

## "Antidépresseurs : les limites d'une méta-analyse"

Stéphane Mouchabac

### 1. Introduction

Des articles remettant en question (ou pondérant) l'effet des antidépresseurs sont régulièrement publiés, réactivant la polémique sur un mode devenu assez stéréotypé, voire manichéen : l'industrie pharmaceutique y est montrée du doigt et accusée de ne publier que ce qui va dans le sens de ses intérêts en "surestimant" l'effet de ses produits. Le clinicien y perd souvent ses repères, entre la puissance de la *médecine fondée sur les preuves* et les réalités de son quotidien.

Cette attitude, parfois plus polémique que scientifique, ne fait pas toujours avancer la pratique et certaines de ces études sont même à prendre avec méfiance, malgré l'impact factor élevé des revues qui les publient et la méthodologie d'allure implacable utilisée.

C'est le cas d'une méta-analyse, publiée en février dernier dans la revue en ligne Plos Medecine, qui avait pour objectif d'évaluer le bénéfice des antidépresseurs dans la dépression (1). Les données utilisées étaient celles qui avaient été soumises à la Food and Drug Administration en vue de l'obtention d'une autorisation de mise sur le marché de plusieurs antidépresseurs "de dernière génération".

Selon ces auteurs, les méta-analyses récentes ne montraient qu'un bénéfice relatif des antidépresseurs par rapport au placebo et ils précisait que lorsque les essais non publiés étaient pris en compte, le bénéfice devenait moins important puisque les différences avec le placebo (tout en étant statistiquement significatif) n'avaient plus de pertinence clinique si l'on se référait à des critères "adaptés" (significativité clinique lorsque l'écart entre les deux groupes est supérieur à un certain nombre de points).

Ainsi, dans leur conclusion, les auteurs affirmaient que les différences avec le placebo étaient fonction de l'intensité initiale de la dépression, allant de l'absence complète de différence pour les dépressions légères et modérées, jusqu'à des différences significatives pour les patients sévèrement déprimés, sachant que le seuil de significativité clinique de cette différence n'était en fait valable que pour les patients ayant les scores initiaux extrêmes.

La presse a alors largement relayé cette publication, "*Les antidépresseurs comme le Prozac sont presque inefficaces, selon une étude anglaise*" titrait le Monde ou "*Les antidépresseurs récents globalement inefficaces, selon une étude britannique*" pour Libération. Si le mot *presque* signifie à *peu près*, cela ne veut pas dire *pas du tout*, or c'est pourtant ce que l'on retiendra du flot

d'information internet (forums, blogs..), où l'on pourra lire par exemple "*leur effet sur la dépression : à peine mieux qu'un vulgaire placebo, C'est la conclusion d'une étude menée par l'université anglaise de Hull*"

Puis une phrase d'Irving Kirsh va être aussi très diffusée : "*La différence d'amélioration entre les patients prenant des placebos et ceux prenant des antidépresseurs n'est pas très importante. Cela signifie que les personnes souffrant de dépression peuvent aller mieux sans traitement chimique*". Et le patient de conclure : "*qu'il paraît que les antidépresseurs ne servent à rien*".

Pourtant la majeure partie des essais sélectionnés montrait une différence significative entre le traitement et le placebo, et la méta-analyse concluait à un écart significatif de 1,8 point entre les deux groupes (différence jugée suffisante par la FDA pour parler d'efficacité).

Alors par quel mécanisme a-t-on transformé un résultat significatif en une absence d'efficacité ? Et surtout que penser de la conclusion des auteurs qui attribuent le lien entre la sévérité initiale de l'épisode dépressif et l'efficacité de l'antidépresseur non pas à une activité pharmacologique de ce dernier, mais à "*une diminution de la réponse au placebo*" ?

Nous proposons donc dans cet article d'analyser comment un désir initial d'objectivité a abouti à des conclusions plus que "limites" du fait d'*a priori* pas toujours neutres (au niveau critères de jugement), d'une méthodologie critiquable et surtout d'interprétations plus que douteuses.

### 2. Une étude pourtant "à haut niveau de preuves"

#### 2.1. Médecine fondée sur les preuves (tableau 1)

La médecine fondée sur le niveau de preuves (EBM ou "evidence based medicine") peut être définie comme "*l'utilisation rigoureuse et judicieuse des meilleures données disponibles lors de prise de décisions concernant les soins à prodiguer à des patients individuels*".

Elle vient s'opposer aux traditionnelles "recettes" ou intuitions pharmacologiques relevant parfois que de la seule croyance du prescripteur.

Niveau de preuve scientifique	Grades de recommandation
<b>Niveau 1 (NP1)</b> *Essai comparatifs randomisés de forte puissance * <b>Méta analyse d'essais comparatifs randomisés</b> *Analyse de décision basée sur des études bien menées	<b>Grade A : Preuve scientifique établie (prouvé)</b>
<b>Niveau 2 (NP2)</b> *Essai comparatifs randomisés de faible puissance *Etudes comparatives non randomisés bien menées *Etudes de cohorte.	<b>Grade B : Présomption scientifique (probable)</b>
<b>Niveau 3 (NP3)</b> Etudes de cas témoin <b>Niveau 4 (NP4)</b> Etudes comparatives comportant des biais importants Etudes rétrospectives Série de cas Etudes descriptives	<b>Grade C : Faible niveau de preuve scientifique</b>

**Tableau 1.** Niveau de preuves selon les recommandations de l'HAS (2).

L'EBM est donc la justification du positivisme scientifique, c'est à dire l'idée que la "réalité" existe en tant que telle, indépendamment des regards de chacun, donc encore moins du regard du prescripteur utilisant sa propre expérience ou du créateur de "recettes". Mais si l'on considère du point de vue épistémologique que la subjectivité peut être "nocive", l'EBM ne doit pas non plus s'imposer comme une autorité qui s'opposerait à la démarche scientifique, c'est-à-dire la remise en question.

