Master 1

Mention Santé Publique

Universités de Rennes 1, Rennes 2, EHESP

Mémoire

UE Projet bibliographique tutoré

Comparaisons multiples et inflation du risque α en épidémiologie

Corentin Choisy

Yemtougoule Sambiani

Thi Chi Phuong Nguyen

Année universitaire 2021-2022

Encadrants : Ronan Garlantézec, Guillaume Collet

RÉSUMÉ

Introduction

Un grand nombre d'études épidémiologiques se basent sur des tests statistiques pour émettre leurs conclusions. Or, la multiplication de ces tests provoque une inflation du risque d'erreur de première espèce. Il existe plusieurs méthodes de correction visant à contrôler cette inflation, avec des objectifs et contraintes différentes. L'objectif de ce projet est de comparer les deux principaux types de correction (FDR et FWER) et formuler des recommandations sur l'utilisation de ces méthodes.

Matériel et Méthodes

Une revue de la littérature méthodologique sur la base *pubmed* a été effectuée, se focalisant dans un premier temps sur les articles de revue, puis sur les articles de revue publiés après 1995, année de publication de l'article original de Y. Benjamini et Y. Hochberg. Trois lecteurs ont ensuite réalisé une sélection des articles sur base des titres et des abstracts. Les articles retenus ont été examinés indépendamment par les trois lecteurs à l'aide d'une grille d'extraction avant une mise en commun des informations extraites.

Résultats

Sur les 158 articles identifiés au cours des recherches (dont 79 uniques), 10 articles ont été retenus pour lecture complète. Les deux principales catégories de méthodes de correction sont basées sur deux mesures distinctes de l'erreur de première espèce: le FWER et le FDR. Les méthodes basées sur le FDR apparaissent comme plus puissantes que les méthodes basées sur le FWER, mais ne permettent pas le contrôle strict du FWER. Ainsi, les deux familles de méthodes s'appliquent à des situations différentes. Une plus grande prise en compte de ces propriétés en perspective avec les objectifs de l'étude est recommandée.

Discussion & Conclusion

L'application d'une correction n'est pas toujours recommandée, mais est globalement souhaitable sur toute famille de tests pour lesquels l'erreur de deuxième espèce n'est pas l'enjeu primaire. Les méthodes basées sur le FWER sont recommandées pour les familles constituées de peu de tests, tandis que celles basées sur le FDR sont globalement indiquées dans toutes les études exploratoires ou pour lesquelles le contrôle du FWER au sens fort n'est pas strictement nécessaire. Cependant, le manque de consensus et l'existence d'autres méthodes alternatives font qu'un approfondissement sur ces méthodes pourrait permettre de préciser les recommandations formulées dans ce projet.

TABLE DES MATIÈRES

TABLE DES ABREVIATIONS

Abréviations générales

FP: *Faux Positif* - Ici, erreur de première espèce

FN: *Faux Négatif*- Ici, erreur de deuxième espèce

Mesures d'erreur de première espèce

FWER: *Family-Wise Error Rate* - Taux d'erreur sur la famille de tests

FWE: *Family-Wise Error* - Erreur sur la famille de tests

FDR: *False Discovery Rate* - Taux de fausses découvertes

FDP: *False Discovery Proportion* - Proportion de fausses découvertes

FCR: *False Coverage Rate* - Taux de fausse couverture

FDX: *False Discovery Exceedence* - Excès de fausses découvertes

Méthodes de correction

BH: Méthode de Benjamini-Hocherg

BY: Méthode de Benjamini-Yekutieli

BL: Méthode de Benjamini-Liu

INTRODUCTION

En tant que science populationnelle de la santé, l'épidémiologie est vouée tant par la nature de ses méthodes que de ses conclusions à reposer sur les statistiques. Ainsi, la viabilité et les limitations d'une étude épidémiologique sont intimement liées à l'intégrité statistique de sa démarche, mais aussi aux limites inhérentes aux statistiques en elles-mêmes.

Dès lors, quand bien même les épidémiologistes ont au fil des années abouti à des conclusions difficilement atteignables avec une approche individuelle de la santé, ces derniers doivent en permanence composer avec la possibilité de commettre une erreur d'inférence, c'est-à-dire d'émettre des conclusions sur la santé d'une population et ses déterminants en étant induits en erreur par le hasard durant l'étude d'un échantillon de cette population. La considération de cette possibilité est donc un élément central de la conception d'une étude épidémiologique, a fortiori lorsque celle-ci a vocation à orienter une prise de décision en santé publique.

L'erreur statistique est décomposable en deux types. Le premier type, l'erreur de première espèce, correspond à l'affirmation à tort d'une association entre un facteur et un événement de santé ou l'élargissement à tort d'une différence entre deux groupes de l'échantillon à une population plus large. Le second, l'erreur de deuxième espèce, correspond à la non-affirmation d'une telle association ou différence alors que cette dernière existe dans la population générale. Si ces deux composantes peuvent revêtir d'une certaine importance dans le cadre de toute étude épidémiologique, l'erreur de première espèce peut s'avérer d'autant plus problématique que l'orientation à tort de politiques de santé publique ou le déploiement à tort de traitements médicaux peut avoir des conséquences fortement délétères.

