Le taux de fausses découvertes dans la littérature sur la préférence de lieu conditionné induite par la nicotine chez la souris

Auteur : Léonard, François
Promoteur(s) : Tirelli, Ezio
Faculté : þÿ F a c u l t é d e P s y c h o l o g i e , L o g o p é d i e e t S c i e n c e s d e l E d u c a t i o n
Diplôme : Master en sciences psychologiques, à finalité spécialisée en neuroscience cognitive et comportement
Année académique : 2019-2020
URI/URL : http://hdl.handle.net/2268.2/10528

LIÈGE université
Psychologie, Logopédie
& Sciences de l'Éducation

Le taux de fausses découvertes dans la littérature sur la préférence de lieu conditionné induite par la nicotine chez la souris

François Léonard

Promoteur : Pr. Ezio Tirelli

Lecteurs : Pr. Étienne Quertemont

&

Pr. Christian Monseur

Année Académique 2019-2020

Mémoire présenté en vue de l'obtention du diplôme de master en psychologie à finalité spécialisée en neuroscience cognitive et comportementale

« Assurons-nous bien des faits, avant de nous inquiéter de la cause. »

Fontenelle

« The evidence is not as strong as it seems to be. »

Ioannidis

Remerciements :

La réalisation de ce mémoire de méta-recherche n'aurait pas été possible sans le soutien et les encouragements du Professeur Ezio Tirelli.

Je remercie aussi Victoria Leclercq et Olivier Bruyère pour les méthodes de sélection et d'extraction apprises en stage.

Je n'aurais jamais pu comprendre la subtilité de la démonstration des formules de TRP et FDR sans l'aide rapide et personnelle du professeur David Colqhoun.

Je tiens aussi à remercier les deux lecteurs de ce mémoire, les professeurs Étienne Quertemont et Christian Monseur, pour leur lecture attentive.

Table des matières

1 Introduction théorique et concepts de base

La science serait en crise. On parle beaucoup de la crise de la reproductibilité, mais il y a de nombreux autres problèmes. Certains concernent les chercheurs eux-mêmes, comme les préjugés conscients et inconscients, les faiblesses théoriques et techniques, les mauvaises habitudes, les pratiques discutables et les fraudes caractérisées. D'autres concernent l'entourage immédiat des chercheurs ou leurs conditions de travail, comme les évaluations centrées sur des critères plus ou moins valides et objectifs (e.g. facteur d'impact, nombre de publications…). D'autres encore sont plus distants, tout en restant dans la sphère scientifique, comme les exigences des revues ou des bailleurs de fonds.

Mais la psychologie nous apprend aussi à étudier les choses autrement. Il ne suffit pas de se mettre en question ou de négocier de meilleures conditions de travail. Il faut aussi regarder plus loin, pour se situer dans un monde qui dépasse largement celui des savants, mesurer précisément ses propres défauts et leurs conséquences, pour ne pas gaspiller dans la lutte contre un défaut véniel ou imaginaire une énergie qui serait mieux investie dans la lutte contre un péché capital, et enfin s'interroger sur ses stratégies et ses investissements, pour ne pas perdre son âme dans des luttes sans espoir.

Mais la science n'a-t-elle pas toujours été en crise ? Parfois elle s'est opposée au pouvoir (comme Gallilée), parfois elle en était trop proche (comme Lysenko). Aujourd'hui l'attention se concentre plutôt sur des cas individuels, comme celui de Stapel (Stapel, 2016), ou sur des problèmes « institutionnels » comme les critères d'évaluation « quantomaniaques ».

Alors la science ne serait-elle pas, comme toujours, en train de rassembler de nouvelles forces contre de nouveaux démons, ou contre de vieux démons que l'on tolérait autrefois et qui deviennent insupportables aujourd'hui grâce aux victoires engrangées précédemment ? Certains d'entre eux, en tous cas, deviennent aujourd'hui vulnérables grâce aux progrès de la connaissance, grâce à de nouvelles collaborations internationales entre individus par l'Internet, et peut-être grâce à la levée du secret militaire (McGrayne, 2011) qui a trop longtemps couvert quelques remarquables progrès statistiques.

Ioannidis (2005) prétend que la plupart des résultats scientifiques publiés seraient faux. Mais qu'en est-il ? Pour en savoir plus, nous tenterons d'évaluer le taux de fausses découvertes dans un domaine particulier de la psychopharmacologie expérimentale, la préférence de lieu conditionné induite par la nicotine chez la souris (PLC_{NIC}).

Nous devrons cependant d'abord aborder quelques bases statistiques et méthodologiques pour situer

le problème et pour comprendre les réponses de Ioannidis et leurs limites. Nous proposerons ensuite un complément qui nous semble raisonnable.

1.1 La question de l'irreproductibilité en sciences

La démarche scientifique se base sur les six étapes du modèle hypothético-déductif classique (Figure 1). Idéalement, on commence par se poser une question et consulter la littérature pour générer des hypothèses de travail. Après quoi on développe un plan expérimental en s'inspirant des travaux de ses prédécesseurs. Une fois le plan expérimental développé, on le met en œuvre et on collecte les données. Ensuite, les données récoltées sont analysées par diverses méthodes statistiques et on utilise les résultats obtenus pour mettre à l'épreuve ses hypothèses. Enfin, on partage sa méthode et ses conclusions pour alimenter la réflexion des autres chercheurs.

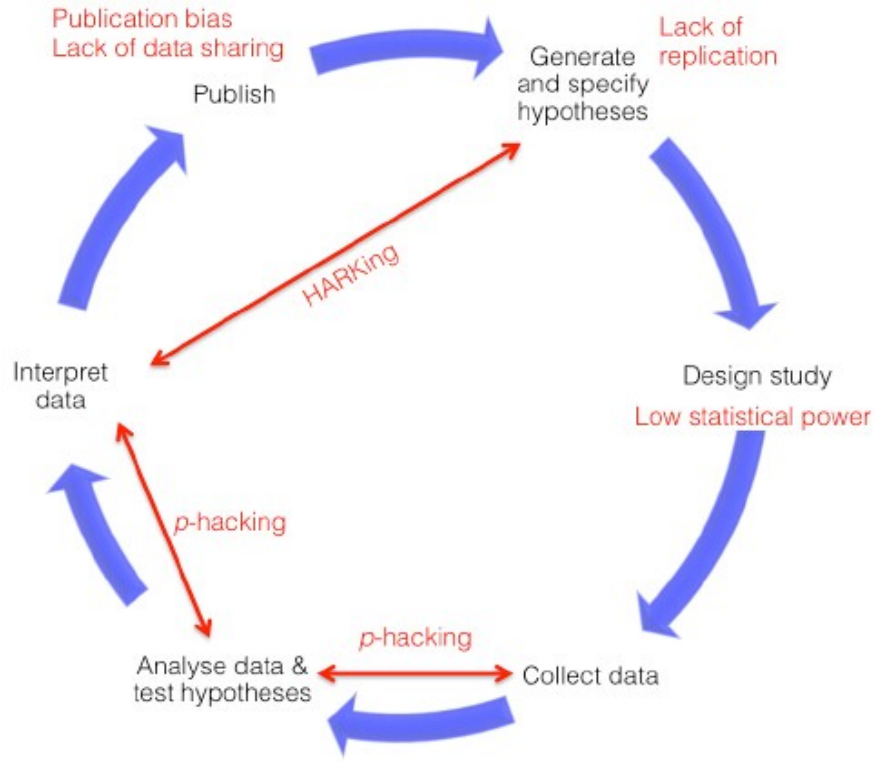

Figure 1: Cycle hypothético-déductif classique de la recherche expérimentale et pratiques de recherche douteuses qui menacent la qualité de la science. Repris de Chambers et al. (2014)

Dans le monde réel, les chercheurs s'écartent parfois de l'idéal et prennent des raccourcis dangereux qui menacent tout le modèle hypothético-deductif.

Bishop (2019) met en évidence quatre de ces raccourcis : le biais de publication, le manque de

puissance, le p-hacking, et le HARKing. Ce sont pour elle les quatre cavaliers de l'apocalypse de l'irreproductibilité en science (Figure 2).

Figure 2: Les quatre cavaliers de l'apocalypse de l'irreproductibilité en sciences. Peinture de Viktor Vasnetsov de 1887. Repris de Bishop (2019)

Le **biais de publication** comprend la tendance à ne publier que les résultats « sensationnels » ou les résultats significatifs qui valident nos hypothèses. Il produit une sur-représentation de ces résultats dans la littérature, ce qui pose de nombreux problèmes.

D'abord, il donne une idée fausse de la réalité en présentant des effets ou des différences exagérés, ou moins pertinents qu'ils ne paraissent.

Il tend aussi à faire disparaître les résultats non significatifs. Comme les journaux préfèrent publier des résultats significatifs, les auteurs hésitent à publier leurs résultats non-significatifs, car la rédaction d'un article représente une somme de travail que l'on préfère s'épargner si les chances de publication sont trop minces. Rosenthal (1979) nomme ce phénomène le « *file drawer problem* ». Mais la stratégie rationnelle du chercheur qui épargne son travail pose à son tour de sérieux problèmes.

Si les études qui ont échoué ne sont pas publiées, elles risquent d'être retentées régulièrement et indépendamment plusieurs fois, sans que personne ne profite de l'expérience des précédents.

Il est possible aussi que plusieurs études individuelles échouent parce que les groupes mobilisés sont trop petits, mais qu'elles soient assez similaires pour que l'on puisse théoriquement les combiner pour produire des groupes « virtuels » capables de mobiliser une plus grande puissance et de détecter des effets plus modestes. Ces combinaisons resteront cependant théoriques si personne ne connaît l'existence de ces études combinables.

Des patients peuvent être privés d'un traitement efficace au profit d'un autre qui n'a fait ses preuves qu'une seule fois dans une étude que personne ne réussit à répliquer (Bishop 2019).

Des croyances absurdes peuvent se répandre dans la communauté scientifique à force de répétition, sans que les faits sur lesquelles elles reposent soient jamais remis en question (Ioannidis, 2008; Young et al., 2008).

Enfin, en favorisant la publication des résultats significatifs, même si leur taille ou leur pertinence sont évidemment surestimées, ou s'ils ont été obtenus par des pratiques discutables, le biais de publication favorise la publication des études de mauvaise qualité au détriment des autres, et récompense les mauvaises pratiques destinées à produire rapidement des résultats significatifs.

Le **p-hacking** commence dès qu'on intervient sur les données ou les analyses en vue d'obtenir une p-valeur significative. Motulsky (2015) énumère neuf pratiques de p-hacking : augmenter la taille de l'échantillon jusqu'à obtenir une valeur significative, analyser seulement une partie des données, ajouter des variables dans le modèle, ajuster les données (e.g. diviser par le poids corporel), transformer les données, enlever les valeurs douteuses, utiliser un autre groupe comme contrôle, changer de variable dépendante, et utiliser un autre test statistique.

On peut y ajouter le « *cherry-picking* » qui consiste à ne rapporter dans l'article que les analyses et les variables qui ont donné des résultats significatifs. Cela évite d'utiliser des seuils de significativité corrigés et permet donc d'augmenter les chances d'obtenir des résultats significatifs.

Le **HARKing** (*Hypothesizing After Results are Know*), consiste à présenter une analyse « post hoc » comme s'il s'agissait d'une hypothèse « *a priori* » (Kerr, 1998). Cela revient aussi à présenter une recherche exploratoire comme une recherche confirmatoire. Les recherches exploratoires ne sont évidemment pas interdites, mais elles doivent répondre à leurs propres règles (Wagenmakers et al., 2012).

La **puissance statistique** se définit comme la probabilité qu'un test statistique rejette l'hypothèse nulle (H_0) quand l'hypothèse alternative (H_1) est vraie. C'est aussi par définition la probabilité de ne

pas commettre une erreur de seconde espèce. Le manque de puissance statistique a pour conséquence secondaire la production d'un grand nombre de résultats peu fiables ou même faux.

Le manque de puissance statistique combiné à la publication préférentielle des résultats significatifs pourrait expliquer au moins en partie la parution d'articles scientifiques de moins en moins fiables, sans qu'il soit nécessaire de postuler la moindre mauvaise intention de la part de quiconque.

1.2 La significativité statistique

La notion de puissance et les problèmes que nous venons d'évoquer reposent sur la notion de significativité statistique. Les statisticiens et les experts en méthodologie répètent que la p-valeur est souvent mal comprise (Cassidy et al., 2019; Gigerenzer, 2004; Gigerenzer et al., 2004; Goodman, 2008). Nous trouvons essentiel de comprendre le raisonnement statistique dont elle dépend et qui s'inscrit dans le modèle « *Null Hypothesis Significance Testing* » (NHST).

Le risque alpha : décider d'après les observations et les lois du hasard seulement

Le NHST prétend donner un moyen objectif de prendre la « meilleure » décision possible à propos d'une hypothèse (H_1) en fonction des résultats d'une expérience et des lois du hasard seulement. Le raisonnement s'effectue en plusieurs étapes.

D'abord, on imagine une expérience capable de montrer que l'hypothèse alternative (H_1) explique mieux la réalité que l'hypothèse selon laquelle toutes les différences seraient due au hasard (H_0). On calcule donc précisément ce qu'il devrait se passer si les phénomènes observés obéissaient simplement aux lois du hasard.

Dans une **deuxième étape,** on fixe un taux d'erreur acceptable. On fixe ainsi un certain risque de se tromper en acceptant indûment l'hypothèse de départ, et l'on définit numériquement ce risque (risque alpha). Le plus souvent on choisit un risque alpha de 5 %. Parfois à cette étape, on fixe un score critique sur la distribution de H_0, correspondant à un résultat suffisamment extrême pour que sa probabilité soit inférieure à 5 % pour un test unilatéral ou à 2,5 % pour un test bilatéral.

À la **troisième étape,** on effectue les mesures.

Ensuite dans une **quatrième étape,** on effectue les statistiques convenues sur les mesures. Et on calcule une p-valeur, qui donne les chances d'obtenir un résultat au moins aussi extrême si H_0 est vraie. Sauf si on a fixé précédemment un score critique. On dit enfin que le résultat est significatif quand la p-valeur obtenue est inférieure au taux d'erreur acceptable, ou quand le score critique est atteint ou dépassé.

Le p-hacking, ou la neutralisation des hypothèses de la significativité statistique

La carrière des auteurs repose souvent sur leurs publications, et les études qui rapportent un effet significatif ont plus de chances d'être publiées. Alors il arrive que certains utilisent des pratiques de recherche discutables pour obtenir ces résultats significatifs dont dépend leur survie académique.

Certains arrondissent leurs p-valeurs pour les abaisser un peu (p=0,054 devient p=0,05). D'autres se trompent dans les données qu'ils fournissent au calculateur et rapportent un résultat qui ne correspond pas à la statistique ou aux degrés de liberté annoncés (Nuijten et al., 2016).

D'autres encore multiplient les études de mauvaise qualité, rapides et peu coûteuses, en comptant sur les erreurs d'échantillonnage pour produire un résultat significatif. Peu importe alors si le phénomène qu'ils prétendent avoir découvert n'existe pas, puisqu'avec un alpha de 0,05 ils ont tout de même une chance sur 20 d'obtenir un résultat significatif. Il leur suffit de multiplier les essais jusqu'à ce que la chance leur sourie.

S'ils utilisent très peu de sujets, ils peuvent encore augmenter leurs chances d'obtenir des résultats positifs par le seul fait des erreurs d'échantillonnage.

Le danger est particulièrement grand quand ils multiplient les analyses intermédiaires ou quand ils interrompent l'expérience dès qu'ils ont obtenu des résultats significatifs (Simmons et al., 2011).

En cas de p-hacking, on devrait trouver une proportion anormale d'études ayant obtenu une p-valeur proche de 5 % (Simonsohn et al., 2014).

1.3 Les tailles d’effet

Même sans p-hacking, l’obtention d’un résultat significatif ne suffit pas à garantir la fiabilité des résultats.

Il faut noter que le raisonnement à la base du concept de significativité statistique n’a pas vraiment précisé l’hypothèse à tester. Il suffit pour décider que les résultats soient égaux ou supérieurs à ce que le hasard peut faire 5 fois sur 100. On « accepte » H_1 parce que l’on n’a pas pu rejeter H_0.

Il est raisonnable de chercher la meilleure décision possible en tenant compte seulement des observations et des lois du hasard quand on ne dispose d’aucune autre information. Mais on a souvent une idée de ce que l’on cherche, et ce n’est alors plus du tout raisonnable de faire comme si l’on n’en savait rien. Envisageons donc maintenant les nouvelles possibilités de calcul et de modélisation qui apparaissent quand on se permet quelques hypothèses supplémentaires sur ce que l’on cherche. En postulant les caractéristiques de la courbe de H_1, à savoir un indice de tendance centrale et un indice de dispersion, on peut calculer la taille d’effet du phénomène étudié. Les tailles d’effet expriment la force de l’association entre deux variables ou l’amplitude de la différence entre les observations du groupe contrôle (H_0) et du groupe expérimental (H_1).

Avec ces nouvelles hypothèses sur la distribution de H_1, on peut définir le risque bêta : la probabilité de tolérer H_0 quand H_1 est vrai. Avec les risques alpha et bêta, on peut définir les quatre inférences et erreurs possibles d’un test statistique.

Inférences et erreurs possibles dans le cadre du NHST

Le Tableau 1 nous montre les inférences et erreurs possibles que l’on peut calculer dans le cadre du NHST, si l’on attribue une moyenne et une variance à H_1.

Les colonnes envisagent la présence ou l’absence de l’effet (H_0 fausse ou vraie). Les lignes envisagent les deux décisions possibles : croire ou non en la présence de l’effet (rejet ou tolérance de H_0). Cela définit quatre cases : les vrais positifs et les faux positifs en première ligne, les faux négatifs et les vrais négatifs en deuxième ligne. Les deux lignes du bas montrent comment calculer la puissance et la confiance à partir de comptages. La probabilité associée à chaque case est indiquée dans le tableau.

La puissance est la probabilité qu’un test détecte un effet (rejette H_0) si cet effet existe (si H_1 est vraie).

Le risque de première espèce est la probabilité qu’un test détecte un effet (rejette H_0) qui n’existe pas (si H_1 est fausse).

Le risque de seconde espèce est la probabilité qu'un test ne détecte rien (tolère H_0) quand l'effet existe (si H_1 est vraie).

La confiance est la probabilité qu'un test ne détecte rien (tolère H_0) quand il n'y a rien à détecter (si H_1 est fausse).

Tableau 1 : Les inférences et erreurs possibles dans le cadre du NHST

	Population	
Échantillon	Présence de l'effet H0 fausse	Absence de l'effet H0 vraie
Rejet de H0 (p-valeur < 0,05)	Vrais positifs (VP) Inférence correcte Probabilité : 1 – β Puissance	Faux positifs (FP) Erreur de type I Probabilité : α Risque de première espèce
Non-rejet de H0 (p-valeur ≥ 0,05)	Faux négatifs (FN) Erreur de type II Probabilité : β Risque de seconde espèce	Vrais négatifs (VN) Inférence correcte Probabilité : 1- α Confiance
	Puissance = 1 – β = VP/(VP+FN)	
		Confiance = 1- α = VN/(FP+VN)

On retrouve les cases principales de ce tableau dans la théorie de la décision statistique, dans la théorie diagnostique (Browner & Newman, 1987), dans la théorie de la détection du signal (Witt, 2019) et dans la Figure 3.

Représentation graphique

Les courbes en cloche de la Figure 3 nous montrent à quoi correspondent graphiquement les cases du Tableau 1 quand on dispose de deux distributions bien définies, qui correspondent par exemple aux résultats que l'on peut attendre d'un test réalisé chez une personne saine ou malade, ou encore aux distributions des résultats attendus selon l'une ou l'autre des deux hypothèses.

Dans chacun des six graphiques on trouve deux courbes normales, qui représentent les distributions des résultats sous les hypothèses nulle (à gauche) et alternative (à droite). Les surfaces sous les deux courbes sont divisées en deux par une ligne pointillée verticale qui représente le seuil de rejet. Ce seuil de rejet a été choisi de telle sorte que la surface rouge à sa gauche représente toujours 5 % de la surface comprise sous la courbe de H_0, soit le risque alpha.

En psychologie comme dans d'autres domaines on a pris depuis longtemps l'habitude de considérer qu'un résultat est « statistiquement significatif » si la probabilité d'observer un résultat au moins

aussi extrême sous l'hypothèse nulle est inférieure à 5 % (risque alpha inférieur à 5 %). C'est ainsi qu'on dit parfois simplement qu'un résultat est « statistiquement significatif » alors qu'il faudrait dire « statistiquement significatif pour un seuil de rejet de 5 % ».

Ce seuil de 5 % est un seuil arbitraire qui a été choisi par commodité il y a longtemps et dont la survivance tient sans doute un peu du mystère.

Avec les possibilités informatiques actuelles, quand on connaît le résultat d'une observation statistique, on peut facilement obtenir la probabilité exacte de trouver un résultat au moins aussi extrême ou « étonnant » sous l'hypothèse nulle. C'est la p-valeur. Dans la figure, le seuil de rejet est repéré par la ligne pointillée verticale.

Les six panneaux de la figure représentent cependant des situations bien différentes, avec des valeurs critiques différentes mais qui correspondent toujours à une même probabilité alpha de commettre une erreur de type I si l'hypothèse nulle est vraie (parce que la surface rouge représente toujours 5 % de la surface définie par la courbe H_0).

La puissance est représentée par les surfaces bleues (prolongées sous les surfaces rouges). On voit qu'elles sont définies par le seuil de rejet et par la courbe représentant la distribution des résultats attendus sous l'hypothèse alternative.

Analysons ces différentes situations.

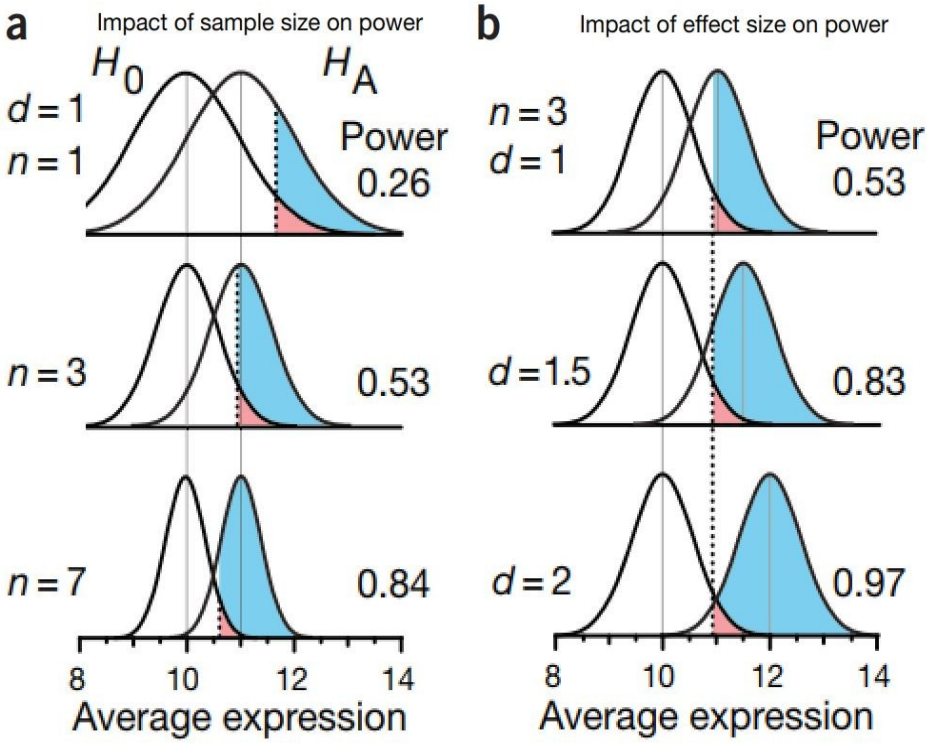

Figure 3: Influence de la taille de l'échantillon et de la taille d'effet sur la puissance. Pour une taille d'effet donnée, la puissance augmente avec la taille de l'échantillon (graphiques de gauche). Pour une taille d'échantillon donnée, la puissance augmente avec la taille d'effet (graphiques de droite). Repris et adapté de Krzywinski & Altman (2013).

Dans la première colonne, la zone de recouvrement des courbes se réduit quand on augmente l'effectif, alors que l'écart entre leurs sommets ne change pas. Quand le nombre de sujets ou d'observations augmente, les variations liées au hasard tendent à s'annuler, et les résultats de l'ensemble se concentrent autour des moyennes. Les courbes de distribution s'élèvent et s'affinent. Les deux courbes se recouvrent moins et la zone bleue occupe une plus grande proportion de la surface comprise sous la courbe H_1. La puissance augmente.

Dans la deuxième colonne, on voit que les deux courbes s'écartent l'une de l'autre sans changer de forme. Cet écart représente la taille d'effet. Il s'agit ici de la différence entre les moyennes des distributions H_0 et H_1. Elle correspond sur le graphique à la distance horizontale qui sépare les sommets des deux courbes.

On voit dans ces trois graphiques que la puissance (toujours représentée par zone bleue prolongée sous la zone rouge) augmente quand la taille d'effet augmente sans que change l'effectif.

Ces graphiques de la Figure 3 montrent chacun une seule combinaison d'effectif, de taille d'effet et de seuil de décision, chacune définissant des rapports différents entre les surfaces représentées, ou entre les performances et les erreurs d'un test sur les valeurs observées. Nous aurons besoin d'autres solutions graphiques, qui permettent de représenter plusieurs distributions simultanément, pour souligner différentes relations entre leurs propriétés. Par exemple, la Figure 6 montre le lien entre la puissance et l'effectif pour différentes tailles d'effet.

Estimer les tailles d'effet à détecter

On comprend maintenant que les hypothèses statistiques à la base du NHST imposent des limites aux différences ou aux effets qui seront détectables par un test statistique (c'est-à-dire susceptibles d'atteindre la limite de signification ou le seuil de rejet). Il faudrait donc avoir une idée de la taille de l'effet à mesurer pour mettre au point une expérience capable de la détecter avec une probabilité de succès suffisante. Mais si l'on connaissait précisément les effets d'un traitement ou d'une manipulation, on n'aurait pas besoin d'expérimentation pour les mesurer.

Pour éviter le problème on peut choisir une taille d'effet minimale avant l'expérience, en décidant quelle différence ou quelle force d'association auront une portée pratique ou théorique suffisante. On peut aussi se donner les moyens de détecter de petites tailles d'effet, pour être certain de pouvoir en détecter de plus grandes aussi.

Le Tableau 2 énumère quelques sources et stratégies utilisables pour déterminer la taille d'effet à détecter.

Tableau 2: Sources utilisables pour choisir la taille d'effet à détecter

Sources utilisables pour choisir la taille d'effet à détecter
Méta-analyses
Catégories prédéfinies, par exemple la classification de Cohen-Sawilowski
Littérature du domaine
Littérature voisine
Méta-analyse intra-laboratoire
Modèles prédictifs
Exigeances curatives
Études exploratoires
Fraction des tailles d'effets précitées

Faute d'idées préalables, il vaut mieux choisir une petite taille d'effet pour être en mesure de détecter aussi les plus importantes.

Utiliser les tailles d'effet conventionnelles de Cohen-Sawilowsky

Il existe de très nombreuses mesures de taille d'effet. Nous utiliserons notamment les d de Cohen, les f de Cohen, et les êta-carrés partiels.

Il faut noter que si les d et f de Cohen correspondent à des écarts entre moyennes, les êta-carrés partiels doivent s'interpréter en termes de pourcentage de variance expliquée.

Le tableau suivant reprend les tailles d'effet de la classification de Cohen-Sawilowsky et leur expression dans ces trois mesures.