Ces preuves proviennent d'études cliniques systématiques, telles que des essais contrôlés randomisés en double aveugle, des méta-analyses, éventuellement des études transversales ou de suivi bien construites. Initialement formalisée pour constituer un ensemble de techniques pédagogiques de lecture et d'évaluation de la qualité scientifique de la littérature médicale, à l'aide de niveaux de pertinence scientifique, l'EBM est appliquée à de nombreux domaines (pédagogie médicale, économie de la santé, aide au jugement clinique ou à la décision thérapeutique).

## 2.2. Justification de la méta-analyse

On trouve selon les auteurs et les pays plusieurs organisations de ces niveaux de preuve, mais quelle que soit la classification la méta-analyse constitue un haut niveau (tableau 1). L'objectif principal de la méta-analyse va donc être de synthétiser les résultats des essais thérapeutiques disponibles sur une question du registre thérapeutique. La méthodologie utilisée répond aussi à des critères spécifiques qui permettront d'assurer une bonne validité à ce travail.

Cucherat et coll., précisent que la méta-analyse est une synthèse *systématique* (car elle implique une recherche exhaustive de tous les essais publiés et non publiés). Elle est aussi *quantifiée* car elle se fonde sur des modèles de calcul statistique permettant une estimation précise de la taille de l'effet du traitement (3).

En synthétisant les résultats des différentes études, on augmente la probabilité de démontrer l'effet (ou non) d'un traitement du fait d'une plus grande puissance statistique et aussi de comparer, en les réunissant, des résultats contradictoires.

Ainsi, si l'effet existe, on est à même de déterminer avec plus de précision l'estimation de la taille de l'effet d'un traitement (nombre de sujets et résultats plus nombreux). On peut alors effectuer des analyses en sous-groupes plus importants (la petite taille des échantillons des essais individuels pouvant être à la base d'un frein statistique) et rechercher des différences dans la taille de l'effet pour ces sous-groupes (ce qui permet de répondre à de nouvelles questions).

La question de la généralisation des résultats est aussi importante dans le cadre d'un essai thérapeutique, or bien souvent, on reproche aux études des biais de recrutement ou bien une sélectivité des sujets inclus dans ces études (qui ne sont pas toujours représentatifs de la population rencontrée par les médecins). Aussi la méta-analyse a le mérite de regrouper différents types de patients qui ressembleront plus aux patients "réels".

Les biais existent et sont variés. Ainsi une méta-analyse qui repose sur des essais moyens méthodologiquement sera elle-même moyenne ; la sélection des essais doit donc être rigoureuse et idéalement reposer sur des essais

randomisés et en double aveugle. Cependant on peut considérer qu'une étude moyenne sera "lissée" par des essais plus consistants.

Les biais de publications, qui concernent bien souvent les résultats négatifs, peuvent entraîner un risque de surestimation de l'efficacité des traitements étudiés. L'obligation récente pour l'industrie de joindre au dossier d'AMM les études négatives réduit en partie ce risque.

Kirsch et ses collaborateurs proposent donc d'analyser les données déposées à la FDA pour quatre antidépresseurs de "nouvelle génération", tous les essais sont donc pris en compte et ont été sélectionnés selon les algorithmes du QUOROM Flow Chart (Quality of Reporting of Meta-analyses) publié en 1999 dans le Lancet (4) (figure 1).

Seuls les essais randomisés, en double aveugle et placebo- contrôlés sont retenus. Les patients ont un diagnostic d'épisode dépressif majeur selon les critères du DSM.

Afin de pouvoir généraliser les résultats à tous les patients, les experts de la FDA demandent un taux d'achèvement pour au moins 70 % des patients dans ces essais à 6 semaines : seulement 4 essais atteignent cet objectif.

Limiter le taux de sorties d'études élimine certains biais, mais la durée des essais (33 essais à 6 semaines, six à 4 semaines, deux à 5 semaines et six à 8 semaines) permet-elle d'obtenir une réponse thérapeutique de qualité, du moins de procéder à des réajustements posologiques efficaces en cas de réponse partielle dont on sait qu'elles sont fréquentes à la fin du traitement d'attaque ?

On peut noter que 37 essais concernent des patients non hospitalisés et seulement deux sont effectués en

hospitalisation : si ces études sont plus proches du patient "tout venant", on peut cependant penser que le faible nombre de patients hospitalisés exclut certaines pathologies "réellement" sévères (par exemple les patients suicidaires, ayant des signes somatiques graves ou des symptômes psychotiques).

Point important à souligner, les doses des traitements ne figurent dans aucun tableau, ce qui constitue une perte d'informativité non négligeable surtout pour des molécules telles que la venlafaxine qui présente des possibilités de posologie très larges associées à un effet-dose.

En recherchant par exemple, les doses de venlafaxine dans les essais sélectionnés on constate qu'elles vont de 150 à 350 mg.