Devant le risque d'erreur de première espèce, les scientifiques se fixent généralement un niveau de risque supposé rendre compte de la probabilité de commettre une telle erreur. Cependant, le contrôle de ce risque d'erreur devient problématique lorsque plusieurs comparaisons sont effectuées au sein d'une même étude. En effet, la réalisation de plusieurs tests statistiques au niveau de risque fixé produit une augmentation de la probabilité de commettre au moins une erreur de première espèce sur l'ensemble des tests réalisés, aussi appelée taux d'erreur sur une famille de tests (FWER). On parle alors d'inflation du risque α.

Pour contrôler ce phénomène, diverses méthodes statistiques dites de correction ont été proposées, visant d'abord à contrôler l'augmentation du FWER. Ces premières méthodes, dont la plus populaire est la méthode de Bonferroni, sont perçues comme satisfaisantes dans leur objectif principal mais critiquées pour leur aspect conservateur, c'est-à-dire l'augmentation du risque d'erreur de deuxième espèce qu'elles provoquent. Pour contrer ce problème, une autre approche basée sur le taux de fausses découvertes (ou FDR) a été proposée, aboutissant en 1995 à la proposition par Y. Benjamini et Y. Hochberg d'une nouvelle méthode de correction visant à contrôler cette autre mesure de l'erreur de première espèce se réclamant moins conservative que les méthodes existantes.

De par leurs différences fondamentales tant sur le principe que l'objectif final de la correction qu'elles apportent, ces méthodes s'adressent ainsi à des applications dans des cadres différents (1). Malgré l'apparent manque de cohérence et de discussion du choix de la méthode (ou du simple fait de corriger ou non) (2), aucune publication majeure n'a formulé de lignes de conduites détaillées et argumentées concernant ce problème à destination des épidémiologistes.

L'objectif de ce projet est donc, à partir d'une revue de la littérature méthodologique et après avoir rappelé les différentes méthodes de corrections basées sur le FWER et le FDR et les éléments de comparaison entre ces méthodes, de formuler des recommandations détaillées sur les méthodes de correction dans les problèmes de comparaisons multiples en épidémiologie sur la base de leurs performances et limites d'application aux différents types d'études épidémiologiques.

MATÉRIEL ET MÉTHODE

La revue de la littérature scientifique s'est concentrée sur des articles de la littérature méthodologique en anglais, sans restriction dans un premier temps puis en se restreignant aux revues et revues systématiques publiées après la publication de l'article original décrivant la méthode de Benjamini-Hochberg en 1995 (3). Les articles traitant de la problématique des comparaisons multiples en santé ont été inclus. De même, les articles présentant les diverses méthodes de correction, ainsi que ceux proposant une comparaison entre ces méthodes, ont également été inclus. Les articles restreignant leur cadre à une application hors des études de santé ont été exclus.

L'ensemble des recherches ont été réalisées sur les bases Pubmed et Google Scholar, en utilisant des équations de recherches ciblant le titre et l'abstract, puis dans un second temps ciblant les termes MeSH de Pubmed. Le détail des équations de recherche utilisées est décrit en *Annexe 1*. Les principaux mots-clés utilisés incluent les deux catégories de méthodes étudiées (*False Discovery Rate* et *Family-Wise Error Rate*) ainsi que des termes décrivant le contexte de la recherche (*epidemiology* et *multiple comparison/ testing*). Une consultation des articles cités par les revues de la littérature a également permis d'identifier des articles n'étant pas ressortis au cours des recherches sur les bases de données. Les premières recherches ont eu lieu le 11 janvier 2022 et les derniers articles ont été identifiés le 18 février 2022.

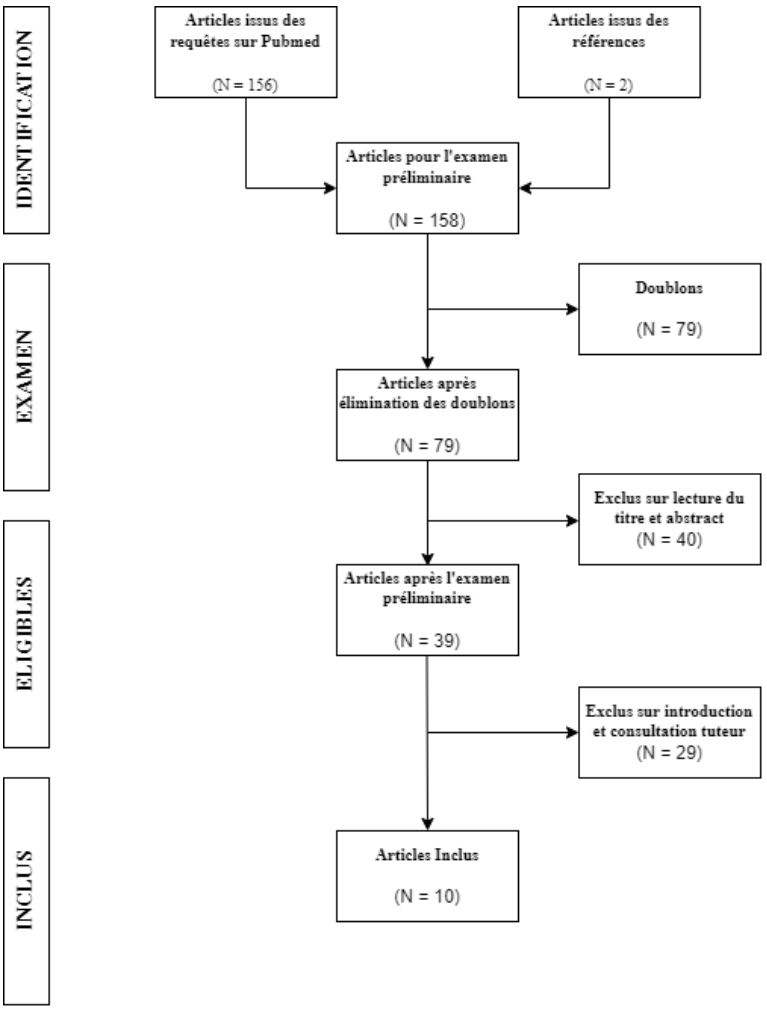
Trois lecteurs (les trois membres du groupe) ont parcouru les titres et les abstracts avant d'effectuer une première sélection après concertation, excluant les articles jugés non pertinents. Cette première étape s'est déroulée après une élimination des doublons à l'aide d'un script bash appliqué au répertoire d'une bibliothèque locale du logiciel Zotero. Une première lecture des abstracts et introductions par ces trois lecteurs des articles restants ont permis de déterminer les articles à inclure, après consultation avec le tuteur du projet. Une attention particulière a été portée sur le journal de publication des articles lors de la sélection, priorisant les articles issus de revues reconnues dans le domaine de l'épidémiologie (*American Journal of Epidemiology, Statistical Methods in Medical Research*)