Tableau 3: Les tailles d'effet de la classification de Cohen-Sawilowsky

Taille d'effet	d de Cohen	f de cohen	êta-carré partiel
très petite	0.01	0.005	0,000025
petite	0.2	0.1	0,009901
moyenne	0.5	0.25	0,0588235
grande	0.8	0.4	0,137931
très grande	1,2	0.6	0,2647059
énorme	2	1	0,5

La surestimation de la taille d'effet[1]

Curieusement, si les effets que l'on a de bonnes chances de détecter avec une puissance faible sont relativement grands, il reste possible d'obtenir de temps à autre des résultats significatifs malgré une puissance faible, quand les hasards de l'échantillonnage sont favorables. Ainsi, si un test statistique détecte un effet significatif malgré une puissance faible, alors la taille de cet effet sera probablement sur-estimée. C'est l'**erreur de type M** (Gelman & Carlin, 2014). Ce risque est réduit quand la puissance est suffisante (au moins 80 %). Le lien entre la puissance et la taille d'effet devient évident si l'on réfléchit en termes de distribution des tailles d'effet.

Si la puissance est faible parce que les effectifs sont trop petits, alors on obtient une grande variance des moyennes théoriquement observables sous H_1. Donc les tailles d'effet observées se répartiront sur un grand intervalle et il y aura une grande probabilité de surestimer la taille d'effet.

1 « Erreur de type M » (Gelman & Carlin, 2014) ; malédiction de l'enchérisseur gagnant : « the winner's curse » (Young et al., 2008).

Si en revanche la puissance est élevée parce que l'on a un grand effectif, la variance des tailles d'effet est réduite puisque le nombre de mesures augmente. Les distributions des valeurs attendues sous H_1 et H_0 sont bien distinctes parce que les courbes sont amincies et se recouvrent moins. Dans ce cas les tailles d'effet mesurées se rapprocheront de la taille d'effet réelle. La probabilité de surestimer la taille d'effet est moindre.

Une explication graphique se trouve dans la Figure 4.

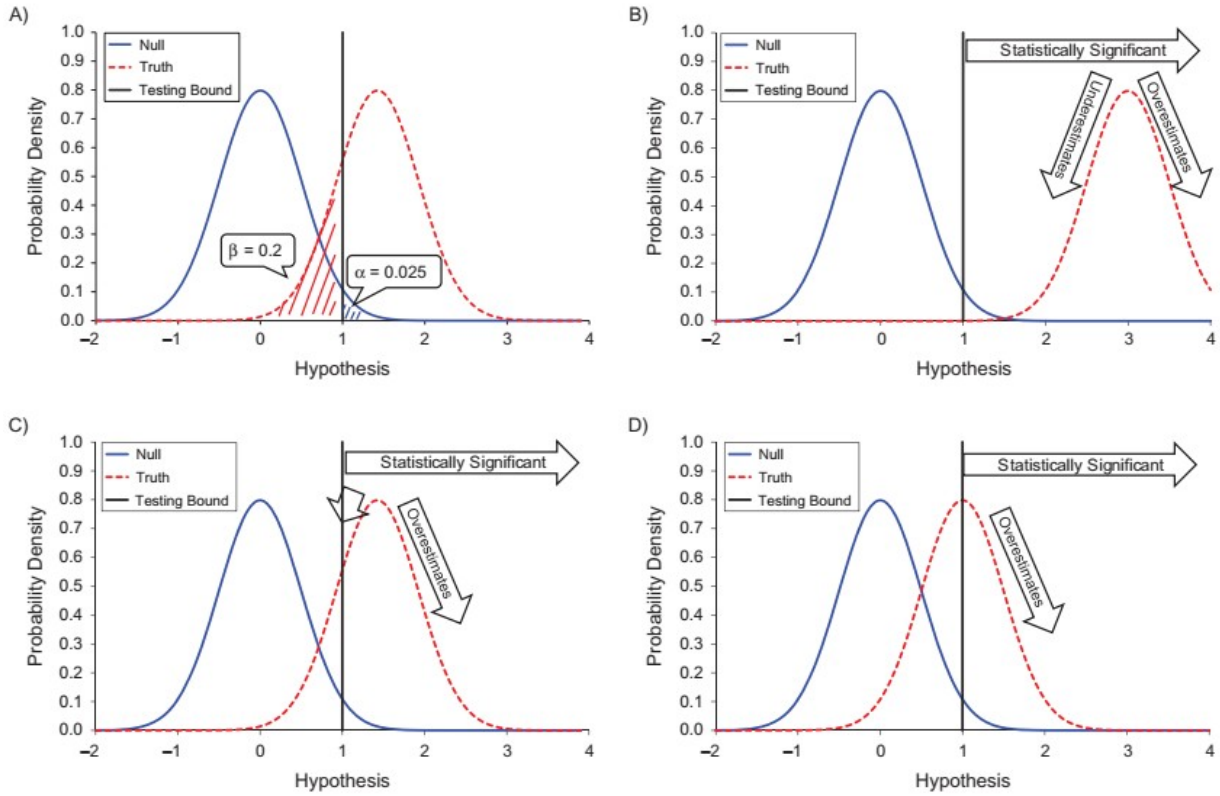

Figure 4: Puissance et risque de surestimation de la taille d'effet. Les courbes bleues représentent la distribution des résultats sous l'hypothèse nulle. Les courbes rouges représentent des distributions réelles de moyenne 1,5 dans les panneaux A et C, 3 dans le panneau B et 1 dans le panneau D. La ligne verticale noire représente le seuil de rejet pour un alpha de 5 %. Quand la taille d'effet augmente (ici l'écartement entre les courbes) les tailles d'effet surestimées sont compensées par les tailles d'effet sous-estimées et on obtient une estimation plus précise de la taille d'effet réelle. Repris de Lash (2017).

Dans chaque graphique, la courbe bleue représente la distribution des résultats attendus sous H_0. La ligne verticale représente le seuil de significativité. La courbe rouge représente les résultats attendus sous H_1 pour une taille d'effet supposée. La puissance est l'aire de la surface située sous la courbe rouge, à droite du seuil de significativité. Le centre de gravité de cette surface représente la moyenne des résultats significatifs que l'on doit attendre. Le panneau A montre les deux courbes

telles qu'on les représente souvent, avec une puissance de 80 %. Le panneau B montre une situation idéale, avec une grande puissance et une grande taille d'effet. Les résultats significatifs sont centrés autour de la taille d'effet réelle. Les flèches indiquent que les résultats sous-estimés compensent les résultats sur-estimés. Le panneau C reprend la situation représentée en A. Les flèches sont déséquilibrées. Il y a une majorité de résultats sur-estimés. Le panneau D montre une petite puissance associée à une petite taille d'effet. Les résultats significatifs y sont tous égaux ou supérieurs à l'effet réel. Il n'y a qu'une flèche, parce qu'aucun résultat sous estimé n'atteint le seuil de significativité.

Le même phénomène est illustré par les résultats de simulations représentés dans la Figure 5.

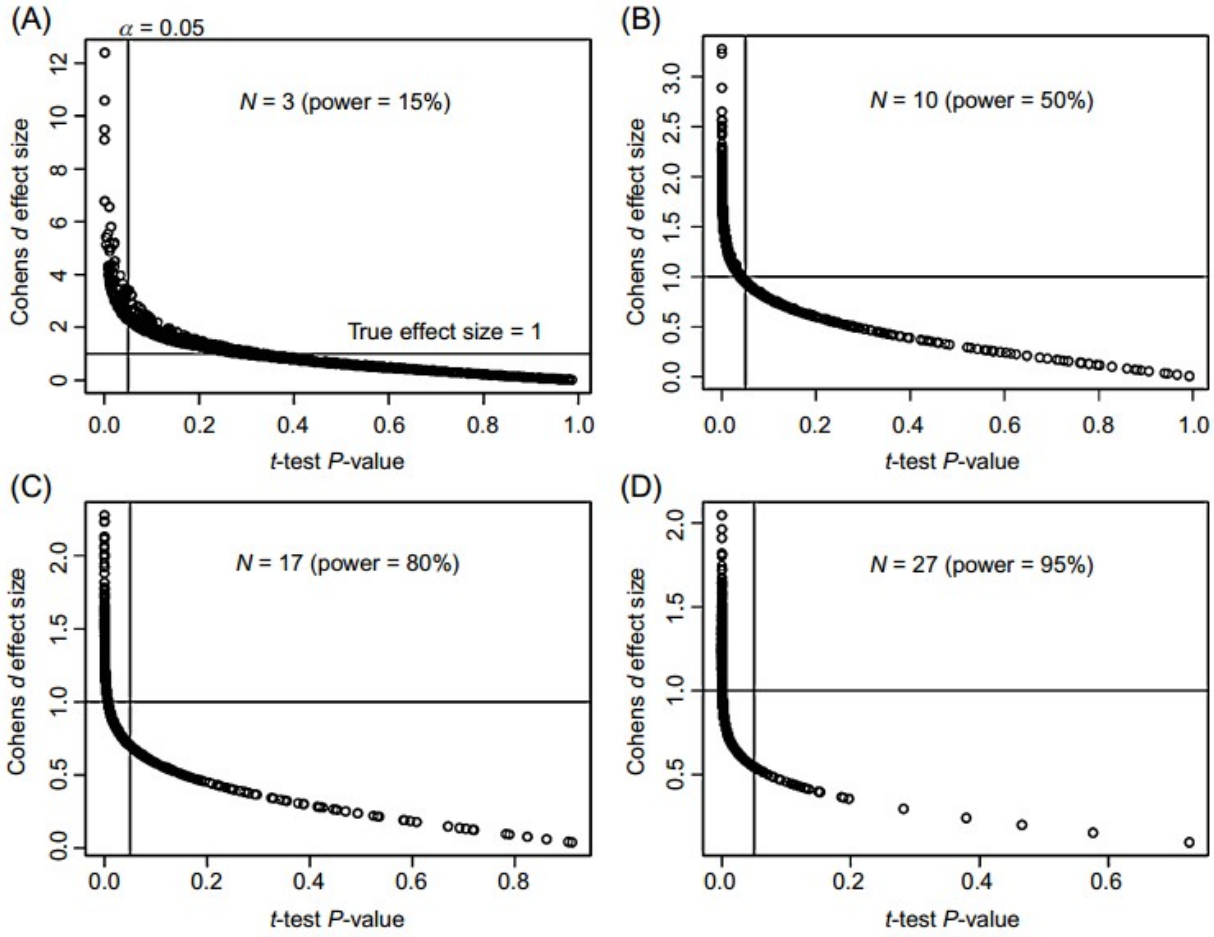

Figure 5: Surestimation de la taille d'effet. Chaque graphique représente les résultats de 1000 simulations (tests t pour échantillons indépendants). Chaque point représente le résultat d'une simulation. Les points à gauche du seuil alpha sont les résultats significatifs. La ligne horizontale représente la taille d'effet de la population. Avec une petite puissance (graphique A), tous les tests significatifs surestiment largement la taille d'effet. La surestimation diminue quand la puissance augmente. Pour les tests avec une puissance élevée (supérieure ou égale à 80%, graphiques C et D), les tailles d'effets observées se répartissent davantage autour de la vraie taille d'effet. Repris de Maiväli, (2015).

Chaque point de la Figure 5 représente les résultats d'une simulation, avec en abscisse la p-valeur donnée par un test t pour échantillons indépendants et en ordonnée la taille d'effet observée. Les deux populations simulées avaient chaque fois un écart-type de 1 et leurs moyennes étaient de 0 pour l'une et de 1 pour l'autre, ce qui donne une taille d'effet de 1 (d de Cohen). Chaque panneau montre les résultats pour une puissance prospective différente (les auteurs ont augmenté la puissance en augmentant l'effectif par groupe). La ligne verticale représente le seuil de significativité alpha (fixé à 5 %). Malheureusement en pratique il s'agit aussi du seuil magique qui donne le signal pour écrire un article qui aura de vraies chances d'être publié (cf. biais de publication). En regardant les valeurs chiffrées, on voit que les tailles d'effets observées se resserrent autour de la vraie taille d'effet quand la puissance augmente. Ainsi, avec une puissance faible de 15 %, la taille d'effet la plus sur-estimée est supérieure à 12, quand avec une puissance élevée de 95 % elle est à peine supérieure à 2. On voit aussi, à l'inverse, qu'il n'y a pas de valeur sous-estimée avec la puissance faible du panneau A, et qu'on en trouve davantage à mesure que la puissance augmente.

La surestimation de la taille d'effet dans la littérature

Divers auteurs ont montré que le phénomène de surestimation de la taille d'effet existe bien dans la littérature.

Par exemple, Dumas-Mallet et al (2017) ont mis en évidence la surestimation de la taille d'effet dans trois domaines de la psychologie clinique : les troubles neurologiques, les troubles psychiatriques, et les troubles somatiques.

Szucs & Ioannidis (2017) ont étudié 26 841 tests statistiques provenant de 3 801 articles. Ils situent les revues sur un graphique particulier qui présente la taille d'effet en abscisse, et en ordonnée les degrés de libertés. Ils y représentent le seuil de significativité et la taille d'effet attendue pour les tests non significatifs. On voit bien ainsi qu'aux puissances basses (petit nombre de degrés de liberté), un test doit avoir une plus grande taille d'effet pour passer le seuil de significativité. Sur le même graphique ils classent les revues selon les médianes de leurs degrés de liberté et de leurs tailles d'effet. Les journaux ont l'air de se ranger par domaine, les revues de psychologie ayant les plus grandes puissances, suivis par les revues d'orientation médicale et enfin par les revues de neurosciences cognitives.

Ioannidis (2008) montre l'inflation de l' « *Odd Ratio* » en fonction de la taille d'échantillon pour 256 méta-analyses avec un résultat significatif répertoriées dans le registre « *Cochrane Library* ». La surestimation est mise en évidence par une courbe de régression « loess » ajoutée au graphique.

Dans une étude portant sur 341 articles provenant de PsycINFO, Kühberger et al (2014) ont trouvé

que la taille d'effet était corrélée négativement avec la taille de l'échantillon, cette corrélation expliquant 18 % de la variance des tailles d'effet.

Comme l'expriment bien Young et al (2008), la surestimation de la taille d'effet dans la littérature pourrait être un produit du système de publication scientifique. En limitant le nombre d'articles publiés dans chacun de leurs numéros, les journaux sont conduits à publier d'abord les articles les plus excitants, donc ceux qui obtiennent les résultats les plus extrêmes. Les résultats moins extrêmes, mais plus conformes à la réalité, doivent attendre leur tour plus longtemps avant d'être publiés. De ce fait il peut se passer des années avant qu'un résultat positif extrême ne soit corrigé (Young et al., 2008).

L'effet de déclin

On parle d'effet de déclin quand un effet perd en magnitude au fur et à mesure des années et des réplications.

Plusieurs études ont observé un tel effet de déclin (Brembs, 2018; Brembs et al., 2013; Gonon et al., 2012; Lodder et al., 2019). Dans le pire des cas, l'effet original peut même disparaître (Brembs et al., 2013).

En psychologie, le projet de réplication de l'Open Science Collaboration (2015) n'a réussi que 36 % de ses tentatives de réplication, et 83 % des études de réplication ont obtenu des tailles d'effet inférieures aux mesures originales.

L'effet de déclin pourrait s'expliquer en partie par la régression à la moyenne. En effet, si l'on obtient des résultats extrêmes avec un échantillon d'une distribution aléatoire, on peut raisonnablement prédire que les résultats des échantillons suivants (ou précédents) seront plus proches de la moyenne. C'est le phénomène de **régression à la moyenne**[2].

L'effet de déclin pourrait aussi se ranger parmi les conséquences du manque de puissance, si la puissance du premier article était insuffisante et a donc surestimé la taille d'effet et si les tentatives de réplication ont mobilisé une puissance plus importante.

Mais on peut aussi imaginer une infinité d'autres explications : la mentalité des sujets change, les chercheurs évaluent la généralité des effets publiés, les habitudes inconscientes des chercheurs et les compétences disponibles dans les labos évoluent, le biais de publication varie selon le goût du moment, etc.

2 C'est le contraire de ce qu'il se passe quand on a affaire à des phénomènes de type « marcheur aléatoire », où les résultats d'un échantillon sont centrés sur la dernière mesure du précédent

1.4 La puissance statistique (1-β)

Obtenir une bonne p-valeur ne suffit pas à garantir la fiabilité des résultats, et nous venons de voir que la taille d'effet n'est pas non plus une garantie, surtout quand la puissance est faible.

La **puissance** est la probabilité d'obtenir un test significatif si l'hypothèse alternative est vraie. En notation probabiliste cela donne :

$$p(test\ significatif | H_1)$$

La puissance dépend de la taille d'effet que l'on veut détecter, de la taille de l'échantillon, du seuil de rejet (alpha), et de l'analyse statistique utilisée.

Elle permet donc de définir l'effectif nécessaire pour qu'un test statistique donné détecte un effet significatif, avec une probabilité donnée (en général 80 %), et pour une taille d'effet donnée elle aussi. Cette puissance est aussi utilisée pour évaluer la puissance atteinte par un test selon les effectifs utilisés, la taille d'effet visée et le seuil de significativité.

Dans la littérature on trouve souvent le terme de puissance prospective dans un sens équivalent. Nous utiliserons donc les termes de « puissance » et « puissance prospective » de manière indifférenciée.

On peut calculer la plus petite taille d'effet détectable avec une puissance prédéfinie (souvent 80 %), quand on connaît l'effectif mobilisé dans l'étude.

Ainsi la Figure 6 montre que si l'effectif est juste suffisant pour détecter un d de Cohen de 0,5 avec une puissance de 80 %, alors il permettra aussi de détecter un d de Cohen de 0,8 avec une meilleure puissance, mais pas un d de Cohen de 0,2 (qui demanderait huit fois plus de sujets). De même si l'effectif est suffisant pour détecter un d de Cohen de 0,2 avec une puissance de 0,8, il suffira aussi pour détecter les plus grandes tailles d'effet avec une puissance presque parfaite.

La même figure montre qu'il faut plus de 25 sujets pour avoir seulement 80 % de chances de détecter une taille d'effet de 0,8, car la ligne bleue croise la ligne pointillée à la verticale de l'effectif 25.

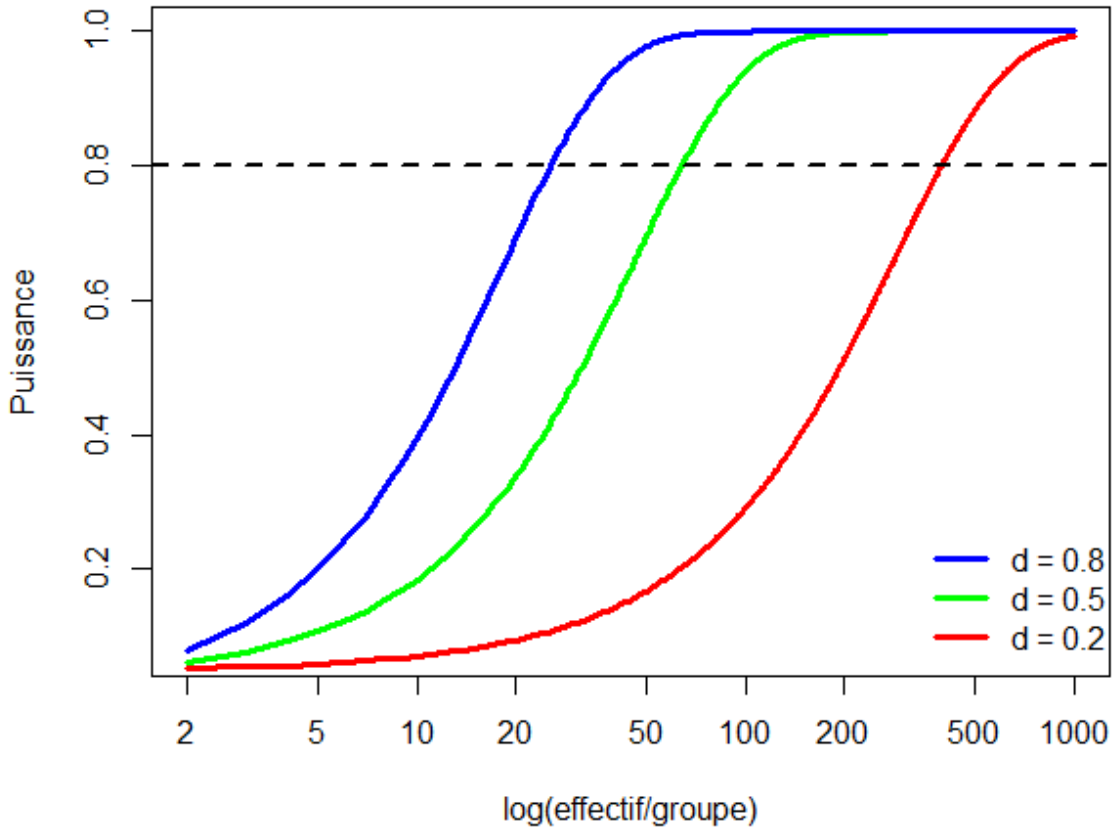

Figure 6: Puissance, effectif et taille d'effet. Impact de l'effectif des groupes expérimentaux sur la puissance des tests t pour échantillons indépendants selon les trois tailles d'effet de la classification de Cohen. À un risque alpha donné, la puissance augmente avec l'effectif jusqu'à atteindre un plafond (puissance maximale de 1). Avec une grande taille d'effet, il faut moins de sujet par groupe pour atteindre une puissance de 0,8, et il en faut d'avantage avec une taille d'effet plus petite. Attention l'axe x est en échelle logarithmique pour couvrir une plus grande gamme d'effectifs. Graphique personnel inspiré de Pezzullo (2013) et réalisé avec R.

La Figure 7 reprend un graphique basé sur des simulations plutôt que sur des formules. Il permet d'estimer la probabilité d'obtenir un résultat significatif quand on connaît le nombre de participants et la taille d'effet que l'on vise. Il montre aussi qu'avec une puissance suffisante pour détecter une petite taille d'effet, on est en mesure de détecter des tailles d'effet plus grandes, et cela avec une quasi-certitude. Notez que les résultats significatifs comprennent aussi les faux positifs (erreurs alpha). Donc cette figure ne représente pas les puissances, mais les puissances augmentées des 5 % d'erreur alpha.

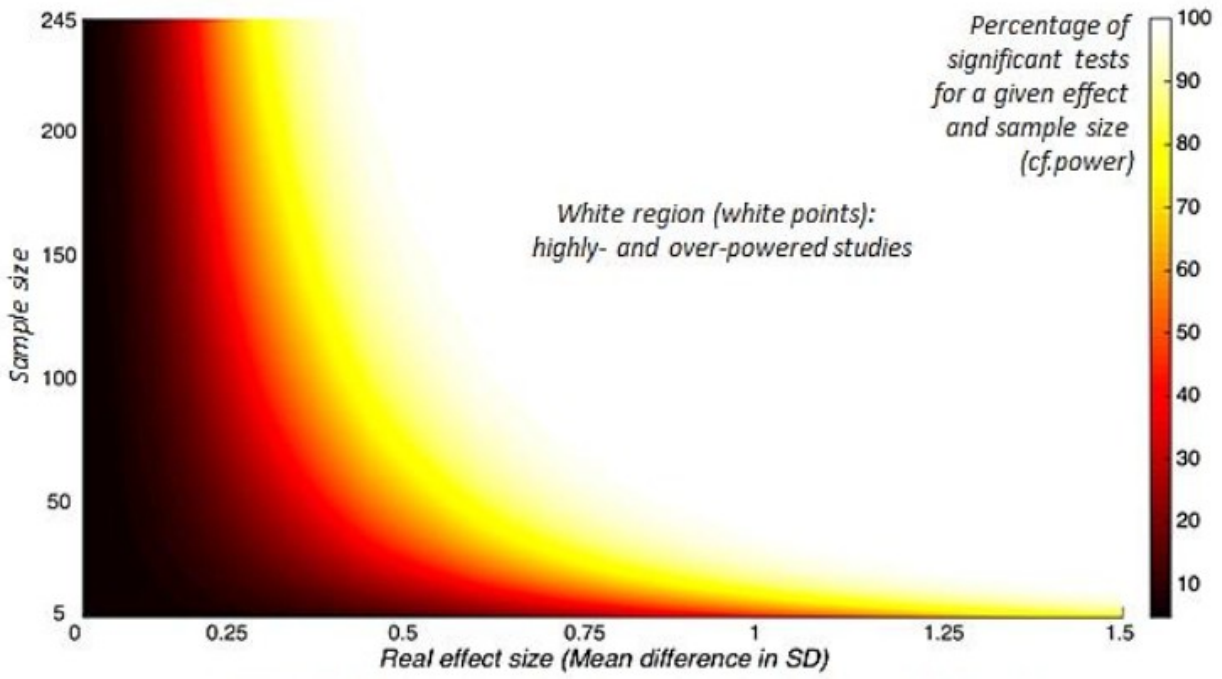

Figure 7: Probabilité de détecter un effet significatif en fonction de la taille d'effet et de l'effectif. L'échelle des couleurs représente les probabilités. Chaque point a été calculé à partir de 100 000 tests t pour échantillons indépendants simulés. Les couleurs représentent la puissance additionnée du risque alpha. Repris et modifié de Wallisch (2015).

La puissance d'un test statistique augmente quand la taille d'effet augmente comme le montre la Figure 8. On y voit trois courbes correspondant à trois effectifs. La courbe correspondant à 25 sujets atteint une puissance de 0,8 pour une taille d'effet de l'ordre de 0,7, alors que la courbe correspondant à 10 sujets n'atteint cette puissance que pour une taille d'effet de l'ordre de 1,3.

Augmenter l'effectif permet donc de détecter des tailles d'effet plus petites avec la même puissance. Il est donc préférable de choisir un grand effectif quand on veut détecter une petite taille d'effet. Si on calcule les effectifs en visant la détection d'une petite taille d'effet et que la taille d'effet réelle est grande, nous serons aussi en mesure de la détecter et nous aurons même une puissance confortable proche de 1.

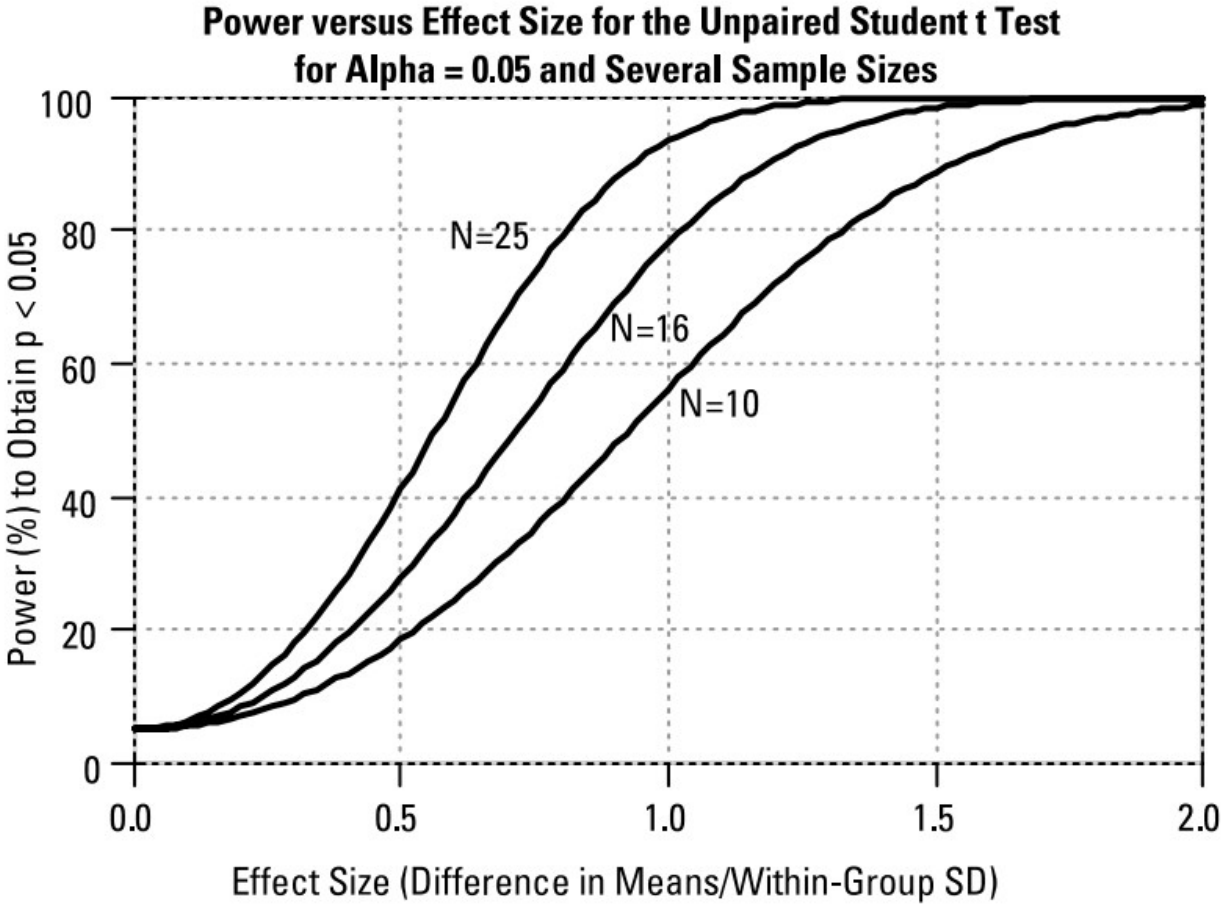

Figure 8: La relation entre la puissance et la taille d'effet. Lorsque la taille d'effet augmente, la puissance augmente aussi et cela d'autant plus vite que le nombre de sujets est important. Illustration pour trois effectifs différents pour un test t pour échantillons indépendants avec un alpha fixé à 5 %. Repris de Pezzullo (2013).

