LES RED FLAGS DES ÉTUDES SCIENTIFIQUES

UN GUIDE POUR EXERCER UN ŒIL CRITIQUE SUR LES ÉTUDES QUANTITATIVES RÉCENTES À PARTIR DE DIFFÉRENTS INDICATEURS DE QUALITÉ

1. Les données et analyses ne sont pas en libre accès
2. Les P valeurs* d'intérêt sont proches de 0.05
3. Il n'y a pas de pré-enregistrement*
4. Il n'y a pas de visualisation de la distribution des données
5. L'article est publié dans une revue prédatrice ou peu filtrante
6. L'article n'a pas été publié - le cas des préprints*
7. Il est impossible d'accéder à la procédure exacte
8. Il n'y a pas de filtre ou de vérification des données

B. Bonus : Tableau des risques dans les études cliniques

BETA VERSION. SI VOUS AVEZ DES SUGGESTIONS, CONTACTEZ ADRIENFILLON@HOTMAIL.FR

* voir glossaire

En 2005, Ioannidis bouleverse le monde biomédical en affirmant qu'au moins la moitié des recherches sont des faux positifs : des études qui indiquent un effet positif alors qu'il n'y en a pas en réalité. Il explique que dans le fonctionnement de la recherche de l'époque, il n'est pas possible de publier un article qui n'a pas réussi à trouver des effets. Mais pour obtenir un emploi et du financement, il faut publier des articles. Ainsi, les chercheurs sont poussés à rapporter des effets, quitte à ce qu'ils soient des faux positifs. Pensez-y : préféreriez-vous faire une étude sur 1000 participants ou 10 études sur 100 participants (ou encore mieux, 33 études sur 30 participants) ? Une étude négative sur 1000 participants serait impubliable. Sur 10 études, par contre, vous aurez une bonne chance qu'au moins une soit positive et publiable, et vous pourrez mettre les résultats des neuf autres dans le tiroir (ce que l'on appelle "l'effet tiroir"). En effet, les tests statistiques sont ainsi construits que même dans le cas d'une absence objective d'effet, il existe une probabilité (généralement 5%) que le test puisse conclure erronément à la présence d'un effet. Ainsi, si vous refaites 20 fois une étude parfaitement, même en l'absence objective d'effet vous obtiendrez en moyenne une étude indiquant tout de même la présence d'un effet.

L'autre problème des études avec peu de participants est l'augmentation du risque de ne pas avoir d'effet statistiquement significatif alors que l'effet existe, des faux négatifs dûs à ce que l'on appelle un "manque de puissance statistique*". En neurosciences par exemple, la puissance statistique moyenne estimée est de 20% (seulement 20% de chance de détecter un effet s'il existe) alors qu'on estime souvent qu'un seuil minimal serait de 80%. En psychologie, en 2016, la puissance statistique serait de 24%. Paradoxalement, l'essentiel des études dans ces littératures devraient être négatives, tandis que l'on y constate une majorité écrasante de résultats positifs. Le problème est éloquent.

En 2012, Simmons, Nelson et Simonsohn publient un article : "La psychologie comme faux-positif : la flexibilité non divulguée de la collecte et de l'analyse des données permet de présenter n'importe quoi comme significatif". Cet article est fondateur dans l'analyse critique des études en psychologie.

Le second séisme apparaît en 2015 avec la publication dans Science de l'essai de réplication de 100 études en psychologie sociale, avec seulement 36% de réussite. Ce manque de réplication a questionné les chercheurs sur la scientificité de la psychologie, mais aussi d'autres domaines de recherche (62% de réplications dans les études en psychologie publiées dans les journaux de Nature et Science, 61% de réplications en économie comportementale, 46% dans les recherches sur le cancer, ...).