Les trois lecteurs ont ensuite réalisé une lecture complète de chaque article, indépendante et sans concertation, en utilisant la grille d'extraction présentée en *Annexe 2* pour extraire les informations importantes. Une synthèse a ensuite été réalisée pour chaque article à l'issue de concertations entre les lecteurs, en remplissant la grille d'extraction avec les informations relevées par les lecteurs puis en rédigeant une synthèse pour chaque article.

Le flowchart a été réalisé à l'aide du logiciel R et du package diagramR. Les tableaux ont été reconstruits à partir de ceux des articles à l'aide des logiciels R et Rmarkdown, sauf précision contraire.

RÉSULTATS

Les premières recherches ont permis l'identification de 158 documents, dont 2 identifiés dans les références d'autres articles. La moitié, soit 79 de ces documents, étaient des doublons. Après lecture des titres et abstracts des 79 articles restants, 40 articles supplémentaires ont été exclus car ils n'étaient pas considérés comme pertinents, soit parce qu'ils se limitaient à un domaine spécifique distinct de l'épidémiologie soit parce qu'ils n'appartenaient pas à la littérature méthodologique. Une première lecture de l'introduction ainsi qu'une consultation avec le tuteur du projet ont permis l'élimination de 29 articles supplémentaires, laissant 10 articles pour la lecture complète. Sur ces 10 articles, 5 proposent une réflexion générale sur le problème des comparaisons multiples et des méthodes de correction, 3 sont dédiés aux méthodes basées sur le FDR, 1 aux méthodes basées sur le FWER et 1 propose une comparaison générale de toutes les méthodes les plus communes. Les 10 articles retenus ont été publiés entre 1995 et 2019 principalement dans des revues d'épidémiologie ou de statistiques.



1. *Inflation du risque a: principe et conséquences*

Dans le contexte d'une étude épidémiologique, le problème de comparaison multiple peut être formulé comme suit (4): Si *n* associations indépendantes associées aux hypothèses nulles H_0^1 ,..., H_0^n sont examinées par un ensemble de *n* tests statistiques, la probabilité qu'au moins une de ces associations soit déclarée statistiquement significative vaut $1 - (1 - \alpha)^n > \alpha$, $\forall n \geq 1$ si H_0^1 ,..., H_0^n sont vraies. Cette probabilité correspond à la probabilité *a priori* (sans avoir connaissance de la structure réelle des hypothèses nulles, c'est-à-dire sans connaissance desquelles sont en réalité vraies ou non) de réaliser au moins une erreur de première espèce donc déclarer au moins une association significative à tort (1). α étant une valeur de probabilité correspondant au risque consenti par l'enquêteur de commettre une erreur de première espèce individuellement sur chaque test, ce phénomène résulte en un dépassement du risque d'erreur consenti (4). La croissance de l'expression de ce risque avec le nombre d'associations testées provoque une convergence vers 100% de la probabilité de commettre au moins une erreur de première espèce, ou *FWER* (2), décrite dans la *Table 1*.

TABLE 1: FWER en fonction du nombre de comparaisons

n	FWER (%)
2	9.75
10	40.13
50	92.31
100	99.41

En général et dans le cadre des études épidémiologiques, ce phénomène est à l'origine de plusieurs conséquences délétères:

- Il n'est pas possible d'identifier quelles hypothèses ont été rejetées à tort ni combien ont été rejetées à tort (2,5).
- L'utilisation des résultats des tests en tant que preuve sans appui scientifique supplémentaire peut conduire à de mauvaises prises de décision (5).
- Dans les structures de tests avec dépendance, l'augmentation du FWER est globalement plus limitée, mais reste problématique (5).

Pour se prémunir face à ce problème et ses conséquences, plusieurs méthodes dites de correction ont été développées. Certaines, basées sur le contrôle du FWER, cherchent à maintenir le FWER sous le niveau α consenti (1,2) tandis que d'autres, principalement basées sur le FDR, visent à limiter les conséquences du problème dans certains contextes et à limiter les inconvénients liés à l'utilisation de corrections basées sur le FWER, mais contrôlent elle-mêmes rarement le FWER sous le seuil consenti (1,3).