Le manque de puissance

Le tableau des inférences NHST a défini les erreurs de type I et les erreurs de type II. Nous avons vu aussi l'erreur de type M, qui surestime la magnitude de l'effet détecté. Nous savons que le manque de puissance augmente le risque d'erreur global.

Comme mentionné plus haut, avec une puissance faible le risque de publier des résultats erronés augmente, et les chances de détecter un effet réel, même trivial, diminuent.

La littérature considère qu'une puissance de 80 % est un minimum. Mais de nombreux travaux montrent que la puissance statistique des articles scientifiques atteint rarement ces 80 %, comme Baldwin, (2017), Button et al (2013), Carneiro et al (2018), Cohen (1962), Fidler et al (2017), Giuffrida (2014), Hofmeister et al (2007), Moher et al (1994), Quintana (2020), Schmidt-Pogoda et al (2019), Sedlmeier & Gigerenzer (1989), Szucs & Ioannidis (2017) et Walum et al (2016).

On pourrait croire qu'un problème aussi commun ne doit pas être bien grave, mais nous l'avons vu, le manque de puissance favorise le p-hacking. Il entraîne un risque de sur-estimation des tailles d'effet et pourrait aussi expliquer, au moins partiellement, l'effet de déclin. Il augmente enfin la probabilité de faux positif (Ioannidis, 2005), et le *« risque de se faire passer pour un idiot »*[3] (Colquhoun, 2014).

On connaît ces problèmes de puissance depuis longtemps puisque Cohen les a déjà soulignés en 1962. Mais il semble que ses avertissements et ceux de ses nombreux collègues n'ont pas eu les effets désirés.

Nous étudierons le manque de puissance d'une part dans la littérature « clinique », qui concerne les sujets humains comme la psychologie sociale, le management, la médecine, la psychologie clinique et les neurosciences cognitives, et d'autre part dans la littérature « préclinique » qui étudie l'écologie et les modèles animaux.

Le manque de puissance dans la littérature « clinique »

Les avertissements de Cohen (1962) n'ont pas eu beaucoup d'effet dans la littérature « clinique ». À l'époque, il montrait que les effectifs mobilisés dans les articles du « *Journal of abnormal and social psychology* » ne permettaient pas de détecter les tailles d'effet petites à moyennes.

Plus tard, Sedlmeier & Gigerenzer (1989) ont trouvé une diminution de la puissance dans les articles du « *Journal of Abnormal Psychology* ». La puissance pour détecter une taille d'effet moyenne qui y était de 0,46 en 1960 y est tombée à 0,37 en 1984.

Vingt-huit ans plus tard, Rossi (1990) a évalué la même revue que Cohen en 1962 en y ajoutant deux autres : le « *Journal of Consulting and Clinical Psychology* » et le « *Journal of personality and Social Psychology* ». Il a montré une légère baisse de la puissance pour les petites tailles d'effet et une augmentation pour les moyennes. Pour les grandes tailles d'effet, il n'a pas observé de modification. En plus, il a énuméré des évaluations de puissance publiées après Cohen (1962) et avant Sedlmeier & Gigerenzer (1989). Toutes observent un manque de puissance.

Walum et al (2016) montrent un manque de puissance dans les études utilisant l'administration intra-nasale d'ocytocine. Dans trois méta-analyses ils trouvent des puissances allant de 12 % à 16 % pour des tailles d'effets de d = 0,21 et d = 0,48. Même cette alarme récente ne semble pas avoir eu d'effet puisque Quintana (2020) a trouvé une puissance moyenne de 12 % et (avec une médiane de 7,8 %) pour détecter une petite taille d'effet, dans le même domaine .

Szucs & Ioannidis, (2017) ont calculé la puissance pour 3 sous-domaines : les neurosciences

3 « to make a fool of yourself »

cognitives, la psychologie, et la médecine. Pour les 3 sous domaines, la puissance pour détecter des tailles d'effets petites à moyennes est inférieure à 80 %. Pour une grande taille d'effet la psychologie et la médecine arrivent tout juste à 80 %, tandis que les neurosciences cognitives n'atteignent que 70 %.

Schmidt-Pogoda et al (2019) ont réalisé une méta-analyse sur les études précliniques et cliniques des ischémies cérébrales. Ils ont trouvé une puissance de 17 % pour une taille d'effet de -0,47. La valeur absolue de cette taille d'effet étant probablement surestimée, car elle provient de leur méta-analyse, leur estimation de la puissance l'est sans doute aussi.

En 2013, Button et al se sont intéressés à la puissance des études de neuro-imagerie. Ils ont rassemblé 48 méta-analyses. Ils ont calculé les puissances de chaque article inclus dans ces méta-analyses en en lui attribuant la taille d'effet calculée dans la méta-analyse qui le mentionnait. La puissance médiane de ces articles était de 8 %.

Le manque de puissance dans la littérature « préclinique »

Les études précliniques ne sont pas épargnées par le manque de puissance. Le Tableau 4 résume les résultats énoncés ci-dessous.

Button et al (2013) ont étudié la puissance des articles inclus dans deux méta-analyses sur les labyrinthes radiaux et aquatiques. Ils ont calculé les puissances en utilisant les tailles d'effet synthétiques déterminées par les méta-analyses. Ils ont trouvé que la puissance était de 18 % (pour un d de Cohen de 0,46) pour le labyrinthe aquatique et de 31 % (pour un d de Cohen de 0,69) pour le labyrinthe radial .

Plus récemment, Carneiro et al (2018) ont calculé la puissance médiane des articles sur le « *fear conditionning* ». Ils ont calculé cette puissance à partir de la taille d'effet observée dans les études les plus puissantes (min. 95 %) du même domaine. Ils trouvent une puissance médiane de 75 %. Ils précisent qu'avec ces puissances « faibles » (et vraisemblablement sur-évaluées), les chercheurs ne sont en mesure de détecter que des tailles d'effet égales ou supérieures à 37,2 % (en différence absolue).

En écologie aussi, on observe une sous-puissance des tests statistiques utilisés. McCarthy & Parris (2001) (cité par Fidler et al (2017)) ont étudié la puissance des articles étudiant la régénération des orteils de grenouilles. Pour les petites tailles d'effet les puissances vont de 6 à 10 %. Pour les tailles d'effet moyennes elles vont de 8 à 21 %, et pour les grandes elles vont de 15 à 60 %.

Smith et al (2012) (cité par Fidler et al., 2017) ont étudié les puissances pour les petites et grandes tailles d'effet dans les articles du journal « *Animal behaviour* ». Ils trouvent des puissances allant de

7 à 8 % et pour les grandes tailles d'effet et de 23 à 26 % pour les grandes.

Hofmeister et al (2007) ont trouvé que la littérature concernant l'analgésie vétérinaire n'avait une puissance suffisante pour détecter une taille d'effet de 20 % d'efficacité du traitement que dans 54 % des cas. Pour un effet du traitement de 80 % (grande taille d'effet), 82 % des études avaient une puissance suffisante.

Giuffrida (2014) a étudié la puissance d'essais randomisés contrôlés (RCT) ayant obtenu des résultats négatifs chez de petits animaux. 39 % seulement des RCT inclus avaient une puissance suffisante (80%) pour détecter des tailles d'effet de 50 %. Pour des tailles d'effet de 25 %, la proportion de RCT suffisamment puissants tombe à 14 %.

Le Tableau 4 donne un aperçu des puissances observées dans la littérature « préclinique » par ces quelques auteurs.

Tableau 4: Puissances dans la littérature « préclinique »

	Puissance	Taille d'effet	T. E. détectable
Button et al 2013	18 %	0,46 d de Cohen	
	31 %	0,69 d de Cohen	
Carneiro et al 2018	75 %	observé	37 % (Diff. Abs.)
McCarthy et Paris 2001	6 % à 10 %	petite	
	8 % à 21 %	moyenne	
	15 % à 60 %	grande	
Smith et al 2012	7 % à 8 %	petite	
	6 % à 10 %	grande	
Hofmeister et al 2007	54 % des études ont une puissance suffisante pour détecter		un traitement avec 20 % d'efficacité
	82 % des études ont une puissance suffisante pour détecter		un traitement avec 80 % d'efficacité
Giufrida 2014	39 % des études ont une puissance suffisante pour détecter	50 %	
	14 % des études ont une puissance suffisante pour détecter	25 %	

On y voit bien que dans la littérature « préclinique », les études atteignent rarement la puissance recommandée de 80 %.

On voit aussi que la manière de rapporter la puissance est loin d'être standardisée, et que de ce fait les comparaisons entre études sont très difficiles.

Le manque de puissance est donc évident dans la littérature, aussi bien « clinique » que « préclinique ».

Macleod et al.(2015) et Ramirez et al. (2017) ont étudié la qualité du compte rendu (« *reporting* ») d'une série articles. Presque aucun d'entre eux ne mentionne de calcul de puissance pour déterminer leurs effectifs, ce qui fait penser que les auteurs ne s'intéressent pas à la puissance. En plus, les rares articles réalisant des analyses de puissance prospective pour déterminer leurs effectifs ne rapportent pas correctement les paramètres utilisés (Cribbie et al., 2019).

Le manque de puissance et le biais de publication

On peut suspecter un biais de publication dans un domaine si la puissance y est inférieure au taux de résultats significatifs (Fidler et al., 2017).

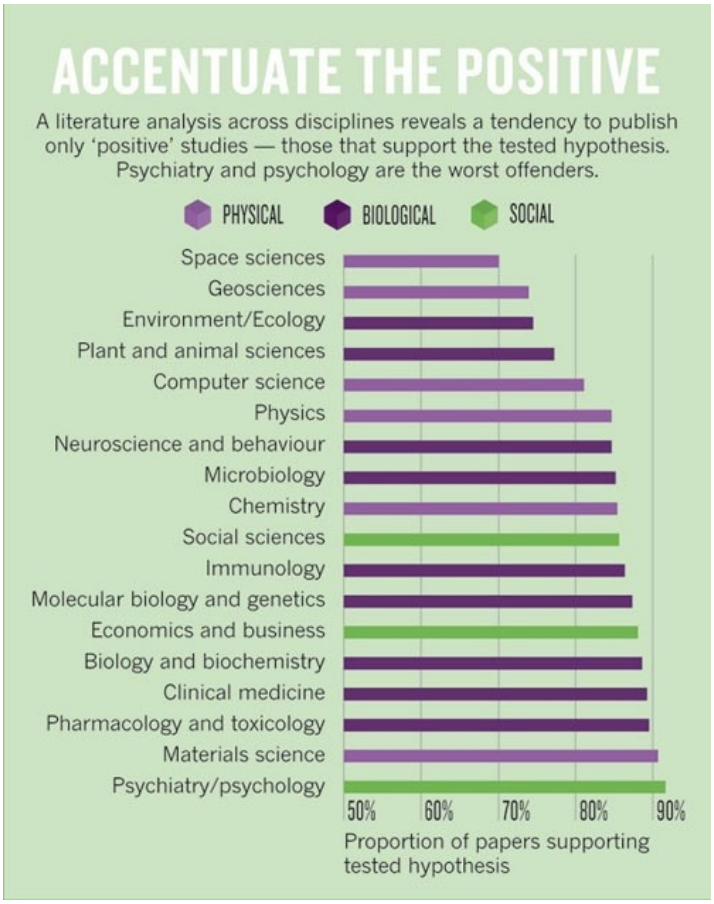

Figure 9: Proportion d'articles confirmant l'hypothèse testée. Repris sans modification de Yong(2012).

La Figure 9 reprend le graphique de Fanelli (2010) adapté par Yong (2012). On peut y comparer la situation dans différentes disciplines.

En psychologie et en psychiatrie, 90 % des résultats publiés sont significatifs contre 70 % en sciences spatiales. Les neurosciences ne se portent guère mieux avec environs 85 % (Fanelli, 2010).

Combiné avec le manque de puissance, le biais de publication menace la validité des méta-analyses en tirant la taille d'effet synthétique vers le haut.

Il est cependant possible de limiter ce risque en estimant ce biais via les « *funnel plots* » et la méthode du « *trim and fill* » développée par Duval & Tweedie (2000). La méthode consiste à estimer le nombre d'études manquantes dans l'échantillon de la méta-analyse et à remplir le « *funnel plot* » en miroir avec le nombre estimé d'études manquantes.

1.5 Les taux de vraies et de fausses découvertes

Les p-valeurs ne sont pas fiables (à cause du p-hacking), les tailles d'effet non plus (à cause de l'erreur de type M), et les puissances mobilisées sont rarement suffisantes, de sorte que les risques d'erreur sont plus importants qu'on ne le croit généralement. Malheureusement ce n'est pas tout, car il reste un élément sérieux que nous n'avons pas encore pris en compte, à savoir la probabilité que l'hypothèse testée soit vraie.

De même qu'on doit interpréter les résultats d'un test médical en tenant compte de la prévalence de la maladie recherchée, on devrait évaluer les chances d'obtenir des résultats correspondant à la réalité en tenant compte de la proportion d'effets réels parmi l'ensemble des effets testables. On s'approcherait ainsi de la vraie question du chercheur, qui ne voudrait pas savoir si les résultats obtenus sont statistiquement significatifs, mais bien si les données qu'il a récoltées rendent son hypothèse plus ou moins vraisemblable.

Enrichissons donc notre réflexion statistique en y incluant la plausibilité des hypothèses ou des résultats espérés. Cette proportion est facile à connaître quand on dispose d'un ensemble fini d'hypothèses avec une proportion connue d'hypothèses vraies, comme quand on veut détecter une maladie dont on connaît la prévalence. On peut alors calculer un taux de vraies découvertes (*True Report Probability* : TRP) qui donne la probabilité qu'un résultat significatif corresponde à une vraie découverte, ce qui s'approche davantage de la vraie question du chercheur que la p-valeur (Ioannidis, 2005; Szucs & Ioannidis, 2017). Comme le dit Colquhoun (2014) « *they make fools of themselves* » en se fiant aux seules p-valeurs, ou même aux autres statistiques que nous avons décrites dans les sections précédentes. Le TRP s'accompagne de son complément, le taux de fausses découvertes (False Discovery Rate : FDR), qui donne la probabilité qu'un résultat significatif soit un faux positif.

En termes statistiques, il faudrait enrichir nos modèles pour tenir compte des chances d'avoir une hypothèse vraie ou d'observer de vrais effets. Ce que Ioannidis a fait en 2005 dans un article désormais classique où il affirme que la majorité des résultats scientifiques publiés sont probablement faux.

Il a construit sa démonstration en empruntant une statistique de la théorie diagnostique : la valeur prédictive positive (VPP), ou Predictive Positive Value (PPV) (en français la valeur prédictive des [résultats] positifs). Szucs & Ioannidis (2017) préfèrent parler de TRP à la place de PPV, car ils ont remplacé la prévalence de la maladie par la plausibilité de l'hypothèse testée.

TRP et test diagnostique

Le TRP ressemble beaucoup à la PPV de la théorie diagnostique, tant par ses formules que par son interprétation. Il estime la probabilité qu'une hypothèse soit vraie à partir de sa vraisemblance et des résultats de l'expérience, comme un médecin doit évaluer la probabilité que son patient soit atteint d'une maladie à partir du résultat et des propriétés d'un test diagnostic et de la fréquence de cette maladie (ou mieux, si possible, de la fréquence de la maladie chez les gens comparables à son patient).

Le Tableau 5 rappelle les concepts du test diagnostique. Le lecteur habitué au vocabulaire des tests diagnostiques pourrait parcourir d'abord le Tableau 6 qui met en évidence la correspondance entre les termes diagnostiques et statistiques.

Tableau 5: Classifications binaires et principales probabilités des tests diagnostique

		statut réel du patient			
		Malade	Sain	Prévalence malades/ population	
Test	Positif	Vrais positifs VP	Faux positifs FP	Positive Predictive Value VP/(VP+FP)	False Discovery Rate FP/(VP+FP)
	Négatif	Faux négatifs FN	Vrais négatifs VN	False Ommission Rate (FOR) FN/(FN+VN)	Negative Predictive Value (NPV) VN/(FN+VN)
		Sensibilité VP/(VP+FN)	False Positive Rate (FPR) FP/(FP+VN)		
		False Negative Rate (FNR) FN/(VP+FN)	Spécificité VN/(FP+VN)		

Tableau 6: Analogies entre le test diagnostique et le test statistique

Test diagnostique	**Test statistique (NHST)**
Absence de la maladie	Hypothèse nulle vraie
Présence de la maladie	Hypothèse nulle fausse hypothèse alternative vraie
Test positif	Effet significatif rejet de l'hypothèse nulle
Test négatif	Effet non-significatif non-rejet de l'hypothèse nulle
Taux de Faux Positifs (1-spécificité)	Risque alpha
Taux de Vrais Positifs (sensibilité)	Puissance
Prévalence de la maladie dans la population nombre de malade / taille de la population	Plausibilité de l'existence de l'effet étudié $p(H_1)/p(H_0)$
Valeur de Prédictive Positive (VPP ou PPV) probabilité a posteriori	Probabilité de vrais résultats (TRP = True Report Probability) probabilité a posteriori
Taux de Fausses Découvertes (TFD ou FDR)	Taux de Fausses Découvertes (TFD ou FDR)

La logique du TRP

Décrivons maintenant la logique probabiliste du TRP[4].

En notation probabiliste le TRP est la probabilité suivante :

$$TRP = p(H_1 | test\ significatif)$$

On peut voir que la notation probabiliste de la puissance utilise les mêmes termes :

$$puissance = p(test\ significatif | H_1)$$

4 Nous répétons ici nos remerciements au professeur David Colqhoun qui nous a aidé (par e-mail) à comprendre la démonstration des formules de TRP et FDR, dont les éléments sont éparpillés dans plusieurs articles différents (Button et al., 2013; Colquhoun, 2019, 2014, 2017; Ioannidis, 2005; Szucs & Ioannidis, 2017).

Si l'on connaît la probabilité de H_1, on peut obtenir le TRP à partir de la puissance en employant le théorème de Bayes :

$$p(A|B)=\frac{p(B|A)\,p(A)}{p(B)}$$

Si on remplace la partie gauche de cette formule de Bayes par la formule du TRP on obtient :

$$p(H_1|\textit{test significatif})=\frac{p(\textit{test significatif}|H_1)\,p(H_1)}{p(\textit{test significatif})}$$

Où le numérateur est la formule de la puissance.

Mais comment obtenir la valeur du dénominateur, à savoir la probabilité d'avoir un test significatif ? C'est possible en utilisant une autre règle des probabilités que voici :

$$p(B)=p(B|A)\,p(A)+p(B|\neg A)\,p(\neg A)$$

Cela nous donne :

$$p(H_1|\textit{test significatif})=\frac{p(\textit{test significatif}|H_1)\,p(H_1)}{p(\textit{test significatif}|H_1)\,p(H_1)+p(\textit{test significatif}|H_0)\,p(H_0)}$$

Voyons maintenant comment effectuer les calculs avec la notion de plausibilité.

Si l'on se souvient que la puissance est la probabilité d'obtenir un résultat significatif si H_1 est vraie et l'erreur de type I est celle d'obtenir un résultat significatif si H_0 est vraie (cfr Tableau 1 : Les inférences et erreurs possibles dans le cadre du NHST), on obtient en remplaçant les termes :

$$TRP=\frac{\textit{puissance}\,p(H_1)}{\textit{puissance}\,p(H_1)+\textit{erreur de type I}\,p(H_0)}$$

Dans cette formule, Ioannidis (2005) divise numérateur et dénominateur par $p(H_0)$, ce qui donne :

$$TRP=\frac{\textit{puissance}\,\frac{p(H_1)}{p(H_0)}}{\textit{puissance}\,\frac{p(H_1)}{p(H_0)}+\textit{erreur de type I}}$$

Cette simplification permet d'utiliser la plausibilité qui est le rapport de chances nommé R. R peut être vu comme la proportion obtenue en divisant le nombre d'hypothèses alternatives vraies par le nombre d'hypothèses nulles vraies. Ou encore comme la probabilité de l'hypothèse alternative divisée par la probabilité de l'hypothèse nulle. Il exprime aussi la plausibilité, et se définit statistiquement comme suit :

$$\frac{p(H_1)}{p(H_0)} = R = plausibilité$$

On obtient donc au final :

$$TRP = \frac{puissance * plausibilité}{puissance * plausibilité + erreur\ de\ type\ I}$$

En version abrégée :

$$TRP = \frac{(1-\beta)R}{(1-\beta)R+\alpha}$$

Le « false discovery rate » (FDR) est le complément du TRP et s'obtient donc directement à partir du TRP via la formule FDR=1-TRP.

On peut voir le FDR et le TRP comme les deux faces d'une même pièce. Le FDR peut être calculé en suivant le même raisonnement que pour le TRP à la différence que le FDR est équivalent à :

$$FDR = p(H_0 | test\ significatif)$$

On obtient donc la formule :

$$FDR = \frac{\alpha}{(1-\beta)R+\alpha}$$

Pour une plausibilité donnée, le TRP donne la probabilité qu'un test significatif soit un vrai positif, et le FDR donne la probabilité qu'un test significatif soit en réalité un faux positif. Ces deux formules montrent que le TRP et le FDR sont influencés par la puissance, le seuil de significativité, et la plausibilité[5].

5 Il existe aussi une formule pour les calculer en tenant compte de l'existence de biais.

Comment utiliser le TRP quand on ne connaît pas la plausibilité

Comme nous ne connaissons pas la plausibilité, nous devons trouver un moyen de contourner le problème pour utiliser malgré tout le TRP.

En ajoutant le FDR (ou le TRP) à la p-valeur lors de la rédaction des articles, on permettrait aux lecteurs de se faire immédiatement une idée de la robustesse de l'effet trouvé (Leek et al., 2017). Mais quelle plausibilité faudrait-il postuler pour calculer ce FDR ? Colquhoun (2014, 2017, 2019) suggère de prendre une plausibilité de 1/1 ou une probabilité *a priori* de 50 %. Ce n'est peut-être pas la meilleure idée, car d'après les résultats de Dreber et al (2015) on risquerait de sur-estimer le TRP et de sous-estimer le FDR. Ces derniers auteurs suggèrent de choisir une plausibilité de 10 %. Cependant la plausibilité qu'ils ont calculée ne s'applique peut-être pas à la psycho-pharmacologie préclinique. Il est peu probable en effet qu'une drogue n'ait aucun effet psychologique. Nous pouvons donc espérer une plausibilité supérieure à 10 %. Pour être fixé, il faudrait organiser un vaste programme de réplication comme le projet « *many labs* » en psychologie sociale (Klein et al., 2014).

C'est pourquoi nous proposons, comme Ioannidis (2005), de contourner le problème avec un graphique qui permet de visualiser l'ensemble des possibilités. Ainsi, la Figure 10 illustre le comportement du TRP et du FDR en fonction de la puissance pour différentes probabilités de l'hypothèse alternative.

L'utilisateur, qui connaît la puissance et la probabilité qu'il attribue à l'hypothèse testée, y trouve immédiatement sur l'ordonnée le taux de fausses découvertes ou la valeur positive prédictive (valeur prédictive des résultats positifs) correspondant.

Alors plutôt que de reporter dans les articles une valeur unique de TRP ou FDR pour une plausibilité arbitrairement choisie, il nous semble plus judicieux de reporter un graphique qui permette au lecteur de situer la puissance de l'étude par rapport à une puissance idéale d'au moins 80 %, selon la plausibilité qu'il voudra accorder à l'hypothèse alternative, avec la méthode que nous venons d'expliquer.

D'autres comme Baldwin, (2017), Button et al., (2013), Nord et al., (2017), Schmidt-Pogoda et al., (2019), Szucs & Ioannidis, (2017), et Walum et al., (2016) ont aussi choisi cette solution.

Des graphiques semblables à ceux de Ioannidis avec en abscisse les plausibilités variant de 0 à 1 se trouvent dans les figures 19, 20, 23 et 24 des résultats. Ces graphiques ne peuvent pas représenter les probabilités de H_1 supérieures à 0,5. Des graphiques étendus avec en abscisse les probabilités de H_1 variant de 0 à 1 se trouvent dans les figures 21, 22, 25 et 26 des résultats. Ces nouveaux graphiques permettent de représenter les plausibilités supérieures à 1.