### 3. Méthodologie

#### 3.1. Les critères de jugement pas si objectifs

Les résultats des différentes études sommés ("poolés") dans une méta-analyse sont souvent très divergents et les outils de mesure peuvent être aussi très différents. Une technique de standardisation est utilisée pour corriger les différences entre les études : les variables continues (par exemple des scores à la HAMD) sont, le plus souvent, exprimées sous forme de moyenne.

Ainsi, pour chaque essai on peut calculer une différence moyenne entre deux groupes (ici placebo vs traitement actif). Il est alors possible de standardiser ces différences moyennes, puis en utilisant un facteur de pondération on va obtenir une différence moyenne pondérée : la somme pondérée de toutes ces différences moyennes pondérées donne l'estimation de l'efficacité "poolée".

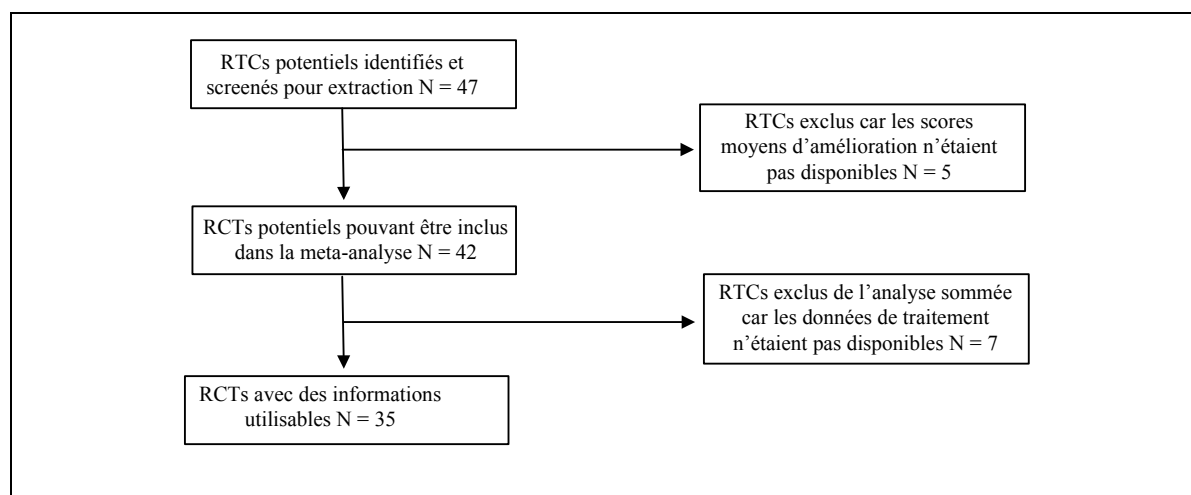


Figure 1. QUOROM Flow Chart, RTC : Randomised Controlled Trials (essais randomisés contrôlés).

Les critères NICE (5) considèrent que: *“les résultats pour lesquels une différence moyenne standardisée est calculée (par exemple lorsque des données de différentes versions d’une échelle sont combinées), une taille d’effet de 0,5 (une taille d’effet moyenne) ou plus est considérée comme significative. Lorsqu’une différence moyenne pondérée est calculée, une différence d’au moins trois points entre les deux groupes est considérée comme significative pour la BDI (Beck Depression Inventory) et la HRSD (Hamilton Rating Scale for Depression). Lorsqu’une taille d’effet est significative au niveau statistique mais non cliniquement et que l’intervalle de confiance exclut des valeurs jugées cliniquement importantes, alors le résultat est considéré comme “peu probablement significatif cliniquement”. Alternativement, si l’intervalle de confiance inclut des valeurs cliniques importantes, le résultat est considéré comme “insuffisant pour déterminer une significativité clinique”.*

L’intention est louable, mais le choix des valeurs de significativité plutôt arbitraire, surtout pour des échelles telles que la Hamilton dont on sait que certains items sont très sensibles aux traitements pharmacologiques non antidépresseurs ou au cadre de l’expérimentation.

E Turner, et R Rosenthal, auteurs de la méta-analyse publiée dans le New England en janvier (6) qui évaluait aussi l’efficacité des antidépresseurs (voir le commentaire de A.Bottéro dans *NPTD* n° 32), soulignent que si leurs résultats sont proches en terme de taille d’effet (0,31 pour les données issues de la FDA), leurs conclusions sont très différentes.

En effet, le choix d’un cut-off à 0,5 revient à transformer des valeurs continues en valeurs binaires : la taille de l’effet moyenne calculée par la méthode précédente peut donc prendre des valeurs de zéro à un, par exemple une valeur de 0,2 peut être considérée comme un effet “léger” et 0,8 un effet “important”. Ainsi, définir un seuil de 0,5 pour la significativité clinique revient à dire que l’effet est soit présent, soit absent. Au delà des considérations habituelles pour savoir ce que l’on fait des valeurs très proches de 0,5 (0,48 est-il vraiment différent de 0,52), ils proposent d’imaginer un volume de 1 unité rempli de 0,31 unité : par la méthode NICE, cela revient à considérer que le ce volume est vide (7)...

C’est pourtant avec ce seuil que les auteurs concluent à l’inefficacité des antidépresseurs dans la majeure partie des cas.