2. *Méthodes de correction basées sur le FWER*

Une première catégorie de méthodes de correction s'attache à contrôler le FWER directement (1,4). Cette catégorie de méthode est reconnue comme la plus populaire parmi les études de santé, représentant 53.4% des articles identifiés par Glickman, Rao et Schultz (2). Le principe général de ces méthodes consiste à appliquer un seuil de significativité "strict" sur les tests individuels pour ramener le FWER sous le seuil consenti α (1,2).

2.1 *Méthode de Bonferroni*

Il s'agit de la méthode au principe le plus basique. Le principe de la méthode de Bonferroni est comme suit (1):

$$Soient\ p_1,\dots,p_n\ les\ p-valeurs\ associées\ à\ n\ tests.$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ aux\ p-valeurs\ vérifiant:\ p \leq \frac{\alpha}{n}\ .$$

Cette méthode, comme toutes les méthodes basées sur le FWER, s'applique aux tests d'hypothèse nulle globale, c'est-à-dire aux situations où plusieurs tests sont réalisés pour tester un même jeu d'hypothèses, cette hypothèse nulle globale étant rejetée lorsqu'au moins une p-valeur passe le seuil ajusté de significativité (2,6). Glickman et al. précisent ainsi que cette méthode indique le rejet de l'hypothèse nulle globale mais ne donne aucune indication sur l'identité des p-valeurs individuelles significatives (2). Elle permet cependant de connaître le nombre minimal de p-valeurs significatives sans pouvoir les identifier (6). L'application de cette méthode ne requiert aucune hypothèse sur la structure de dépendance (1).

Cette méthode est reconnue pour être particulièrement conservative, c'est-à-dire qu'elle provoque une forte perte de puissance (2,3). Il s'agit de la méthode commune basée sur le FWER ou le FDR la plus conservative (1). Cependant, cette perte de puissance est

minimisée lorsque l'*effect size* est grand, la plupart des p-valeurs passant le seuil de 5% étant suffisamment petites pour passer le seuil de significativité ajusté lorsque la différence attendue entre les groupes est grande (6). Bien que les p-valeurs individuelles soient testées sur un seuil ajusté, le test de l'hypothèse nulle globale est bien réalisé à un niveau d'erreur de 5% (2).

2.2 *Méthode de Šidák*

Cette méthode applique un seuil de significativité progressif, dérivée d'une autre méthode dite de Šidák avec seuil statique (1):

$$Soient\ p_1,\dots,p_n\ les\ p-valeurs\ associées\ \text{à}\ n\ tests.$$

$$Trier\ p_1,\dots,p_n\ par\ ordre\ croissant,\ numéroter\ p_{(1)}<\dots<p_{(n)}$$

$$Identifier\ k\ l'indice\ de\ la\ première\ p-valeur\ ne\ vérifiant\ pas:\ p_{(k)}\leq(1-\alpha)^{\frac{1}{n-k+1}}.$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ \text{à}\ p_{(1)},\dots,p_{(k-1)}$$

Cette méthode s'applique aux mêmes situations que la méthode de Bonferroni, mais requiert l'existence d'une structure de dépendance particulière dite positive orthant (1). Sous ces conditions (dépendance positive orthant ou indépendance), la méthode de Šidák est réputée moins conservative que la méthode de Bonferroni (7).

Contrairement à la méthode de Bonferroni et au même titre que toutes les méthodes à seuil progressif, cette méthode ne permet pas le calcul d'un intervalle de confiance d'inférence simultanée, qui peut s'avérer être un outil utile (7). De plus, cette méthode permet également l'identification des p-valeurs individuelles à considérer significatives (1,7).

2.3 *Méthode d'Hölm*

La méthode d'Hölm ou Bonferroni-Hölm est une méthode à seuil progressif dérivée de la méthode de Bonferroni (7):

$$Soient\ p_1,\dots,p_n\ les\ p-valeurs\ associées\ \text{à}\ n\ tests.$$

$$Trier\ p_1,\dots,p_n\ par\ ordre\ croissant,\ numéroter\ p_{(1)}<\dots<p_{(n)}$$

$$Identifier\ k\ l'indice\ de\ la\ première\ p-valeur\ ne\ vérifiant\ pas:\ p_{(k)}\leq\frac{\alpha}{n-k+1}.$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ \text{à}\ p_{(1)},\ldots,p_{(k-1)}$$

Étant dérivée de la méthode de Bonferroni, la méthode d'Hölm s'applique sous les mêmes conditions, et ne requiert donc pas l'indépendance ni une structure de dépendance particulière (7). Cette méthode est considérée comme une amélioration de la procédure de Bonferroni du fait d'un petit gain en puissance sans conditions de dépendance supplémentaire (1,8).

Cette méthode présente également l'avantage d'identification des p-valeurs individuelles à rejeter et ne permet pas le calcul de l'intervalle de confiance d'inférence simultanée, au même titre que la méthode de Šidák (7).

2.4 *Méthode d'Hochberg*

Cette méthode est une méthode à seuil progressif développée par Y. Hochberg en 1988, dérivée de la procédure d'Hölm (1):

$$Soient\ p_1,\ldots,p_n\ les\ p-valeurs\ associées\ \text{à}\ n\ tests.$$

$$Trier\ p_1,\ldots,p_n\ par\ ordre\ croissant,\ numéroter\ p_{(1)}<\ldots<p_{(n)}$$

$$Identifier\ k\ l'indice\ de\ la\ dernière\ p-valeur\ vérifiant:\ p_{(k)}\leq\frac{\alpha}{n-k+1}\ .$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ \text{à}\ p_{(1)},\ldots,p_{(k)}$$

Cette méthode se base donc sur une autre utilisation des seuils d'Hölm dans le but de produire une procédure moins conservative que la procédure d'Hölm (1,8). Son application requiert cependant des conditions de dépendance plus restrictives, proches de l'indépendance (7).

