Quelques réactions à un article de Hoogman *&al.*(2017)

Le problème de l'évaluation d'un système de mesure

Marc Bourdeau

Professeur associé, École Polytechnique de Montréal

Ceci est une ébauche, à ne pas citer nommément.

Lien avec l'article de : Hoogman &al. « Subcortical brain volume differences […] », Lancet Psychiatry 2017. On trouvera d'autres références que celles en fin de texte sous ce lien.

Les auteurs se demandent si les différences volumétriques du cerveau sous-cortical pourraient entrer dans une explication du TDAH.

Les facteurs qui joueraient éventuellement sur cette différence : le sexe, l'âge, la gravité, les médicaments utilisés par les sujets, et sans doute pas mal d'autres (dont l'origine ethnique des sujets, etc.), ne sont pas considérés, sauf l'âge. On fait ensuite des études avec des tests T, même pas appariés à ce que je comprends. Je serais curieux de voir les études originales : des Anova multi-factorielles, à plusieurs facteurs, enfin j'espère, mais peu de sujets.

On rapporte les *d* de Cohen associés aux différences entre sujets sains (témoins) et sujets affectés de TDAH pour une méga-analyse (à ne pas confondre avec une méta-analyse) qui a regroupé les données individuelles (ce qui en soit pose un problème d'homogénéïté).

Normalement on calcule des *d* de Cohen sur des paires de groupes, donc d'une Anova à un facteur (deux niveaux). Le protocole d'analyse statistique (les analyses de covariance en particulier) et les résultats de ces analyses sont trop lacunaires pour qu'on ait une idée claire, de la qualité du regroupement.

Une limite peut-être importante sont les tailles des groupes regroupés ici. Ainsi, la médiane des âges de la totalité des sujets serait aux environs de 14 ans. Il serait étonnant de trouver beaucoup de sujets plus vieux que, disons, 25 ans, et probablement 18 ans… Aucun histogramme ne permet de porter un jugement cette caractéristique de cohorte (il ne s'agit pas d'un échantillon).

Et pourtant, les graphiques de la page 6 font état de tous les âges dont les points de l'abscisse sont « empiétants » (regroupements de 5 années décalés d'un an d'abscisse en abscisse)! c'est la première fois que je vois ce genre de graphique.

Quant à faire des remarques sur ces graphiques, le dernier en bas à droite de la page, pour le volume intra-cortical a des ordonnées exprimées on ne sait trop en quelles unités, des mm^3 pour les autres graphiques mais on voit pour le dernier des puissance -6 d'une unité de longueur, le -6 est à peine visible sur ma copie, est-ce à dire en cm^3 ? Mais alors le volume intracrânien serait de l'ordre de 2 cm^3… Tout cela est probablement évident pour les spécialistes du domaine.

Pendant qu'on est sur la questions de ces mesures, on pourrait se demander quelle est la précision du système de mesure? C'est en fait la première question à se poser quand on prend des mesures. Ce système comprend 27 instruments d'imagerie (27 sites), des IRM (MRI en anglais), des techniciens mesureurs sur chaque site, et des mesures uniques (i.e. non répétées) de parties

vivantes de sujets. Quelle en est la précision? Normalement, les entreprises qui produisent ces IRM (instruments de résonance magnétique) fournissent des spécifications pour les précisions… Et quantifient les précisions en termes des pourcentages de variation pour chaque facteur de variation : instruments (sites) × techniciens × répétitions. : ce sont-là les résultats d'une étude dite de répétabilité & reproductibilité (en anglais : *R&R gage studies, cf.* Wheeler, 2006).

Disons, pour faire simpliste, qu'on mesure des petits cubes de matière (inanimée), de l'ordre de 5000mm^3 grandeurs qu'on retrouve sur les figures de la page 6, donc d'environ C=17mm de côté ($V=C^3$) avec une précision de 0,5mm sur chaque dimension (erreurs de répétition). Les erreurs sur chacune des dimensions se transmettent à l'erreur sur l'erreur du volume de façon simple. Ce qui n'est pas le cas dans les mesures de volumes des organes vivants par MRI.

Sur le volume l'erreur est de l'ordre de 400mm^3 ($dV=3C^2dC$, la dérivée de V par rapport à C de $V=C^3$). Juste pour l'erreur du système de mesure, on ne parle pas de celle issue des techniciens qui joue peut-être un rôle, de l'erreur de répétition, et de réplication des systèmes de mesure (en anglais « *R&R gage studies* »). Que dire lorsqu'on mesure des organes du cerveau tels que ceux mentionnés dans le texte ?..