Par-dessus ce constat déjà dramatique vient se greffer l'observation que les études rétractées sont plus citées que les rétractions. Pire, les études rétractées continuent d'être utilisées dans les revues systématiques et les méta-analyses. Le taux de rétraction augmente, passant de 45 par mois en 2010 à 300 par mois en 2023, ce qu'on estime toujours bien trop peu. L'acquisition de financements selon le nombre d'études publiées a encouragé la mise en place d'un marché de la publication : des entreprises créeent de faux articles faux qu'elles vendent à des chercheurs en demande, les fameux "paper mills" (fermes à articles). On estime que 2% de toute la littérature scientifique provient de "paper mills", ce qui correspondait environ à 400 000 articles publiés ces dernières années (70 000 seulement en 2022). L'utilisation de l'IA comme ChatGPT permet de publier encore plus vite en automatisant les processus d'analyse, d'écriture, voir en créant des jeux de données faux mais parfaitement réalistes.

Selon Dorothy Bishop, aujourd'hui, la réponse de la communauté scientifique est complètement inadéquate. Il y a peu de tentative de vérification automatique de fraude : la science est perçue comme une poursuite gentleman où nous devrions supposer que tout le monde possède des intentions honorables. Même quand des cas clairs de méconduites sont mis en évidence, il peut se passer des mois, voire des années pour rétracter un article.

Pourtant, ces cas de fraude existent bel et bien. Vous pourrez très bien tomber sur une étude dont les résultats ont été entièrement inventés par les chercheurs, ou générés par IA. Bien sûr, le plus souvent les études ont réellement été conduites mais contiennent des erreurs et des méconduites de recherche qui empêchent d'accorder du crédit aux arguments avancés par les auteurs. Basé sur cette description du terrain de la recherche en psychologie, nous en sommes aujourd'hui à devoir tous nous interroger :

Comment se fier à une étude scientifique ?

Ce guide a pour objectif de permettre aux personnes (un peu) motivées à lire un article scientifique de pouvoir décoder si cet article paraît fiable ou non. L'objectif est de s'adresser au plus grand nombre, même des personnes n'ayant que peu de connaissances en psychologie, en statistique ou en science de manière plus générale. Cependant, il est nécessaire de connaître ce qu'est une moyenne, un écart-type et de vouloir lire des tableaux et des figures. Il n'est tout simplement pas possible d'exercer un œil critique sur un article scientifique si on ne va pas au moins regarder les données descriptives.

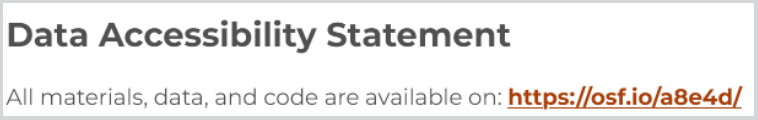
Les indicateurs que nous proposons, les "Red Flags", ne sont que des indicateurs que l'étude pourrait être de mauvaise qualité. Il existe des articles fiables en présence de ces indicateurs, tout comme il existe des articles non-fiables en leur absence : une investigation exhaustive d'article réclame énormément de temps et de compétences. Ce ne sont que des indicateurs qui nous paraissent les plus pertinent au moment de rédiger ce guide.

Ces indicateurs ont pour avantage d'être simples à vérifier, suffisamment généraux pour s'appliquer à l'ensemble de la psychologie, voir à d'autres sciences médicales et sociales, et sont basés scientifiquement. Nous accompagnons chaque indicateur de quelques références. Pour les chercheurs, il en existe d'autres qui dépendent du niveau d'expertise dans le domaine ou la méthodologie scientifique. Nous avons aussi volontairement réduit le nombre d'indicateurs pour avoir un guide facile d'accès.

Pour nous, tous ces indicateurs n'ont pas la même importance. Ainsi, nous avons évalué un degré d'importance de chaque indicateur à prendre en compte pour estimer la qualité de l'étude, et cocher plusieurs indicateurs peut renforcer l'idée que l'étude a un niveau de qualité insuffisant pour lui accorder sa confiance. Enfin, ces indicateurs prennent tout leur sens pour évaluer les études publiées actuellement. Il est normal que les études publiées quand le pré-enregistrement n'existait pas ne soient pas pré-enregistrées. Pour ces études, il est important de vérifier qu'elles aient été répliquées ou que la théorie ait été vérifiée par une méta-analyse de qualité.