### 3.2. Quels patients dans ces études ?

De nombreux auteurs soulignent que le mode de recrutement des essais sélectionne une catégorie de sujets pouvant être non représentative : screening dans les journaux ou par “bouche à oreille”, indemnités ou rémunération des participants. On ne peut donc exclure certains niveaux “limites” d’intensité dépressive, d’autant

que la pression des inclusions s’accompagne parfois d’une plus grande “souplesse” dans l’évaluation clinique.

Peut on aussi comparer des sujets, dont l’éventualité de recevoir une substance non active est acceptée, à ceux rencontrés en pratique clinique quotidienne ? Ces patients sont plus à même de s’améliorer facilement en dehors de toute prescription d’une substance active dans l’essai.

L’existence d’idées suicidaires pose aussi problème, puisque éthiquement, le risque est considéré comme trop important pour que des patients présentant ces items reçoivent un placebo, surtout en ambulatoire. Or, souvent, l’idéation suicidaire est un critère d’exclusion qui de ce fait élimine de l’échantillon toute une catégorie de patients pour lesquels la discussion bénéfique semble plus évidente et surtout un degré de sévérité psychométriquement plus élevé puisque ces items ne seront pas quantifiables dans l’échantillon.

On comprend que dans le cadre des dépressions dites légères, pour lesquelles le rapport bénéfice risque des antidépresseurs est souvent discuté, la faible différence (même si elle est significative) entre le placebo et l’antidépresseur est certainement liée à d’autres effets non mesurés directement ou bien non pris en compte dans l’interprétation des résultats.

Par exemple, la notion d’évolution naturelle de la maladie dépressive n’est quasiment jamais considérée dans l’analyse des résultats. En effet, le diagnostic selon les critères de DSM implique une durée minimale (“au moins quinze jours”) de présence des symptômes, mais il est rare que dans les critères d’exclusion d’une étude, on demande en plus que cette durée ne soit pas à l’inverse “trop ancienne”.

On sait ainsi que l’évolution naturelle de la maladie se fait dans environ 60 à 70 % des cas vers la guérison en 6 à 8 mois, certes avec une probabilité de rechute ou de récurrence plus élevée, des séquelles cognitives, fonctionnelles et sociales, mais si l’on se réfère aux critères du DSM ou de la CIM, le patient n’est plus malade et psychométriquement il est possible de quantifier cette donnée (HDRS < 7 par exemple). Imaginons alors que l’on inclut un patient “en queue de dépression”, son score initial à la baseline sera certes bas, mais suffisant pour l’inclusion. La dépression sera alors qualifiée de légère, mais dans le cadre de l’étude, l’amélioration observée sous placebo est-elle due à un effet de celui-ci, ou tout simplement à l’évolution spontanée de la maladie ? Maladie qui peut d’autant évoluer favorablement que le patient participant au protocole bénéficie d’un soutien et d’une accessibilité médicale presque “inhabituelle”.

De même, les visites régulières et le cadre souvent sécurisé vont influencer de manière indirecte bon nombre de paramètres de la dépression (par exemple le pôle

anxieux dont on sait qu'il "côte" de manière importante dans les échelles de dépression). L'intérêt pour leur santé inhérent au principe d'un essai est en soi un élément rassurant donc thérapeutique sur certains aspects de la dépression que nous avons déjà souligné (l'anxiété).

### 3.3. Des choix statistiques discutés et des résultats discutables

Les auteurs analysent les données de deux manières : ils se servent d'un premier modèle, utilisant les différences moyennes standardisées et d'un deuxième comparant les moyennes arithmétiques pondérées (taille de l'effet de la méta-analyse).

De plus, la présence d'une hétérogénéité statistique entre différentes études doit être évaluée et les modèles à effet mixtes sont utilisés pour estimer la magnitude de la taille de l'effet.

Le modèle à effet fixe (*fixed effects model*), estime si les études incluses dans la méta-analyse montrent que le traitement produit un effet sur la moyenne. S'il n'y a pas d'hétérogénéité statistique démontrée, un modèle d'effets fixes, qui présuppose qu'il n'y a qu'une seule valeur sous-jacente pour l'effet constaté, peut être utilisé. Suivant ce modèle, une variation des effets observés est liée au hasard.

Le modèle à effet aléatoire (*random effects model*) permet de savoir, sur la base des études qui sont évaluées, s'il est possible de considérer que le traitement produira un résultat. Si une hétérogénéité statistique est démontrée entre les études d'une méta-analyse, on doit utiliser ce deuxième modèle. Ce modèle statistique considère que les effets divergents observés dans les études sont liés au hasard, mais aussi à des variations réelles entre les études. L'hypothèse d'un modèle d'effet aléatoire est qu'il existe une "population" d'effets éventuels avec une répartition précise autour d'un effet global moyen.

Le modèle à effet randomisé est sur le plan du calcul plus "pointu" que le modèle à effet fixe, mais si les données sont homogènes alors ils sont équivalents. Comme les analyses brutes montraient la même tendance, les auteurs ont préféré ne présenter que les résultats liés à l'effet fixe (supposant que les essais sont homogènes, ce qui reste à vérifier).

La moyenne pondérée de l'amélioration sur l'échelle de Hamilton était de 9,6 points pour le groupe antidépresseur et 7,8 pour le groupe placebo. Cette différence de 1,8 points, bien que significative (et remplissant les critères de la FDA pour lesquels une différence de 1 est suffisante) ne l'est pas sur le plan clinique selon les critères de la NICE (> 3 points).

