Asymptotically, the joint distribution of $T_1, \ldots, T_K$ is a multivariate normal distribution with zero mean and a certain covariance matrix. Thus Liquet and Commenges [14] propose that the p_{value} (associated with the test T_{max}) defined in (2) using numerical integration [18] be calculated. They called their method the "exact method".

Categorical transformations In the context of a categorical transformation in m_k classes, the score test testing $\mathcal{H}_0 :$ "$\boldsymbol{\beta_k} = \mathbf{0}_{m_k - 1}$" (with $\boldsymbol{\beta_k} \in \mathbb{R}^{m_k - 1}$) follows asymptotically a χ^2 distribution with $m_k - 1$ degrees of freedom and is defined as:

$$T_k = U_k^T I_k^{-1} U_k;$$

where U_k and I_k are respectively the score function and the Fisher information matrix under the null hypothesis [19]. To compute the p_{value} defined in (3), it is necessary to know the joint distribution of $\mathbf{T} = (T_1, \ldots, T_K)$. Some studies have defined the distribution of the multivariate χ^2 [20,21]. However, even though the correlation between the different tests could be easily estimated, it has not been possible, as far as we know, to obtain the joint distribution of $\mathbf{T} = (T_1, \ldots, T_K)$. To overcome this problem, we propose approximating the p_{value} (defined in (3) by the minimum p_{value} procedure) using a resampling method (defined in the next section) which also accounts for the correlation between the test statistics.

Significance level correction

Bonferroni method

One of the most common corrections in multiple testing is the Bonferroni method. It has been described by several authors in various applications [7,11,22]. It allows an upper bound of the significance level of the minimum p_{value} procedure to be computed as:

$$p_{value} = P(P_{min} \leq p_{min}) \leq K \times p_{min}$$

where K is the number of tests. This method is very simple and does not require any assumption about the correlation between the different tests. It can therefore be applied directly to the different possible codings of an explanatory variable. However, this only provides an upper bound of the p_{value}, which may be very conservative if the correlation between tests are high and the number of transformation are large.

Resampling based methods

We propose the use of resampling based methods [23,24] with the aim of building a reference distribution for the test statistics. These procedures have the advantage of taking into account the dependence of the test statistics for evaluating the correct significance level of the minimum p_{value} procedure (or the maximum test procedure). The principle of resampling procedures is to define new samples from the probability measure defined under $\mathcal{H}_0 :$ "$\boldsymbol{\beta_k} = \mathbf{0}_{m_k - 1}$".

Permutation test procedure Permutation methods can be used to construct tests which control the Type-I error rate [25]. In our context, the algorithm of the permutation procedure is defined as follows:

1. Apply the minimum p_{value} procedure to the original data for the K transformations considered. We note p_{min} the realization of the minimum of the p_{value};
2. As under $\mathcal{H}_0$, the X_i variable has no effect on the response variable, a new dataset is generated by permuting the X_i variable in the initial dataset;
3. Generate B new datasets $s_b^*, b = 1, \ldots, B$ by repeating B times the step 2;
4. For each new dataset, apply the minimum p_{value} procedure for the transfomation under consideration. We note p_{min}^{*b} the smallest p_{value} for each new dataset.
5. The p_{value} defined in (3) is then approximated by:

$$\widehat{p_{value}} = \frac{1}{B} \sum_{b=1}^{B} I_{\left\{p_{min}^{*b} < p_{min}\right\}},$$

where $I_{\{\cdot\}}$ is an indicator function.

However, it is important to note that exchangeability need to be satisfied [25-30]. This condition is much more restrictive than it appears at first sight. In fact, Commenges [29] and Commenges and Liquet [25] showed that the permutation test approach for the score test is robust if the model has only one intercept under the null hypothesis, or if $\mathbf{X_i}$ are independent of Z_i for all i in the context of a linear model and the proportional hazards model. This issue applies in our context. Thus we investigated, the robustness of the permutation method when the exchangeability assumptions is violated.

Parametric bootstrap procedure In 2000, Good [31] explained: "Permutations test hypotheses concerning distributions; bootstraps test hypotheses concerning parameters. As a result, the bootstrap implies less stringent assumptions". Therefore, an alternative way may be to use

resampling method based on bootstrap [32], which give us an asymptotic reference distribution. This procedure could be defined by the following algorithm:

1. Apply the minimum p_{value} procedure to the original data for the K transformations being considered. We note p_{min} the realization of the minimum of the p_{value};
2. Fit the model under the null hypothesis, using the observed data, and obtain $\hat{\gamma}$, the maximum likelihood estimate (MLE) of γ;
3. Generate a new outcome Y_i^* for each subject from the probability measure defined under $\mathcal{H}_0$. For example, for a logistic model (where $a(\phi) = 1$, $b(\theta_i) = log(1 + e^{\theta_i})$, and $\mu_i = \mathbb{E}(Y_i) = e^{\theta_i}/(1 + e^{\theta_i})$), we generate Y_i^* according to:

$$P(Y_i^* = 1|\mathbf{Z_i}) = \frac{e^{\hat{\gamma}\mathbf{Z_i}}}{1 + e^{\hat{\gamma}\mathbf{Z_i}}}.$$

   Repeat this for all the subjects to obtain a sample noted $s^* = \{Y_i^*, \mathbf{Z_i}, X_i\}$
4. Generate B new datasets $s_b^*, b = 1, \ldots, B$ by repeating B times the step 3;
5. Apply for each new dataset, the minimum p_{value} procedure for the transformation considered. We note p_{min}^{*b} the smallest p_{value} for each new dataset.
6. Then, the p_{value} defined in (3) is then approximated by:

$$\widehat{p_{value}} = \frac{1}{B}\sum_{b=1}^{B} I_{\left\{p_{min}^{*b} < p_{min}\right\}}.$$

Results

Simulation study

The aim of this simulation study was to assess the performance of the two resampling methods to correct the significance level. Three different scenarios of transformations were investigated: dichotomous transformations, categorical transformations with three classes, and categorical transformations with different numbers of classes. To shorten the simulation study section we have not presented the results for the Box-Cox transformations. For each simulation case, the control of the Type-I error and the power of the developed methods were evaluated. For all simulations, the data come from a logistic model (where $a(\phi) = 1$, $b(\theta_i) = log(1 + e^{\theta_i})$, and $\mu_i = \mathbb{E}(Y_i) = e^{\theta_i}/(1 + e^{\theta_i})$) consisting of two explanatory variables: Z, an adjustment variable, and X, the variable of interest. We considered the following models:

$$Logit(P(Y_i = 1|Z_i, X_i(k))) = \theta_i(k) = \gamma_0 + \gamma Z_i + \boldsymbol{\beta}\mathbf{X}_i(k); \quad (4)$$

where Z_i and X_i are independent and were generated according to a standard normal distribution and the vector $\mathbf{X}_i(k)$ was a transformation of a continuous variable X_i. The sample size was set to be 100. We used 1000 replications for each simulation and 1000 samples for the resampling methods.

Dichotomous transformations

We only considered dichotomous transformations to explore a shape effect of the variable of interest. To obtain the best transformation, several cutpoints c_k may be tested. When epidemiological references are not available, a strategy based on the quantile of the continuous variable is most commonly applied. In this simulation we used the median for one dichotomous transformation. For two dichotomous transformations we used the first tercile as the first cutpoint, and the second tercile as the second cutpoint, and so on. This strategy is summarized in Table 1.

Firstly, we investigated the Type-I error rate. For a replication, the rejection criterion of the null hypothesis ($\boldsymbol{\beta_k} = 0$) was a p_{value} less than 0.05. Thus, for a simulation of 1 000 replications, the empirical Type-I error rate was the proportion of tests where the p_{value} was less than 0.05. Figure 1(a) shows the evolution of the Type-I error rate for dichotomous transformations. The *naive* method, without correction of the multiple testing, increases the Type-I error rate with the number of codings tried. For ten codings this error rate reached 0.27. The error rate calculated by the Bonferroni method decreased with the number of cutpoints. This correction was therefore too conservative whereas the exact method and resampling methods gave a Type-I error rate close to the nominal 0.05 value.