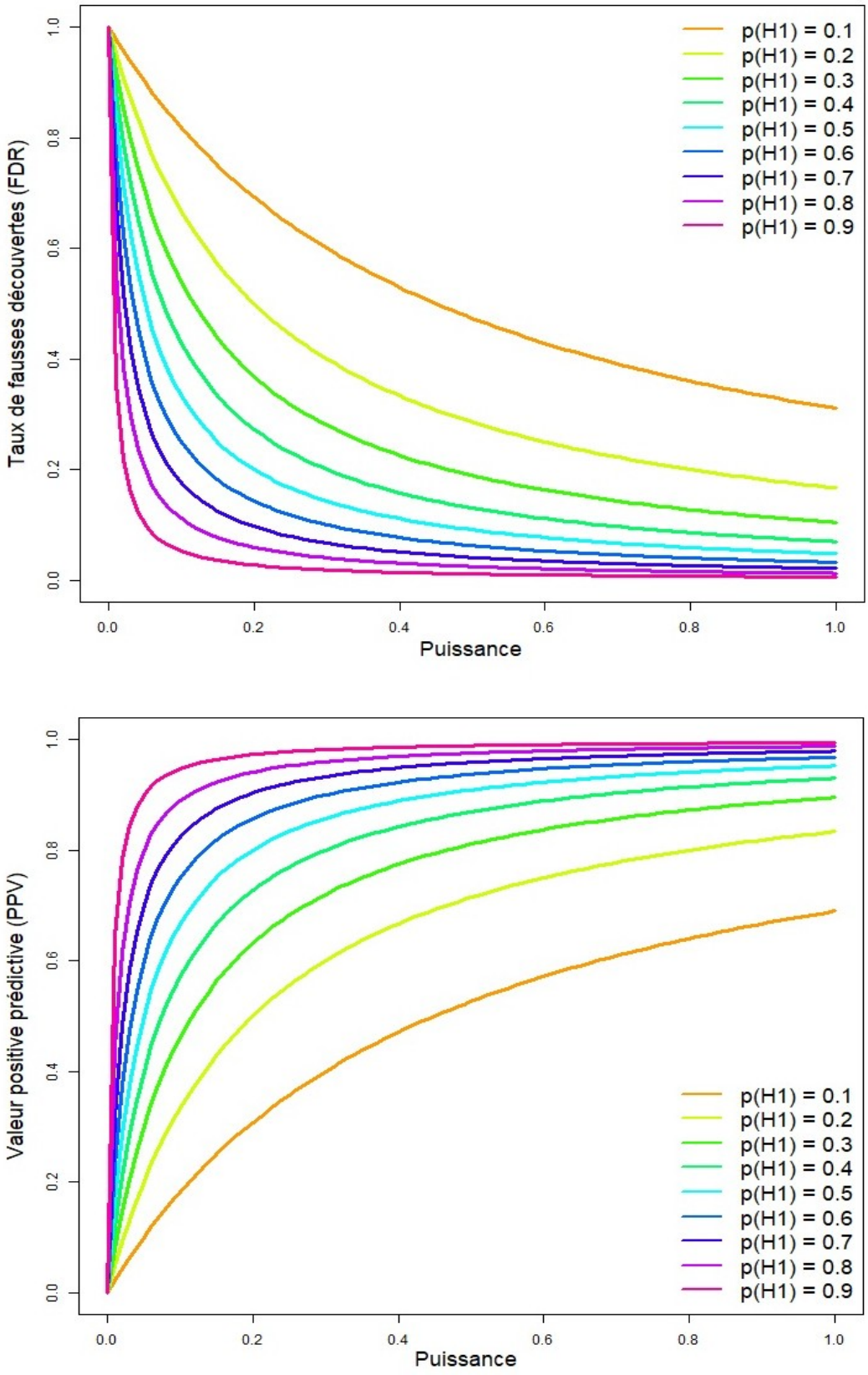

Figure 10: Comportement des courbes de FDR et de TRP (PPV) en fonction de la puissance et de 9 probabilités que l'hypothèse alternative soit vraie. L'on constate que le FDR augmente fortement et que le TRP chute aux puissances basses. Inversement, plus notre puissance est importante moins l'on produit de faux résultats. Graphiques inspirés de ceux de Ioannidis (2005) et réalisé avec R.

2 Les objectifs du présent travail

Nous n'avons pas trouvé de littérature sur la fiabilité des résultats, ni même sur les puissances mobilisées en psychopharmacologie préclinique.

Par ailleurs nous disposons maintenant des outils nécessaires pour analyser la puissance et les risques de faux positifs dans beaucoup de domaines. Il reste à prouver la faisabilité de cette analyse en l'appliquant sur un domaine réduit. Nos objectifs principaux sont de :

1. Calculer les FDR et TRP pour un sous-domaine complet.
2. Présenter efficacement les résultats finaux et intermédiaires sous forme graphique.
3. Développer des techniques d'extraction efficaces et sécurisées.
4. Automatiser les calculs autant que possible.
5. Évaluer les outils et les procédures avant de les appliquer sur des ensembles de données plus vastes.

De ces objectifs principaux dérivent des questions particulières susceptibles de recevoir des réponses objectives et vérifiables, et qui seront présentées dans l'ordre suivant :

1. Définir un sous-domaine avec des articles suffisamment nombreux pour pouvoir calculer des statistiques descriptives, mais pas trop pour être analysables par une équipe très réduite. Cette première question a été vite résolue, parce que le laboratoire a travaillé dans un domaine répondant à ces critères : la préférence de lieu conditionné induite par la nicotine chez la souris (PLC_{NIC}).
2. Observer la distribution des p-valeurs dans ce sous-domaine, puisque les faux positifs et les vrais positifs ne concernent que les résultats significatifs.
3. Calculer des tailles d'effet observées (en êta-carrés partiels) et les mettre en graphique pour observer une éventuelle surestimation de la taille d'effet associée aux puissances faibles.
4. Calculer la puissance prospective des tests de la PLC_{NIC} à partir des 6 tailles d'effet conventionnelles de la classification de Cohen-Sawilowsky (Sawilowsky, 2009).
5. Calculer le taux de fausses découvertes (FDR) et le taux de vraies découvertes (TRP) des études utilisant la PLC_{NIC} à la manière de Ioannidis (2005) et pour diverses probabilités de H_1.

Nous décrivons ce sous-domaine immédiatement, pour ne pas surcharger les sections suivantes.

La préférence de lieu conditionné (PLC)

La procédure de préférence de lieu conditionné permet d'évaluer les effets hédonistes ou aversifs de diverses drogues (Prus et al., 2009). Elle utilise soit une arène divisée en 2 compartiments (A et B) pour les designs à **choix forcé**, soit une arène avec un compartiment intermédiaire pour les designs à **choix non forcé** (voir figure 11). L'un des deux compartiments (A ou B) est associé à une drogue par conditionnement, et l'autre pas. On mesure le temps que passe l'animal dans chaque compartiment et l'on en déduit sa préférence ou son aversion éventuelle pour le compartiment associé à la drogue.

Dans la **procédure non-biaisée**, le compartiment associé à la drogue est choisi aléatoirement pour chaque souris. On doit utiliser au moins deux groupes d'animaux pour contrebalancer une éventuelle préférence des animaux pour l'un des deux compartiments. Dans cette procédure les expérimentateurs ne tiennent pas compte des préférences individuelles des souris pour un compartiment ou l'autre avant le conditionnement.

Dans la **procédure biaisée**, on tient compte de la préférence individuelle de chaque souris pour l'un ou l'autre compartiment. Les souris reçoivent la drogue dans le compartiment le « moins préféré » si l'on teste une drogue pour ses effets hédonistes, ou dans le compartiment préféré si l'on veut étudier des effets aversifs.

Le bas de la figure illustre la différence entre procédures biaisée et non-biaisée pour une procédure à choix non-forcé.

On mesure le temps passé par chaque souris dans chaque compartiment. Certaines études utilisent ces temps bruts tandis que d'autres utilisent des scores de préférence, comme la différence entre le temps passé dans chaque compartiment avant et après le conditionnement ou le pourcentage de temps passé dans chaque compartiment.

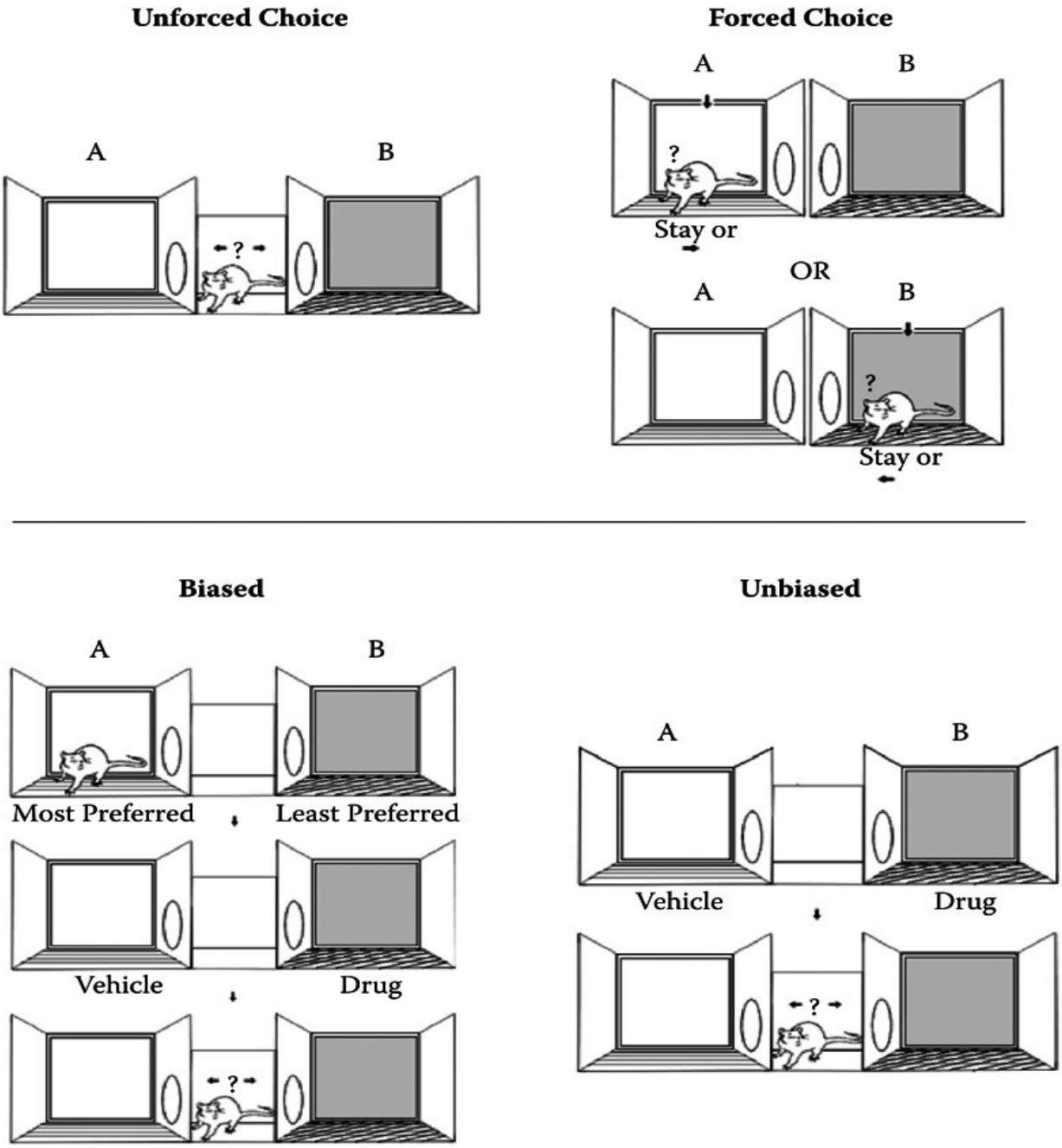

Figure 11: Préférence de lieu conditionné. En haut : différence entre les arènes à choix forcé et les arènes à choix non-forcé. Dans les procédures à choix non-forcé, les souris en phase de test sont placées dans le compartiment central. Dans les procédures à choix forcé les souris en phase de test sont placées dans l'un des deux compartiments (associé ou non à la drogue). En bas: différence entre les procédures biaisées et non-biaisées dans une procédure à choix non-forcé. Repris de Prus et al dans Buccafusco JJ (2009).

3 Méthode

Nous étudierons la distribution des p-valeurs, la taille d'effet et son évolution, la puissance mobilisée, et le risque de faux positifs, dans une sélection d'articles sur la PLC_{NIC} répertoriés dans PubMed .

3.1 Recherche des articles dans la littérature

La recherche de la littérature relative à la PLC_{NIC} s'est basée uniquement sur la base de données PubMed qui est la principale base de données pour les recherches cliniques et précliniques. Il est donc fort probable qu'elle regroupe la grande majorité des articles utilisant la PLC_{NIC}.

La base de données a été interrogée le 27/08/2019 avec la chaîne de caractères « *conditioned place preference nicotine mice* ». Cette recherche a donné 160 résultats. On a ensuite appliqué le filtre « *other animals* » à cette recherche. Cela a abouti à 139 résultats, sans restriction de date.

La recherche des articles peut être traduite de la manière suivante en utilisant les termes du Mesh de Medline : « *(("conditioning (psychology)"[MeSH Terms] OR ("conditioning"[All Fields] AND "(psychology)"[All Fields]) OR "conditioning (psychology)"[All Fields] OR "conditioned"[All Fields]) AND place[All Fields] AND preference[All Fields] AND ("nicotine"[MeSH Terms] OR "nicotine"[All Fields]) AND ("mice"[MeSH Terms] OR "mice"[All Fields])) AND "animals"[MeSH Terms:noexp]* ».

3.2 Processus de sélection des articles

L'ensemble des résultats de PubMed a été exporté dans un fichier au format csv puis ouvert dans LibreOffice Calc (v6). Nous en avons tiré un nouveau fichier qui a servi pour l'extraction. Le Tableau 7 décrit les colonnes de ce fichier.

Tableau 7: Données de sélection des articles

Colonnes originales de PubMed	
Nom	Type
titre	texte libre
auteurs	texte libre
année	nombre entier
journal	texte libre
URL	lien cliquable
Colonnes ajoutées	
identifiant	nombre allant de 1 à 139
inclusion	booléen (oui = 1 ; non = 0) ;
raison d'exclusion	texte libre
commentaires	texte libre

Les critères d'inclusion étaient ceux du Tableau 8

Tableau 8 : Critères d'inclusion des articles

Critères d'inclusion
être rédigé en anglais ou en français
utiliser des souris (toutes souches)
tenter de produire une préférence de lieu à la nicotine
évaluer l'effet de la nicotine, et ou l'effet de la dose de nicotine, et ou l'effet de la souche
utiliser au moins une statistique F ou une statistique t
reporter les degrés de liberté choisis ou la p-valeur obtenue
reporter le mode d'administration de la nicotine et la ou les doses utilisées.

L'inclusion des articles s'est décidée sur base de la lecture de la méthode et des résultats des articles, sauf quand une lecture plus partielle suffisait à établir une cause d'exclusion. Cette sélection a été réalisée uniquement par moi-même. Une double sélection par une personne supplémentaire avec une mise en commun des inclusions et discussion en cas de désaccord n'a pas pu être mise en place.

Plusieurs tests publiés dans un même article pouvaient être inclus dans l'échantillon, pour autant qu'ils répondent aux exigences précitées.

3.3 Extraction des données

L'extraction des données a été réalisée manuellement dans les mêmes conditions que la sélection.

Les données extraites des articles ont été encodées dans LibreOffice Calc (v6), avec le système d'exploitation Linux (version Ubuntu 18.04 Bionic Beaver). Le fichier était régulièrement enregistré sous un nouveau nom, pour limiter le risque de pertes de données en cas de panne ou de fausse manœuvre. Il y avait des copies de sauvegarde sur un disque dur externe.

Les colonnes du fichier d'extraction sont décrites dans le Tableau 9 .

Tableau 9:Colonnes de la grille d'extraction

Nom	Type	Commentaire
identifiant du test	Texte calculé	identifiant de l'article + « - » + numéro du test
identifiant de l'article	entier	
numéro du test	entier	
inclusion	booleen	
commentaires	texte libre	
auteurs	texte libre	copié-collé du tableau de sélection
abréviation du journal	texte libre	
année de publication	entier	
CPP choice	Booleen (Forced/Unforced)	design à choix forcé ou non
CPP bias	Booleen (Biased/Unbiaised)	design tenant compte de la préférence de chaque souris (Biased) ou non (Unbiased)
CPP : preference score	texte libre	Mode de calcul du score de préférence
Test	texte libre	type de test effectué (effet de la nicotine, de la dose…). Cette variable est un aide mémoire. Elle n'est pas utilisée pour le traitement des données
Dose	texte libre	dose de nicotine injectée en mg/kg. Lorsque le test évalue les effets de différentes doses cette variable prend la valeur « multiple »
Souche de souris utilisée	texte libre	
Sexe des souris utilisées	catégoriel	Mâle, femelle, mixte
Souris/groupe	entier	nombre de souris dans chaque groupe. Il s'agit souvent d'une valeur calculée à partir des

		degrés de liberté, parce que les auteurs reportent très rarement les effectifs par groupe. En cas d'effectifs inégaux, on retient la moyenne par groupe
Stat utilisée	catégorielle (t / F / unspecified)	cette variable contient le type de statistique qui est utilisée (t ou F). Si le test reporte uniquement une p-valeur, cette variable prend la valeur « unspecified »
F reporté	nombre	valeur de la statistique F
df num	entier	degrés de liberté du numérateur. Si ce degré de liberté n'est pas fourni, il est recalculé à partir du nombre de groupes dans l'ANOVA
df den	entier	degrés de liberté du dénominateur. S'il n'est pas reporté, il est recalculé à partir des effectifs et du plan expérimental
p-valeur reportée	nombre	
p-valeur calculée	nombre	seulement s'il n'y a pas de p-valeur reportée
t reporté	nombre	
df	entier	degrés de liberté. Si le df n'est pas fourni dans l'article, il est déduit à partir de l'effectif plan expérimental
p-valeur reportée	nombre	
p-valeur calculée	nombre	Seulement s'il n'y a pas de p-valeur reportée

3.4 Analyse des données

Calcul des valeurs manquantes

Pour chaque test qui ne les mentionnait pas, nous avons calculé la statistique (F ou t), les degrés de liberté, les tailles d'échantillons, et la p-valeur.

Calcul des tailles d'effet

Pour les tests dont la taille d'effet était calculable (c'est-à-dire quand nous disposions des données nécessaires), nous avons calculé les pourcentages de variance expliquée, en utilisant les êta-carrés

partiels. Ils ont l'avantage d'être applicables pour les tests t comme pour les tests F, et pour les plan-expérimentaux intra-sujet comme inter-sujets. On les trouve avec les formules suivantes :

$$\eta_p^2 = \frac{F * df_{effet}}{F * df_{effet} + df_{erreur}} \qquad \eta_p^2 = \frac{t^2}{t^2 + df_{erreur}}$$

Il faut noter que l'êta-carré partiel s'interprète en termes de proportion de variance expliquée, alors que le d de Cohen s'interprète comme une distance standardisée. Il faut donc effectuer une conversion pour obtenir les êta-carré partiels correspondant aux tailles d'effets de la classification de Cohen-Sawilowsky. Cette conversion s'effectue avec la formule suivante, qui donne l'êta-carré à partir du f de Cohen :

$$\eta_p^2 = \frac{f^2}{f^2 + 1}$$

Le lecteur intéressé trouvera les chiffres obtenus dans la Figure 16 .

Nous avons développé une fonction de calcul dans R pour obtenir les êta-carrés partiels à partir des données disponibles.

Calcul des puissances

Pour chaque test, nous avons calculé avec R, via le package WebPower, les puissances pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky, à savoir d = 0.01, d = 0.2, d = 0.5, d = 0.8, d = 1.2, et d = 2 pour les tests t, ou leur équivalent en f de Cohen pour les tests F.

Nous calculerons aussi une corrélation entre les différentes puissances et l'année de publication.

Visualisation des tailles d'effet

Avec R et sa librairie ggplot2, nous montrerons la relation entre les tailles d'effet observées et les puissances calculées, et nous ajouterons une droite de régression avec son intervalle de confiance à 95 %.

Nous calculerons aussi la corrélation entre les tailles d'effet observées d'une part et d'autre part les puissances calculées avec les tailles d'effet hypothétiques de la classification de Cohen-Sawilowski.

Les graphiques de distribution des puissances

Les calculs de puissances s'effectuent différemment pour les tests t et pour les tests F et seront présentés successivement de deux manières, d'une part avec des graphiques en violons et boites à moustaches qui nous montrent la médiane, l'intervalle interquartile, le minimum, le maximum, et la

distribution des données, et d'autre part sous forme de tableaux résumés.

Les tableaux résumés donneront les médianes ainsi que les premiers et troisièmes quartiles des puissances calculées, pour l'ensemble des tests considérés et pour les six tailles d'effet envisagées.

Les tailles d'effet minimales détectables par les tests t

La taille des échantillons mobilisés pour les tests t limite les tailles d'effet que le test peut détecter avec une puissance de 0,8. Nous calculerons successivement la taille d'effet minimale détectable avec une puissance de 80 % par les tests t pour échantillons indépendants et par les tests t pour échantillons appariés.

Calcul des TRP et FDR

Le calcul des TRP et FDR se base sur les puissances médianes. Nous reporterons les puissances médianes, car il est peu probable que les puissances suivent une distribution normale, puisque l'étendue des valeurs possibles est limitée. Les calculs seront effectués sur R et les résultats enregistrés dans un fichier Rmarkdown (disponible sur demande). Le TRP et le FDR étant des fonctions à 3 paramètres (puissance, seuil de significativité et plausibilité), nous ne reporterons pas une valeur mais bien un graphique représentant les résultats pour l'ensemble des plausibilités allant de 0 à 1 comme décrit dans Ioannids (2005). Les courbes de TRP et FDR pour la puissance médiane seront entourées par leur intervalle interquartile et comparées à la courbe correspondant à une puissance de 80 %.

Nous proposerons une autre série de graphiques représentant les résultats selon la probabilité de H_1.

4 Résultats

Les données sont disponibles sur demande au format Rdata.

4.1 La sélection des études

La réponse de PubMed donnait 139 références. Parmi ces 139 références, 8 ont été exclues parce qu'elles ne correspondaient pas à des articles empiriques (livres, revue de questions, articles d'opinion). Un article a été exclu parce que rédigé en japonais.

Parmi les 130 articles restants, 8 ont été exclus parce qu'ils avaient utilisé des rats et non des souris et 34 parce qu'ils ne reportaient pas de PLC_{NIC} ou parce qu'ils ne comportaient pas de comparaison solution saline versus nicotine. Enfin 41 ont été exclus parce qu'ils ne fournissaient pas les données nécessaires aux calculs de puissance. Il nous restait donc 48 articles.

Le Tableau 10 résume la sélection des articles. On voit que 41 articles sur 88 (46,7 %) ne rapportaient pas assez de données pour permettre les calculs.

Tableau 10 : Cheminement de la sélection des articles

Critère	Éliminés	Restants
Trouvés par pubMed	/	139
Articles de recherche	8	131
En français ou en anglais	1	130
Sur des souris	8	122
Avec PLC induite par la nicotine et comparaison salin/nicotine	34	88
Rapportant les données nécessaires aux calculs	**41**	**47**

Les 48 articles inclus contenaient au total 109 tests statistiques utilisables pour nos analyses.

4.2 Les caractéristiques des études incluses

Journaux

L'ensemble des articles inclus provient de 20 journaux, dont les plus prestigieux du domaine. Le Tableau 11 donne la liste des journaux représentés dans l'échantillon retenu.

Tableau 11: Liste des journaux présents dans l'échantillon d'articles retenus

Journal
Neuropharmacolgy
Psychopharmacology (Berl)
Pharmacology Biochimy and Behavior
Progress in Neurpsychopharmacology & biological psychiatry
Journal of Neuroscience
Behavioural Neuroscience
Genes Brain and Behaviour
Journal of Pharmacology and Experimental Therapeutics
Neuropsychopharmacolgy
PloS One
Behavioral brain research
European Journal of pharmacology
Human Molecular Genetics
Journal of Nippon Medecin School
Neurotoxicology and theratology
Proceedings of the National Academy of Sciences of the United States of America
Learning and Memory
Cell report
Drug and Alcohol Dependence
Neuroscience Letters

Articles inclus

Le Tableau 12 reprend les auteurs, les revues, les années de publication des articles retenus ainsi que les types de PLC (forcé ou non, biaisé ou non) utilisés.

Tableau 12 : Caractéristiques des études incluses dans les analyses

Id	Auteurs	Journal	Année	Choix	Biais
1-1	Canseco-Alba A, Schanz N, Sanabria B, Zhao J, Lin Z, Liu QR, Onaivi ES.	Behav Brain Res	2019	Forcé	Biaisé
4-1	Briggs SB, Hafenbreidel M, Young EJ, Rumbaugh G, Miller CA.	Learn Mem	2018	Non forcé	Non biaisé
5-1	Bagdas D, Alkhlaif Y, Jackson A, Carroll FI, Ditre JW, Damaj MI.	Neuropharmacology	2018	Non forcé	Non biaisé
6-1	Parker RL, O'Neill HC, Henley BM, Wageman CR, Drenan RM, Marks MJ, Miwa JM, Grady SR, Lester HA.	PLoS One	2017	Non forcé	Biaisé
8-1	Peng C, Engle SE, Yan Y, Weera MM, Berry JN, Arvin MC, Zhao G, McIntosh JM, Chester JA, Drenan RM.	PLoS One	2017	Non forcé	Biaisé
10-1	Xia L, Nygard SK, Sobczak GG, Hourguettes NJ, Bruchas MR.	Cell Rep	2017	Forcé	Non biaisé
13-1	Jackson A, Bagdas D, Muldoon PP,	Neuropharmacol	2017	Non forcé	Non biaisé

Id	Auteurs	Journal	Année	Choix	Biais
	Lichtman AH, Carroll FI, Greenwald M, Miles MF, Damaj MI.	ogy			
14-1	Liu Y, Harding M, Dore J, Chen X.	Prog Neuropsychopharmacol Biol Psychiatry	2017	Non forcé	Non biaisé
30-1	Kutlu MG, Ortega LA, Gould TJ.	Behav Neurosci	2015	Forcé	Biaisé
33-1	Bagdas D, Muldoon PP, Zhu AZ, Tyndale RF, Damaj MI.	Neuropharmacology	2014	Non forcé	Non biaisé
34-1	Ise Y, Mori T, Katayama S, Suzuki T, Wang TC.	J Nippon Med Sch	2014	Forcé	Unspecified
36-1	Bernardi RE, Spanagel R.	Behav Brain Res	2014	Non forcé	Biaisé
39-1	Kotagale NR, Walke S, Shelkar GP, Kokare DM, Umekar MJ, Taksande BG.	Behav Brain Res	2014	Non forcé	Biaisé
40-1	Harenza JL, Muldoon PP, De Biasi M, Damaj MI, Miles MF.	Genes Brain Behav	2014	Non forcé	Non biaisé
41-1	Titomanlio F, Perfumi M, Mattioli L.	Psychopharmacology (Berl)	2014	Forcé	Non biaisé
45-1	Bernardi RE, Spanagel R.	Drug Alcohol Depend	2013	Non forcé	Biaisé
50-1	Jackson KJ, Wang JB, Barbier E, Damaj MI, Chen X.	Neurosci Lett	2013	Forcé	Non biaisé
51-1	Ignatowska-Jankowska BM, Muldoon PP, Lichtman AH, Damaj MI.	Psychopharmacology (Berl)	2013	Non forcé	Non biaisé
63-1	Jackson KJ, McLaughlin JP, Carroll FI, Damaj MI.	Psychopharmacology (Berl)	2013	Non forcé	Non biaisé
65-1	Smith JS, Schindler AG, Martinelli E, Gustin RM, Bruchas MR, Chavkin C.	J Neurosci	2012	Non forcé	Non biaisé
67-1	Lee AM, Messing RO.	Proc Natl Acad Sci U S A	2011	Forcé	Non biaisé
69-1	McGranahan TM, Patzlaff NE, Grady SR, Heinemann SF, Booker TK.	J Neurosci	2011	Non forcé	Non biaisé
71-1	Cahir E, Pillidge K, Drago J, Lawrence AJ.	Neuropsychopharmacology	2011	Non forcé	Biaisé
75-1	Neugebauer NM, Henehan RM, Hales CA, Picciotto MR.	Pharmacol Biochem Behav	2011	Non forcé	Biaisé
81-1	Damaj MI, Grabus SD, Navarro HA, Vann RE, Warner JA, King LS, Wiley JL, Blough BE, Lukas RJ, Carroll FI.	J Pharmacol Exp Ther	2010	Non forcé	Non biaisé
83-1	Jackson KJ, Marks MJ, Vann RE, Chen X, Gamage TF, Warner JA, Damaj MI.	J Pharmacol Exp Ther	2010	Non forcé	Non biaisé

Id	Auteurs	Journal	Année	Choix	Biais
90-1	Jackson KJ, Walters CL, Miles MF, Martin BR, Damaj MI.	Neuropharmacology	2009	Non forcé	Non biaisé
92-1	Trigo JM, Zimmer A, Maldonado R.	Neuropharmacology	2009	Non forcé	Non biaisé
95-1	Brunzell DH, Mineur YS, Neve RL, Picciotto MR.	Neuropsychopharmacology	2009	Non forcé	Non biaisé
98-1	Mineur YS, Brunzell DH, Grady SR, Lindstrom JM, McIntosh JM, Marks MJ, King SL, Picciotto MR.	Genes Brain Behav	2009	Non forcé	Non biaisé
99-1	Portugal GS, Gould TJ.	Pharmacol Biochem Behav	2009	Forcé	Biaisé
101-1	TammimÃ¤ki A, Chistyakov V, Patkina N, Skippari J, Ahtee L, Zvartau E, MÃ¤nnistÃ¶ PT.	Eur J Pharmacol	2008	Forcé	Biaisé
102-1	Rauhut AS, Hawrylak M, Mardekian SK.	Pharmacol Biochem Behav	2008	Forcé	Biaisé
104-1	Merritt LL, Martin BR, Walters C, Lichtman AH, Damaj MI.	J Pharmacol Exp Ther	2008	Non forcé	Non biaisé
106-1	Kota D, Martin BR, Damaj MI.	Psychopharmacology (Berl)	2008	Non forcé	Non biaisé
108-1	Zhu H, Lee M, Agatsuma S, Hiroi N.	Hum Mol Genet	2007	Non forcé	Non biaisé
109-1	Sahraei H, Aliabadi AA, Zarrindast MR, Ghoshooni H, Nasiri A, Barzegari-Sorkheh AA, Yari M, Zardooz H, Hossein-Mardi L, Faraji N, Shams J.	Eur J Pharmacol	2007	Forcé	Non biaisé
110-1	Nolley EP, Kelley BM.	Neurotoxicol Teratol	2007	Forcé	Biaisé
113-1	CastaÃ±Ã© A, Soria G, Ledent C, Maldonado R, Valverde O.	Neuropharmacology	2006	Non forcé	Non biaisé
114-1	Korkosz A, Zatorski P, Taracha E, Plaznik A, Kostowski W, Bienkowski P.	Prog Neuropsychopharmacol Biol Psychiatry	2006	Forcé	Biaisé
116-1	Grabus SD, Martin BR, Brown SE, Damaj MI.	Psychopharmacology (Berl)	2006	Non forcé	Non biaisé
117-1	Walters CL, Brown S, Changeux JP, Martin B, Damaj MI.	Psychopharmacology (Berl)	2006	Non forcé	Non biaisé
122-1	Berrendero F, MendizÃ¡bal V, Robledo P, Galeote L, Bilkei-Gorzo A, Zimmer A, Maldonado R.	J Neurosci	2005	Non forcé	Non biaisé
123-1	Sahraei H, Falahi M, Zarrindast MR, Sabetkasaei M, Ghoshooni H, Khalili M.	Eur J Pharmacol	2004	Non forcé	Non biaisé

Id	Auteurs	Journal	Année	Choix	Biais
131-1	Berrendero F, Kieffer BL, Maldonado R.	J Neurosci	2002	Non forcé	Non biaisé
133-1	Castañé A, Valjent E, Ledent C, Parmentier M, Maldonado R, Valverde O.	Neuropharmacology	2002	Non forcé	Non biaisé
139-1	Risinger FO, Oakes RA.	Pharmacol Biochem Behav	1995	Forcé	Non biaisé

Designs

On voit dans le tableau 13, qui donne un comptage des designs de PLC biaisés ou non et forcés ou non utilisés dans les différents articles, que le design non-biaisé et non forcé domine largement.