En utilisant la différence moyenne standardisée, on obtenait alors une amélioration de 1,24 pour le groupe traitement et 0,92 pour le placebo (la magnitude de ces effets étant donc très importantes), mais on obtient une différence entre les deux groupes de 0,32 ce qui, encore une fois, ne répond pas aussi aux critères du NICE (< 0.5).

Pour certains, la méthode de calcul de la différence des moyennes du changement est mauvaise, car l'idéal est de calculer la différence pour chaque étude (amélioration moyenne du groupe antidépresseur moins amélioration moyenne du groupe placebo) et non de calculer les valeurs séparément pour chaque bras.

Donc, la taille de l'effet doit être calculée directement dans chaque étude et ensuite moyennée. En effet, ces essais peuvent montrer des différences d'amélioration pour de multiples raisons (par exemple les variations de la taille des échantillons dans le bras expérimental, régression à la moyenne, amélioration spontanée).

Or, par ce choix méthodologique ces variables sont confondues : on risque alors de perdre de l'informativité, car une diminution importante de l'HRSD peut être associée à une variance elle aussi plus importante, ce qui n'apparaît plus dans ce type d'analyse.

Aussi, comme le précise R. Waldmann, une meilleure estimation du bénéfice d'un antidépresseur vs placebo est calculée par  $1/((1/n)+(1/pn))$  ou  $n$  est le nombre de patients traités par antidépresseur et  $pn$  par placebo (8).

Avec cette formule, il recalcule l'estimation du bénéfice et obtient 3,23 pour les études publiées et 2,64 lorsqu'il inclut les 35 essais, soit un biais de publication d'environ 0,6 sur l'échelle de Hamilton. Dans le premier cas (essais publiés) le bénéfice répond aux critères du NICE et dans le deuxième cas (tous les essais) il s'en rapproche !

Comme nous avons pu le constater, les auteurs ont cherché à évaluer le degré d'hétérogénéité des données des études en utilisant les indices : le test Q de Cochran et  $I^2$  (indice de Altman et Higgins qui varie de 0 à 100 : une valeur  $I^2 < 25$  indique une hétérogénéité faible, des valeurs comprises entre 25 et 50 une hétérogénéité modérée et une valeur  $> 50$  une hétérogénéité importante (9).

Les auteurs trouvent des valeurs significativement élevées pour les deux tests dans chaque groupe ( $Q_{\text{antidépresseur}} = 51,8$  et  $Q_{\text{placebo}} = 74,59$ , et  $I^2_{\text{antidépresseur}} = 34,18$  et  $I^2_{\text{placebo}} = 54,47$ ).

Ils annoncent alors que les changements de moyenne exposés dans les études n'apportent qu'une description faible des résultats et que des modèles avec modérateurs sont indiqués.

Face à ce constat d'hétérogénéité, plusieurs attitudes auraient été possibles (10)

- soit renoncer à faire une méta-analyse (pour ensuite proposer une synthèse méthodique)
- soit exclure les études source d'hétérogénéité, à l'aide d'une analyse de sensibilité, tout en recherchant quelle composante de ces études en est à l'origine
- soit, rechercher les interactions entre les résultats observés et une ou plusieurs covariables, par une analyse en sous-groupes, par une modélisation de l'effet sur les données résumées des études ou en ayant recours à des modèles de type aléatoire.

Les analyses avec des modèles à effet fixe et aléatoires vont donc prendre en compte :

- le type de molécule utilisée, mais seulement 4 molécules sont représentées, ce qui devra limiter en soi la généralisation
- la durée de traitement : comme nous l'avons vu la majeure partie des essais dure 6 semaines, ce qui est court en regard des dernières études qui montrent la pertinence d'un traitement d'attaque de 2 mois (par exemple Star-D), et surtout qui ne permet pas d'évaluer le bénéfice sur la prévention de la rechute et des récurrences
- la sévérité initiale (sachant que la HRSD n'est pas la plus pertinente pour ce critère).

Un premier modèle montre que le type de molécule et la durée du traitement n'influencent pas les résultats qui varient juste en fonction de la sévérité initiale ; les modèles utilisés confirment ces tendances.

Dans un premier temps, on observe une courbe en U inversé, pour laquelle les scores initiaux les plus bas ou les plus élevés n'éprouvaient qu'un faible bénéfice alors que les sujets situés entre les deux avaient le bénéfice maximal.

Certains auteurs pensent que les méta-régressions comme celles utilisées par Kirsh ne sont pas adaptées à l'utilisation de données ordinales telles que celles obtenues avec la HDRS

La pente de la courbe du placebo diminuait alors que celle du traitement était positive. Ce résultat ne devait pas à priori leur convenir, puisqu'après quelques corrections (élimination d'un essai comportant deux sous groupes avec des scores très bas et très sévères) les auteurs constatent que l'aspect curviligne disparaissait et que si la pente du placebo baissait, celle des antidépresseurs n'augmentait plus.

Le seuil d'efficacité selon les critères NICE apparaissait pour des valeurs à la HRSD supérieures à 28.

Pour Kirsh et coll., c'est devant ce type de graphique "*qu'il paraît clair que l'augmentation de la différence est due à une diminution de l'amélioration dans le groupe placebo, plus qu'à une augmentation dans le groupe traitement*" !!!

#### 4. Le placebo à bon dos !!

Les auteurs font donc comme si les essais de cette méta-analyse étaient à priori destinés à évaluer l'efficacité du placebo, voir de les considérer comme un essai de non infériorité...du placebo.