When information on the shape of the effect of the explanatory variable was unknown we investigated the power of the methods applied above. We studied the power for *a threshold effect model* with a cutpoint value at the first tercile. Figure 1(b) gives the power as a function of the number of cutpoints tried. The power of the exact and resampling methods are quite similar to one another, and higher than the Bonferroni method. The difference between these methods and Bonferroni method

Table 1 Strategy for dichotomous transformations: values of the cutpoints c_k according to the number of transformations (q_α represents the quantile of order α)

Number of transformations	c_1	c_2	c_3	...	c_9
1	$q_{1/2}$				
2	$q_{1/3}$	$q_{2/3}$			
3	$q_{1/4}$	$q_{2/4}$	$q_{3/4}$		
⋮	⋮	⋮	⋮	⋮	⋮
9	$q_{1/10}$	$q_{2/10}$	$q_{2/10}$	...	$q_{9/10}$

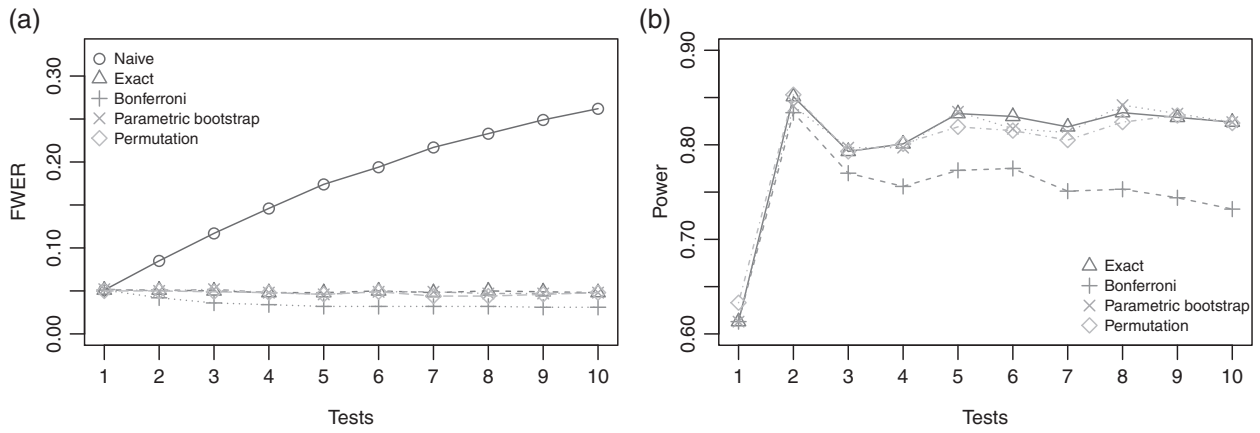


Figure 1 (a) Type-I error rate for various numbers of cutpoints tried according to different methods; (b) power for *a threshold effect model* at the first tercile. ($\gamma_0 = -2.5, \gamma = 1, \beta = 2$).

increases with the number of cutpoints. We also observed that the power was highest at two cutpoints (two transformations). This result, was in fact, expected since we used the first and second terciles respectively as cutpoints for each dichotomous transformation. Power increased again when trying five and eight codings due to the fact that one of these codings corresponded to the first tercile. To conclude, the simulation study with dichotomous transformations showed that the resampling methods provide similar results for the Type-I error rate control and the power as those seen with the exact method.

Categorical transformations with same number of classes

We considered here only categorical transformations with three classes. In this situation, the choices of the two cutpoints (noted c_k^1 and c_k^2) defining the categorical variables into three classes are also subjective. For this simulation study, our strategy was to attempt to find the most favorable transformation into three classes. This consisted of using the tercile of the variable for one transformation with two cutpoints ($c_1^1 = q_{1/3}$ and $c_1^2 = q_{2/3}$); for two transformations we add to the previous choice a transformation with the first quartile and the third quartile for the two cutpoints ($c_2^1 = q_{1/4}$ and $c_2^2 = q_{3/4}$). The global strategy until we obtain 10 transformations in three classes is presented in Table 2.

We investigated the Type-I error rate. Figure 2(a) shows the evolution of the Type-I error rate for categorical transformations in three classes. The results are similar to those we observed for dichotomous transformations. The Bonferroni correction was still too conservative, while resampling methods gave a Type-I error rate close to the nominal 0.05 value.

Next we considered the power of the different methods when the simulated model was specified with a categorical transformation of the continuous variable in three classes defined by cutpoints at the first and third quartile. The two resampling methods gave similar results with a higher power than the Bonferroni method (see Figure 2(b)). The power was highest for two transformations. This result was also expected because, with the strategy presented in Table 2, the transformation into three classes with cutpoints at the first and third quartile is used.

Various categorical transformations

In this last simulation, we presented a more realistic situation where different kinds of transformations were used to investigate the effect of the variable of interest. We proposed trying different categorical transformations and varying the number of classes. The most natural method is to use a dichotomous transformation at the median for one transformation. For two transformations, we added the previous coding and a categorical transformation in

Table 2 Strategy for the categorical transformations in three classes: values of the cutpoints (c_k^1 and c_k^2) for all transformations

Number of transformations	c_k^1	c_k^2
1	$q_{1/3}$	$q_{2/3}$
2	$q_{1/4}$	$q_{3/4}$
3	$q_{1/4}$	$q_{1/2}$
4	$q_{1/2}$	$q_{3/4}$
5	$q_{2/5}$	$q_{4/5}$
6	$q_{1/5}$	$q_{3/5}$
7	$q_{3/5}$	$q_{4/5}$
8	$q_{1/5}$	$q_{2/5}$
9	$q_{1/5}$	$q_{4/5}$
10	$q_{2/5}$	$q_{3/5}$

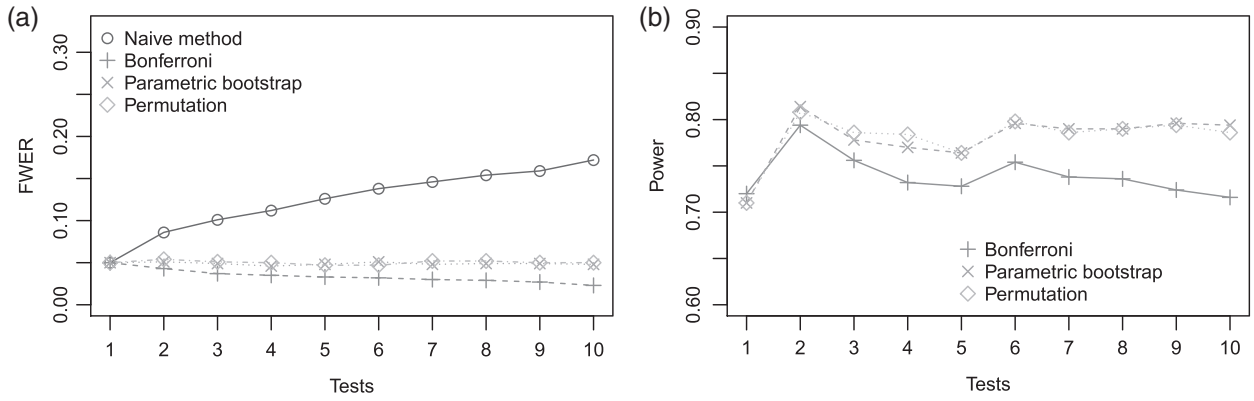


Figure 2 (a) Type-I error rate for various numbers of categorical in three classes according to different methods, (b) and power for an effect of a categorical transformation in three classes defined by cutpoints at the first and third quartile. ($\gamma_0 = -1.25, \gamma_1 = 1, \beta^T = (2, 1.8)$).

three classes based on the tercile. For three transformations, we added the two previous codings and a categorical transformation in four classes based on the quartile, and so on. The strategy proposed in this simulation is presented in Table 3.

The results for the Type-I error rate were similar to the previous simulation case (not shown here). We then studied the power of the different methods when the simulated model is specified with a categorical transformation of the continuous variable in five classes defined by cutpoints at the quintile. We can see in Figure 3, that, in this situation, the parametric bootstrap method seems slightly more powerful than the permutation method. The resampling methods were also more powerful than the Bonferroni method. Finally, as expected, we can see that the power was highest for four transformations, where one of the transformations used corresponded to a categorical transformation with quintiles as cutpoints.

Robustness of resampling methods

We investigated the robustness of the resampling methods when the exchangeability assumption is violated. The data came from the model defined in (4) with two dependent variables $X_i(k)$ and Z_i. The dependency between $X_i(k)$ and Z_i (formalized by the correlation ratio(η^2)) was specified by the following model:

$$Z_i = \beta^* \mathbf{X}_i(k) + \epsilon_i; \tag{5}$$

where $X_i(k)$ is the binary coding of the X_i variable with a cutpoint at the median. The coefficient β^* was computed according to η^2 and the variance of $X_i(k)$ variable.
We tested three different binary codings with cutpoints at the first, the second and the third tercile. The strategy is used for various values of the correlation ratio (η^2) from 0 to 0.6.

The robustness of the permutation method when the exchangeability assumption is violated was evaluated with respect to the results of the exact method. For different correlation ratios (η^2) we evaluated the control of the Type-I error, the power, the Mean Square Error (MSE) of the estimated p_{value} (p_{value} from the exact method was used as a reference), and the rate of good decision (same decision as for the exact method). These results are presented in Figure 4 and show the good behavior of the permutation method since the Type-I error is controlled at the level 0.05, the power is the same for all the methods, the rate of good decision is always greater than 0.97, and the MSE is very low. Moreover, the distributions of the estimated p_{value} are quite similar for different methods (not shown).