Tout comme les deux méthodes précédentes, la méthode d'Hochberg ne permet pas le calcul d'intervalles de confiance mais permet l'identification des p-valeurs individuelles significatives (7).

3. *Méthodes de correction basées sur le FDR*

Face aux problèmes de conservativité des méthodes de correction basées sur le FWER, Y. Benjamini et Y. Hochberg ont proposé en 1995 une méthode basée sur une autre mesure de l'erreur de première espèce dans les comparaisons multiples: le *False Discovery Rate* ou FDR (3). Cette mesure peut être décrite comme suit:

$$FDR = E[FPR]$$

Ainsi, le FDR peut être décrit comme l'espérance de la proportion de fausses découvertes. Contrôler le FDR sous un niveau de 5% revient ainsi à assurer qu'en moyenne, au plus 5% des hypothèses nulles rejetées le sont à tort (3).

Cette autre mesure de l'erreur de première espèce, bien que n'assurant pas par son contrôle un contrôle du FWER (3), peut être considérée comme la mesure la plus pertinente dans plusieurs situations où le contrôle du FWER n'est pas strictement nécessaire, ou suffisante quand il est trop pénalisant (2,3). De plus, Y. Benjamini et D. Yekutieli ont mis en avant le fait que le contrôle de cette mesure dans plusieurs familles de tests séparément garantit son contrôle sur l'ensemble de ces familles combinées, ce qui n'est pas le cas pour le FWER (9).

3.1 *Méthode de Benjamini-Hochberg*

Il s'agit d'une méthode à seuil progressif basée sur le FDR, proposée par Y. Benjamini et Y. Hochberg en 1995 (3):

$$Soient\ p_1,\dots,p_n\ les\ p-valeurs\ associées\ \text{à}\ n\ tests.$$

$$Trier\ p_1,\dots,p_n\ par\ ordre\ croissant,\ numéroter\ p_{(1)}<\dots<p_{(n)}$$

$$Identifier\ k\ l'indice\ de\ la\ dernière\ p-valeur\ vérifiant:\ p_{(k)}\leq\frac{k\alpha}{n}\ .$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ \text{à}\ p_{(1)},\dots,p_{(k)}$$

Cette méthode contrôle le FDR mais ne contrôle le FWER au sens faible, c'est-à-dire dans la situation théorique où toutes les hypothèses nulles sont vraies (3). Elle est également considérée comme applicable à tous les problèmes de comparaison multiple (et non pas seulement aux problèmes d'hypothèse nulle globale) sous condition de dépendance strictes proches de l'indépendance (1,3)**.**

La procédure de Benjamini & Hochberg est reconnue comme uniformément plus puissante et versatile que la totalité des méthodes communes basées sur le FWER (2,3) par une marge relativement importante (1,3)**.** Cette propriété vaut à cette méthode une indication particulière lorsque l'erreur de deuxième espèce est considérée comme plus problématique

que l'erreur de première espèce ou lorsque la connaissance de la proportion estimée de fausses découvertes est jugée utile (2,3).

2.2 *Méthode de Benjamini-Liu*

Cette méthode proposée par Y. Benjamini et W. Liu en 1998 propose une alternative à la méthode de Benjamini-Hochberg sous condition d'indépendance (1):

$$Soient\ p_1,\dots,p_n\ les\ p-valeurs\ associées\ \text{à}\ n\ tests.$$

$$Trier\ p_1,\dots,p_n\ par\ ordre\ croissant,\ numéroter\ p_{(1)}<\dots<p_{(n)}$$

$$Identifier\ k\ l'indice\ de\ la\ première\ p-valeur\ ne\ vérifiant\ pas:$$

$$p_{(k)} \leq 1-(1-min(1,\frac{n}{n-j+1}\alpha))$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ \text{à}\ p_{(1)},\dots,p_{(k-1)}$$

Cette procédure est conçue dans une approche complémentaire à la procédure de Benjamini-Hochberg, plus puissante que cette dernière lorsque le nombre d'hypothèses testées est petit et que l'on s'attend à ce qu'une large proportion des hypothèses nulles soient fausses (1). En pratique, cette situation est plus rare que les situations communes dites de sporadicité (1).

2.3 *Méthode de Benjamini-Yekutieli*

Il s'agit d'une méthode proposée comme une extension des méthodes contrôlant le FDR sous dépendance arbitraire (1,9):

$$Soient\ p_1,\dots,p_n\ les\ p-valeurs\ associées\ \text{à}\ n\ tests.$$

$$Trier\ p_1,\dots,p_n\ par\ ordre\ croissant,\ numéroter\ p_{(1)}<\dots<p_{(n)}$$

$$Identifier\ k\ l'indice\ de\ la\ dernière\ p-valeur\ vérifiant:\ p_{(k)} \leq \frac{k}{n\times c(n)}\alpha\ .^{(a)}$$

$$Rejeter\ les\ hypothèses\ nulles\ associées\ \text{à}\ p_{(1)},\dots,p_{(k)}$$

Cette méthode est considérée comme uniformément moins conservative que les méthodes corrigeant le FWER, tout en restant plus conservative que la méthode de Benjamini-Hochberg (1,9). Cependant, l'absence de conditions de dépendance précises la rendent applicable à plus de situations pratiques (9).