L'un des 48 articles n'a pas spécifié s'il avait utilisé un design forcé ou non.

Tableau 13 : Caractéristiques des procédures de PLC

	Non-biaisé	Biaisé	Non-spécifié	Total
Non-forcé	26	7	0	33
Forcé	6	7	1	14
Total	32	14	1	48

Dates de publication

La Figure 12 montre la répartition, par année de publication, des 139 études fournies pas PubMed.

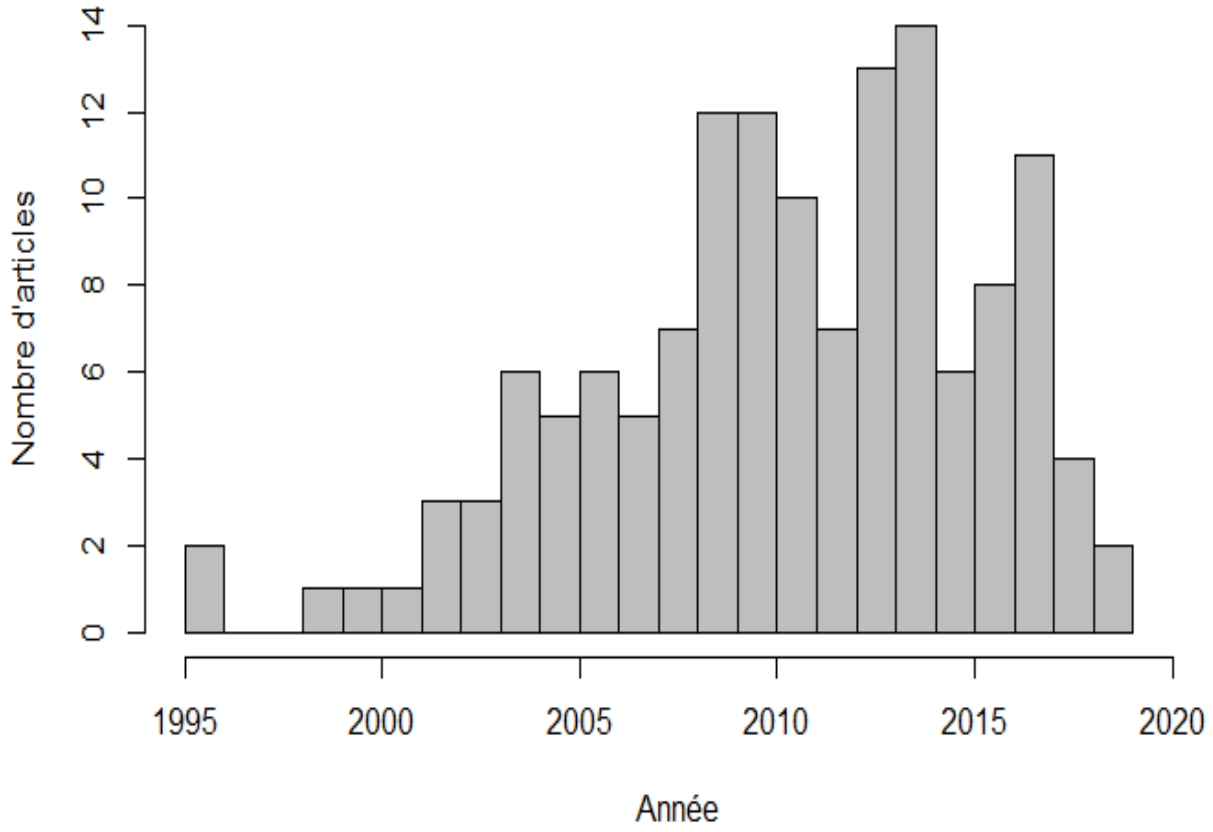

Figure 12: Répartition par année de l'ensemble des 139 articles trouvés sur PubMed (inclus et exclus de nos analyses)

Elles se répartissent sur 24 années (de 1995 à 2019) et montrent l'intérêt croissant puis décroissant accordé à la PLC_{NIC}.

La répartition, par année de publication, des 39 articles retenus pour nos analyses est représentée dans la Figure 13.

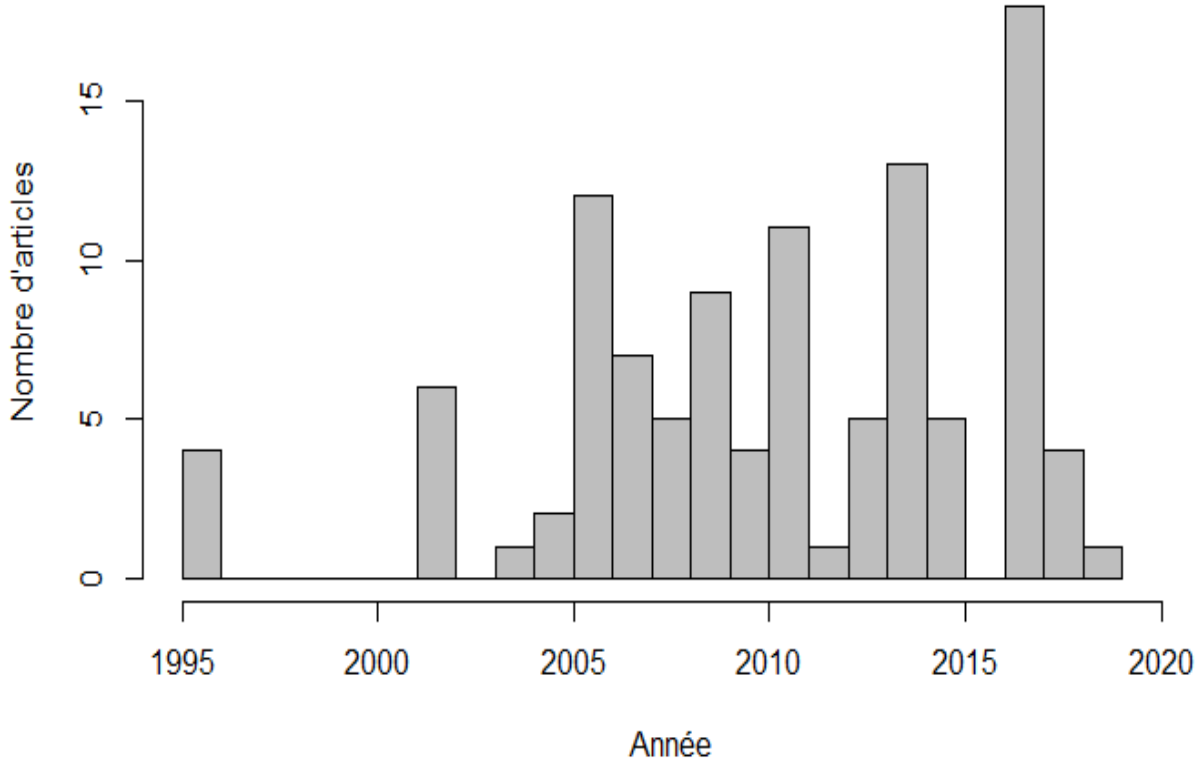

Figure 13: Répartition des études incluses dans nos analyses par année de publication

Aucun article publié dans les années 1996 à 2002 ou en 2015 n'a rencontré nos critères d'inclusion. Le maximum d'études incluses pour une année est de 7 (2013).

Les analyses suivantes concernent les 109 expériences incluses, et non plus les articles.

4.3 La distribution des p-valeurs

Pour l'ensemble des tests retenus, la plupart des p-valeurs sont inférieures à 0,05 comme le montre la Figure 14.

Si l'on se concentre sur les tests significatifs (Figure 15) on remarque qu'il n'y a pas de pic juste en dessous de 0,05. Nous ne trouvons donc pas ce signe de p-hacking. On voit aussi qu'il y a une grande quantité de p-valeurs très significatives. La distribution des p-valeurs n'indique donc pas la présence d'un biais de publication. Elle ne l'exclut pas non plus, naturellement, puisque l'absence d'un indice ne prouve pas l'absence du « délit ».

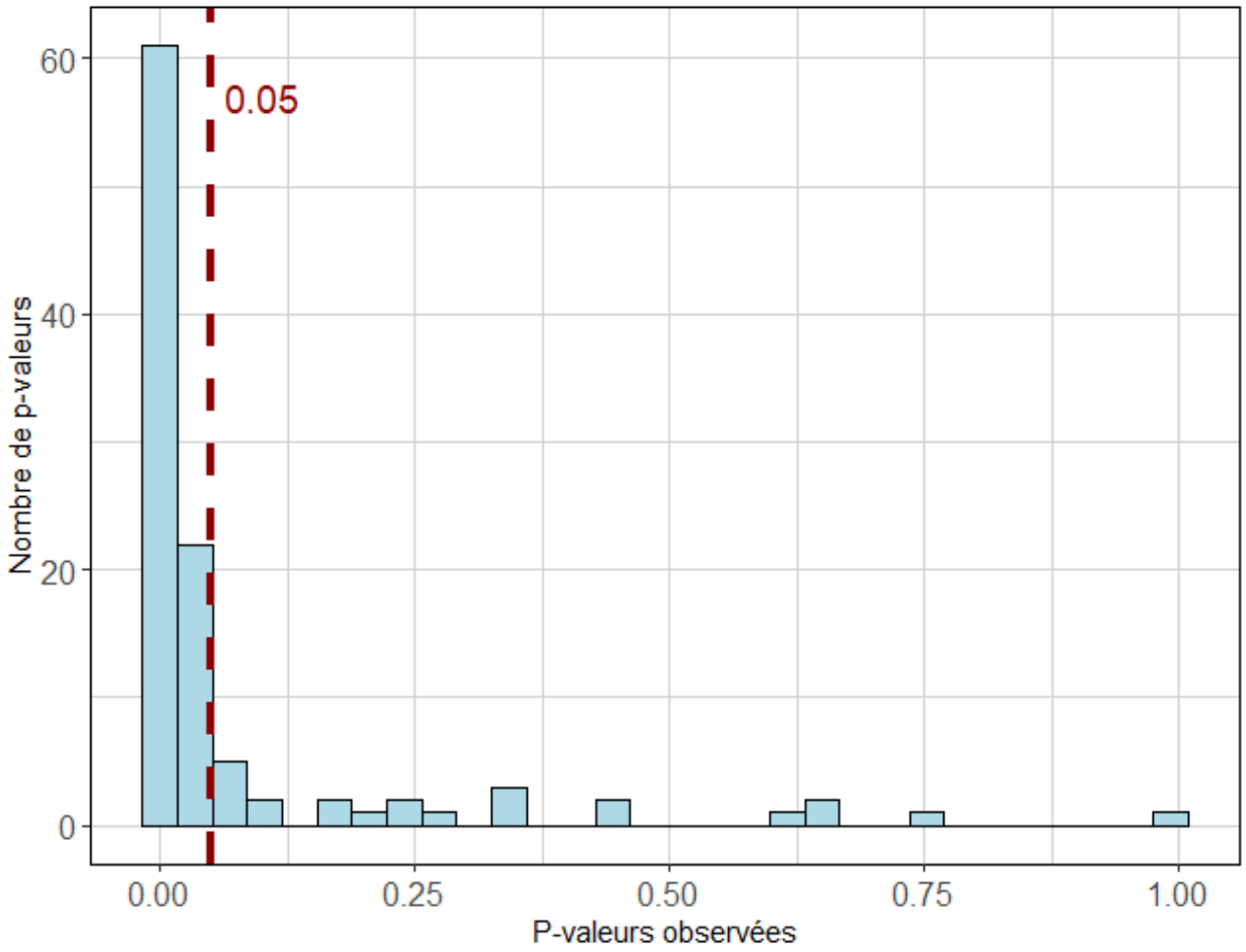

Figure 14: Distribution des p-valeurs pour l'ensemble des tests inclus (t et F). La ligne pointillée rouge indique la position du seuil de rejet à 5 %. La majorité des p-valeurs se trouvent sous se seuil.

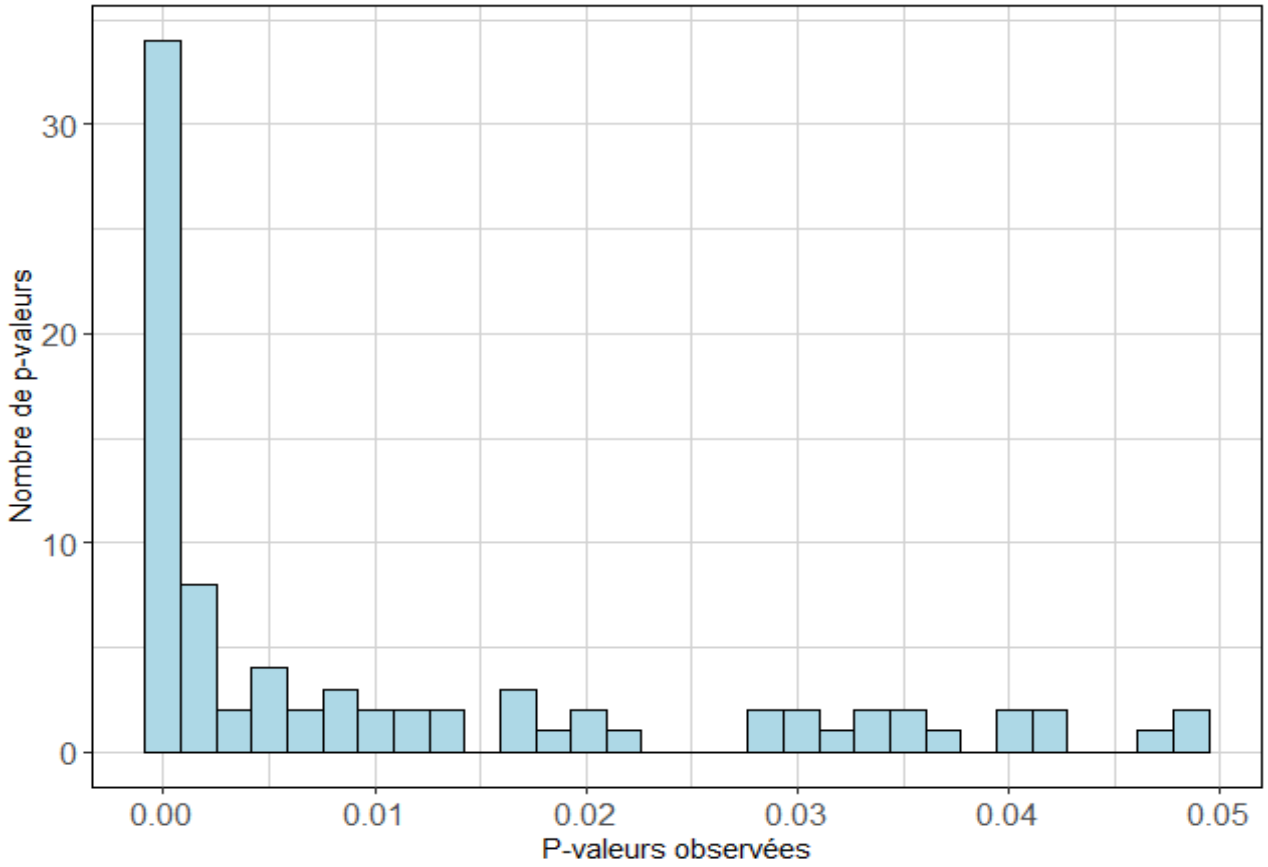

Figure 15: Distribution des p-valeurs significatives pour un seuil alpha de 5% pour les tests t et F inclus dans l'échantillons.

4.4 Les tailles d'effet

La surestimation de la taille d'effet

Pour mettre en évidence le phénomène de la surestimation de la taille d'effet, nous avons calculé les êta-carrés partiels pour les 80 tests significatifs (F et t) de notre population. Les êta-carrés partiels obtenus varient de 0,05196 à 0,94440 avec une médiane de 0,29518 et une moyenne de 0,36572.

Nos résultats montrent bien (Figure 16) la surestimation de la taille d'effet observée quand on utilise des tests manquant de puissance, et cela pour toutes les tailles d'effets sauf la plus grande (la taille d'effet énorme avec un êta-carré partiel de 0,5).

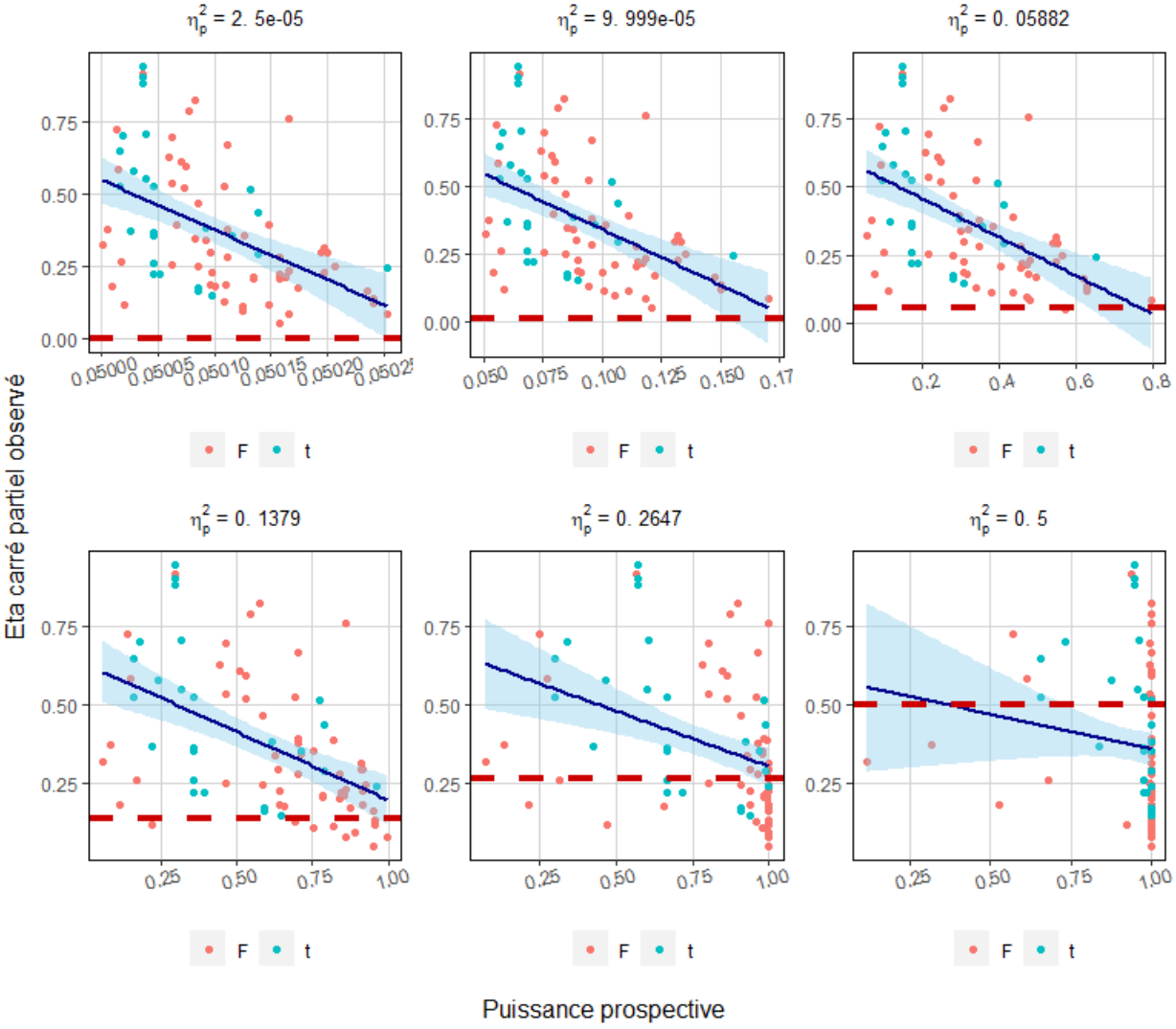

Figure 16 : Mise en évidence de la surestimation de la taille d'effet pour les tests F et t significatifs de l'échantillon (n = 80). Chaque point représente les résultats d'un test significatif. En abscisse on a la puissance prospective calculée à partir de l'effectif, et en ordonnée on a la taille d'effet observée. Chaque nuage de points représente les résultats obtenus par les calculs de puissance en supposant la taille d'effet particulière indiquée en haut du chaque graphique et représentée par la ligne pointillée rouge. En bleu on a les droites de régression et leurs intervalles de confiance.

Un groupe hors-norme

Un article (Jackson et al., 2017) a obtenu des tailles d'effets pour les tests t indépendants très supérieures à celles des autres (groupe de points bleus extrêmes dans la Figure 16).

Augmentation de la taille d'effet au cours du temps

On trouve une corrélation positive de 0,302 [0,118 ; 0,467] (r de Pearson et intervalle de confiance à 95 %) entre la taille d'effet observée et l'année de publication. Elle va dans le sens d'une surestimation grandissante de la taille d'effet avec un coefficient de détermination expliquant 9,12 % [1,39 % ; 21,81 %] de la variance.

4.5 Les puissances

Les puissances calculées sont négativement corrélées avec l'année de parution (voir Tableau 14). Cette corrélation suggère que la puissance des tests du domaine baisse avec l'année de parution.

Tableau 14 : Corrélation des puissances avec l'année de parution

Puissance pour une taille d'effet	Corrélation avec l'année de parution (r de Pearson avec leur intervalle de confiance à 95 %)
très petite	-0,413 [-0,559 ; 0,242]
petite	-0,409 [-0,556 ; -0,238]
moyenne	-0,416 [-0,561 ; -0,245]
grande	-0,409 [-0,556 ; -0,238]
très grande	-0,302 [-0,465 ; -0,119]
énorme	-0,109 [-0,293 ; 0,083]

À partir d'ici, nous envisagerons séparément la puissance des 70 tests F et des 39 tests t extraits des 39 articles retenus.

Les puissances des tests F

Les articles inclus contiennent 70 tests F.

La Figure 17 montre l'ensemble des puissances prospectives calculées pour les tests F. Elle combine des « violons » et des boîtes à moustaches.

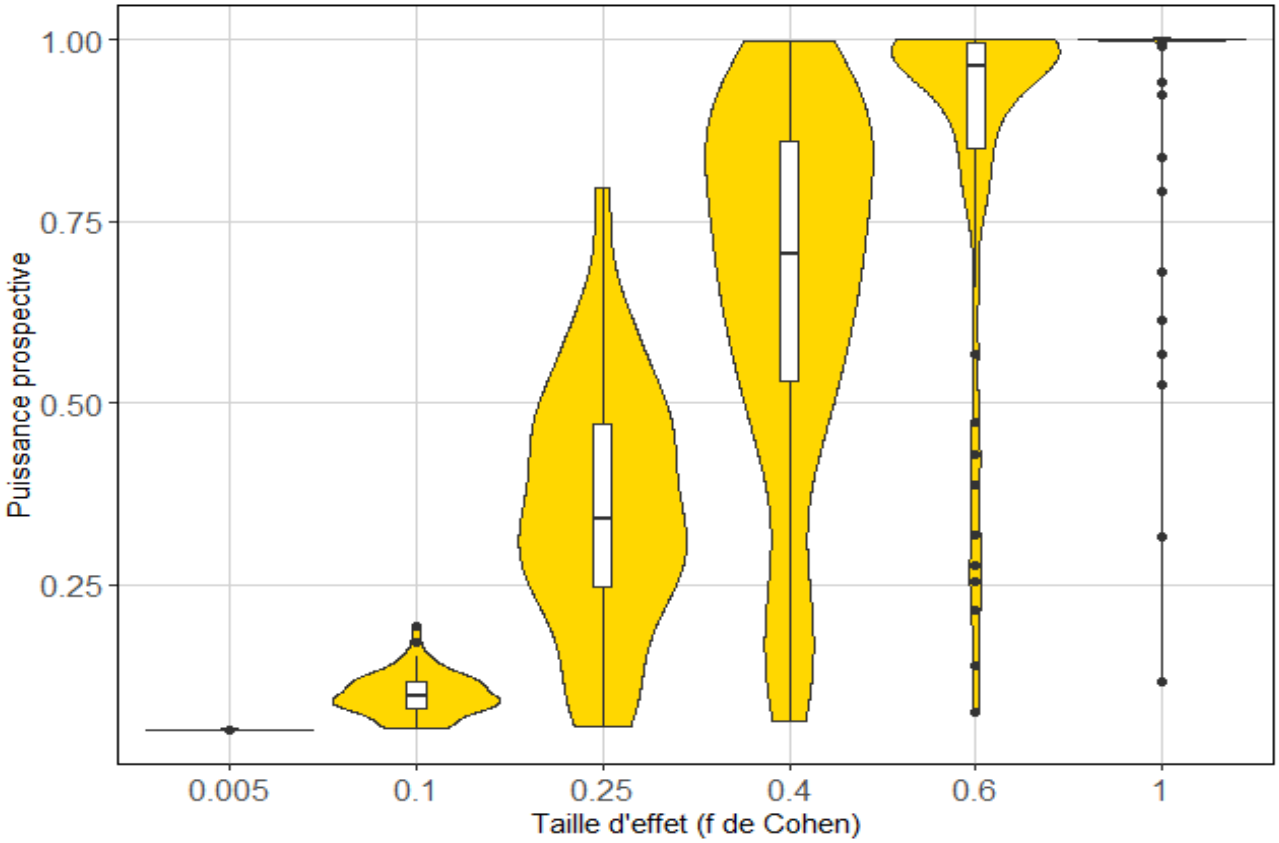

Figure 17: Distributions des puissances prospectives pour les 70 tests F. Chaque ensemble violon-boîte à moustaches représente la distribution des puissances calculées pour une taille d'effet de la classification de Cohen-Sawilowsky (en f de Cohen). Les extrémités inférieures et supérieures de la boîte à moustache représentent les premier et troisième quartiles. La ligne horizontale qui divise le rectangle représente la médiane. La moustache supérieure se prolonge jusqu'à la plus grande donnée inférieure à Q3 + 1,5(Q3-Q1). La moustache inférieure se prolonge jusqu'à la plus petite donnée comprise dans l'intervalle Q1 - 1,5(Q3-Q1).