Table 3 Strategy for different categorical transformations: values of the cutpoints for all transformations

Number of transformations	c_k^1	c_k^2	...	c_k^9	c_k^{10}
1	$q_{1/2}$				
2	$q_{1/3}$	$q_{2/3}$			
⋮	⋮	⋮	⋮	⋮	⋮
9	$q_{1/10}$	$q_{2/10}$	...	$q_{9/10}$	
10	$q_{1/11}$	$q_{2/11}$	...	$q_{9/11}$	$q_{10/11}$

Example: revisiting the PAQUID cohort example

In order to find the real association between the two variables of interest in the example described at the end of Background section, we applied our newly developed approach which combined different kinds of transformations. Liquet and Commenges [14] have proposed seven dichotomous and five Box-Cox transformations. However, their method did not allow for categorical transformations. We proposed to add, to the seven dichotomous and

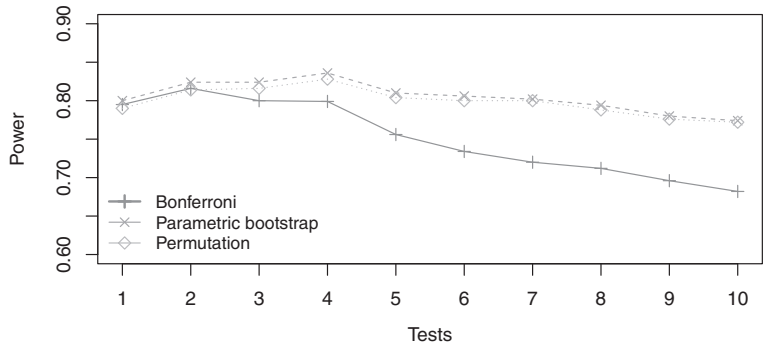


Figure 3 Power for an effect of a categorical transformation in five classes defined by cutpoints at the quintile. ($\gamma_0 = -2.5, \gamma_1 = 1, \beta^T = (-1.3, -0.8, 1.4, 1.7)$).

five Box-Cox transformations for this application, four codings in three classes and four codings in four classes. The best transformation appeared to be the dichotomous transformation of HDL-cholesterol with a cutpoint at the third quartile, as already found by Bonarek et al. [16]. The Bonferroni correction gave a p_{value} equal to 0.140, thus not significant for an α level at 0.05. The p_{value}, which is given by both resampling based methods is 0.038. To conclude, it is important to chose a powerful method of correction, because in this context the p_{value} with no correction given by Bonarek et al. [16] was very optimistic (0.007), and the Bonferroni correction was very conservative, yielding an

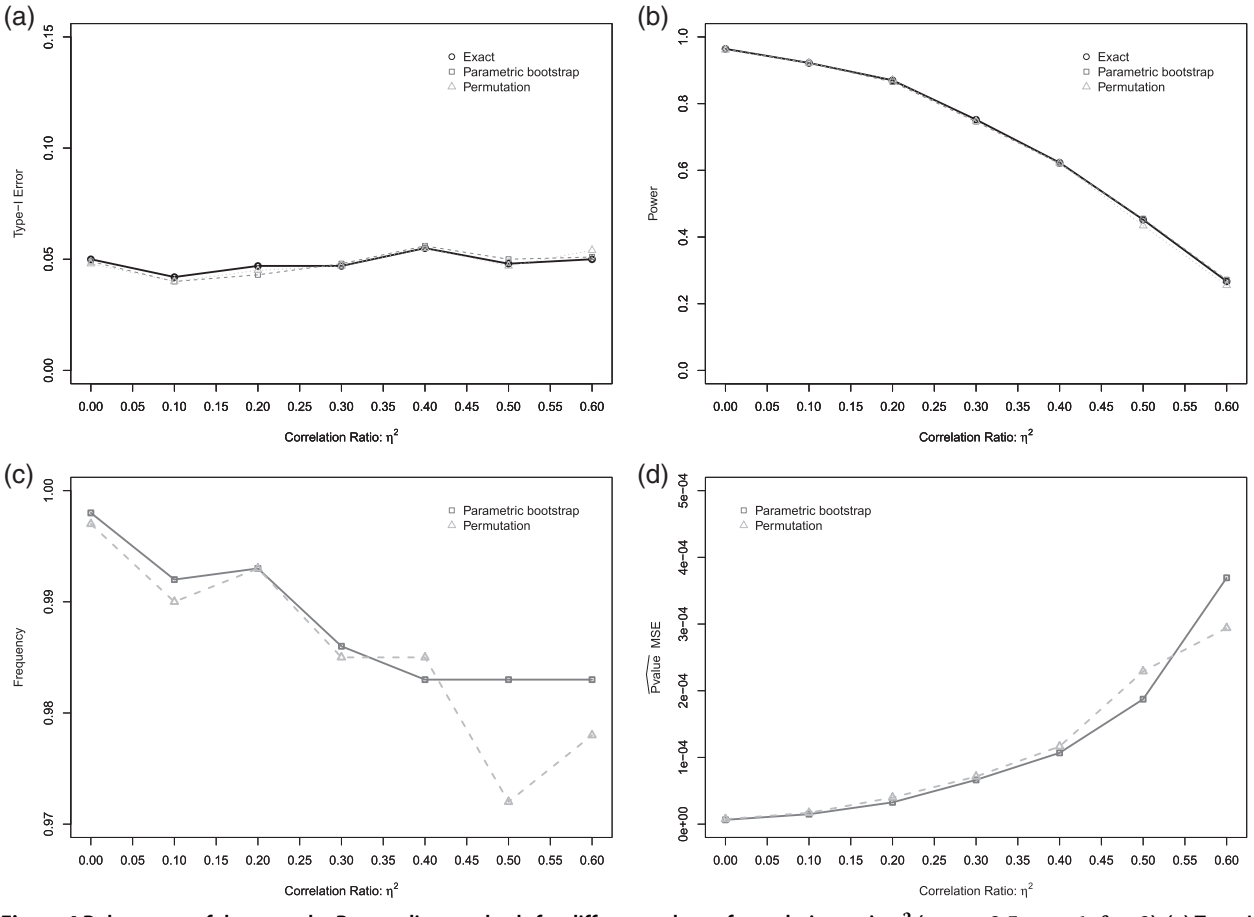


Figure 4 Robustness of the p_{value} by Resampling methods for different values of correlation ratio η^2 ($\gamma_0 = -2.5, \gamma_1 = 1, \beta = 2$): (a) Type-I error; (b) Power; (c) Correct decision rate; (d) Mean Square Error (MSE) of the $\widehat{p_{value}}$.

incorrect conclusion. The proposed approach based on the resampling procedure gave a result which was still significant and more realistic than the uncorrected p_{value}.

Discussion

In this paper, we have considered the problem of correction of significance level for a series of several codings of an explanatory variable in a Generalized Linear Model with several adjusting variables. The methods developed, based on resampling methods, enable us to consider categorical transformations as more flexible in order to explore the unknown shape of the effect between an explanatory and a dependent variable. The simulation studies presented above show, firstly, that the resampling method provides similar results for the Type-I error rate control and the power as those found with the exact method proposed by Liquet and Commenges [14] for dichotomous and Box-Cox transformations. Secondly, in the situation of categorical transformations, these simulations demonstrate the good performance of our proposed approaches. Finally we observed the robustness estimation of the p_{value} by the resampling methods. These methods can be easily generalized to other models, such as the proportional hazards model, and to potentially extend the work of Hashemi and Commenges [15] in the same context.

Conclusion

To conclude, the methods developed, based on resampling, demonstrate good performances, and we have implemented different methods and different strategies of coding in an R package called CPMCGLM M (for Correction of the Pvalue after Multiple Coding in a Generalized Linear Model).

Appendix

The package `CPMCGLM` has been developed in R, an open source statistical software available at http://www.r-project.org. The methods presented in this paper are available in the main function `CPMCGLM()` for Probit, Logit, Linear, and Poisson models. Briefly, the user can specify the transformations tested: Box-Cox, dichotomous or categorical transformations. Two options are possible for defining the cutpoints of the dichotomous and the categorical transformations: the user can either specify them, or the program will automatically use the strategy based on the quantile presented in the simulation study.