4. *Comparaison des performances des méthodes*

Parmi les études prises en compte, 2 proposent une comparaison quantitative des méthodes en se basant sur des simulations. Farcomeni (1) propose une comparaison des diverses méthodes présentées dans deux situations: une première simulation de 100000 tests dont 90000 avec une une hypothèse nulle vraie et une seconde simulation sur 100 tests dont 90 avec une hypothèse nulle vraie, présentées comme représentatives de situations de test réalistes par l'auteur. Les résultats de ces simulations sont décrits par les *Tables 2* et *3*.

TABLE 2: Comparatif des méthodes pour n= 100000 tests et 90000 hypothèses nulles vraies, à alpha=0.05 (1)

Méthode	Nombre de FP moyen	Nombre de FN moyen	FWE	FDR
Sans correction	4497.59	3334.47	1	0.04
Bonferroni	0.047	9089.72	0.045	0
Sidak	0.048	9081.76	0.046	0
Holm	0.047	9087.17	0.045	0
Hochberg	0.047	9087.17	0.046	0
BH	203.92	5669.27	1	0.045
BL	0.049	9079.70	0.047	0
BY	10.15	7291.05	1	0.037

TABLE 3: Comparatif des méthodes pour n= 100 tests et 90 hypothèses nulles vraies, à alpha=0.05 (1)

Méthode	Nombre de FP moyen	Nombre de FN moyen	FWE	FDR
Sans correction	4.49	3.27	0.989	0.386
Bonferroni	0.044	6.60	0.044	0.013
Sidak	0.050	6.48	0.050	0.014
Holm	0.048	6.48	0.048	0.015
Hochberg	0.048	6.52	0.048	0.014
BH	0.264	5.53	0.221	0.044
BL	0.042	6.44	0.042	0.010
BY	0.030	6.65	0.030	0.008

L'auteur s'appuie sur ces simulations pour mettre en avant les contrastes entre les méthodes (1):

- Les méthodes contrôlant la même mesure de l'erreur de première espèce (FWER ou FDR) présentent des performances assez similaires, à l'exception de la méthode de Benjamini-Liu.
- Pour 100000 tests, les méthodes contrôlant le FWER tendent à ne pas rejeter 90% des hypothèses nulles fausses, contre environ 50% pour les méthodes basées sur le FDR. Cet écart rend compte du gain de puissance offert par ces dernières.

- L'écart de puissance et de contrôle du nombre de faux positifs est croissant avec le nombre de tests. En particulier, le gain présumé en puissance entre les méthodes est négligeable pour un faible nombre de tests dans une configuration sporadique, mais ce gain augmente avec le nombre de tests et la sporadicité.

Ces conclusions sont appuyées par les simulations de Benjamini et Hochberg (3), qui confirment que la méthode proposée est uniformément plus puissante que les méthodes d'Hochberg et Bonferroni sur toutes les configurations de sporadicité et nombres de tests Ces simulations mettent également en avant les situations dans lesquelles ces 3 méthodes de correction sont les plus puissantes et moins puissantes:

- La puissance des méthodes de correction est minimale lorsque le plus grand nombre d' hypothèses nulles sont fausses.
- La puissance des méthodes de correction est maximale lorsque le plus grand nombre d' hypothèses nulles sont vraies.

En pratique, ces propriétés se traduisent dans des études épidémiologiques par un gain significatif de puissance lorsqu'une large proportion des p-valeurs sont très basses, et par des comportements similaires lorsque la distribution des p-valeurs est uniforme (2):

TABLE 4: Comparatif entre les méthodes de Bonferroni et BH selon la distribution des p-valeurs, alpha=0.05 (2)

Etude	Distribution des p-valeurs	n	Sans Ajustement	Bonferroni	BH
Marx et al.	Uniforme	28	2	0	0
Bombardier et al.	Basse	55	27	0	26

Ces situations correspondent, selon l'auteur, aux configurations théoriques lorsqu'une association existe effectivement (forte proportion de p-valeurs basses) et lorsqu'il n'existe pas d'association (distribution uniforme).

5. *Autres considérations méthodologiques*

Face au problèmes posés par les corrections, qui imposent toutes un contrôle moindre du FWER ou une perte en puissance, plusieurs auteurs ont recommandé une approche plus critique du problème des comparaisons multiples (4,5,8), certains pointant du doigt un

manque perçu de discussion sur le choix de la méthode et les justifications de la correction dans les articles scientifiques du domaine de la santé (4).

Une première approche communément recommandée dans la gestion du problème des comparaisons multiples au niveau de la conception des études consiste à limiter le problème en limitant au maximum le nombre de tests effectués (8), ce qui passe par une restriction des objectifs de l'étude à des questions précises biologiquement pertinentes (4). Cette proposition rejoint l'appel de S. N. Goodman à une science guidée par les preuves (5), arguant que le niveau de preuve externe sur une association (c'est-à-dire les preuves ou indications prédatant l'étude) doivent guider la détermination des objectifs d'une étude et qu'une découverte statistique, même solide, se doit d'être mise en perspective de sa vraisemblance biologique et des preuves antérieures, impliquant qu'une évaluation séparée des associations est recommandable et que l'utilisation de critères externes (*Bradford Hill*…) est nécessaire à une interprétation scientifiquement saine des p-valeurs. Une telle approche est également considérée comme associée à une capacité à recueillir des données de meilleure qualité sur les expositions, à l'inverse d'études évaluant de multiples associations et disposant ainsi de données de moins bonne qualité sur les multiples expositions d'intérêt (4).