Il paraît pertinent de reprendre quelques notions sur le placebo, ses mythes et ses réalités.

##### 4.1. Définition

Du latin "placere" (plaire) le terme placebo a donc pour traduction "je plairai". Son apparition dans le vocabulaire médical ("méthode banale de remède") date de 1785, dans le *Mot herby's New Medical Dictionary*, puis sa

définition se complète dans l'édition de 1803 du *New Medical Dictionary*, où on l'on propose que le placebo soit "*un épithète donnée à tout remède présent pour faire plaisir au patient plutôt que pour lui être utile*".

Le placebo aurait donc un "effet" et dans de nombreuses situations (prise en charge de la douleur, de l'asthme, de l'hypertension, mais aussi en psychiatrie pour certaines manifestations aiguës telles que l'anxiété ou les troubles du sommeil), et il a été donné pour "éviter" l'usage d'autres substances jugées plus "nocives" (le cas typique est celui de la morphine).

Le terme *placebo* entre dans les dictionnaires médicaux français à la fin des années 50, mais il va être surtout proposé comme une "*Substance neutre que l'on substitue à un médicament pour contrôler ou susciter les effets psychologiques accompagnant la médication*" ou "*Préparations pharmaceutiques (pilules, cachets, potions, etc.) dépourvues de tout principe actif et ne contenant que des produits inertes*".

P. Pichot définit dès 1961 l'effet placebo de cette manière : "*l'effet placebo est, lors de l'administration d'une drogue active, la différence entre la modification constatée et celle imputable à l'action pharmacologique de la drogue*".

Il est alors courant d'entendre que l'effet placebo améliorerait l'évolution de ces pathologies (tant au niveau subjectif qu'objectif) dans des proportions pouvant aller de 30 à 40 % (voir plus dans certaines conditions cliniques). Mais on a tendance à se baser sur des caractéristiques évaluées dans des travaux souvent anciens, réalisés dans des conditions méthodologiques peu rigoureuses.

La taille des échantillons, l'absence de double insu, les analyses statistiques parfois peu académiques constituent des biais évidents qui n'ont pourtant pas empêché de faire entrer ces chiffres (35 %) dans les esprits et les pratiques. L'effet placebo n'est plus seulement un outil pour "plaire au patient", mais aussi au prescripteur qui lui attribue les mêmes des propriétés.

##### 4.2. L'effet placebo : des faits, des mythes

###### 4.2.1. L'article de Beecher et ses conséquences

L'article princeps de 1955, écrit par Henry K Beecher est basé sur un protocole expérimental sur la douleur post-opératoire, réalisé en double aveugle, où les patients recevaient soit de la morphine, soit du sérum physiologique (11).

La réduction de la douleur sous placebo (mais aussi les effets secondaires ou "négatifs") était donc attribuée à l'effet placebo seul (sans tenir compte des autres paramètres pouvant influencer cette réponse, surtout dans le cadre de la prise en charge de la douleur). Il considère alors qu'un tiers des effets thérapeutiques observés dans les essais cliniques était donc du au placebo.

Pour renforcer son propos, son article était complété par une section comportant une méta-analyse de quinze essais, dans lesquels on considérait que les patients étaient améliorés de manière satisfaisante (ce qui ne veut

pas dire "significative"). L'effet global est alors estimé à 35,2 +/- 2,2.

Ce chiffre (qui deviendra un *golden standard*) correspond simplement à une fréquence des placebo-répondeurs au sein des essais, mais en aucun cas à un renseignement sur la taille réelle de cet effet.

Par la suite, en retrouvant la même grandeur dans un protocole évaluant un effet placebo "chirurgicale" (fausse intervention chirurgicales par des incisions mimant une ligature des artères mammaires, moyen utilisé dans les années 60 pour la prévention des douleurs angineuses), il généralise ses données à toutes les méthodes médicales. Pourtant, on constatera que l'amplitude de cet effet est tout sauf "fixe" et que l'on retrouve des valeurs allant de 0 à 100 %.

Ce papier est le plus cité dans la littérature de l'effet placebo et jusqu'en 1997 peu d'auteurs s'étaient attachés à en étudier la validité.

Selon Kienle et Kiene, pour démontrer son effet, un placebo doit obéir à trois critères fondamentaux (12) :

- le patient doit recevoir le placebo (administration perceptible quelle que soit la voie)
- la variation de l'événement doit être un effet du placebo, c'est-à-dire qu'il ne doit pas apparaître en l'absence de l'administration du placebo
- l'événement mesuré doit être lié à la maladie ou au symptôme pris en compte.

Après avoir examiné plus de 800 études de la littérature leur constat était plutôt inquiétant puisque la majorité des travaux ne pouvait conclure à l'existence d'un effet réel du fait de méthodologies médiocres, et surtout qu'un nombre important de facteurs pouvait créer "un faux effet". Sur ce principe les auteurs reprennent les études utilisées par Beecher dans sa méta-analyse et vérifient que ces facteurs ont bien été contrôlés ou pris en compte dans l'analyse des résultats. 14 études sur 15 ont pu être analysées selon ces critères (études en double aveugle), et ils constatent que Beecher a attribué rétrospectivement l'amélioration des symptômes à l'administration du placebo.