The main function provides the best codings according to the maximum test and minimum p_{value} procedures. For this coding, the different methods of correction of the Type-I error rate presented in this paper are provided. We present an illustration of the `CPMCGLM` function on a simulated dataset:

```
data(data_sim)
result<-CPMCGLM(formula=
Weight  Age+as.factor(Sport)+Desease
+Height,
family="gaussian",link="identity",
data=data_sim, varcod="Age",
nb.dicho = 4, nb.categ = 4, nboxcox =
3, N = 10000)

result

Call:
CPMCGLM(formula = Weight   Age +
as.factor(Sport) + Desease +
Height, family = "gaussian", link =
"identity",
data = data_sim,varcod = "Age",
nb.dicho = 4,
nb.categ = 4, nboxcox = 3, N = 10000)

Generalized Linear Model Summary
Family: gaussian
Link: identity
Number of subject: 100
Number of adjustment variable: 4

Resampling
N: 10000

Best coding
Method: Dichotomous transformation
Value of the order quantile cutpoints: 0.6
Value of the quantile cutpoints: 26.4834

Corresponding adjusted pvalue:

           Adjusted pvalue
naive            0.0191
bonferroni       0.2096
bootstrap        0.0686
permutation 0.0656
exact: Correction not available for
these codings
```

Competing interests
Both authors declare that they have no competing interests.

Authors' contributions
BL and JR developed the methodology, the R code, performed the simulation and the analysis on the dataset as well as wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
We would like to thank Luc Letenneur (ISPED, University of Bordeaux) for making the data available and the Danone Research Clinical Study Platform.

Author details
[1]University Bordeaux, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France. [2]INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France. [3]MRC Biostatistics Unit, Institute of Public Health, Cambridge, CB2 0SR, UK. [4]Danone Research, Avenue de la Vauve, Route départementale 128, Palaiseau Cedex 91767, France.

Received: 9 January 2013 Accepted: 17 May 2013
Published: 8 June 2013

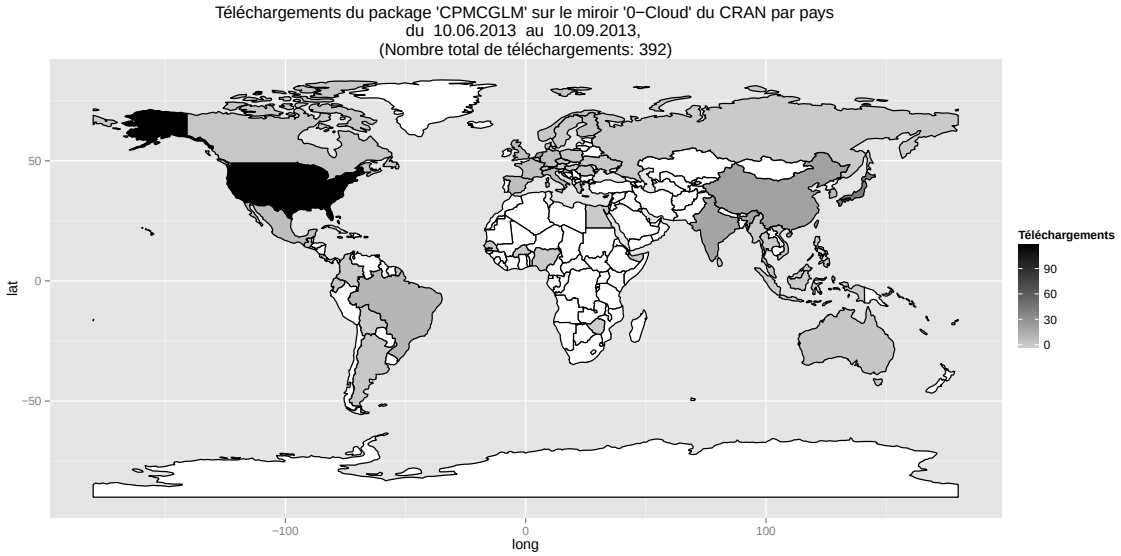
References

1. Bennette C, Vickers A: **Against Quantiles: categorization of continuous variables in epidemiologic research, and its discontents.** *BMC Med Res Methodol* 2012, **12:**21–25.
2. Altman D, Lausen B, Sauerbrei W, Schumacher M: **Dangers of using optimal cutpoints in the evaluation of prognostic factors.** *J Natl Cancer Inst* 1994, **86**(11):829–835.
3. Royston P, Altman D, Sauerbrei W: **Dichotomizing continuous predictors in multiple regression: a bad idea.** *Stat Med* 2006, **25:**127–141.
4. Harrell FE, Lee KL, Mark DB: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.** *Stat Med* 1996, **15**(4):361–387.
5. Miller RG: *Simultaneous statistical inference. 2nd ed*. New York - Heidelberg: Berlin: Springer- Verlag. XVI 299; 1981. figs. DM 44.00.
6. Westfall PH: **Improving power by dichotomizing (even under normality).** *Stat Biopharm Res* 2011, **3**(2):353–362.
7. Simes R: **An improved bonferroni procedure for multiple tests of significance.** *Biometrika* 1986, **73**(3):751–754.
8. Sidak Z: **Rectangular confidence regions for the means of multivariate normal distributions.** *J Am Stat Assoc* 1967, **62:**626–633.
9. Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Stat* 1979, **6:**65–70.
10. Hommel G: **A stagewise rejective multiple test procedure based on a modified Bonferroni test.** *Biometrika* 1988, **75:**383–386.
11. Hochberg Y: **A sharper Bonferroni procedure for multiple test procedure.** *Biometrika* 1988, **75:**800–802.
12. Efron B: **The length heuristic for simultaneous hypothesis tests.** *Biometrika* 1997, **84:**143–157.
13. Liquet B, Commenges D: **Correction of the P-value after multiple coding of an explanatory variable in logistic regression.** *Stat Med* 2001, **20:**2815–2826.
14. Liquet B, Commenges D: **Computation of the p-value of the minimum of score tests in the generalized linear model, application to multiple coding.** *Stat Probability Lett* 2005, **71:**33–38.
15. Hashemi R, Commenges D: **Correction of the p-value after multiple tests in a Cox proportional hazard model.** *Lifetime Data Anal* 2002, **8:**335–348.
16. Bonarek M, Barberger-Gateau P, Letenneur L, Deschamps V, Iron A, Dubroca B, Dartigues J: **between cholesterol, apolipoprotein E polymorphism and dementia: a cross-sectional analysis from the PAQUID study.** *Neuroepidemiology* 2000, **19:**141–48.
17. McCullagh P, Nelder J: *Genaralized Linear Models*. 2edition. New York: Chapman & Hall; 1989.
18. Genz A: **Numerical computation of multivariate normal probabilities.** *J Comput Graphical Stat* 1992, **1:**141–149.
19. Cox D, Hinkley D: *Theoretical Statistics*. London: Chapman & Hall; 1994.
20. Royen T: **Expansions for the multivariate chi-Square distribution.** *J Multivariate Anal* 1991, **38:**213–232.
21. Dagupsta N, Spurrier J: **A class of multivariate χ^2 distributions with applications to comparison with a control.** *Commun Stat- Theory Methods* 1997, **26:**1559–1573.
22. Worsley K: **An improved Bonferroni inequality and applications.** *Biometrika* 1982, **69:**297–302.
23. Westfall PH, Young S: *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*. New York: NY Wiley; 1992. xvii, 340 p.
24. Yu K, Liang F, Ciampa J, Chatterjee N: **Efficient *p*-value evaluation for resampling-based tests.** *Biostatistics* 2011, **12**(3):582–593.
25. Commenges D, Liquet B: **Asymptotic distribution of score statistics for spatial cluster detection with censored data.** *Biometrics* 2008, **64**(4):1287–1289.
26. Romano J: **On the behavior of randomization tests without a group invariance assumption.** *J Am Stat Assoc* 1990, **85**(411–412):686.
27. Xu H, Hsu J: **Applying the generalized partitioning principle to control the generalized familywise error rate.** *Biom J* 2007, **49:**52–67.
28. Kaizar E, Li Y, Hsu J: **Permutation multiple tests of binary features do not uniformly control error rates.** *J Am Stat Assoc* 2011, **106**(495):1067–1074.
29. Commenges D: **Transformations which preserve exchangeability and application to permutation tests.** *J Nonparametric Stat* 2003, **15**(2):171–185.
30. Westfall PH, Troendle JF: **Multiple testing with minimal assumptions.** *Biom J* 2008, **50**(5):745–755.
31. Good P: *Permutation Tests: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New-York: Springer-Verlag; 2000.
32. Efron B, Tibshirani R: *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. London: Chapman and Hall/CRC; 1994.

doi:10.1186/1471-2288-13-75
Cite this article as: Liquet and Riou: Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology* 2013 **13**:75.

Téléchargements du package 'CPMCGLM' à travers le monde

Afin de se rendre compte de l'utilisation du package 'CPMCGLM' disponible sur le CRAN, nous avons réalisé une carte du monde du nombre de téléchargements par pays, et sur une durée de trois mois, du package sur le miroir '0-cloud' du CRAN. Cette carte est représentée dans la Figure ci-dessous :



6 Conclusion générale

Les méthodes développées au cours de cette thèse peuvent être résumées autour de deux axes. Le premier concerne le calcul de taille d'échantillon et l'analyse de données de co-critères de jugement principaux. Le second, quant à lui, est plus spécifique et porte sur la correction du degré de signification lors de la recherche d'un codage dit ≪optimal≫. Ce chapitre va donc naturellement se focaliser sur les principales conclusions tirées lors de ce travail de thèse, mais également évoquer les principales perspectives scientifiques qu'il ouvre.