Dans le cadre d'études réalisant malgré tout de multiples comparaisons évaluant plusieurs associations, une évaluation de la qualité de la preuve au-delà de la significativité statistique sur l'association d'intérêt est recommandée. En effet, la capacité d'une étude à retrouver des associations exposition/événement bien connues scientifiquement peut apporter une garantie sur la qualité des données, puisque l'aboutissement dans l'analyse à une conclusion contraire à un fait scientifique établi serait une indication du contraire (4).

De plus, plusieurs auteurs supportent également la mise en perspective des choix d'action sur les comparaisons multiples avec les objectifs et la portée de l'étude, encourageant une dichotomie entre les études à vocation exploratoires et confirmatoires (8)en particulier en épidémiologie(2).

DISCUSSION

1. *Faut-il toujours corriger ?*

Le choix d'appliquer une correction, tout comme celui de la correction à appliquer, constitue un compromis entre le contrôle du risque d'erreur de première espèce et la puissance. Le critère principal déterminant le choix d'appliquer ou non une correction est donc le rapport entre l'importance d'évitement des erreurs de première et deuxième espèce dans le contexte de l'étude.

Ainsi, il semble recommandable, dans le cadre d'une étude cherchant impérativement à éviter les erreurs de deuxième espèce, de se dispenser d'une correction conformément aux arguments avancés par plusieurs auteurs (2,4,8)en relation avec la perte de puissance engendrée par les corrections. De même, les familles de tests dans lesquels l'évitement des erreurs de première espèce n'est pas un impératif sont également des cadres dans lesquels la correction n'est pas indispensable, en particulier pour les données descriptives des études dans lesquelles la déclaration à tort de différences peut avoir des conséquences plus limitées que sur des analyses à vocation étiologiques (10).

Par ailleurs, une autre dichotomie sur le choix d'appliquer une correction pourrait être faite sur la nature de l'étude et le nombre de tests effectués. Ainsi, bien qu'une correction soit recommandable dans le cadre d'une étude confirmatoire réalisant un petit nombre de tests, elle peut s'avérer dispensable dans le cadre d'études exploratoires posant des hypothèses *post-hoc*, tout en restant utile dans ces études si le nombre de tests venait à entraver l'interprétation des résultats (par exemple lorsque des milliers de tests seraient déclarés significatifs sans correction).

Finalement, les cas dans lesquels l'évitement de l'erreur de première espèce est impératif apparaissent comme les cas dans lesquels les corrections sont le plus fortement recommandées. En particulier, les familles des tests sur une hypothèse nulle globale (par exemple les analyses en sous-groupes), qui sont les cas pour lesquels les corrections basées sur le FWER ont été développées, sont des situations dans laquelle l'application d'une correction semble obligatoire pour l'intégrité des résultats, car la présence d'une seule erreur

de première espèce entraînerait le rejet de l'hypothèse nulle globale, invalidant ainsi toute la conclusion sur cette famille.

2. *Recommandations sur le choix de la méthode de correction*

Le choix de la méthode de correction est également subordonné à la balance d'importance entre les différents types d'erreurs statistiques et aux objectifs de l'étude. On peut ainsi distinguer différents cadres d'application des méthodes.

Dans le cadre d'études confirmatoires, l'application des méthodes basées sur le FWER sur les familles de tests sur lesquelles elles sont applicables est recommandée pour les raisons évoquées précédemment (6). On recommandera ainsi l'utilisation de méthodes plus contemporaines et moins conservatives comme la méthode d'Hochberg ou la méthode d'Hölm dans le cas de statistiques de test non indépendantes.

Pour des études à vocation exploratoire, il apparaît préférable de s'orienter vers des corrections basées sur le FDR. En effet, l'importance de la puissance dans ce type d'étude ainsi que le plus grand intérêt sémantique du contrôle de la proportion d'erreurs par rapport à celui du FWER dans ce cadre font que des méthodes comme celles de Benjamini-Hochberg ou Benjamini-Yekutieli (pour les statistiques de tests dépendantes) sont plus indiquées (2,3). Cette recommandation est extensible aux études omiques dont la nature et le grand nombre de tests rendent également les méthodes basées sur le FDR plus pertinentes.

Pour les études dans lesquelles la non-détection d'une association est délétère, l'utilisation de méthodes basées sur le FDR semble également plus pertinente du fait de la meilleure puissance qu'elles permettent. Ainsi, on peut recommander l'application de ces corrections dans les études de pharmacoépidémiologie, dans lesquelles une perte de puissance trop importante pourrait mener à la non-détection de potentiels effets indésirables.

3. *Limites et approfondissements*

Bien que la revue effectuée et les recommandations formulées semblent couvrir l'ensemble des cas d'application en épidémiologie, plusieurs approfondissements supplémentaires visant à affiner ces recommandations peuvent être proposés.

3.1 *Autres méthodes*

Malgré l'exclusion des méthodes de correction visant à contrôler d'autres mesures de l'erreur de première espèce (FDX, FCR) du cadre de cette revue, plusieurs sources font référence à ces méthodes plus contemporaines (1,7). Étendre cette revue vers une prise en compte plus exhaustive de la littérature méthodologique pourrait ainsi permettre de formuler des recommandations plus détaillées et adaptées aux besoins des épidémiologistes. Une extension incluant les ressources de la littérature au-delà du domaine biomédical pourrait également permettre de mieux identifier de potentielles approches et méthodes émergentes.