1 Les données et analyses ne sont pas en libre accès

À quoi faire attention ?



Vérifier si les données, sous format ouvert, ont été mises à disposition par les auteurs. Parfois, les recherches utilisent des données externes comme provenant de grosses bases de données. Dans ce cas, il faut vérifier si elles sont accessibles.
Vérifier si l'analyse des données est ouverte et accessible, par exemple comme un script R, Python, ou lisible par un logiciel open source comme Jasp ou Jamovi.

Pourquoi faire attention ?

Un article scientifique doit amener des preuves solides que ses arguments sont fondés. Un article sans données et analyse n'est fondamentalement pas différent d'une publicité commerciale. Beaucoup d'articles indiquent que les données sont disponibles sous requête "data available upon request". C'est faux. Hussey (2023) a demandé à obtenir les données de 52 articles et n'a reçu les données que de 14 (27%) des articles. En majorité, les données disponibles sous requête ne sont pas disponibles. Dans un contexte ou un chercheur comme Dieterich Stapel a pu inventer les données de plus de 143 articles, ou dans lequel deux équipes de chercheurs indépendants ont pu fabriquer des données de deux expérimentations d'une même étude (Data Colada #109), il n'est pas possible de croire les chercheurs sur parole.

Degré d'importance

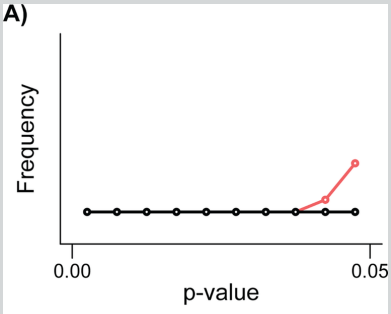
Modéré.
Beaucoup de chercheurs ne rendent pas leurs données et analyses disponibles alors qu'elles sont correctes, principalement à cause d'un manque de formation et de sensibilisation à ces problèmes. D'un autre côté, il existe des études avec des données et analyses disponibles, mais mauvaises, car les peer-reviewers ne les vérifient pas et qu'aucune procédure n'est mise en place pour automatiser l'analyse de la qualité des données et des analyses.

Références

- Data Colada 109 (2023). Data Falsificada (Part 1): "Clusterfake". http://datacolada.org/109
- Enserink, M. (2012). Final Report on Stapel Also Blames Field As a Whole. In Science (Vol. 338, Issue 6112, pp. 1270–1271). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.338.6112.1270
- Hussey, I. (2023, May 8). Data is not available upon request. https://doi.org/10.31234/osf.io/jbu9r

2 Les p-valeurs* d'intérêt sont proches de 0.05

À quoi faire attention ?



Head et al., 2015

Dans le résumé de l'article ou dans les résultats principaux, il est anormal d'avoir une surreprésentation de p valeurs entre 0.01 et 0.05 par rapport à des p valeurs supérieures à 0.05 ou inférieures à 0.01.

Pourquoi faire attention ?

Les chercheurs décident généralement de déclarer un résultat significatif si une p-valeur est inférieure à p = 0.05. Utiliser cette valeur comme critère a amené à des biais de publications* dont la manifestation est une distribution anormale des p-valeurs, juste en dessous de 0.05.

Ainsi, si on regarde les p-valeurs d'intéret dans un article, on ne devrait jamais observer de nombreuses p-valeurs proches de 0.05 dans le cas où il y a réellement un effet. Ce type de résultat provient plus vraisemblablement de biais de publications (manque de puissance* ou p-hacking*).

Degré d'importance

Extrêmement préoccupant

La distribution des p-valeurs est probablement la première chose à regarder dans un article pour s'assurer de la confiance dans les résultats. Si les p-valeurs sont majoritairement très proches de 0.05, il ne faut vraisemblablement pas accorder trop de confiance dans les découvertes faites dans cet article.