6.1 Co-critères de jugement principaux

A l'heure actuelle, une autorisation de mise sur le marché ou une demande d'allégation santé nécessite, de la part des industriels, une étude approfondie concernant les modes d'action multifactoriels du produit. Cela implique la prise en compte de critères de jugement principaux multiples au sein des essais cliniques réalisés.

Lorsque de tels critères sont utilisés, les scientifiques doivent s'assurer du bien fondé de chaque critère principal qu'ils prévoient d'intégrer à l'étude, afin de ne pas surestimer leur nombre. Ce point permettra d'être en adéquation avec les recommandations de l'ICH, et les recommandations des différentes autorités de santé. Cette réglementation souhaite respecter le principe de parcimonie scientifique.

C'est donc dans le contexte de co-critères de jugement principaux que nous avons développé les méthodologies sur lesquelles nous allons revenir dans la suite de cette section.

6.1.1 ≪At least one win≫

Nous nous sommes tout d'abord interessés au cas où les promoteurs concluent au succès de l'essai si au moins un critère de jugement principal est significatif. Dans la littérature, à l'image de l'article de Dmitrienko et al. [Dmitrienko et al., 2012], ces critères de jugement sont désignés par l'anglicisme ≪at least one win≫.

Dans ce contexte, nous avons développé une méthode individuelle permettant une correction de la multiplicité lors du calcul de taille d'échantillon et de l'analyse des données. Cette méthode est basée sur le principe des tests d'Union Intersection et n'est valable que pour des variables continues. Nous avons pu voir précédemment que cette méthode est plus puissante que les méthodes classiquement utilisées. Cela s'explique majoritairement par la prise en compte de la loi conjointe des statistiques de test, et de leurs corrélations. Pour définir la loi conjointe des statistiques de tests, nous nous sommes basés sur une approximation par une loi multivariée de Student de type-I, conseillée par Hasler et Hothorn [Hasler and Hothorn, 2011]. En parallèle de cette méthode, nous nous sommes interessés à une méthode dite ≪globale≫. Cette méthode se base sur un modèle multivarié et permet d'éviter une correction de la multiplicité. Au vu des simulations effectuées, cette méthode paraît plus puissante que les méthodes individuelles pour des corrélations faibles $(\rho \leq 0.3)$ et fortes $(\rho \geq 0.6)$. Ce phénomène ne fait que s'amplifier quand il est possible d'ajuster nos résultats sur d'autres variables et que la distribution conjointe n'est pas connue. Cependant, cette méthode possède aussi une limite conséquente puisqu'elle ne nous permet pas une conclusion individuelle quant aux critères de jugement. Une solution pour pallier cette limite pourraît être d'utiliser une méthode basée sur le principe de ≪closed testing≫, et d'obtenir à terme un résultat individuel.

6.1.2 ≪At least r win≫

Nous nous sommes ensuite interessés aux critères que nous avons nommés par l'anglicisme ≪at least r win≫, qui correspondent aux critères pour lesquels le succès de l'essai est conclu si au moins r critères de jugement principaux sont significatifs ($r \in \{1, \ldots, m\}$). L'objectif de ce travail consistait à travailler sur la généralisation du calcul de taille d'échantillon pour ce type de critères, et ce pour les procédures ≪stepwise≫ ou ≪single step≫. Malgré le fait que nous ayons obtenu des formules de puissance générales, nous avons dû nous restreindre aux critères de jugement continus. En effet,

nous ne connaissons la forme analytique de la loi conjointe des statistiques que pour les critères de jugement continus. En dépit de cette limite, ce travail est novateur et devrait pouvoir être utile au sein de la recherche clinique. En effet, ce travail se base sur les procédures les plus communément utilisées dans le domaine, et il est disponible sous le logiciel R. Ces deux points devraient donc permettre de parfaitement intégrer notre travail au sein des pratiques cliniques actuelles. De plus, il est important de noter que le package Sample regroupe l'ensemble des méthodes que nous avons développées sur la problématique des ≪co-primary endpoints≫.

6.1.3 Perspective

Une des perspectives possible à ces travaux consiste à généraliser l'ensemble des méthodes développées à des critères de jugement catégoriels, ou mixtes (quantitatifs et qualitatifs). La difficulté consiste donc à trouver la loi conjointe des statistiques. Une des solutions pourrait consister à utiliser une méthode par rééchantillonage permettant d'estimer la loi conjointe.

6.2 ≪Optimal≫ Coding

L'autre axe de recherche consistait à généraliser le travail de Liquet et Commenges concernant la correction du degré de signification lors de la recherche d'un codage ≪optimal≫ dans un modèle linéaire généralisé [Liquet and Commenges, 2005]. Ce travail concernait des codages binaires et des transformations continues (i.e. Box-Cox), pour lesquels les statistiques du score utilisées peuvent être simplifiées afin que leurs lois conjointes puissent suivre une distribution Gaussienne multivariée.
Dans notre contexte, cette simplification n'est pas possible, et aucune forme analytique de la loi conjointe n'est disponible. Pour pallier cette difficulté, nous avons choisi d'utiliser des méthodes de rééchantillonage permettant d'estimer une distribution conjointe de référence des tests statistiques. Les simulations ont montré le bon comportement de ces méthodes dans le cadre général. Cependant, l'utilisation des méthodes par permutation nécessite la validité de certaines hypothèses. Il est par exemple important, comme nous l'expliquons dans l'article, de vérifier l'hypothèse d'échangeabilité pour ces méthodes. Si cette hypothèse n'est pas vérifiée il est alors conseillé de privilégier la méthode par bootstrap paramétrique qui nécessite moins d'hypothèses [Good, 2000, p. 171], et qui, dans ce cas de figure, donnera des résultats plus fiables.

L'ensemble des méthodes présentes dans le papier de Liquet et Commenges [Liquet and Commenges, 2005], ainsi que les méthodes de rééchantillonage sont développées au sein du package CPMCGLM, disponible sur le site du CRAN qui référence l'ensemble des packages R.

Perspective

Hashemi et Commenges [Hashemi and Commenges, 2002] ont généralisé le travail de Liquet et Commenges [Liquet and Commenges, 2005] aux modèles à risques proportionnels. Ici, nous pensons qu'il pourrait être également intéressant de généraliser nos méthodes à ce type de modèles, et de les ajouter au sein du package R que nous avons développé.

6.3 Apports industriels

L'ensemble des travaux réalisés durant la thèse a eu pour objectif de répondre aux problématiques industrielles inhérentes à la problématique des tests multiples.
Afin de les rendre accessible en interne au plus grand nombre, j'ai dispensé pendant toute la durée de ma thèse des formations sur cette problématique, sur les principales méthodes présentes dans la littérature, ainsi que sur les différentes méthodes que nous avons développées au cours de la thèse. Ces formations étaient réalisées en deux temps, une partie portant sur la théorie et l'autre plus pratique sur l'utilisation de ces méthodes sous les logiciels statistiques (SAS et/ou R).
En parallèle de ces formations, la méthode développée dans le chapitre 2 a été utilisée dans un essai clinique pour le calcul de taille d'échantillon, ainsi que pour son analyse. J'ai donc participé au sein de cet essai à l'écriture du Plan d'Analye Statistique, à l'analyse, ainsi qu'à la rédaction du rapport d'étude.
Mon rôle au sein de la biométrie chez Danone a aussi été une fonction de support méthodologique principalement sur les problématiques des tests multiples.
In fine, cette thèse a permis, au sein du centre de recherche Danone, de sensibiliser les scientifiques à la problématique des tests multiples, et de développer ou de leur proposer des méthodes adaptées à leurs problématiques scientifiques.

7 Bibliographie

[Armitage, 1998] Armitage, P. (1998). Multivariate t distribution. In *Encyclopedia of biostatistics*. Wiley.

[Bach et al., 2011] Bach, J.-F., Bach, J., Jérome, D., D'artemare, B., et al. (2011). Du bon usage de la bibliométrie pour l'évaluation individuelle des chercheurs. Technical report, Institut de France, Académie des sciences.

[Bauer et al., 1986] Bauer, P., Hackl, P., Hommel, G., and Sonnemann, E. (1986). Multiple testing of pairs of one-sided hypotheses. *Metrika*, 33 :121–127.

[Bauer et al., 1998] Bauer, P., Röhmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17(18) :2133–2146.

[Begun and Gabriel, 1981] Begun, J. and Gabriel, K. (1981). Closure of the newman-keuls multiple comparisons procedure. *Journal of the American Statistical Association*, 76(374) :241–245.