Si on mesure la surface de tranches du cerveau dûment multipliée par la séparation entre les tranches (on pense à un pain tranché dont on veut connaître le volume), et qu'on additionne le 'volume' de chaque tranche ainsi obtenue pour connaître le volume total de certains organes, l'effet des erreurs de mesures sur les surfaces, S, des tranches espacées d'une certaine distance, E, sur l'erreur d'estimation du volume, V, d'un organe n'est pas difficile à calculer. Les variations dS (surface des tranches), dE (espacement des tranches) sur la variation présumée du volume dV du volume d'un organe précis n'est pas bien difficile à obtenir.

Naturellement, il faudrait savoir avec plus de détails comment on obtient les estimations volumétriques issues des IRM (les auteurs mentionnent les algorithmes FreeSurfer en 2 versions), qui seraient décrits dans un appendice que je n'ai pas Il n'est pas clair qu'une étude R&R soit simple à réaliser en IRM !

Alors que dire des conclusions qu'on en tire sur les différences de volumes entre les deux groupes (cas et témoins), et on sait que les différences sont terribles pour la propagation des erreurs sur chaque terme, surtout pour les petites valeurs...

Rappelons que les *d* de Cohen[1] sont des différences entre paires de moyennes normées par un écart type conjoint/regroupé (en anglais, on dit *pooled).* Ils expriment donc les différences de moyennes en unités d'écart type regroupé. Les *d* sont sans unité. Il y a des balises pour juger de la grandeur des effets d'une groupe à l'autre. Pour une illustration intéressante.

Kline (2013, p. 140) rapporte que les erreurs ont tendance à sous-estimer les variances des mesures (ce qui est évident), et donc à surestimer les *d* de Cohen.

[1] Il y a plusieurs sortes de *d* de Cohen suivant les mesures d'écarts types utilisés. Les auteurs sont ici muets.

Une petite recherche sur Google rapporte nombres d'articles sur les problèmes des erreurs issues des IRM notamment tels ceux des *T1-weighted MRI* utilisés ici :

https://www.theguardian.com/science/neurophilosophy/2015/apr/09/bold-assumptions-fmri

Dont la citation: «*The article focuses on the problems plaguing functional neuroimaging research - small sample sizes, low statistical power, and lack of replicability - and how researchers in the field are facing up to them.* »:

Et encore:

https://www.braindecoder.com/post/bold-assumptions-why-brain-scans-are-not-always-what-they-seem-1069949099

http://appliedradiology.com/articles/radial-t1-weighted-magnetic-resonance-imaging-background-clinical-applications-and-future-directions (Applied Radiology, 2016)

Dont la citation: « *However, it is still prone to several technical shortcomings, one of the most important being **its*** **sensitivity to physiologic and respiratory motion** (c'est nous qui soulignons). *This concern becomes of utmost significance when imaging certain body regions in debilitated patients, the elderly, and children, who cannot perform adequate breath holds. » (et les sujets atteints de TDAH*…)

Un peu partout dans le texte, les innombrables auteurs (Hoogman *&al*) mentionnent que la puissance de cette étude est élevée, cela est dû aux grandes tailles échantillonnales de deux groupes (les données proviennent de 27 sites). Et même à un endroit, on mentionne la puissance élevée du *d* de Cohen. Je ne connais pas la puissance associée à un *d* de Cohen (je ne trouve pas dans Google). Pour mémoire, la puissance, par définition, c'est la probabilité (qu'il faut préciser) pour détecter des écarts non nuls (lesquels) à une hypothèse nulle. Tout ça m'échappe ici. Aucune précision n'est apportée dans le texte (à moins que je n'aie lu trop rapidement). Mais chaque étude individuelle est peu puissante, qu'en est-il sur le regroupement?

Les auteurs mentionnent qu'ils font une méga-analyse, sensée corriger certains défauts d'une méta-analyse, telle surtout, l'effet du biais entraîné par l'effet de la `filière ronde', celui de non publication des résultats non significatifs. Les auteurs ne développent pas. La méga-analyse décrite provient d'articles déjà publiés. Et il n'est question nulle part (?) des études de R&R qui fondent la qualité des appareils de mesure.

Je n'aime pas beaucoup les méta-analyses, dont la validité est souvent très faible. Et je ne connais pas les méga-analyses. Ce terme semble venir du regroupement des sujets (« *pooling* ») d'analyses indépendantes, après quelques précautions pour harmoniser les résultats qui sont rapportés dans l'annexe que je n'ai pas.