Evolution naturelle de la maladie
Amélioration spontanée
Fluctuations des symptômes
Habituation aux symptômes
Régression à la moyenne (tendance qu'ont les données expérimentales à se rapprocher de la moyenne lorsque l'on procède à des mesures répétées, la régression statistique sera proportionnelle au degré d'anormalité de la mesure initiale).
Effet des traitements adjuvants
Biais liés à l'observateur
Conditions de changement du traitement (placebo ou substance active selon l'intensité des symptômes)
Biais liées aux instruments de mesure et au recueil des variables
Mauvaise définition des critères d'efficacité de la substance active
Réponses sans rapport avec variable étudiée ou douteuses (ex : euphorie dans l'hypomanie spontanée, amélioration liée à l'arrêt d'un traitement ayant des effets secondaires désagréables)
Persistance d'effets toxiques de substances données antérieurement
Biais liés aux patients
Réponse de politesse envers un médecin qui "cherche à l'aider"
Subordination expérimentale (le sujet dit ce qu'il pense être la réponse attendue)
Réponses conditionnées (apprentissage lié à la prise de médicament et un effet attendu)
Défaut de jugement d'origine psychotique ou névrotique
Pas de placebo donné
Psychothérapie
Médecine traditionnelle, sans prise de traitement.
Case report non vérifiables ou non critiquables.
Attribution erronée au placebo d'effets secondaires négatifs comme preuve de son activité (qui peuvent être en fait des symptômes de la pathologie, voir d'une comorbidité)

**Tableau 2.** Facteurs pouvant créer l'impression d'un effet placebo, d'après Kienle et Kiene (12).

Dans certains cas les auteurs de ces essais n'avaient attribué aucune activité au placebo et Beecher aurait interprété de manière erronée certains résultats.

Kienle et Kiene proposent donc une série de facteurs de confusion (cf tableau 2) et ont recherché dans chaque essai si ceux-ci pouvaient avoir influencé les résultats. Dans 14 études (la dernière n'étant pas interprétable) l'existence d'un effet pour le placebo ne pouvait être démontré, puisque les effets observés étaient tous indissociables dans l'absolu de l'évolution naturelle des symptômes et sujets à de nombreuses erreurs potentielles de mesure.

#### 4.2.2. L'effet placebo : un biais méthodologique ?

L'article de Hrobjartsson et Gotzche, publié dans le New England Journal of Medicine en 2001, proposa de clarifier cette notion à l'aide d'une méta-analyse très bien menée (13).

L'objectif n'étant pas de s'intéresser à l'utilisation du placebo dans les études contrôlées, mais à l'effet clinique qui leur est attribué dans le traitement de certaines maladies.

Les études retenues étaient celles qui comportaient un bras placebo (défini par les auteurs comme traitement contrôle, dont l'aspect était similaire au traitement de l'étude mais dénué d'activité spécifique), mais surtout elles devaient comporter un bras de patients "sans traitement". L'évaluation d'un potentiel "effet placebo" était ainsi basée sur la comparaison entre les résultats de ces deux groupes (placebo vs "pas de traitement").

De nombreux paramètres étaient évalués, tels que la nature du placebo (pharmacologique : e.g. comprimé, physique : e.g. manipulation ou psychologique : simple conversation) ou bien si les problèmes cliniques rapportés par les patients pouvaient être observés par un tiers et si les résultats objectifs reposaient sur des mesures de laboratoire ou bien un examen de mandant la coopération des patients.

Les études incluant des volontaires sains, ainsi que des sujets rémunérés n'étaient pas retenues.

Les résultats étaient regroupés soit de manière "binaire" (ex : fumeurs vs non fumeurs), soit continue (nombre de cigarettes), la première manière étant préférée. De plus pour éviter de comparer des données avec un taux de sorties d'étude trop élevé ou de données manquantes, les résultats "immédiats" étaient préférés (c'est à dire les mesures effectuées juste après le traitement) par rapport aux données prospectives.

32 essais pour les résultats binaires ont été sélectionnés (n = 3795) et 82 pour les variables continues (n = 4730). Les pathologies étudiées étaient très hétérogènes (un quarantaine), telles que l'asthme, hypertension, troubles lipidiques, dépression, schizophrénie, troubles anxieux et phobiques, ménopause, maladie de parkinson, douleurs,

syndrome du canal carpien, infections virales ou bactériennes, tabagisme, obésité...

Pour l'analyse des résultats binaires les auteurs ont calculé le risque relatif (RR) d'un résultat non désiré, à partir du nombre de sujet ayant un résultat inattendu dans chaque groupe (placebo et aucun traitement). Un RR inférieur à 1 indiquait un effet bénéfique du placebo, sauf si l'intervalle de confiance (IC) englobait cette valeur.

Pour les variables continues, la comparaison reposait sur la différence des valeurs moyennes d'événements non attendus dans chaque groupe divisée par la déviation standard "poolée". Une valeur de -1 signifie que la moyenne dans le groupe placebo était de 1 déviation standard sous la moyenne du groupe sans intervention, indiquant un effet bénéfique.

Concernant les résultats binaires, le placebo n'a montré aucune activité par rapport à l'absence d'intervention (RR = 0.95, IC : 0.88-1.02), une analyse en prenant en compte la nature subjective ou objective des variables mesurées ne montrait pas d'effet du placebo.

L'hétérogénéité de certains résultats (effet placebo plus prononcé dans certains essais) était due selon eux, à la petite taille des échantillons.

L'évaluation d'un effet potentiel dans des conditions "binaires" spécifiques (tabac, nausées, dépression) ne permettait pas de mettre un quelconque effet placebo dans ces domaines cliniques.