[Benjamini and Braun, 2002] Benjamini, Y. and Braun, H. (2002). John w. tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6) :1576–1594.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1) :289–300.

[Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of statistics*, 29(4) :1165–1188.

[Bennette and Vickers, 2012] Bennette, C. and Vickers, A. (2012). Against quantiles : categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, 12 :21–25.

[Berger, 1982] Berger, R. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4) :295–300.

[Bernhard et al., 2004] Bernhard, G., Klein, M., and Hommel, G. (2004). Global and multiple test procedures using ordered p-values : a review. *Statistical Papers*, 45(1) :1–14.

[Boge et al., 2009] Boge, T., Rémigy, M., Vaudaine, S., Tanguy, J., Bourdet-Sicard, R., and Van der Werf, S. (2009). A probiotic fermented dairy drink improves antibody response to influenza vaccination in the elderly in two randomised controlled trials. *Vaccine*, 27(41) :5677–5684.

[Bretz et al., 2003] Bretz, F., Hothorn, L., and Hsu, J. (2003). Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statistics in medicine*, 22(6) :847–858.

[Bretz et al., 2010] Bretz, F., Hothorn, T., and Westfall, P. (2010). *Multiple comparisons using R*. Chapman & Hall/CRC.

[Bretz et al., 2009] Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine*, 28(4) :586–604.

[Bretz et al., 2011] Bretz, F., Maurer, W., and Hommel, G. (2011). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in medicine*, 30(13) :1489–1501.

[Burman et al., 2009] Burman, C., Sonesson, C., and Guilbaud, O. (2009). A recycling framework for the construction of bonferroni-based multiple tests. *Statistics in Medicine*, 28(5) :739–761.

[Chen et al., 2009] Chen, C., Cohen, A., and Sackrowitz, H. (2009). Admissible, consistent multiple testing with applications including variable selection. *Electronic Journal of Statistics*, 3 :633–650.

[Chen et al., 2011] Chen, J., Luo, J., Liu, K., and Mehrotra, D. (2011). On power and sample size computation for multiple testing procedures. *Computational Statistics & Data Analysis*, 55(1) :110–122.

[Chippaux, 2004] Chippaux, J. (2004). *Pratique des essais cliniques en Afrique*. Didactiques (Bondy). IRD.

[Chow et al., 2008] Chow, S., Shao, J., and Wang, H. (2008). *Sample size calculations in clinical research*, volume 20. Chapman & Hall.

[Chuang-Stein et al., 2007] Chuang-Stein, C., Stryszak, P., Dmitrienko, A., and Offen, W. (2007). Challenge of multiple co-primary endpoints : a new approach. *Statistics in medicine*, 26(6) :1181–1192.

[Cook and Farewell, 1996] Cook, R. and Farewell, V. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(1) :93–110.

[Craiu and Sun, 2008] Craiu, R. V. and Sun, L. (2008). Choosing the lesser evil : trade-off between false discovery rate and non-discovery rate. *Statistica Sinica*, 18 :861–879.

[D'Agostino Sr, 2000] D'Agostino Sr, R. (2000). Controlling alpha in a clinical trial : the case for secondary endpoints. *Statistics in medicine*, 19(6) :763–766.

[Dalal and Mallows, 1992] Dalal, S. and Mallows, C. (1992). Buying with exact confidence. *The Annals of Applied Probability*, 2(3) :752–765.

[Davison et al., 2011] Davison, B., Cotter, G., Sun, H., Chen, L., Teerlink, J., Metra, M., Felker, G., Voors, A., Ponikowski, P., Filippatos, G., et al. (2011). Permutation criteria to evaluate multiple clinical endpoints in a proof-of-concept study : lessons from pre-relax-ahf. *Clinical Research in Cardiology*, 100(9) :745–753.

[Dmitrienko et al., 2012] Dmitrienko, A., D'Agostino, R. B., and Huque, M. F. (2012). Key multiplicity issues in clinical drug development. *Statistics in medicine*, 32(7).

[Dmitrienko et al., 2003] Dmitrienko, A., Offen, W., and Westfall, P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine*, 22(15) :2387–2400.

[Dmitrienko et al., 2009] Dmitrienko, A., Tamhane, A., and Bretz, F. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/Crc Biostatistics Series. Chapman & Hall/CRC.

[Dudoit and Van der Laan, 2008] Dudoit, S. and Van der Laan, J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer.

[Dunn, 1958] Dunn, O. (1958). Estimation of the means of dependent variables. *The Annals of Mathematical Statistics*, 29(4) :1095–1111.

[Dunn, 1961] Dunn, O. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293) :52–64.

[Dunnett, 1955] Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272) :1096–1121.

[Dunnett and Tamhane, 1993] Dunnett, C. and Tamhane, A. (1993). Power comparisons of some step-up multiple test procedures. *Statistics & probability letters*, 16(1) :55–58.

[Dunnett and Tamhane, 1992] Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, 87(417) :162–170.

[Editorship, 2010] Editorship, B. (2010). Uniform requirements for manuscripts submitted to biomedical journals : Writing and editing for biomedical publication. Technical report, International Committee of Medical Journal Editors.

[Efron, 2008] Efron, B. (2008). Simultaneous inference : When should hypothesis testing problems be combined ? *The Annals of Applied Statistics*, 2(1) :197–223.

[EMA, 2002] EMA (2002). Cpmp points to consider on multiplicity issues in clinical trials. Technical report, European Agency fo the Evaluation of Medicinal Products.

[EMA, 2009] EMA (2009). Chmp guideline on clinical development of fixed combination medicinal products. Technical report, European Agency of the Evaluation of Medicinal Products.

[Ferkingstad et al., 2008] Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008). Unsupervised empirical bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 2(2) :714–735.

[Finner, 1988] Finner, H. (1988). Abgesschlossene multiple spannweitentests. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypothesenprüfung - Multiple Hypotheses Testing*, pages 10–32. Heidelberg : Springer.

[Finner et al., 2006] Finner, H., Giani, G., and Straßburger, K. (2006). Partitioning principle and selection of good treatments. *Journal of Statistical Planning and Inference*, 136(7) :2053–2069.

[Finner and Straßburger, 2002] Finner, H. and Straßburger, K. (2002). The partitioning principle : a powerful tool in multiple decision theory. *The Annals of statistics*, 30(4) :1194–1213.

[Finner and Straßburger, 2006] Finner, H. and Straßburger, K. (2006). On δ-equivalence with the best in k-sample models. *Journal of the American Statistical Association*, 101(474) :737–746.

[Foulkes, 2009] Foulkes, A. (2009). *Applied statistical genetics with R : for population-based association studies*. Springer Verlag.

[Gabriel, 1969] Gabriel, K. (1969). Simultaneous test procedures - some theory of multiple comparisons. *Ann. Math. Stat.*, 40 :224–250.

[Games, 1971] Games, P. (1971). Multiple comparisons of means. *American Educational Research Journal*, 56(293) :531–565.

[Genovese and Wasserman, 2002] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :499–517.

[Good, 2000] Good, P. (2000). *Permutation tests*. Wiley Online Library.

[Grouin et al., 2005] Grouin, J., Coste, M., and Lewis, J. (2005). Subgroup analyses in randomized clinical trials : statistical and regulatory issues. *Journal of Biopharmaceutical Statistics*, 15(5) :869–882.

[Hashemi and Commenges, 2002] Hashemi, R. and Commenges, D. (2002). Correction of the p-value after multiple tests in a cox proportional hazard model. *Lifetime Data Analysis*, 8 :335–348.

[Hasler and Hothorn, 2011] Hasler, M. and Hothorn, L. (2011). A dunnett-type procedure for multiple endpoints. *The International Journal of Biostatistics*, 7(1) :1–15.

[Hayter and Hsu, 1994] Hayter, A. and Hsu, J. (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association*, 89(425) :128–136.

[Hochberg, 1988] Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4) :800–802.

[Hochberg and Tamhane, 1987] Hochberg, Y. and Tamhane, A. (1987). *Multiple comparison procedures*. Wiley series in probability and mathematical statistics : Applied probability and statistics. Wiley.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2) :65–70.

[Hommel, 1988] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2) :383–386.

[Hommel, 1989] Hommel, G. (1989). A comparison of two modified bonferroni procedures. *Biometrika*, 76(3) :624–625.

[Hommel and Hoffmann, 1988] Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypothesenprüfung - Multiple Hypotheses Testing*, pages 154–161. Heidelberg : Springer.

[Hotelling, 1931] Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3) :360–378.

[Hung and Wang, 2009] Hung, H. and Wang, S. (2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics*, 19(1) :1–11.

[ICH, 1999] ICH (1999). Ich topic e9 : Notes for guidance on statistical principles for clinical trials. Technical report, International Conference on Harmonization.