On peut observer un effet plafond pour les tailles d'effet grandes à énormes (de 0,6 à 1 en f de Cohen), et un effet plancher pour les petites et très petites tailles d'effet (de 0,005 et 0,1). Nous observons aussi qu'un certain nombre de tests (représentés par les points) ont une puissance particulièrement basse même pour les tailles d'effets très grandes et énormes (0,6 et 1 en f de Cohen).

La puissance médiane n'atteint le seuil de 0,80 que pour deux tailles d'effet (très grande et énorme).

Elle varie de 0,09 à 0,704 pour les trois puissances de la classification de Cohen (petite, moyenne et grande, f = 0,1, 0,25, et 0,4).

Le Tableau 15 présente la distribution des puissances calculées pour les six tailles d'effet de la classification de Cohen-Sawilowsky. Il ne donne que les médianes et les quartiles parce que les distributions ne sont pas normales comme le montre la Figure 17.

Tableau 15 : Médianes, premiers et troisièmes quartiles des distributions de la puissance calculée pour les six tailles d'effet de la classification de Cohen-Sawilowsky, pour les 70 tests F retenus.

Taille d'effet	(f de Cohen)	Puissance médiane	Premier quartile	Troisième quartile
très petite	0.005	0.05011	0.05007	0.05016
petite	0.1	0.09504	0.07958	0.115
moyenne	0.25	0.3408	0.2461	0.4701
grande	0.4	0.7041	0.5292	0.8591
très grande	0.6	0.9629	0.8508	0.9952
énorme	1	1	0.991	1

L'analyse du tableau montre que le troisième quartile de la distribution des puissances calculées avec un f de 0,4 dépasse le seuil recommandé de 0,8, ce qui signifie que 25 % à 50 % des tests F inclus dans l'échantillon seraient en mesure de détecter correctement une grande taille d'effet. On voit aussi que 75 % au moins des tests F atteignent une puissance suffisante pour détecter les tailles d'effet très grandes et énormes.

Les puissances des tests t

Les articles inclus contiennent 37 tests de Student (tests t). Certains concernent des échantillons indépendants et d'autres concernent des échantillons appariés. Nous les avons analysés ensemble.

Les puissances calculées

Comme avec les tests F, nous avons calculé les puissances pour les six tailles d'effet de la classification de Cohen-Sawilowsky. Elles sont représentées dans la Figure 18.

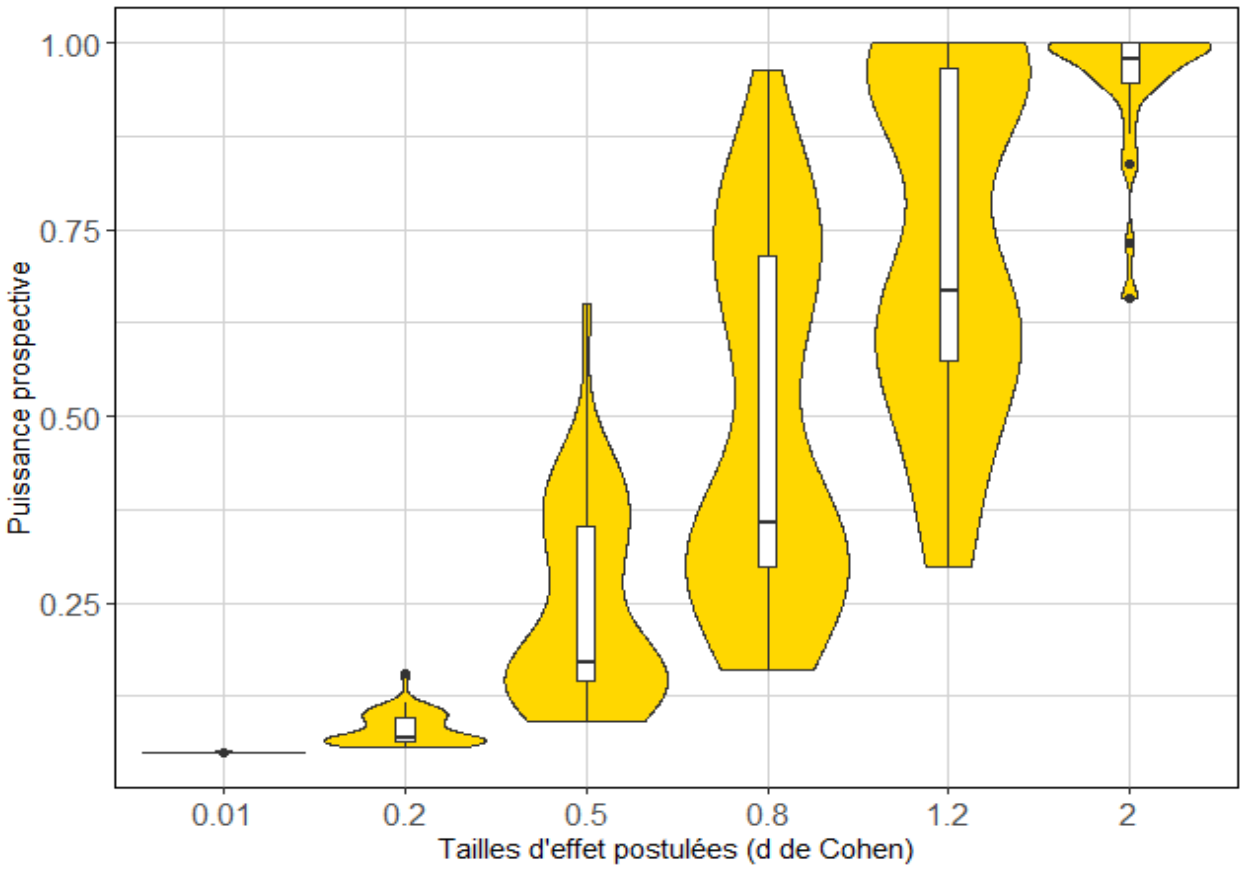

Figure 18 :Distributions des puissances prospectives pour les 37 tests t de Student retenus. Chaque ensemble violon-boîte à moustaches représente la distribution des puissances calculées pour une taille d'effet de la classification de Cohen-Sawilowsky (en d de Cohen). Les extrémités inférieures et supérieures de la boîte à moustache représentent les premiers et troisième quartiles. La ligne horizontale qui divise le rectangle représente la médiane. La moustache supérieure se prolonge jusqu'à la plus grande donnée inférieure à Q3 + 1,5(Q3-Q1). La moustache inférieure se prolonge jusqu'à la plus petite donnée comprise dans l'intervalle Q1 - 1,5(Q3-Q1).

Elle montre que la distribution des puissances pour les tests t diffère autant de la courbe normale que celle des tests F. Elle apparaît même clairement bimodale pour les tailles d'effet de 0,8 et 1,2, ce qui pourrait correspondre à la différence de puissance entre les tests pour échantillons indépendants et les tests pour échantillons appariés, la puissance ayant tendance à être plus grande pour les tests à mesures répétées.

On observe aussi moins de valeurs extrêmes pour les tests t que pour les tests F.

Le Tableau 16 présente la distribution des puissances prospectives calculées pour les six tailles d'effet de la classification de Cohen-Sawilowsky.

Tableau 16 : Médianes, premiers et troisièmes quartiles des distributions de la puissance calculée pour les six tailles d'effet de la classification de Cohen-Sawilowsky, pour les 37 tests t retenus.

Taille d'effet	d de Cohen	Puissance médiane	Premier quartile	Troisième quartile
très petite	0.01	0.05005	0.05004	0.05012
petite	0.2	0.06847	0.06487	0.0972
moyenne	0.5	0.1696	0.1459	0.3528
grande	0.8	0.3579	0.2993	0.7137
très grande	1,2	0.6667	0.5737	0.9651
énorme	2	0,9781	0.9459	1

La médiane des puissances calculées atteint le seuil de 0,8 pour toutes les tailles d'effet inférieures à 2 (d de Cohen). Le troisième quartile est inférieur à 0,8 pour les tailles d'effet jusqu'à 0,8. Il vaut 0,96 pour une taille d'effet de 1,2, et 1 pour une taille d'effet de 2.

L'analyse du tableau montre que le troisième quartile de la distribution des puissances calculées avec un d de Cohen de 1,2 dépasse le seuil recommandé de 0,8 ce qui signifie que 25 % des tests t inclus dans l'échantillon seraient en mesure de détecter correctement une très grande taille d'effet. On voit aussi que 75 % (premier quartile) au moins des tests t atteignent une puissance suffisante pour détecter les tailles d'effet énormes.

Les puissances médianes obtenues pour les tests t sont plus basses que celles obtenues pour les tests F. Une explication pourrait être que la majorité des comparaisons entre 2 groupes sont réalisées pour des comparaisons *post-hoc*.

Les tailles d'effet minimales détectables par les tests t

Comme décrit dans la méthode, nous avons décidé de baser nos calculs sur la taille médiane des échantillons plutôt que sur leur moyenne.

Le Tableau 20 nous a donné les effectifs nécessaires pour détecter différentes tailles d'effet, ou les tailles d'effet minimales détectables compte tenu de la taille des échantillons, avec une puissance de 0,8.

Pour les tests t à échantillons indépendants l'effectif médian par groupe est de huit souris. D'après le Tableau 20, cet effectif est suffisant pour détecter des tailles d'effet énormes avec une puissance de 80 %. Le calcul précis de l'analyse de puissance montre que l'on ne peut détecter que les d de Cohen supérieurs à 1,507, ce qui tombe effectivement entre les tailles d'effet très grandes et les tailles d'effet énormes.

Pour les tests t à échantillons appariés, on a un nombre de paires médian de 6,5. Comme pour les tests t indépendants, cet effectif est suffisant pour détecter des tailles d'effet énormes avec une puissance de 80 % (Tableau 20). Le calcul précis de l'analyse de puissance montre que l'on ne peut détecter que les d de Cohen supérieurs à 1,346 (d de Cohen). Vu que nous avons rarement des demi souris, nous avons refait les calculs pour 6 et 7 souris. Nous obtenons des tailles d'effets minimales détectables de 1,435 et 1,273. Ces deux chiffres tombent aussi au niveau des tailles d'effet de la catégorie « très grandes ».

4.6 Les taux de vraies et de fausses découvertes

Les taux de vraies et de fausses découvertes (FDR et TRP) n'ont de sens que pour les tests significatifs. Nous avons donc recalculé les puissances pour les seuls tests significatifs pour une probabilité alpha de 5 %. Ensuite nous avons calculé les courbes FDR et les courbes TRP.

Comme pour la puissance nous avons analysé de manière séparée les tests F et les tests t.

Pour les tests F (version Ioannidis)

Parmi les 70 tests F retenus, 57 sont significatifs au seuil alpha de 5 %. Nous avons donc une proportion de 81 % de significatifs et 19 % de non-significatifs. Cela pourrait être une indication de la présence d'un biais de publication dans le domaine. Mais cela peut aussi être un signe que l'effet recherché existe vraiment.

Nous avons recalculé les puissances médianes et leur IQR pour les seuls tests F significatifs (Tableau 17). Les résultats sont très semblables à ceux que l'on obtient pour l'ensemble des tests F, et les seules valeurs qui diffèrent sont les médianes pour les tailles d'effet de 0,1, 0,25 et 0,4, qui sont à peine supérieures. Tous les premiers et troisièmes quartiles sont identiques à ceux du Tableau 15 pour l'ensemble des tests.

*Tableau 17 : Médianes, premiers et troisièmes quartiles des distributions de la puissance calculée pour les six tailles d'effet de la classification de Cohen-Sawilowsky, pour les 57 tests F **significatifs** pour un alpha de 0,05 alpha retenus. En gras les chiffres qui diffèrent de ceux obtenus sur l'ensemble des tests retenus (Tableau 15).*

Taille d'effet	(f de Cohen)	Puissance médiane	Premier quartile	Troisième quartile
très petite	0.005	0.05011	0.05007	0.05016
petite	0.1	**0.09564**	0.07958	0.115
moyenne	0.25	**0.3445**	0.2461	0.4701
grande	0.4	**0.7046**	0.5292	0.8591
très grande	0.6	0.9629	0.8508	0.9952
énorme	1	1	0.991	1

Les FDR calculés pour les tests F significatifs sont présentés dans la Figure 19. On y observe que la proportion estimée de fausses découvertes diminue quand on suppose une taille d'effet plus importante, et aussi quand la plausibilité est plus élevée. La puissance calculée en supposant une grande taille d'effet (de f = 0,4) donne une courbe de FDR très proche de celle qu'on obtient avec une puissance de 80 %.

Les TRP calculés pour les tests F significatifs sont présentés dans la Figure 20. Pour les petites tailles d'effets (f de 0,01), on obtient une puissance calculée proche de 0,096, ce qui donne un TRP d'environ 60 % pour une plausibilité de 1, et des TRP bien plus petits pour les plausibilités moindres. On est donc bien en dessous du TRP de 95 % que l'on imagine habituellement avec un alpha de 0,05. Pour une plausibilité de 0,5 et des tailles d'effets petites, moyenne et grande, nous avons des TRP de 48,885 %, 77,505 %, et 87,572 %. Enfin pour une plausibilité de 10 % et les mêmes tailles d'effet, nous avons des TRP de 16,057 %, 40,795 %, et 58,494 %.

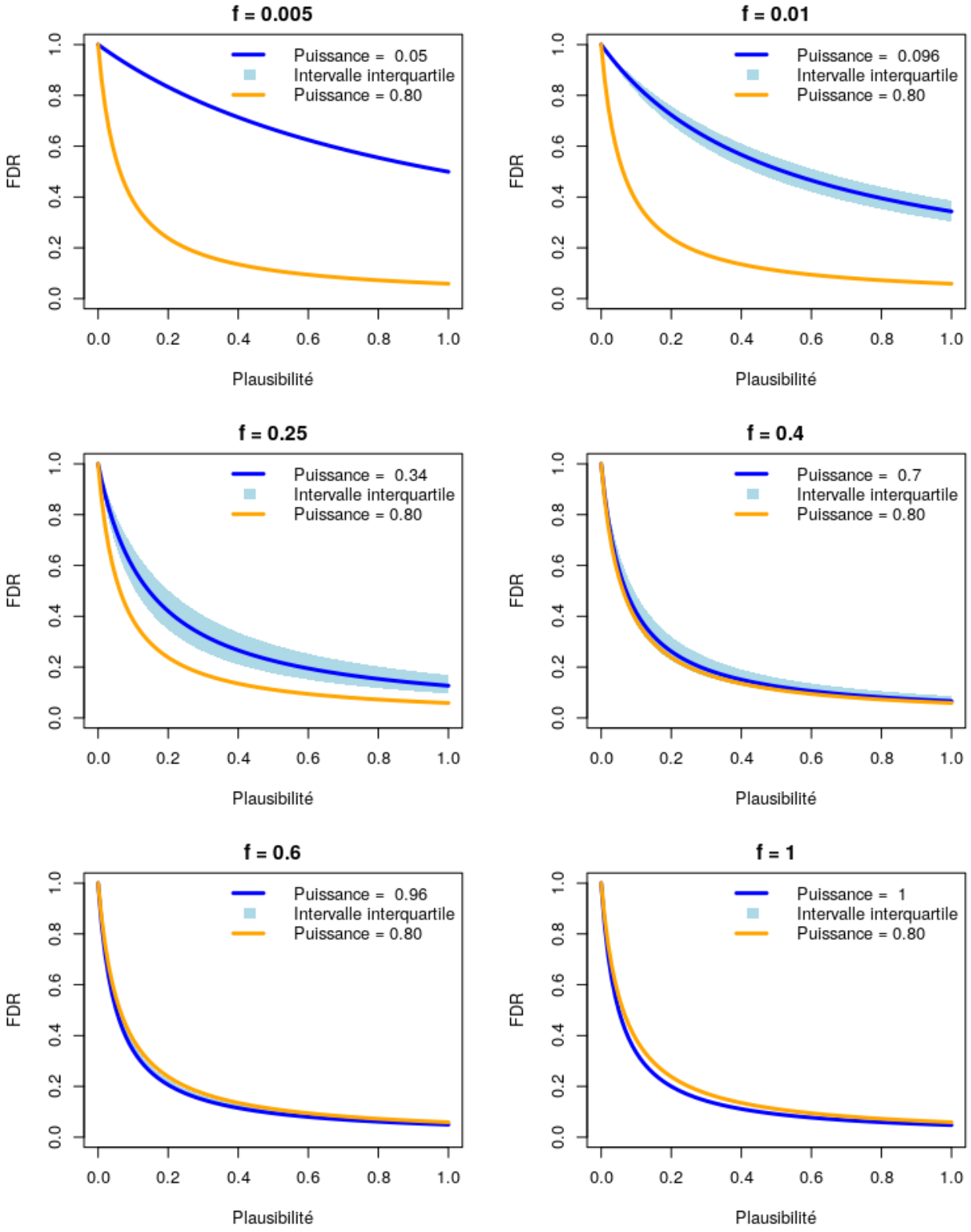

Figure 19: Courbes de FDR calculées pour les tests F pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. La courbe orange est le FDR théorique pour une étude de puissance 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (f de Cohen : 0,005, 0,01, 0,25, 0,4, 0,6, 1). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

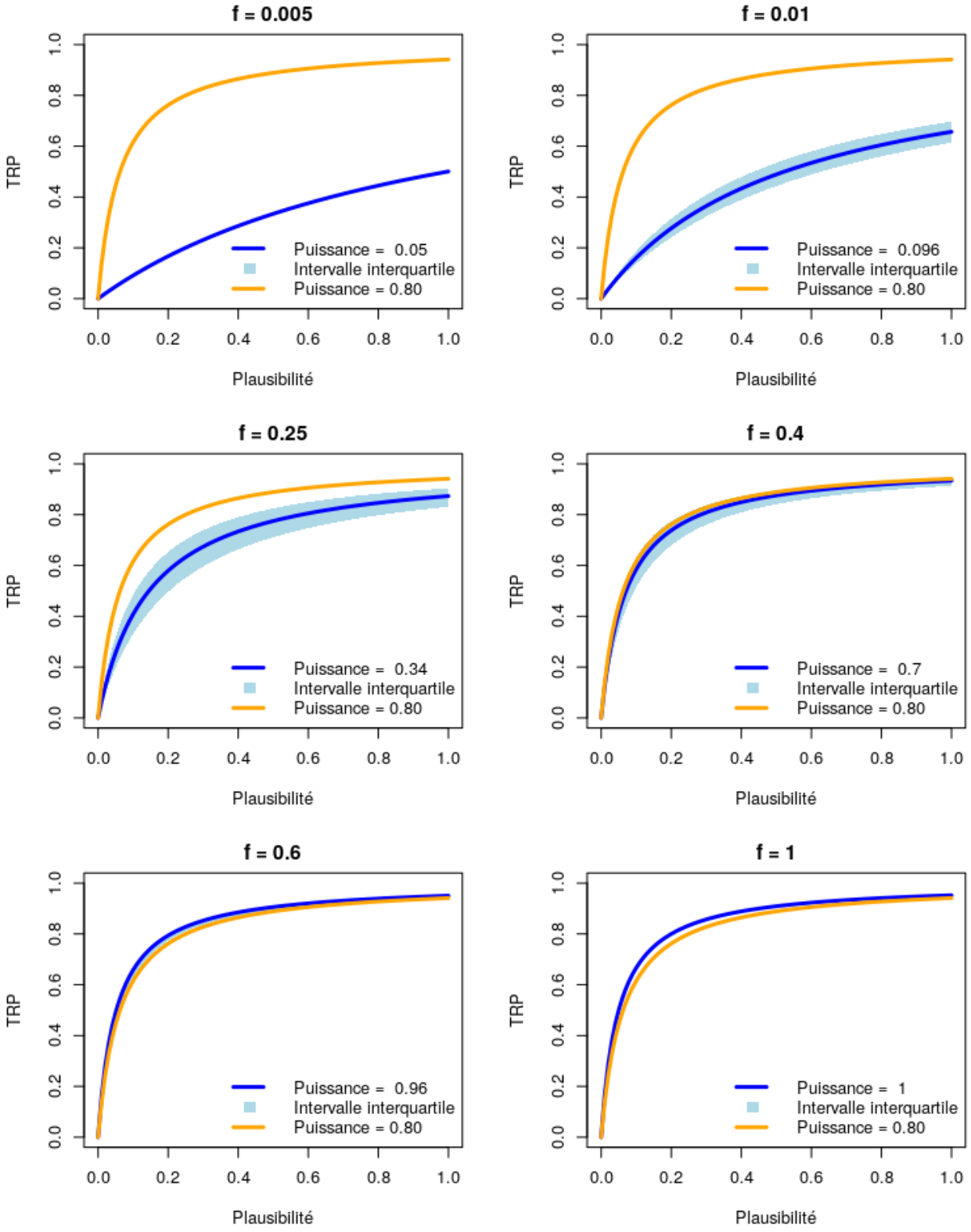

Figure 20: Courbes de TRP calculées pour les tests F pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. La courbe orange est le TRP théorique pour une étude de puissance 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (f de Cohen : 0,005, 0,01, 0,25, 0,4, 0,6, 1). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

Pour les tests F (version étendue)

Les graphiques précédents présentent toutes les valeurs possibles pour une plausibilité allant de 0 à 1. Nous ajoutons des graphiques similaires qui présentent toutes les valeurs possibles pour une probabilité de H_1 allant de 0 à 1. On voit que la courbe de FDR descend jusque 0 dans tous les cas quand la probabilité de H_1 atteint 1. Inversement, la courbe de TRP monte jusque 1 dans tous les cas quand la probabilité de H_1 atteint 1.

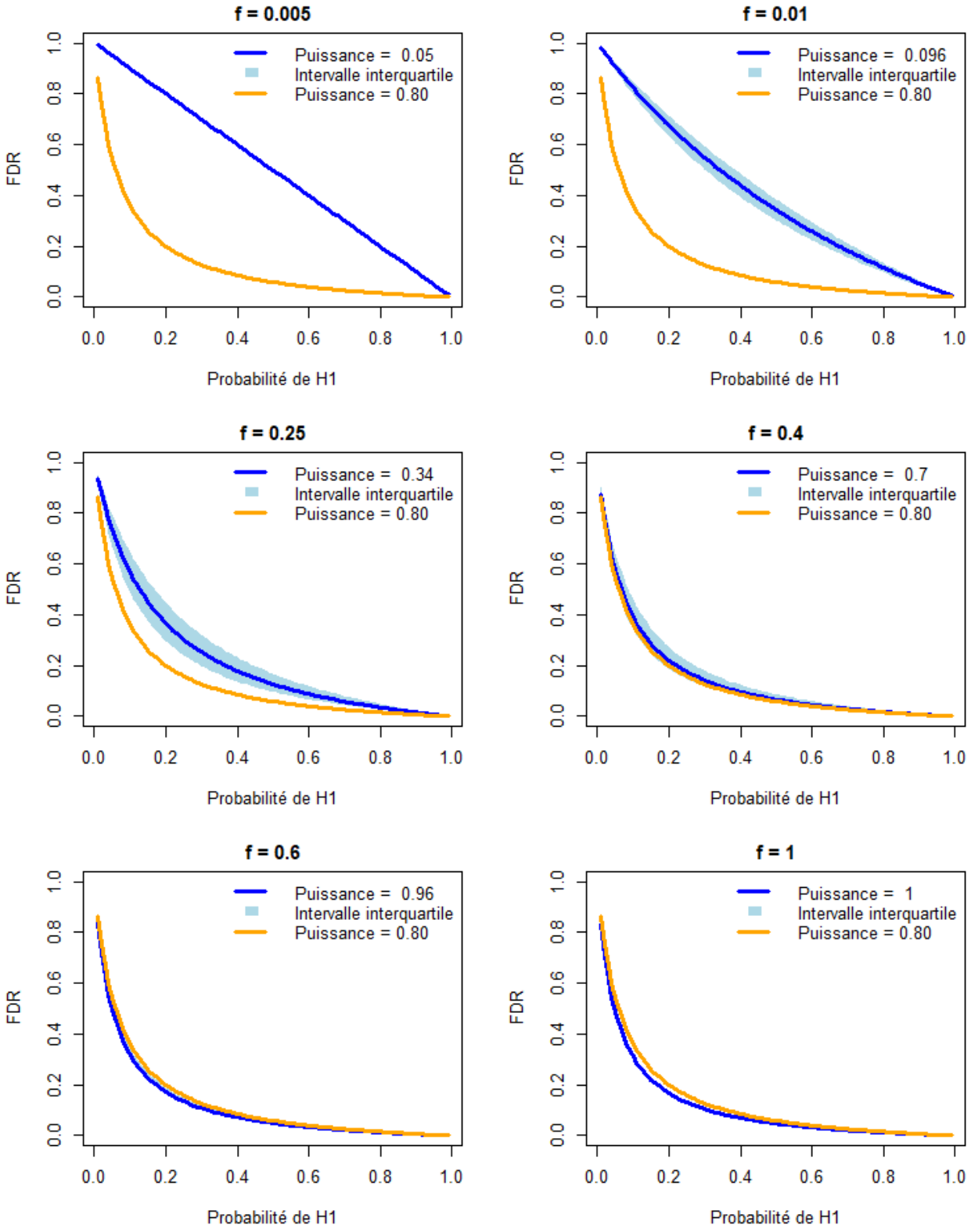

Figure 21 : Version étendue des courbes de FDR calculées pour les tests F pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. L'abscisse représente la probabilité de H_1 variant de 0 à 1. La courbe orange est le FDR pour une étude virtuelle qui a atteint une puissance de 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (d de Cohen : 0,01, 0,2, 0,5, 0,8, 1,2, 2). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

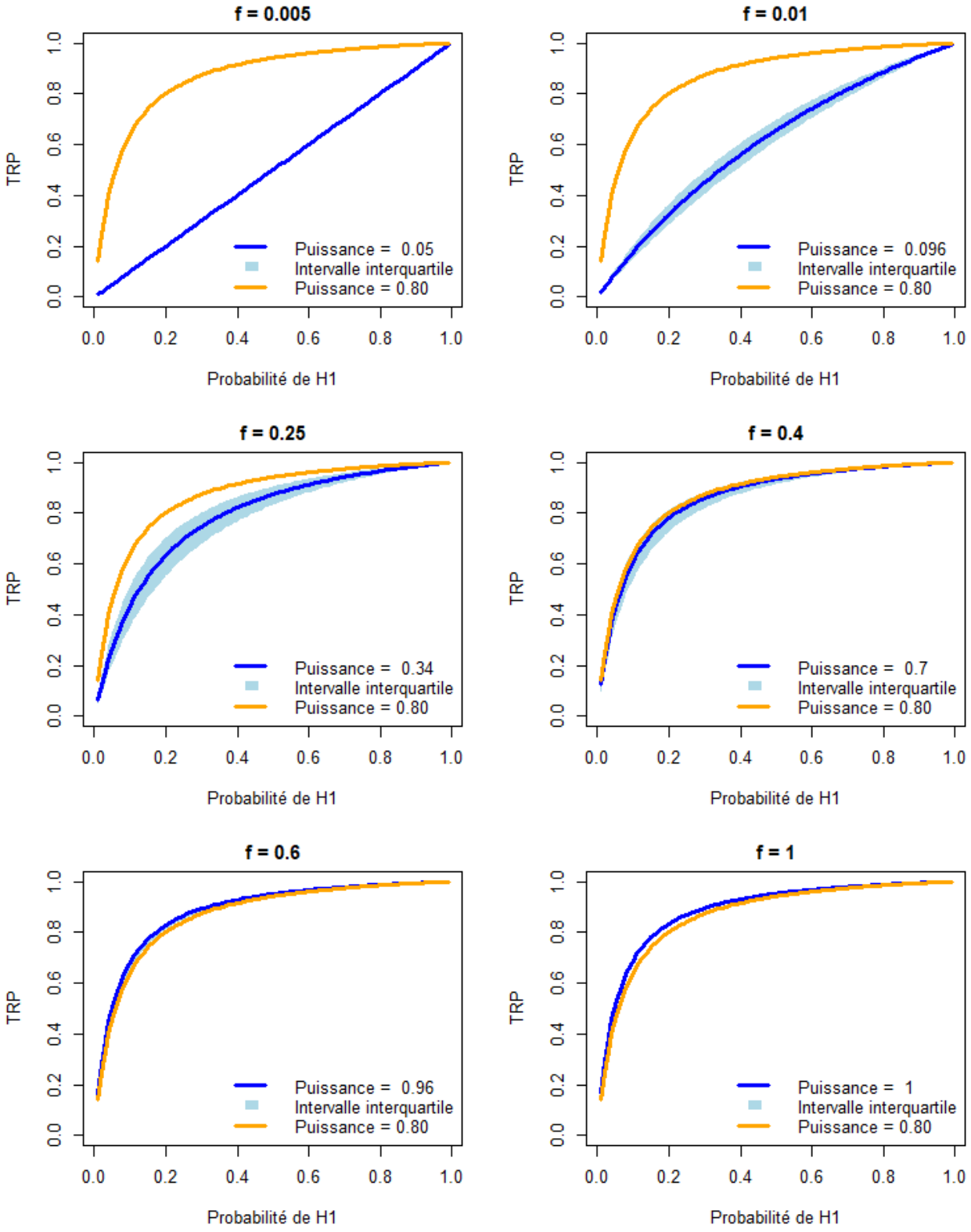

Figure 22: Version étendue des courbes de TRP calculées pour les tests F pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. L'abscisse représente la probabilité de H_1 variant de 0 à 1. La courbe orange est le TRP pour une étude virtuelle qui a atteint une puissance de 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (d de Cohen : 0,01, 0,2, 0,5, 0,8, 1,2, 2). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

Pour les tests t (version Ioannidis)

Nous avons 26 tests t significatifs sur les 37 retenus. Ce qui nous fait 72,22 % de tests significatifs dans l'échantillon des tests t. Les puissances médianes sont les mêmes pour les tests significatifs que pour l'ensemble des tests**.** Les premiers quartiles sont identiques. Les valeurs pour le troisième quartile sont plus basses presque partout. Les plus grandes différences concernent les tailles d'effets de 0,2, 0,5, et 0,8 (voir tableau 10).