3.2 *Contradictions et utilisation de la méthode de Bonferroni*

Durant l'écriture de cette revue, plusieurs contradictions ont été relevées entre les différentes ressources de la littérature concernant le cadre d'utilisation des méthodes assimilées à la méthode de Bonferroni. En effet, alors que certains auteurs (2,3)mettent en avant la restriction de ces méthodes aux tests d'hypothèse nulle globale, d'autres ne mentionnent pas cette différence fondamentale avec les méthodes basées sur le FDR dans leurs comparaisons (1,6). De ce fait, l'interprétation de cette revue de la littérature est conditionnée par la reconnaissance de cette différence lorsqu'elle n'est pas évoquée, ce qui n'est pas une garantie. Cependant, cette situation met également en avant l'importance de l'établissement de lignes de conduites communément reconnues face aux potentielles difficultées rencontrées par des chercheurs dans l'identification des enjeux de chaque méthode de correction.

CONCLUSION

Les comparaisons multiples posent des problèmes majeurs dans le cadre des études épidémiologiques, tant par le risque d'erreur de première espèce supplémentaire introduit que par la perte de puissance engendrée par les méthodes permettant sa correction.

A partir d'une revue non exhaustive de la littérature méthodologique, nous avons pu identifier les principales méthodes de correction permettant de contrôler l'inflation du risque α et comparer leurs cadres d'utilisation et leurs performances. La revue a également permis l'identification de lignes de conduite et problématiques supplémentaires à prendre en compte dans le cadre des comparaisons multiples en épidémiologie.

Ainsi, nous avons pu formuler plusieurs recommandations à destination des épidémiologistes concernant la correction de l'inflation du risque α en épidémiologie, principalement orientées par deux axes principaux: les objectifs de l'étude et son échelle. Cependant, la non-exhaustivité de la revue et sa restriction aux deux principales mesures de l'erreur de première espèce rendent de plus amples approfondissements sur ces plans souhaitables, non seulement car la présence de méthodes alternatives plus contemporaines (basées sur le FDX, FCR) peuvent amener à faire évoluer et préciser les recommandations formulées, mais aussi car la présence d'un consensus, même partiel, sur le problème des comparaisons multiples ne semble pas encore acquise.

ANNEXE 1: EQUATIONS DE RECHERCHE

Equation	**Bonferroni and adjustement**	**Bonferroni and adjustement and methods**	**Bonferroni and FWER**	**"Data Interpretation, Statistical" [MAJR] AND Bonferroni**	**Bonferroni and adjustment and method [MESH]**	**“False Positive Reactions"[MeSH] AND 'Bonferroni'**
Résultats dans les revues	32	25	1	9	4	5
Equation	**"Data Interpretation, Statistical" [MAJR] AND 'bonferroni'**	**"Data Interpretation, Statistical" [MAJR] AND 'multiple comparison'**	**"False Positive Reactions" [MeSH] AND 'FDR'**	**"False Positive Reactions" [MeSH] AND 'FWER'**	**"False Positive Reactions" [MeSH] AND 'multiple comparison'**	**"False Negative Reactions"[MeSH] AND 'p-value correction'**
Résultats dans les revues	9	37	3	2	11	1
Equation	**"Data Interpretation, Statistical" [MeSH] AND 'p-value correction'**	**"Data Interpretation, Statistical" [MeSH] AND 'family-wise error'**	**"False Positive Reactions" [MeSH] AND 'multiple testing' AND 'methodology' AND 'p-value'**	**"False Positive Reactions" [MeSH] AND 'correction' AND 'methodology' AND 'p-value'**		
Résultats dans les revues	6	6	3	2		

ANNEXE 2: GRILLE D'EXTRACTION

Titre / Auteurs / Dates	
Méthodes considérées	
Méthodologie (comment la méthode est étudiée)	
Contexte d'étude de la méthode	
Discussions sur la méthode	
Valeurs à retenir/ indicateurs	
(Remarques)	
Limites	
Justification de l'inclusion	
Conclusion	

BIBLIOGRAPHIE

1. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res. août 2008;17(4):347‑88.
2. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol. août 2014;67(8):850‑7.
3. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289‑300.
4. Savitz DA, Olshan AF. Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data. Am J Epidemiol. 1 nov 1995;142(9):904‑8.
5. Goodman SN. Multiple comparisons, explained. Am J Epidemiol. 1 mai 1998;147(9):807‑12; discussion 815.
6. VanderWeele TJ, Mathur MB. SOME DESIRABLE PROPERTIES OF THE BONFERRONI CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD? Am J Epidemiol. mars 2019;188(3):617‑8.
7. Tamhane AC, Gou J. Advances in p-Value Based Multiple Test Procedures. J Biopharm Stat. 2018;28(1):10‑27.
8. Streiner DL, Norman GR. Correction for multiple testing: is there a resolution? Chest. juill 2011;140(1):16‑8.
9. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann Stat. 2001;29(4):1165‑88.
10. Savitz DA, Olshan AF. Describing Data Requires No Adjustment for Multiple Comparisons: A Reply from Savitz and Olshan. Am J Epidemiol. 1 mai 1998;147(9):813‑4.