Cela étant noté, on constate (Tableau 2 et 3) à une exception près les *d* de Cohen sont presque tous inférieurs à 0,2 (en grandeur), seuil qui est qualifié de faible dans la littérature. Les examens de leurs intervalles de confiance 95%IC sont très intéressants eux aussi (rapportés toujours aux Tab. 2 & 3), presque la moitié comprend la valeur d=0, qui signifie une absence totale d'effet. Les *d* inférieurs à 0,1 sont médiocres. Mais pourquoi, la Fig. 1 ne montre-t-elle pas les valeurs de *d* par classe d'âge avec ces intervalles de confiance pour les valeurs de *d* elles-mêmes et non les erreurs types ce qui n'a pas le sens intuitif des IC? L'erreur type s'explique mal, puisque, si je

comprends bien les auteurs dans chaque cas, la méga-analyse ne produit qu'un seul *d*. Les *d* restent faibles pour les cas où les probabilités de dépassement $p_{dép}$ sont infimes (en anglais *p-value*). Les auteurs ont raison de souligner que cette étude n'est pas propice pour poser des hypothèses sur les différences de volumes d'un groupe à l'autre… Ce que confirme amplement les graphiques de la page 6. Les auteurs prétendent (p.8), sans référence à l'appui, avoir la puissance pour détecter des *d* jusqu'à d=0,08. Cette question reste à creuser, mais n'est pas cruciale puisque les intervalles de confiance à 95% pour les *d* les amènent rarement dans des zones confortables

Conclusions (provisoires)

Ce texte rapporte trop peu d'éléments, finalement, d'une étude fort complexe pour le recueil des données et la constitution de la matrice de données à analyser, ainsi que la validation de la précision attendue. Mais l'analyse des données, une fois la base de données constituée et validée, serait normalement assez directe. On s'attendait à beaucoup de graphiques à cet égard. On n'a qu'une seule figure, et encore pose-t-elle des problèmes. Normalement, selon les directives qui ont cours, mais ne sont pas souvent respectées (Ellis, 2013), on devrait trouver quelque part les données et suffisamment d'indications pour reproduire les analyses rapportées (APA, 2010; Bailar & Mosteller, 1989; Wilkinson, 1999 — surtout le Bailar & Mosteller qui sert de modèle à presque tous les guides de publication). Ce n'est pas le cas ici. Le plus gros problème à mon humble avis, est l'absence des études de validation des systèmes de mesure.

Références

APA, 2010, *Publication manual of the American Psychological Association*, 6e éd., Washington DC: American Psychological Association.

John C. Bailar the 3rd & Frederick Mosteller, 1988, « Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations ». *Annals of internal Medicine,* 108(2), 266-273.

Jacob COHEN, 1988, *Statistical power for the behavioral sciences*, 2e éd., Hilldale NJ: Lawrence Erlbaum Associates.

Paul D. ELLIS, 2013, *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results.* Cambridge UK: Cambridge University Press.

Rex B. KLINE, 2013, *Beyond significance testing. Statistics reform in the behavioral sciences*, 2e éd., Washington DC: American Psychological Association.

Xiaofeng Steven LIU, 2014, *Statistical power analysis for the social and behavioral sciences. Basic and advanced techniques*, New York NY: Routledge.

L. Wilkinson & Task Force on statistical inference, 1999, « Statistical methods in psychology journals: guidelines and expectations », *American Psychologist*, 54(8), 594-604.

Donald J. Wheeler, 2006, *EMP III. Evaluating the measurement process & using imperfect data.* Knoxville TN: SPC Press.

Magn Reson Imaging. 1991;9(4):589-95.

Sources of error in the quantitative analysis of MRI scans.

Plante E[1], Turkstra L.

Author information

Abstract

The increasing use of quantitative analysis of MRI scans in the literature has produced a need to identify potential sources of bias in such analysis procedures. Six sources of bias are demonstrated in this paper. These include bias attributable to partial volume effects, head tilt, plane of view, use of noncontiguous slices, contrast/intensity manipulations, and magnetic inhomogeneities. The magnitude of bias for each source varied according to whether a hemisphere or regions within a hemisphere were measured, with regional effects typically exceeding hemisphere effects.

PMID:

1779731

- Format: Abstract

Send to

Med Biol Eng Comput. 1993 Nov;31(6):600-6.

Processing MRI data for electromagnetic source imaging.

Wieringa HJ[1], Peters MJ.

Author information

Abstract

Estimation of the source of activity in the brain from electro- and magneto-encephalographic measurements is becoming increasingly common. Structural information could assist in improving the calculation of the sources as well as providing the context of the source location. Magnetic resonance images are very useful for this purpose, but they still need to undergo various processing steps. The paper describes in detail a practical method for full automatic processing of MRI images of a head, including segmentation of the images and triangulation of the surfaces.

PMID:

8145586