Tableau 18 : Puissances médianes, premiers et troisièmes quartiles pour les tests t ***significatifs****. En gras les chiffres qui diffèrent de ceux obtenus sur l'ensemble des tests t retenus (Tableau 16).*

Taille d'effet	Taille d'effet (d de Cohen)	Puissance médiane	Premier quartile	Troisième quartile
très petite	0.01	0.05005	0.05004	0.05011
petite	0.2	0.06847	0.06487	**0.08925**
moyenne	0.5	0.1696	0.1459	**0.3043**
grande	0.8	0.3579	0.2993	**0.6378**
très grande	1,2	0.6667	0.5737	**0.9334**
énorme	2	0,9781	0.9459	**0,9999**

Les figures suivantes montrent les courbes de FDR (Figure 23) et de TRP (Figure 24) pour les tests t significatifs.

On remarque d'abord que les courbes oranges calculées pour la puissance recommandée de 80 % ne sont pas comprises dans les intervalles interquartiles des courbes calculées sauf pour les tailles d'effet supérieure à 0,8, et cela autant pour les FDR que pour les TRP.

Nous pouvons aussi remarquer que les bornes inférieures des IQR des tests t sont plus basses que celle des tests F. La quantité de faux positifs est tout aussi inquiétante que pour les tests F. Si l'on prend une taille d'effet (d) de 0,2 et une plausibilité de 1 nous obtenons un FDR de l'ordre de 40 %. Lorsque l'on prend une plausibilité de 0,5, on obtient un FDR de 59,359 %. Pour une taille d'effet de 0,5, une plausibilité de 1 donne un FDR supérieur à 20 %, et une plausibilité de 0,5 un FDR de 37,097 %.

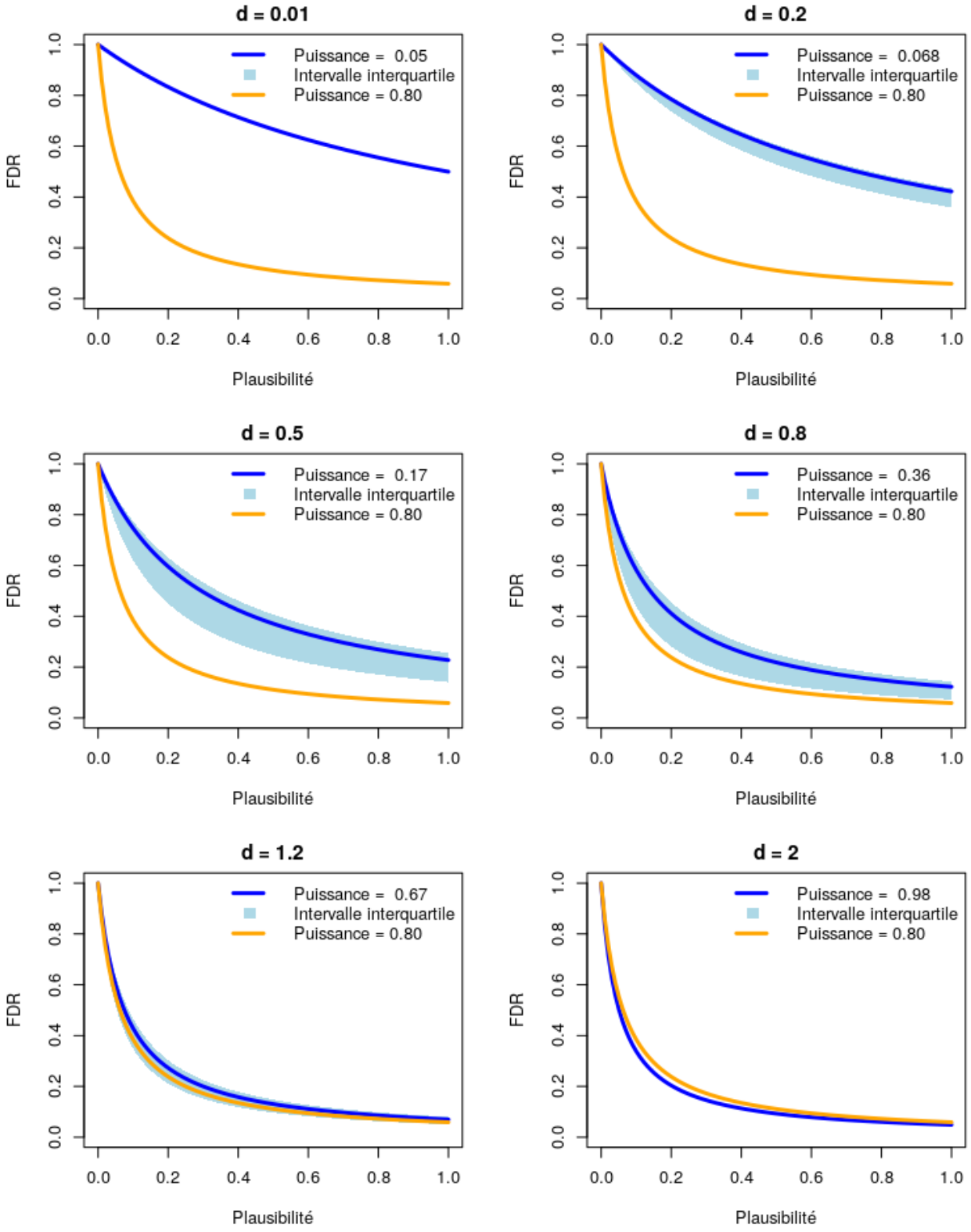

Figure 23 : Courbes de FDR calculées pour les tests t pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. La courbe orange est le FDR théorique pour une étude de puissance 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (d de Cohen : 0,01, 0,2, 0,5, 0,8, 1,2, 2). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

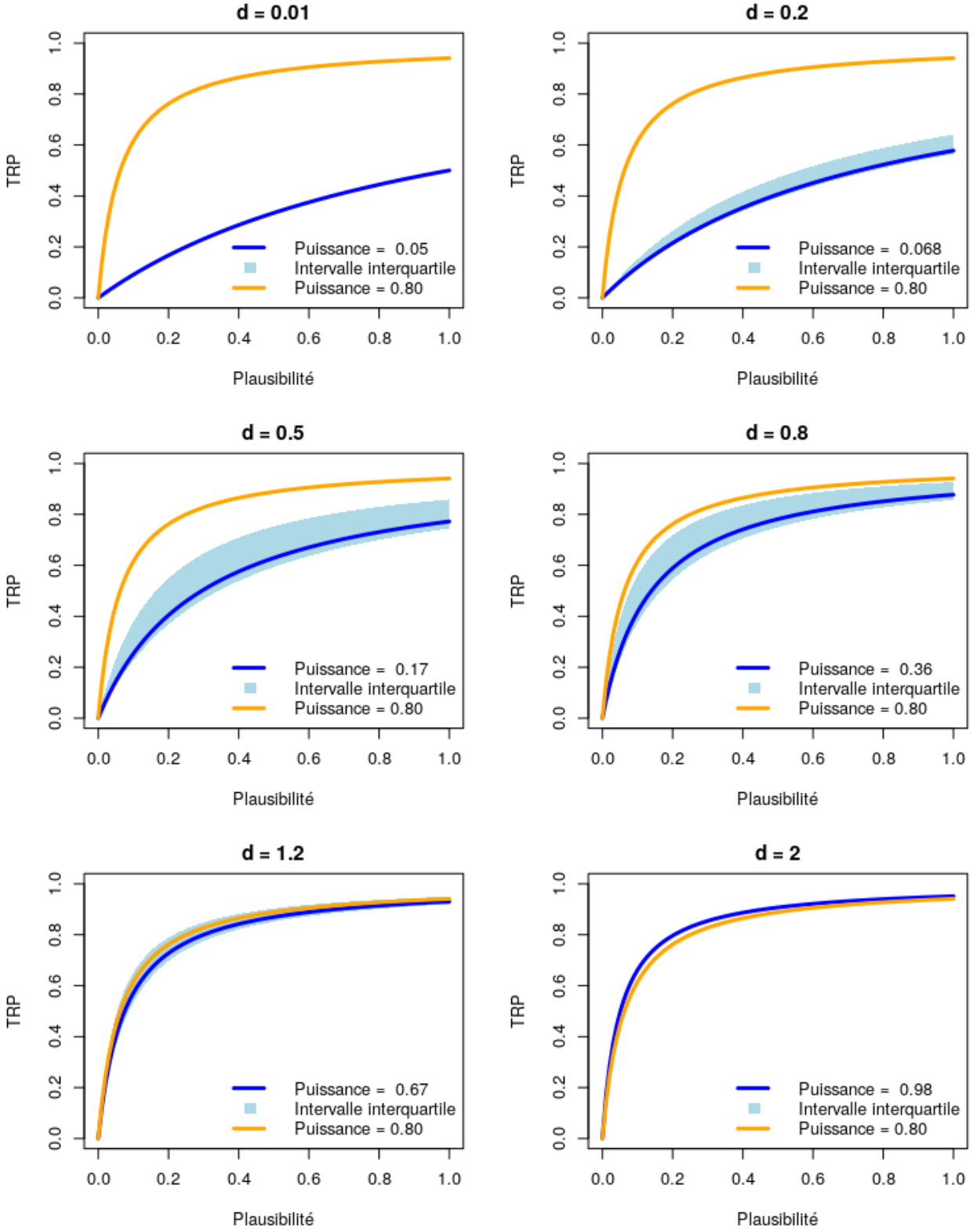

Figure 24: Courbes de TRP calculées pour les tests t pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. La courbe orange est le TRP théorique pour une étude de puissance 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (d de Cohen : 0,01, 0,2, 0,5, 0,8, 1,2, 2). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

Pour les tests t (version étendue)

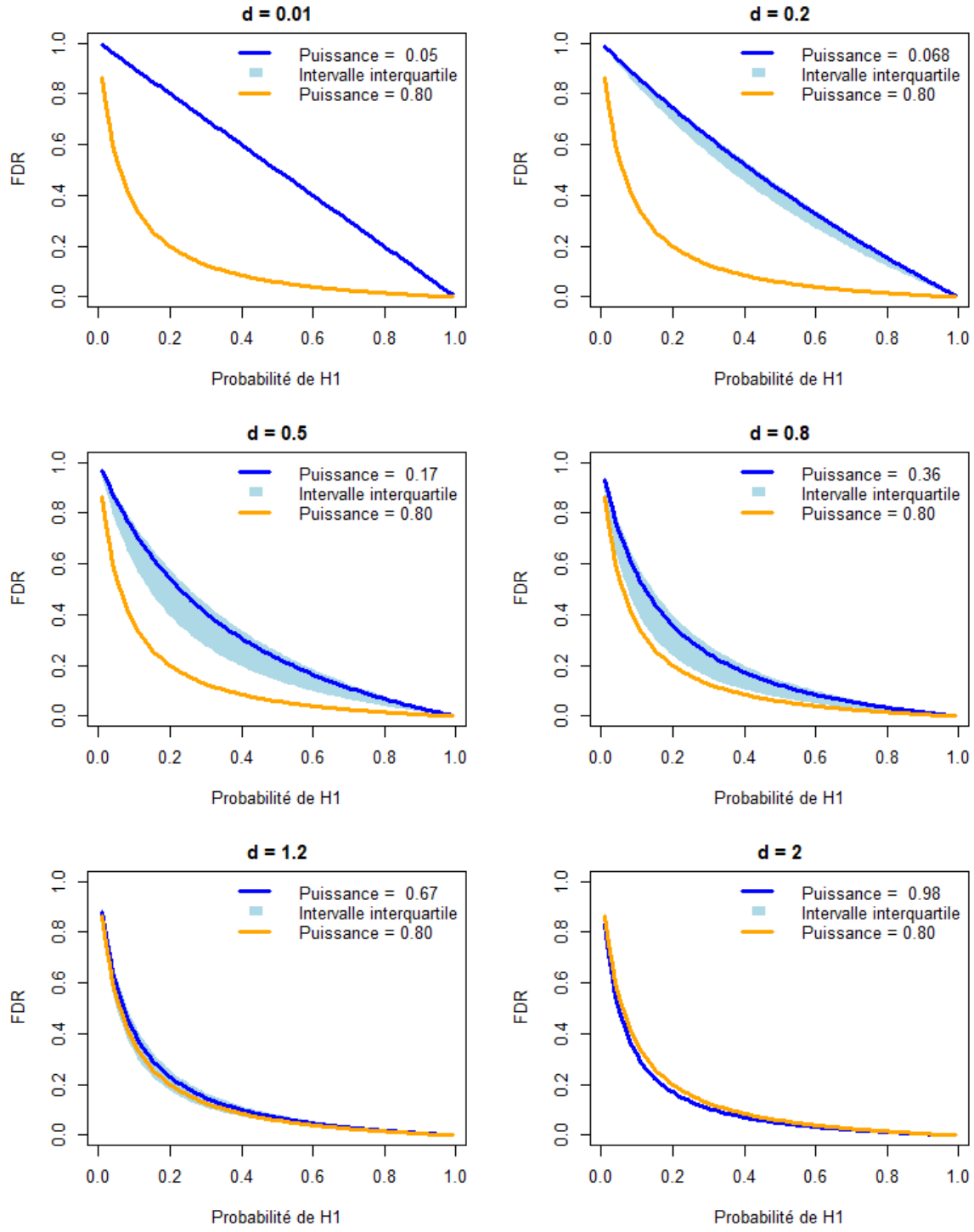

Figure 25: Version étendue des courbes de FDR pour les tests t calculées pour les tests t pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. L'abscisse représente la probabilité de H1 variant de 0 à 1. La courbe orange est le FDR pour une étude virtuelle qui a atteint une puissance de 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (d de Cohen : 0,01, 0,2, 0,5, 0,8, 1,2, 2). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

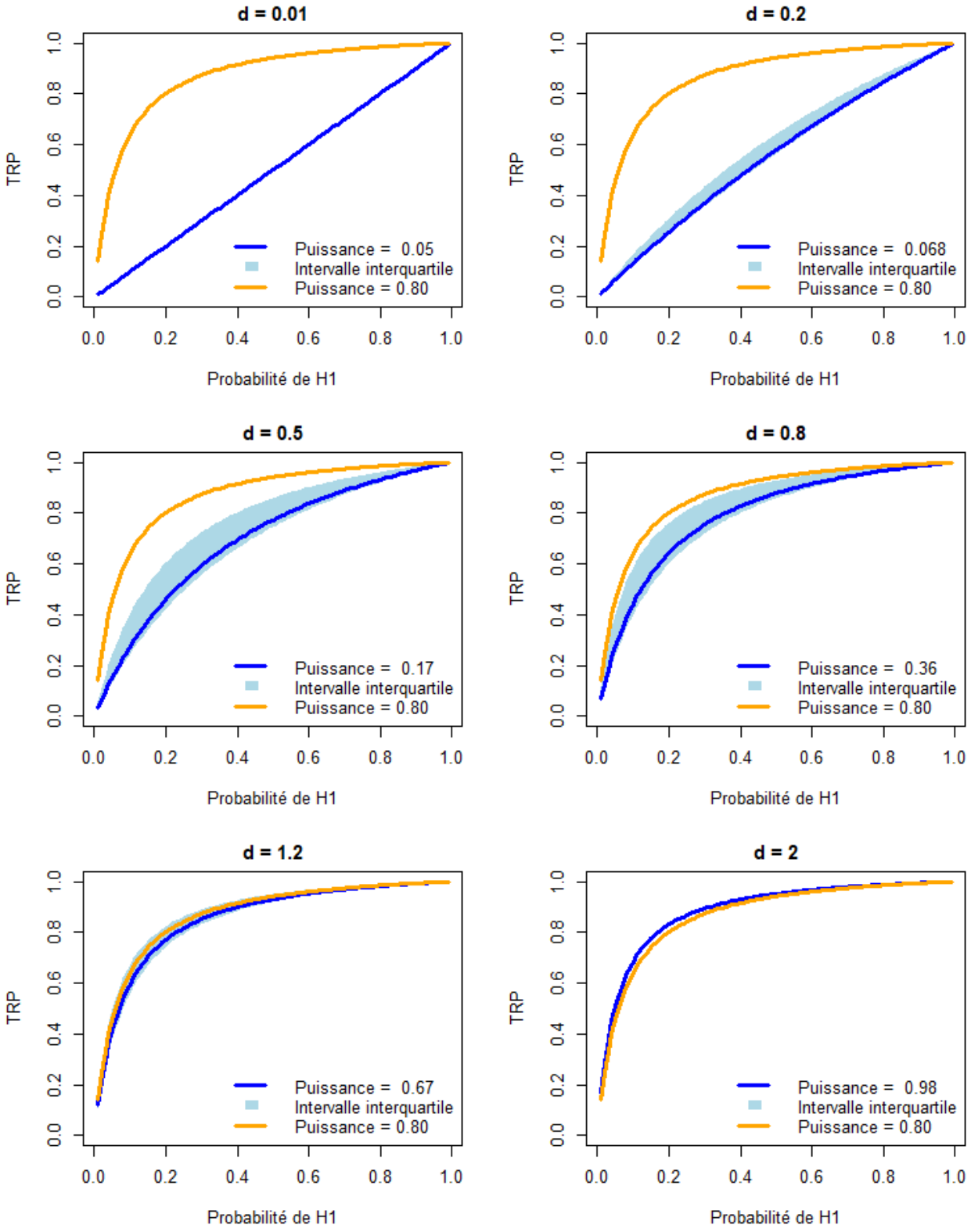

Figure 26: Version étendue des courbes de TRP pour les tests t calculées pour chacune des tailles d'effet de la classification de Cohen-Sawilowsky. L'abscisse représente la probabilité de H1 variant de 0 à 1. La courbe orange est le TRP pour une étude virtuelle qui a atteint une puissance de 0,8. La courbe bleue est la puissance médiane calculée pour la taille d'effet considérée (d de Cohen : 0,01, 0,2, 0,5, 0,8, 1,2, 2). L'aire bleu clair représente l'intervalle interquartile des valeurs de la distribution des puissances.

5 Discussion

Notre démarche s'inscrit dans la volonté d'évaluer les productions scientifiques et d'identifier des problèmes d'ordre méthodologiques. Cette volonté rentre dans les deux premières phases de la méta-recherche selon Hardwicke et al (2019), à savoir identifier et quantifier un problème. À notre connaissance, aucune étude ne s'est intéressée au manque de puissance en psychopharmacologie préclinique expérimentale. Le but principal de ce mémoire était de savoir si les études utilisant la préférence de lieu conditionné à la nicotine chez la souris avaient une puissance prospective suffisante pour détecter diverses tailles d'effet.

Ensuite nous voulions savoir l'impact de cette puissance sur le FDR et le TRP.

La littérature nous amène aussi à nous questionner sur le phénomène de sur-estimation de la taille d'effet.

5.1 La sélection des études et l'extraction des données

Parmi les 88 articles de recherche répertoriés dans PubMed, écrits en français ou en anglais, traitant de la PLC_{NIC}, et comprenant une comparaison saline vs nicotine, 47 (53,4 %) seulement rapportaient les chiffres dont nous avions besoin pour nos calculs. Il faudrait s'assurer que nos critères d'inclusion n'étaient pas trop sévères, et qu'il n'était pas possible d'extraire les données nécessaires par d'autres méthodes. Par exemple en contactant systématiquement les auteurs.

D'autres comparaisons entre les articles inclus et exclus seraient nécessaires pour savoir si les articles exclus par manque de données sont comparables aux autres ou si le manque de chiffres rapportés s'accompagne d'autres différences. Par exemple, les articles inclus pourrait mieux rendre compte de leurs méthodes expérimentales et de leurs statistiques. Cela pourrait être étudié avec la grille ARRIVE développée par Kilkenny et al (2010) qui a été mise à jour cette année par Sert et al (2020). On sait par ailleurs que la qualité du compte rendu des articles laisse souvent à désirer (Cribbie et al., 2019; Kilkenny et al., 2009; Macleod et al., 2015; Ramirez et al., 2017).

La distribution des dates de publication des études sélectionnées par PubMed semble montrer un pic d'intérêt pour le domaine en 2013. Cette tendance est moins évidente pour les seuls articles retenus, mais le petit nombre d'observations combiné au faible pourcentage d'articles retenus ne permet pas de tirer de conclusion forte. Il faut aussi tenir compte du fait que les articles ont été sélectionnés en août 2019 et que les chiffres de cette année-là sont incomplets.

Les erreurs d'encodage et de calcul sont toujours possibles. Pour limiter ces risques, nous avons repassé l'ensemble des articles en revue en vérifiant les critères d'inclusion à l'occasion d'un autre

travail[6]. Ce n'est pas une stratégie raisonnable pour un travail plus important. Il aurait été préférable qu'une bonne âme réalise une double extraction.

5.2 La distribution des p-valeurs

La distribution des p-valeurs peut montrer des traces de p-hacking lorsque les valeurs juste inférieures à 0,05 sont surreprésentées (si on a un pic juste en dessous de 0,05). Elle peut aussi suggérer un biais de publication si la majorité des p-valeurs sont inférieures à 0,05 et si les valeurs supérieures sont très peu représentées. Une telle distribution peut cependant s'expliquer aussi par la présence d'un effet réel : il est alors logique de trouver peu de p-valeurs supérieures 0,05 (Simonsohn et al., 2014).

La plupart des p-valeurs observées sont en dessous de 0,05. Il n'y a pas de pic juste en dessous de 0,05. Nous ne trouvons donc pas de signe de p-hacking. On voit aussi qu'il y a une grande quantité de p-valeurs très significatives (Figure 15). La distribution des p-valeurs n'indique donc pas la présence d'un biais de publication.

5.3 La surestimation de la taille d'effet

On sait que les faibles puissances créent un risque de surestimation de la taille d'effet. Nos résultats en Figure 16 montrent le lien que nous observons entre la taille d'effet observée et la puissance dans les tests que nous avons étudiés et cela pour toutes les tailles d'effets sauf la plus grande (la taille d'effet énorme avec un êta-carré partiel de 0,5).

Les puissances ont une corrélation négative avec l'année de publication, alors que les tailles d'effet ont une corrélation positive avec l'année de publication. Cette relation est logique si on pense que les tailles d'effets minimales détectables augmentent quand la puissance diminue. Notre corrélation explique 15,8 % de la variance totale, ce qui correspond aux résultats de Kühberger et al (2014) qui trouvaient un chiffre de 18 % [CI : 4,9 % ; 30 %]. Nous pouvons donc généraliser leurs résultats à la PLC_{NIC}.

Analyse d'un article hors norme

L'article de Jackson et al (2017) rapporte des tailles d'effet plus grandes que les autres dans la même gamme de puissances.

Ont-ils développé quelques « trucs et ficelles » qu'ils ne reportent pas ? Ou il y a-t-il d'autres différences que nous ne soupçonnons pas entre cette étude et les autres ? Quoi qu'il en soit ces auteurs ont produit 13 articles sur les 47 retenus soit 27,66 %. Une étude approfondie pour les

6 Le codage d'une grille ARRIVE modifiée en cours de développement.

comparer aux autres serait peut-être intéressante.

5.4 Les puissances

Les puissances sont négativement corrélées avec l'année de parution, peut-être parce que les budgets sont réduits et qu'une puissance supérieure n'apparaît pas nécessaire aux yeux des chercheurs.

Puissances et biais de publication

On peut suspecter un biais de publication dans un domaine étudié si la puissance y est inférieure au taux de résultats significatifs (Fidler et al., 2017).

Notre étude suggère une proportion de 77,57 % de résultats positifs. Cela place le domaine parmi les « meilleurs élèves », juste après « *environment/ecology* » et avant « *plant and animal science* » dans le classement de la Figure 9.

Pour les tests F nous avons 81,143 % de tests significatifs. Avec les puissances médianes que nous avons calculées, nous aurions donc un biais de publication si la vraie taille d'effet se situait en dessous d'un f de 0,4 ou d'un d de 0,8 (conversion : f = d/2). Pour les tests t, on a 72,22 % de tests significatifs. Avec les puissances médianes calculées plus haut, on se trouverait face à un biais de publication si la vraie taille d'effet était inférieure à un d de 2.

Il convient cependant de prendre ces conclusions avec prudence, puisque près de la moitié des articles n'ont pas pu être analysés parce qu'ils ne fournissaient pas les données nécessaires. Une analyse complémentaire des articles qui n'ont pas rencontré nos critères d'inclusion permettrait peut-être d'y voir plus clair.

Les puissances calculées

Tests F

Dans les articles que nous avons étudiés, la puissance calculée pour les tests t et les tests F n'atteint le seuil recommandé de 80 % qu'à partir des grandes voire des très grandes tailles d'effets.