[Johnson, 2008] Johnson, A. (2008). Publish and be wrong. *The Economist 9th Oct*.

[Johnson and Kotz, 1972] Johnson, N. and Kotz, S. (1972). *Distributions in statistics : continuous multivariate distributions*. Wiley series in probability and mathematical statistics : Applied probability and statistics. Wiley.

[Julious and McIntyre, 2012] Julious, S. and McIntyre, N. (2012). Sample sizes for trials involving multiple correlated must-win comparisons. *Pharmaceutical Statistics*, 11(2) :177–185.

[Julious and Owen, 2006] Julious, S. and Owen, R. (2006). Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical statistics*, 5(1) :29–37.

[Keuls, 1952] Keuls, M. (1952). The use of the studentized range in connection with an analysis of variance. *Euphytica*, 1(2) :112–122.

[Khurana et al., 2009] Khurana, V., Teo, C., Kundi, M., Hardell, L., and Carlberg, M. (2009). Cell phones and brain tumors : a review including the long-term epidemiologic data. *Surgical neurology*, 72(3) :205–214.

[Laska and Meisner, 1989] Laska, E. and Meisner, M. (1989). Testing whether an identified treatment is best. *Biometrics*, 45(4) :1139–1151.

[Lehmann, 1986] Lehmann, E. (1986). *Testing statistical hypotheses*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley.

[Lehmann and Romano, 2005] Lehmann, E. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3) :1138–1154.

[Lind, 1772] Lind, J. (1772). *A Treatise on the Scurvy : In Three Parts, Containing an Inquiry Into the Nature, Causes, an Cure, of that Disease, Together with a Critical and Chronological View of what Has Been Published on the Subject*. S. Crowder.

[Liquet and Commenges, 2001] Liquet, B. and Commenges, D. (2001). Correction of the p-value after multiple coding of an explanatory variable in logistic regression. *Statistics in medicine*, 20(19) :2815–2826.

[Liquet and Commenges, 2005] Liquet, B. and Commenges, D. (2005). Computation of the p-value of the maximum of score tests in the generalized linear model ; application to multiple coding. *Statistics & probability letters*, 71(1) :33–38.

[Liu et al., 2007] Liu, Y., Hsu, J., and Ruberg, S. (2007). Partition testing in dose–response studies with multiple endpoints. *Pharmaceutical Statistics*, 6(3) :181–192.

[Logan and Tamhane, 2004] Logan, B. and Tamhane, A. (2004). On o'brien's ols and gls tests for multiple endpoints. *Lecture Notes-Monograph Series*, 47 :76–88.

[Marcus et al., 1976] Marcus, R., Eric, P., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3) :655–660.

[Maurer and Mellein, 1988] Maurer, W. and Mellein, B. (1988). On new multiple tests procedures based on independent p-values and the assessment of the power. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypothesenprüfung - Multiple Hypotheses Testing*, pages 48–66. Heidelberg : Springer.

[Millot, 2009] Millot, G. (2009). Comprendre et réaliser les tests statistiques a l'aide de r. *De Boeck*, page 180.

[Naik, 1975] Naik, U. (1975). Some selection rules for comparing p processes with a standard. *Communications in Statistics-Theory and Methods*, 4(6) :519–535.

[Neuhäuser, 2006] Neuhäuser, M. (2006). How to deal with multiple endpoints in clinical trials. *Fundamental & clinical pharmacology*, 20(6) :515–523.

[Newman, 1939] Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, 31(1/2) :20–30.

[O'Brien, 1984] O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4) :1079–1087.

[O'Neill, 1997] O'Neill, R. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials*, 18(6) :550 – 556.

[Pédrono et al., 2009] Pédrono, G., Thiébaut, R., Alioum, A., Lesprit, P., Fritzell, B., Lévy, Y., and Chêne, G. (2009). A new endpoint definition improved clinical relevance and statistical power in a vaccine trial. *Journal of Clinical Epidemiology*, 62(10) :1054–1061.

[Peto, 1991] Peto, R. (1991). Vitamins and iq. *BMJ : British Medical Journal*, 302(6781) :906.

[Pocock et al., 1987] Pocock, S., Geller, N., and Tsiatis, A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, 43(3) :487–498.

[Poulain, 2011] Poulain, J.-P. (2011). *Sociologie de l'alimentation*. Puf.

[Rafter et al., 2002] Rafter, J., Abell, M., and Braselton, J. (2002). Multiple comparison methods for means. *Siam Review*, 44(2) :259–278.

[Rom, 1990] Rom, D. (1990). A sequentially rejective test procedure based on a modified bonferroni inequality. *Biometrika*, 77(3) :663–665.

[Romano and Shaikh, 2006] Romano, J. P. and Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, 34(4) :1850–1873.

[Roquain and van de Wiel, 2009] Roquain, E. and van de Wiel, M. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3 :678–711.

[Roy, 1953] Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2) :220–238.

[Roy and Bose, 1953] Roy, S. and Bose, R. (1953). Simultaneous confidence interval estimation. *The Annals of Mathematical Statistics*, 24(4) :513–536.

[Royston et al., 2006] Royston, P., Altman, D., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression : a bad idea. *Statistics in medicine*, 25(1) :127–141.

[Sankoh et al., 2003] Sankoh, A., D'Agostino, R., and Huque, M. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine*, 22(20) :3133–3150.

[Sankoh et al., 1997] Sankoh, A., Huque, M., and Dubey, S. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in medicine*, 16(22) :2529–2542.

[Sankoh et al., 1999] Sankoh, A., Huque, M., Russell, H., and D'Agostino, R. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug information journal*, 33(1) :119–140.

[Sarkar, 2002] Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30(1) :239–257.

[Sarkar et al., 2008] Sarkar, S., Zhou, T., and Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling fdr and fnr from a bayesian perspective. *Statist. Sinica*, 18(3) :925–945.

[Sarkar, 1998] Sarkar, S. K. (1998). Some probability inequalities for ordered mtp2 random variables : a proof of the simes conjecture. *The Annals of Statistics*, 26(2) :494–504.

[Sarkar, 2008] Sarkar, S. K. (2008). On the simes inequality and its generalization. *IMS Collections Beyond Parametrics in Interdisciplinary Research : Festschrift in Honor of Professor Pranab K. Sen*, 1 :231–242.

[Sarkar and Chang, 1997] Sarkar, S. K. and Chang, C.-K. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92(440) :1601–1608.

[Scheffé, 1999] Scheffé, H. (1999). *The Analysis of Variance*. Wiley Classics Library. John Wiley & Sons.

[Schulz and Grimes, 2005a] Schulz, K. and Grimes, D. (2005a). Epidemiology 4 multiplicity in randomised trials i : endpoints and treatments. *Lancet*, 365 :1591–95.

[Schulz and Grimes, 2005b] Schulz, K. and Grimes, D. (2005b). Epidemiology 5 multiplicity in randomised trials ii : subgroup and interim analyses. *Lancet*, 365 :1657–61.

[Seeger, 1968] Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics*, 10(3) :586–593.

[Seignalet, 2004] Seignalet, J. (2004). *L'alimentation ou la troisième médecine*. F.-X. de Guibert.

[Senn and Bretz, 2007] Senn, S. and Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, 6(3) :161–170.

[Shaffer, 1986] Shaffer, J. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395) :826–831.

[Šidák, 1967] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318) :626–633.

[Simes, 1986] Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3) :751–754.

[Soric, 1989] Soric, B. (1989). Statistical" discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406) :608–610.

[Sozu et al., 2010] Sozu, T., Sugimoto, T., and Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in medicine*, 29(21) :2169–2179.

[Sozu et al., 2011] Sozu, T., Sugimoto, T., and Hamasaki, T. (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics*, 21(4) :650–668.

[Stefansson et al., 1988] Stefansson, G., Kim, W., and Hsu, J. (1988). On confidence sets in multiple comparisons. *Statistical Decision Theory and Related Topics IV*, 2 :89–104.

[Storey, 2002] Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :479–498.

[Storey, 2003] Storey, J. (2003). The positive false discovery rate : A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6) :2013–2035.

[Strassburger and Bretz, 2008] Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the holm procedure and other bonferroni-based closed tests. *Statistics in medicine*, 27(24) :4914–4927.

[Strassburger et al., 2007] Strassburger, K., Bretz, F., and Finner, H. (2007). Ordered multiple comparisons with the best and their applications to dose–response studies. *Biometrics*, 63(4) :1143–1151.

[Strassburger et al., 2004] Strassburger, K., Bretz, F., and Hochberg, Y. (2004). Compatible confidence intervals for intersection union tests involving two hypotheses. *Lecture Notes-Monograph Series*, 47 :129–142.

[Sun and Cai, 2007] Sun, W. and Cai, T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479) :901–912.