Pour les variables continues, (que les auteurs estiment moins pertinentes, car impliquant nécessairement l'attribution d'un "Cut-off" pouvant être arbitraire), ils observaient des différences significatives pour les données subjectives, mais pas pour les données objectives. De la même manière, la taille de l'effet diminue proportionnellement à l'augmentation de la taille de l'échantillon et l'effet placebo semble plus spécifique dans la douleur (il n'est pas retrouvé pour des pathologies telles que l'obésité, l'asthme, l'HTA, l'insomnie ou l'anxiété...).

De plus on n'observe pas de différence selon les types de placebo (pharmacologique, physique ou psychologique) qui n'ont individuellement, selon les résultats des auteurs, aucun effet.

Ils concluaient alors que devant l'absence d'évidence en faveur d'un effet réel du placebo, celui ci ne devait pas être utilisé en dehors des essais thérapeutiques et absolument pas dans une logique thérapeutique.

Hrobjartsson et Gotzche ont reçu un courrier important critiquant des aspects de leur analyse. Pour certains, l'hétérogénéité des situations cliniques observées allait à l'encontre du principe même de la méta-analyse qui était de comparer des données comparables (mais pour les auteurs, ce type d'approche "plus large", reste adaptée pour mesurer l'effet placebo ou l'homéopathie). De même le fait que certains facteurs pouvant produire un

effet placebo (régression à la moyenne, réponse de politesse, mesures répétées) affectent aussi le "groupe traitement" faisait dire à d'autres que l'on ne mesurait pas l'effet du placebo, mais que l'on comparait en fait deux groupes de placebos (mais la définition du placebo étant claire dès le départ, cette assertion reste peu recevable dans l'absolu).

#### 4.2.3. *Que penser de la conclusion de Kirsh ?*

Comme nous l'avons déjà évoqué, que penser de la conclusion des auteurs qui stipule que *"la relation entre la sévérité initiale et l'efficacité des antidépresseurs est attribuable à la diminution de la réponse au placebo chez les patients les plus sévères, plus qu'à une augmentation de la réponse au traitement"*.

Un essai randomisé double aveugle avec bras placebo n'est pas mis en œuvre pour dire si un traitement est bénéfique dans certaines situations (par exemple les dépressions sévères) et que le placebo ne l'est pas (ce que suggère cette conclusion), mais seulement qu'un effet augmente lorsque la dépression est plus sévère.

**Face à une différence significative on doit conclure à un effet de la molécule et non l'inverse.** Car considérant la définition d'un placebo, de son effet (avec tout le niveau d'incertitude scientifique qui l'entoure) et des objectifs des essais randomisés, la conclusion de Kirsh et ses collaborateurs est douteuse sur le plan scientifique.

De même, pour la dépression moyenne, la généralisation des résultats repose en fait sur une seule étude !

Des indicateurs d'efficience, tels que la qualité de vie, les facteurs d'observance, la tolérance interviennent dans l'évaluation de l'indication d'un traitement, en plus de son efficacité mesurée psychométriquement. Dans leur discussion, ces notions ne sont pas abordées.

Nous avons vu que de nombreuses interprétations viennent entacher la volonté d'objectivité initiale des auteurs (qui ne sont pas a priori des prescripteurs car non médecins).

De par ces conclusions, la méta-analyse de Kirsh est un peu comme une formule 1 qui démarre bien sa course et sort de piste avant la ligne d'arrivée...

## REFERENCES

1. Kirsch I., Deacon B.J., Huedo-Medina T.B., Scoboria A., Moore T.J., Johnson B.T. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008 ; 5 : e45.
2. HAS. Guide d'analyse de la littérature et gradation des recommandations. Janvier 2000.
3. Cucherat et al. Lecture critique des meta-analyses. <http://w.w.w.spc.univ-lyon1.fr/lecture-critique/metaanalyse/frame1.htm>
4. Moher D. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*. 1999 Nov 27 ; 354(9193) : 1896-900.
5. National Institute for Health and Clinical Excellence. Depression: management of depression in primary and secondary care. Clinical practice guideline CG23. 2004.
6. Turner E.H., Matthews A.M., Linardatos E., Tell R.A., Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*, 358, 252-260.
7. Turner E.H., Rosenthal R. Editorial: efficacy of antidepressant. *BMJ* 2008 ; 336 : 516-517 (8 March).
8. Waldmann R. : <http://rjwaldmann.blogspot.com/2008/03/just-cant-let-it-go.html>
9. Chevalier P., van Driel M., Vermeire E. Hétérogénéité dans les synthèses méthodiques et méta-analyses : *Minerva* 2007 ; 6(10) : 160-160.
10. Higgins J.P.T., Green S. *Cochrane handbook for systematic reviews of interventions*, 2008, p 132.
11. Beecher HK The powerful placebo. *J Am Med Assoc*. 1955 Dec 24;159(17):1602-1606.
12. Kienle GS, Kiene H. 1997. The powerful placebo effect: fact or fiction ? *J Clin Epidemiol*. 50 : 1311-1318.
13. Hróbjartsson A, Gøtzsche PC. Is the placebo powerless ? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med*. 2001 May 24;344(21):1594-602. Review. Erratum in: *N Engl J Med* 2001 Jul 26 ; 345(4) : 304.

**Mots clés :** antidépresseurs, méta-analyse