Le Tableau 15 et la Figure 17 nous montrent très clairement que les tests F déployés dans les articles retenus n'atteignent la puissance recommandée d'au moins 80 % que pour des tailles d'effets supérieures à un f de 0,6 (très grande tailles d'effet). La puissance médiane que nous avons observée pour un alpha de 5 % et une grande taille d'effet n'est que de 0,7041. Ce qui est similaire à ce qu'ont trouvé Szucs & Ioannidis (2017) qui obtenaient une puissance médiane de 0,73 pour les grandes tailles d'effet dans la littérature « clinique ».

Button et al (2013) n'ont inclus que deux méta-analyses concernant les études « précliniques ». Pour ces analyses, ils ne donnent leurs résultats que pour la taille d'effet moyenne. Pour cette taille d'effet moyenne, notre résultat de 34 % de puissance pour les tests F est similaire à leur puissance de 31 % pour les études animales utilisant le labyrinthe radial. Pour les études animales utilisant le labyrinthe aquatique, ils obtiennent une puissance de 18 % (toujours en postulant une taille d'effet moyenne), ce qui est proche des 16,96 % que nous retrouvons avec les tests t utilisés pour la PLC_{NIC}.

Nous pensons donc avoir généralisé les résultats de Button et al (2013) et de Szucs & Ioannidis (2017) dans un domaine particulier de la psychopharmacologie expérimentale préclinique.

Tests t

Les puissances des tests t sont plus faibles que celles des tests F pour des tailles d'effet postulées similaires. Cela pourrait s'expliquer par le fait que les tests t et leurs équivalents (par exemple les tests de Dunnet) sont plus souvent réalisés lors d'analyses *post-hoc* pas toujours planifiées et qui ne sont pas les objectifs principaux de la recherche. Il pourrait donc être intéressant d'effectuer deux analyses, l'une excluant les tests *post-hoc* et l'autre leur étant réservée.

Comparaison avec la littérature « préclinique »

Nous pouvons comparer nos résultats avec les autres en les résumant dans un tableau semblable au Tableau 4 qui donnait un aperçu des puissances observées dans la littérature « préclinique ».

Tableau 19 : Puissances médianes de nos tests inclus

	Puissance médiane	Taille d'effet	d de Cohen
Tests F	5 %	très petite	0.01
	10 %	petite	0.2
	34 %	moyenne	0.5
	70 %	grande	0.8
	96 %	très grande	1,2
	100 %	énorme	2
Tests t	5 %	très petite	0.01
	7 %	petite	0.2
	17 %	moyenne	0.5
	36 %	grande	0.8
	67 %	très grande	1,2
	99 %	énorme	2

Le Tableau 19 montre les puissances médianes que nous avons calculées pour les tests t et f, pour les tailles d'effet de la classification de Cohen-Sawilowsky. Nos résultats n'atteignent la puissance idéale de 80 % dans 50 % des cas au moins que pour les tailles d'effet très grandes et énormes. Et pourtant ils pourraient se comparer favorablement avec ceux que nous avons trouvés dans la littérature « préclinique ». Les comparaisons restent pourtant hasardeuses, d'une part parce que la proportion d'articles étudiés est réduite, et d'autre part parce que nous ne comparons pas précisément les mêmes chiffres (médianes contre moyennes ou pourcentage d'études assez puissantes, tailles d'effet conventionnelles ou tailles d'effet mesurées ou pourcentage d'efficacité du traitement).

Les comparaisons resteront difficiles sur base des seules données fournies dans les articles tant qu'il n'y aura pas de consensus sur les meilleurs chiffres à rapporter.

Tailles d'effet minimales détectables avec une puissance de 80 % pour les tests t

Les tailles d'effet minimales qui sont détectables avec une puissance de 80 % pour les tests t dépendent de l'effectif mobilisé dans chaque groupe. Le Tableau 20 nous donne les effectifs nécessaires pour détecter les six tailles d'effet de la classification de Cohen-Sawilowsky avec une puissance de 0,8.

Tableau 20 : Nombre de sujets nécessaires dans chaque groupe pour détecter les 6 tailles d'effets de la classification de Cohen-Sawilowsky pour les tests t pour échantillons indépendants et pour échantillons appariés

Taille d'effet	d de Cohen	Échantillon nécessaire pour un test t à échantillons indépendants	Échantillon nécessaire pour un test t apparié
très petite	0,01	156 979	78 491
petite	0,2	394	199
moyenne	0,5	64	34
grande	0,8	26	15
très grande	1,2	12	8
énorme	2	6	5

Nous avons calculé les effectifs médians mobilisés dans nos échantillons de tests t appariés et indépendants. Avec ces effectifs, les tests ne sont capables de détecter avec une puissance de 80 % que des tailles d'effet très grandes ou énormes (au moins égales à 1,346 pour les tests appariés et 1,507 pour les tests indépendants).

Ces résultats sont un peu moins bons que ceux de Button et al (2013) qui trouvent des chiffres de 1,26 pour le labyrinthe aquatique et de 1,2 pour le labyrinthe radial.

5.5 FDR et TRP

Les auteurs que nous avons trouvés et qui ont calculé des TRP et FDR (Baldwin, 2017; Button et al., 2013; Quintana, 2020; Schmidt-Pogoda et al., 2019; Szucs & Ioannidis, 2017; Walum et al., 2016) en viennent tous à la même conclusion pour divers domaines de recherches : la proportion de résultats fiables est bien trop faible.

Certains se sont reposés sur des tailles d'effets déduites de méta-analyses. D'autres ont utilisé des tailles d'effet calculées sur leur échantillon. Nous avons envisagé toutes les tailles d'effet de la classification de Cohen-Sawylowsky. Cela peut donner l'impression de résultats moins précis, mais cela permet de comparer la puissance sans se soucier des tailles d'effets. Et cela évite de baser les calculs sur une taille d'effet potentiellement surestimée.

Les puissances prospectives mobilisées pour les tests significatifs sont très semblables à celles des tests non significatifs. Ce qui tend à faire croire que les tests significatifs n'étaient pas moins puissants que les autres.

Les intervalles interquartiles des courbes de FDR et de TRP que nous avons calculés pour les différentes tailles d'effet n'englobent donc pas la courbe idéale associée à une puissance de 80 %, sauf pour les tailles d'effet grandes ou supérieures pour les tests F, et très grandes ou supérieures pour les tests t.

Les résultats du FDR et du TRP calculés à partir des puissances médianes et présentés dans graphiques à la manière de Ioannidis nous montrent que le taux de fausses découvertes du domaine de la PLC_{NIC} est plus élevé que les 5 % que l'on a tendance à imaginer.

Nous pensons donc avoir retrouvé dans nos études sur la PLC_{NIC} les résultats prédits par Ioannidis (2005).

Nous pensons cependant que le modèle proposé par Ioannidis pourrait être étendu pour tenir compte des choix des chercheurs, car ils peuvent préférer les hypothèses qui présentent de bonnes chances de réussite. Il y a là selon nous un facteur supplémentaire capable d'influencer favorablement les TRP et FDR. Ce facteur peut être pris en compte très simplement en étendant les graphiques de Ioannidis vers la droite pour prendre en compte des plausibilités supérieures à 1. Mais cette extension pourrait être infinie, car la plausibilité est définies jusqu'à l'infini. Une solution est donc de représenter les probabilités de H_1.

C'est pourquoi nous avons ajouté des graphiques similaires basés sur la probabilité de H_1 qui, eux,

couvrent l'ensemble des possibilités, y compris celles qui correspondent à des rapports de chances plus favorables que 1/1.

Avec la PLC_{NIC}, nous ne sommes pas du tout dans une situation comparable à celle des laboratoires qui cherchent une ou plusieurs molécules efficaces parmi un ensemble de molécules. Les études que nous avons envisagées sont pour la plupart des modifications d'études antérieures qui ont déjà donné de bons résultats avec ce paradigme. Leurs chances de réussite sont donc vraisemblablement plus élevées. Nos graphiques étendus permettent de tenir compte de ce phénomène pour calculer les FDR ou les TRP en considérant des probabilités de H_1 supérieures à 0,5.

Dans d'autres domaines aussi on peut espérer que les auteurs, qui sont au courant du fait que les résultats négatifs sont rarement publiés, ne se lancent pas trop souvent dans des études qui n'auraient pas de bonnes chances de donner des résultats positifs, et que les comités d'éthique vérifient aussi que les études autorisées ne représentent pas de gaspillage.

Rappelons encore que près de la moitié des articles n'ont pas pu être analysés parce qu'ils ne fournissaient pas les données nécessaires et qu'il faut considérer ces conclusions avec prudence.

5.6 Objectifs généraux

Nous voulions montrer par l'exemple qu'il était **possible** de calculer les FDR et TRP pour un sous-domaine complet. Nous l'avons fait pour un très petit domaine, mais nous n'avons pu évaluer qu'une partie des articles rassemblés (47/88=53,4 %) parce que les autres ne rapportaient pas les données nécessaires aux calculs.

Nous avons pu **présenter** les résultats finaux et intermédiaires sous forme graphique, mais l'efficacité des graphiques FDR et TRP pourrait être discutée parce qu'il a fallu représenter séparément chacune des puissances de la liste de Cohen-Sawilosky et que l'ensemble occupe toute une page.

L'un des objectifs de ce travail était de mettre au point des méthodes d'extraction et d'analyse des données susceptibles d'être appliquées à des projets plus ambitieux. Nous évoquerons donc ici quelques observations à prendre en compte pour améliorer nos processus à l'avenir.

Nous voulions des techniques d'extraction **efficaces** et sécurisées. Cet objectif reste à travailler : encoder les résultats manuellement dans un fichier de tableur que l'on modifie sans cesse crée trop de risques d'erreurs d'encodage, de modifications accidentelles, ou de perte de données.

Nous n'avons pas pu **automatiser** les calculs comme nous le souhaitions, parce que les fichiers exportés par le tableur ne répondaient pas aux exigences de notre programme d'analyse statistique. Il a donc fallu créer manuellement, à partir du fichier d'extraction, plusieurs fichiers de données

analysables par R. C'était une perte de temps et cela introduisait des risques d'erreur. Mais d'un autre côté cela a aussi supprimé la tentation d'analyser des données partielles.

Nous constatons donc que l'**objectif principal est atteint**, mais que la méthode de travail peut être optimisée par quelques changements simples et raisonnables.

1. Modifier les titres des colonnes du tableur de telle sorte qu'ils soient utilisables par R sans modification.
2. Protéger les cellules exportables du tableur de telle sorte que l'on ne puisse plus y introduire de données invalides.
3. Standardiser l'encodage des plans expérimentaux et écrire une fonction de reconnaissance des plans pour choisir le bon calcul de puissance à réaliser, ce qui éviterait de réaliser tous les calculs manuellement dans un script R et diminuerait le risque d'erreur d'encodage.
4. Établir un système de sauvegarde des données mieux sécurisé, avec des fichiers redondants et physiquement indépendants, et un système de récupération et de traçage des erreurs (en sauvegardant avec la date du jour dans le nom de fichier, par exemple).
5. Adapter la programmation des analyses statistiques, pour pouvoir les appliquer sans modification à des jeux de données complets, sans avoir à extraire manuellement les informations pertinentes.

Il faudrait aussi, même si c'est sans doute bien plus complexe, programmer des routines capables d'importer, de contrôler et de combiner plusieurs jeux de données d'extraction sans les mélanger, pour pouvoir comparer le travail de plusieurs encodeurs, par exemple, ou pour vérifier la cohérence de plusieurs encodages réalisés à des moments différents.

6 Conclusion générale

Ioannidis (2005) affirme que la plupart des découvertes scientifiques publiées sont fausses. Il explique le problème par le manque de puissance, par l'action de nombreux biais, et par une faible plausibilité.

Nous l'avons dit : le manque de puissance favorise le p-hacking et il entraîne un risque de surestimation des tailles d'effet. Il augmente aussi la probabilité de faux positif. Combiné avec la publication préférentielle des résultats positifs, il pourrait même favoriser la survivance d'idée fausses dans des domaines scientifiques entiers (Akerlof & Michaillat, 2018).

Nous trouvons effectivement un manque de puissance, et un taux trop élevé de résultats positifs dans la littérature utilisant la PLC_{NIC}. Mais faut-il pour autant douter de tous ses résultats ?

Nous avons montré comment la réflexion statistique pouvait s'enrichir à mesure qu'on y intègre des données ou des éléments plus larges, depuis la taille d'effet jusqu'à la proportion d'hypothèses vraies parmi l'ensemble des hypothèses envisageables. Ioannidis base ses conclusions sur des proportions tirées de la génétique et de la recherche thérapeutiques, où il faut tester quantité d'hypothèses équiprobables (parce qu'un petit nombre seulement de gènes ou de molécules auront les effets désirés, et que l'on ne sait pas trop où chercher). Nous pensons que ce modèle pourrait être élargi pour prendre en compte les stratégies de construction et de choix des hypothèses de recherches. Le chercheur en psychopharmacologie comportementale ne choisit pas forcément ses hypothèses au hasard dans une liste de possibilités équiprobables. Il construit au contraire des hypothèses vraisemblables à partir de théories et de connaissances préalables, et en tenant compte des résultats précédemment publiés. Et souvent il peut se reposer sur des protocoles éprouvés par son équipe et qu'il modifie petit à petit pour en préciser les limites[7]. On doit donc raisonnablement espérer un très important biais de sélection des hypothèses de recherche, peut-être capable de neutraliser les biais énumérés par Ioannidis.

Les problèmes de réplication existent. Ils sont certainement liés à des phénomènes statistiques tels que la surestimation des tailles d'effet dans des études qui manquent de puissance. Mais les explications alternatives ne manquent pas, et certaines sont au moins aussi vraisemblablement pertinentes que la négligence statistique, comme les erreurs de design ou de contrebalancement, la description incomplète des dispositifs expérimentaux, la dérive du vocabulaire, ou même pourquoi pas des modifications insoupçonnées des animaux ou des circonstances.

7 Par exemple, on connaît une dose qui produit une dépendance à coup sûr. On monte une expérience avec trois groupes : un groupe « contrôle », un groupe « dépendance certaine », et un groupe « dose intermédiaire ». Les deux groupes extrêmes garantissent l'obtention d'un résultat significatif.

Cela laisse beaucoup de questions à investiguer, depuis la modélisation des biais liés à la construction et à la sélection des hypothèses de recherche, jusqu'à la qualité du compte rendu dans les articles scientifiques, en passant par le choix de statistiques réellement adaptées aux questions de recherche.

Il serait intéressant de continuer à étudier le sujet et d'aborder des domaines plus vastes : les modèles de la dépendance à l'alcool ou à d'autres substances par exemple, ou encore le bien-être des animaux de laboratoire, ou même la qualité du compte rendu des articles.

Par ailleurs, on pourrait aussi promouvoir une vraie réflexion statistique chez les chercheurs et dans la population générale en développant des outils plus commodes et mieux adaptés aux possibilités informatiques actuelles.

7 Bibliographie

Akerlof, G. A., & Michaillat, P. (2018). Persistence of false paradigms in low-power sciences. *Proceedings of the National Academy of Sciences*, *115*(52), 13228-13233. https://doi.org/10.1073/pnas.1816454115

Baldwin, S. A. (2017). Improving the rigor of psychophysiology research. *International Journal of Psychophysiology*, *111*, 5-16. https://doi.org/10.1016/j.ijpsycho.2016.04.006

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435-435. https://doi.org/10.1038/d41586-019-01307-2

Brembs, B. (2018). Prestigious Science Journals Struggle to Reach Even Average Reliability. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00037

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact : Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00291

Browner, W. S., & Newman, T. B. (1987). *Are All Significant P Values Created Equal? 257*(18), 5.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure : Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi.org/10.1038/nrn3475

Carneiro, C. F. D., Moulin, T. C., Macleod, M. R., & Amaral, O. B. (2018). Effect size and statistical power in the rodent fear conditioning literature – A systematic review. *PLOS ONE*, *13*(4), e0196258. https://doi.org/10.1371/journal.pone.0196258

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing Grade : 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233-239. https://doi.org/10.1177/2515245919858072

Cohen, J. (1962). The statistical power of abnormal-social psychological research : A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145-153. https://doi.org/10.1037/h0045186

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216-140216. https://doi.org/10.1098/rsos.140216

Colquhoun, D. (2017). The reproducibility of research and the misinterpreteation of p-values. *Royal Society Open Science*, 22. https://doi.org/10.1098/rsos.171085

Colquhoun, D. (2019a). The False Positive Risk : A Proposal Concerning What to Do About *p* -Values. *The American Statistician*, *73*(sup1), 192-201. https://doi.org/10.1080/00031305.2018.1529622

Colquhoun, D. (2019b). A response to critiques of 'The reproducibility of research and the misinterpretation of *p* -values'. *Royal Society Open Science*, *6*(11), 190819. https://doi.org/10.1098/rsos.190819

Cribbie, R., Beribisky, N., & Alter, U. (2019). *A Multi-faceted Mess : A Review of Statistical Power Analysis in Psychology Journal Articles* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3bdfu

Chambers, C., Feredoes, E., Muthukumaraswamy, S., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules : Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4-17. https://doi.org/10.3934/Neuroscience.2014.1.4

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343-15347. https://doi.org/10.1073/pnas.1516179112

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science : A review of three human research domains. *Royal Society Open Science*, *4*(2), 160254. https://doi.org/10.1098/rsos.160254

Duval, S., & Tweedie, R. (2000). A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis. *Journal of the American Statistical Association*, *95*(449), 89-98. https://doi.org/10.1080/01621459.2000.10473905

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., & Gordon, A. (2017). Metaresearch for Evaluating Reproducibility in Ecology and Evolution. *BioScience*, *67*(3), 282-289. https://doi.org/10.1093/biosci/biw159

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations : Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641-651. https://doi.org/10.1177/1745691614551642

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual : What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. In D. Kaplan, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (p. 392-409). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n21

Giuffrida, M. A. (2014). Type II error and statistical power in reports of small animal clinical trials. *Journal of the American Veterinary Medical Association*, *244*(9), 1075-1080. https://doi.org/10.2460/javma.244.9.1075

Gonon, F., Konsman, J.-P., Cohen, D., & Boraud, T. (2012). Why Most Biomedical Findings Echoed by Newspapers Turn Out to be False : The Case of Attention Deficit Hyperactivity Disorder. *PLOS ONE*, *7*(9), e44275. https://doi.org/10.1371/journal.pone.0044275

Goodman, S. (2008). A Dirty Dozen : Twelve P-Value Misconceptions. *Seminars in Hematology*, *45*(3), 135-140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Hofmeister, E. H., King, J., Read, M. R., & Budsberg, S. C. (2007). Sample size and statistical power in the small-animal analgesia literature. *Journal of Small Animal Practice*, *48*(2), 76-79. https://doi.org/10.1111/j.1748-5827.2006.00234.x

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), 6. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated: *Epidemiology*, *19*(5), 640-648. https://doi.org/10.1097/EDE.0b013e31818131e7

Jackson, A., Bagdas, D., Muldoon, P. P., Lichtman, A. H., Carroll, F. I., Greenwald, M., Miles, M. F., & Damaj, M. I. (2017). In vivo interactions between α7 nicotinic acetylcholine receptor and nuclear peroxisome proliferator-activated receptor-α : Implication for nicotine

dependence. *Neuropharmacology*, *118*, 38-45. https://doi.org/10.1016/j.neuropharm.2017.03.005

Kerr, N. L. (1998). HARKing : Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196-217.

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving Bioscience Research Reporting : The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology*, *8*(6), 5. https://doi.org/10.1371/journal.pbio.1000412

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F. W., Cuthill, I. C., Fry, D., Hutton, J., & Altman, D. G. (2009). Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. *PLoS ONE*, *4*(11), e7824. https://doi.org/10.1371/journal.pone.0007824

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability : A “Many Labs” Replication Project. *Social Psychology*, *45*(3), 142-152. https://doi.org/10.1027/1864-9335/a000178

Krzywinski, M., & Altman, N. (2013). Power and sample size. *Nature Methods*, *10*(12), 1139-1140. https://doi.org/10.1038/nmeth.2738

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology : A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Lash, T. L. (2017). The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. *American Journal of Epidemiology*, *186*(6), 627-635. https://doi.org/10.1093/aje/kwx261

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B., & Goodman, S. N. (2017). Five ways to fix statistics. *Nature*, *551*(7682), 557-559. https://doi.org/10.1038/d41586-017-07522-z

Lodder, P., externe, L. vers un site, fenêtre, celui-ci s’ouvrira dans une nouvelle, Ong, H. H., Grasman, R. P. P. P., & Wicherts, J. M. (2019). A comprehensive meta-analysis of money

priming. *Journal of Experimental Psychology: General*, *148*(4), 688-712. http://dx.doi.org/10.1037/xge0000570

Macleod, M. R., Lawson McLean, A., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., Hirst, T., Hemblade, R., Bahor, Z., Nunes-Fonseca, C., Potluru, A., Thomson, A., Baginskitae, J., Egan, K., Vesterinen, H., Currie, G. L., Churilov, L., Howells, D. W., & Sena, E. S. (2015). Risk of Bias in Reports of In Vivo Research : A Focus for Improvement. *PLOS Biology*, *13*(10), e1002273. https://doi.org/10.1371/journal.pbio.1002273

Maiväli, Ü. (2015). Study Design. In *Science Methodology and What Can Go Wrong* (p. 111-157). Elsevier. https://doi.org/10.1016/B978-0-12-418689-7.00003-X

McCarthy, M., & Parris, K. (2001). Identifying effects of toe clipping on anuran return rates : The importance of statistical power. *Amphibia-Reptilia*, *22*(3), 275-289. https://doi.org/10.1163/156853801317050070

McGrayne, S. B. (2011). *The Theory That Would Not Die : How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.

Moher, D., Dulberg, C. S., & Wells, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA, 272*(2), 122-124.

Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *British Journal of Pharmacology, 172*(8), 2126-2132. https://doi.org/10.1111/bph.12884

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716-aac4716. https://doi.org/10.1126/science.aac4716

Pezzullo, J. (2013). *Biostatistics For Dummies* (1 edition). For Dummies.

Prus, A. J., James, J. R., & Rosecrans, J. A. (2009). Conditioned Place Preference. In J. J. Buccafusco (Éd.), *Methods of Behavior Analysis in Neuroscience* (2nd éd.). CRC Press/Taylor & Francis. http://www.ncbi.nlm.nih.gov/books/NBK5229/

Quintana, D. S. (2020). *Most oxytocin administration studies are statistically underpowered to reliably detect (or reject) a wide range of effect sizes* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kzp4n

Ramirez, F. D., Motazedian, P., Jung, R. G., Di Santo, P., MacDonald, Z. D., Moreland, R., Simard, T., Clancy, A. A., Russo, J. J., Welch, V. A., Wells, G. A., & Hibbert, B. (2017). Methodological Rigor in Preclinical Cardiovascular Studies : Targets to Enhance Reproducibility and Promote Research Translation. *Circulation Research*, *120*(12), 1916-1926. https://doi.org/10.1161/CIRCRESAHA.117.310628

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rossi, J. S. (1990). Statistical power of psychological research : What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*(5), 646-656. https://doi.org/10.1037/0022-006X.58.5.646

Sawilowsky, S. S. (2009). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597-599. https://doi.org/10.22237/jmasm/1257035100

Schmidt-Pogoda, A., Bonberg, N., Koecke, M. H. M., Strecker, J.-K., Wellmann, J., Bruckmann, N.-M., Beuker, C., Schäbitz, W.-R., Meuth, S. G., Wiendl, H., Minnerup, H., & Minnerup, J. (2019). Why most acute stroke studies are positive in animals but not in patients. *Annals of Neurology*, *n/a*(n/a). https://doi.org/10.1002/ana.25643

Sedlmeier, P., & Gigerenzer, G. (1989). *Do Studies of Statistical Power Have an Effect on the Power of Studies?* 8.

Sert, N. P. du, Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Hurst, V., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Würbel, H. (2020). Reporting animal research : Explanation and elaboration for the ARRIVE guidelines 2.0. *PLOS Biology*, *18*(7), e3000411. https://doi.org/10.1371/journal.pbio.3000411

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve : A key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534-547. https://doi.org/10.1037/a0033242

Smith, J. S., Schindler, A. G., Martinelli, E., Gustin, R. M., Bruchas, M. R., & Chavkin, C. (2012). Stress-Induced Activation of the Dynorphin/ -Opioid Receptor System in the Amygdala Potentiates Nicotine Conditioned Place Preference. *Journal of Neuroscience*, *32*(4), 1488-1495. https://doi.org/10.1523/JNEUROSCI.2980-11.2012

Stapel, D. (2016). *Faking Science : A True Story of Academic Fraud*.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632-638. https://doi.org/10.1177/1745691612463078

Wallisch, P. (2015). Brighter than the sun : Powerscape visualizations illustrate power needs in neuroscience and psychology. *arXiv:1512.09368 [q-bio]*. http://arxiv.org/abs/1512.09368

Walum, H., Waldman, I. D., & Young, L. J. (2016). Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies. *Biological Psychiatry*, *79*(3), 251-257. https://doi.org/10.1016/j.biopsych.2015.06.016

Witt, J. K. (2019). Insights into Criteria for Statistical Significance from Signal Detection Analysis. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.871

Yong, E. (2012). Replication studies : Bad copy. *Nature News*, *485*(7398), 298. https://doi.org/10.1038/485298a

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. *PLoS Medicine*, *5*(10), e201. https://doi.org/10.1371/journal.pmed.0050201

Résumé :

Depuis quelques années un manque de puissance a été mis au jour dans divers domaines de recherche. Cela a abouti à de sérieux doutes quant à la reproductibilité de beaucoup de résultats scientifiques. À notre connaissance, aucune étude n'a évalué ce problème en psychopharmacologie expérimentale préclinique. Nous avons choisi d'étudier le sous-domaine de la préférence de lieu conditionné induite par la nicotine chez la souris (PLC_{NIC}).

Nous avons identifié les articles le concernant sur PubMed et nous en avons extrait les tailles des échantillons, le type de tests statistiques (F ou t), leurs résultats, leurs degrés de liberté, et leurs p-valeurs. À partir de ces valeurs, nous avons calculé la puissance prospective médiane pour 6 tailles d'effets (classification de Sawilowsky, 2009). Le taux de vraies découvertes (*True Report Probability* : TRP) a été calculé à partir des puissances médianes, de l'erreur de type I fixée à 5 %, et de la plausibilité que nous avons fait varier de 0 à 1.

Des 139 articles trouvés sur PubMed, 48 ont rencontré nos critères d'inclusion. Ces 48 articles contenaient 109 tests statistiques utilisables pour notre projet. Dans cet échantillon de tests 77,57 % sont significatifs. Les puissances médianes pour les tests F pour les petites, moyennes, et grandes tailles d'effet sont respectivement de 9,5 %, 34,1 % et 70,4 %. Pour les tests t nous avons trouvé 6,8 %, 17 %, et 35,8 %. Aucun de ces chiffres n'atteint le seuil recommandé de 80 %. Pour une plausibilité de 10 %, nous trouvons des TRP pour les tests F pour une petite, moyenne et grande taille d'effet de 16 % 40,8 %, et 58,5 % soit des taux de fausses découvertes de 84 %, 59,2 %, et 41,5 %. Pour les tests t les TRP pour les mêmes tailles d'effets sont de 12 %, 25,3 %, et 41,7 % soit des taux de fausses découvertes de 88 %, 74,7 %, et 58,3 %, bien supérieurs aux 5 % maximum que l'on croit garantis par la probabilité alpha.

Nous retrouvons donc, dans le sous-domaine de la PLC_{NIC}, le manque de puissance et l'augmentation de faux positif associée qui ont déjà été observés dans d'autres disciplines.

Nous avons aussi étendu les graphiques de Ioannidis (2005) pour représenter les TRP qu'on obtient avec des rapports de chance supérieurs à 1/1 qui pourrait donner des TRP plus favorables.