[Tamhane and Dunnett, 1999] Tamhane, A. and Dunnett, C. (1999). Stepwise multiple test procedures with biometric applications. *Journal of Statistical Planning and Inference*, 82(1) :55–68.

[Tamhane et al., 1998a] Tamhane, A., Liu, W., and Dunnett, C. (1998a). A generalized step-up-down multiple test procedure. *Canadian Journal of Statistics*, 26(2) :353–363.

[Tamhane et al., 1998b] Tamhane, A., Liu, W., and Dunnett, C. (1998b). A generalized step-up-down multiple test procedure. *Canadian Journal of Statistics*, 26(2) :353–363.

[Tang and Zhang, 2007] Tang, W. and Zhang, C. (2007). Empirical bayes methods for controlling the false discovery rate with dependent data. *Lecture Notes-Monograph Series*, 54 :151–160.

[Tong, 1980] Tong, Y. (1980). *Probability inequalities in multivariate distributions*. Probability and mathematical statistics. Academic Press.

[Tong, 1990] Tong, Y. (1990). *The multivariate normal distribution*. Springer series in statistics. Springer-Verlag.

[Toothaker, 1991] Toothaker, L. (1991). *Multiple comparisons for researchers.* Sage Publications, Inc.

[Troendle, 1996] Troendle, J. (1996). A permutational step-up method of testing multiple outcomes. *Biometrics*, 52(3) :846–859.

[Tröhler, 2005] Tröhler, U. (2005). Lind and scurvy : 1747 to 1795. *Journal of the Royal Society of Medicine*, 98(11) :519–522.

[Tukey and Braun, 1994] Tukey, J. and Braun, H. (1994). *The Collected Works of John W. Tukey : Multiple Comparisons, 1948-1983*. Chapman & Hall.

[Vale, 2008] Vale, B. (2008). The conquest of scurvy in the royal navy 1793–1800 : a challenge to current orthodoxy. *The Mariner's Mirror*, 94(2) :160–175.

[Victor, 1982] Victor, N. (1982). Exploratory data analysis and clinical research. *Methods of Information in Medicine*, 21(2) :53–54.

[Westfall and Krishen, 2001] Westfall, P. and Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99(1) :25–40.

[Westfall and Tobias, 2007] Westfall, P. and Tobias, R. (2007). Multiple testing of general contrasts. *Journal of the American Statistical Association*, 102(478) :487–494.

[Westfall et al., 2011] Westfall, P., Tobias, R., and Wolfinger, R. (2011). *Multiple comparisons and multiple tests using SAS*. SAS Publishing.

[Westfall and Troendle, 2008] Westfall, P. and Troendle, J. (2008). Multiple testing with minimal assumptions. *Biometrical Journal*, 50(5) :745–755.

[Westfall and Young, 1993] Westfall, P. and Young, S. (1993). *Resampling-Based Multiple Testing : Examples and Methods for P-Value Adjustment*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. Wiley.

[Westfall, 2005] Westfall, P. H. (2005). Combining p values. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd.

[Westfall and Young, 1989] Westfall, P. H. and Young, S. S. (1989). P value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84(407) :780–786.

[Whitehead, 1991] Whitehead, R. (1991). Vitamins, minerals, schoolchildren, and iq. *BMJ : British Medical Journal*, 302(6776) :548.

[Wright, 1992] Wright, S. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, 48(4) :1005–1013.

[Yekutieli and Benjamini, 1999] Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2) :171–196.

[Yoon et al., 2011] Yoon, F., Fitzmaurice, G., Lipsitz, S., Horton, N., Laird, N., and Normand, S. (2011). Alternative methods for testing treatment effects on the basis of multiple outcomes : Simulation and case study. *Statistics in Medicine*, 30(16) :1917–1932.

[Yu et al., 2011] Yu, K., Liang, F., Ciampa, J., and Chatterjee, N. (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics*, 12(3) :582–593.

8

Valorisations scientifiques et enseignement

8.1 Publications

[1] Lafaye de Micheaux, P., Liquet, B., Marque, S., Riou,J. (2013) Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. *Journal of Biopharmaceutical Statistics,* accepté pour publication le 21 Octobre 2012.

Contribution Personnelle : Développement méthodologique, Programmation du Code R, Simulations, Applications, Rédaction du premier draft de l'article, Création du package associé.

[2] Delorme, P., Lafaye de Micheaux, P., Liquet, B., Riou,J. Type II generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination. *Biometrika,* Soumis le 30 Octobre 2013 à Biometrika.

Contribution Personnelle : Validation du développement méthodologique, Participation à la programmation du code R, Applications, Rédaction du premier draft de l'article, Création du package associé.

[3] Liquet, B., Riou,J. (2013). Correction of the significance level when attempting multiple transformations of an explanatory variable in Generalized Linear Models. *BMC Medical Research Methodology,*13(1) : 75-85.

Contribution Personnelle : Développement méthodologique, Programmation du Code R, Simulations, Application, Rédaction du premier draft de l'article, Création du package associé.

8.2 Communications scientifiques orales en congrès

[1] Liquet, B., Riou,J. * (2011). Méthodes de correction du degré de signification pour une recherche de codage optimal dans un modèle linéaire généralisé. *43 èmes Journées de la Statistique*, Tunis (Tunisie)

[2] Liquet, B., Riou,J.* (2011). Correction of the significance level after multiple coding of an explanatory variable in generalized linear model. *7 th International Conference on Multiple Comparison Procedures*, Washington (Etats Unis)

[3] Liquet, B., Riou,J.* (2012). Package CPMCGLM : Correction de la pvaleur engendrée par la recherche d'un codage d'une variable explicative dans un modèle linéaire généralisé *Premières Rencontres R*, Bordeaux (France).

[4] Lafaye de Micheaux, P.*, Liquet, B., Marque, S., Riou,J. (2012). Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. *XXVI th International Biometric Conference*, Kobe (Japon).

[5] Delorme, P., Lafaye de Micheaux, P., Liquet, B., Riou,J.* (2013). Calcul de taille d'échantillon dans le cadre de critères de jugement multiples avec un contrôle de la ≪r-Power≫ et du ≪gFWER≫. *45 èmes Journées de la Statistique*, Toulouse (France)

[6] Delorme, P., Lafaye de Micheaux, P., Liquet, B., Riou,J.* (2013). Package Sample : Sample size determination and data analysis in the context of continuous co-primary endpoints in clinical trials. *Deuxièmes Rencontres R*, Lyon (France)

[7] Delorme, P., Lafaye de Micheaux, P., Liquet, B., Riou,J.* (2013). Type II generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination. *8 th International Conference on Multiple Comparison Procedures*, Southampton (Angleterre)

8.3 Communications scientifiques affichées en congrès

[1] Delorme, P., Lafaye de Micheaux, P., Liquet, B., Riou,J. (2012) Power and sample size computation for a control of the ≪r-Power≫ *NZSA 2012 Conference* Dunedin (New Zealand)

∗. Orateur

8.4 Autres communications scientifiques orales

[1] Elfakir, A.*, Riou, J.* (2011). Multiplicity Adjustment strategy in Actimel studies and Meta-Analyses. *Danone Research, Biometrics Conference*, Wageningen (The Netherlands)

[2] Riou, J.* (2012). Multiple Testing Framework and Common Procedures. *Danone Research, Biometrics Conference*, Palaiseau (France)

[3] Riou, J.* (2013). Co-Primary Endpoints in Clinical Trials : How to manage Multiple Testing ? *Danone Research, Life Science Seminar*, Palaiseau (France)

[4] Riou, J* (2013). Type II generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination. *Novartis*, Basel (Switzerland)

[5] Riou, J.* (2013). Multiple Testing and Sample Size Determination in clinical trials. *Séminaire en Statistiques de l'Université de Franche-Comté*, Besançon (France)

8.5 Packages R

[1] Diakite,A., Liquet, B., Riou,J.† (2012) CPMCGLM : Correction of the significance level after multiple coding of an explanatory variable in generalized linear model.
http://cran.r-project.org/web/packages/CPMCGLM/index.html

[2] Delorme, P., Lafaye de Micheaux, P., Liquet, B., Riou,J.† (2013) Sample : Sample size determination and data analysis in the context of multiple continuous endpoints in clinical trials. Version beta actuellement disponible auprès des auteurs.

8.6 Enseignement

[1] 2011-2013 : Enseignant Vacataire, Université de Bordeaux Segalen, Master2 Recherche mention Sciences de l'éducation, UE : Méthodologie en Recherche (Production et Analyse), 120h (Cours Magistraux et Travaux Dirigés).

†. Référent

[2] 2013 : Enseignant Vacataire, Université de Bordeaux Segalen, Master2 mention Santé Publique spécialité Biostatistique, UE : Apprentissage automatique (Multiplicité des Tests), 6h (Cours Magistraux et Travaux Dirigés).