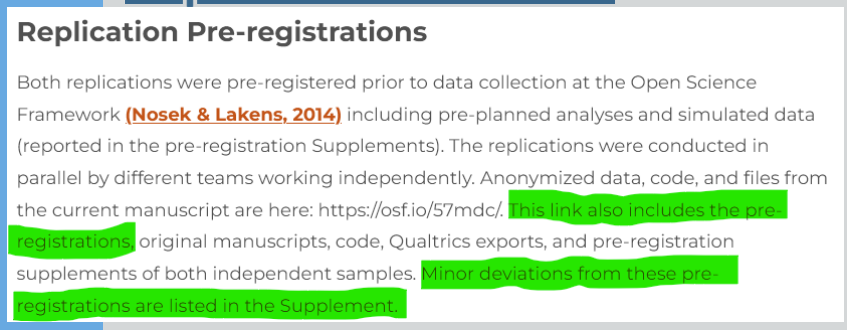
En plus de l'analyse des p-valeurs, des chercheurs plus expérimentés peuvent aussi se trouver vers l'analyse Z-curve qui, étant plus complexe, semble être aussi plus prometteuse dans l'estimation des risques de faux positifs.

Références

- Bartoš, F., & Schimmack, U. (2020, January 10). Z-Curve 2.0: Estimating Replication Rates and Discovery Rates. https://doi.org/10.15626/MP.2021.2720
- Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." Journal of Experimental Psychology: General 143 (2): 534–47. doi:10.1037/a0033242.

3 Il n'y a Pas de pré-enregistrement* / Pas de Registered Report

À quoi faire attention ?



Replication Pre-registrations

Both replications were pre-registered prior to data collection at the Open Science Framework **(Nosek & Lakens, 2014)** including pre-planned analyses and simulated data (reported in the pre-registration Supplements). The replications were conducted in parallel by different teams working independently. Anonymized data, code, and files from the current manuscript are here: https://osf.io/57mdc/. This link also includes the pre-registrations, original manuscripts, code, Qualtrics exports, and pre-registration supplements of both independent samples. Minor deviations from these pre-registrations are listed in the Supplement.

Dans le résumé ou dans la partie méthode d'un article, il est possible de vérifier si les auteurs ont indiqué avoir fait un pré-enregistrement. Le pré-enregistrement se fait généralement sur OSF, Aspredicted, etc.

Un préenregistrement possède nécessairement une date exacte qui doit être antérieure à la collecte des données, des hypothèses claires et une méthodologie claire d'analyse de ces données.

Pourquoi faire attention ?

Il y a peu de standardisation en sciences sociales, ce qui amène à avoir énormément de flexibilité dans la collecte de données, l'analyse et l'interprétation des résultats. Les chercheurs ont un degré de liberté extrêmement important (Simmons et al., 2011) qui peut amener à des biais de publication et du p-hacking. Le seul moyen de prouver la rigueur d'une procédure est d'obtenir une preuve que les hypothèses et les analyses statistiques ont été prévues avant la récolte des données grâce à un horodatage. C'est l'objectif du pré-enregistrement (Nosek et al., 2018).

Il est important de vérifier le pré-enregistrement, car de nombreux articles sont pré-enregistrés, mais ne renseignent pas correctement d'importantes déviations. Une analyse (Claesen et al. 2021) de 27 articles indique que 89% d'entre eux n'ont pas correctement renseigné leurs déviations. Les éditeurs et peer-reviewers ne regardent pas les pré-enregistrements (3% le font, Syed, 2023), il faut donc le faire soi-même. Il existe également des Registered Report qui sont des articles entièrement pré-enregistrés. Généralement, le terme Registered Report est indiqué dans le titre et/ou dans le résumé de l'article.

Degré d'importance

Modéré

Un article sans pré-enregistrement est douteux, car il est difficile de croire que les analyses et les hypothèses n'aient pas été créées après avoir sélectionné des résultats significatifs. De manière générale, il est possible d'avoir une confiance dans les Registered Reports, et une confiance dans les études pré-enregistrées, si un coup d'œil au pré-enregistrement permet de nous assurer que l'enregistrement est correct : il y a les hypothèses et les analyses correctement renseignées.

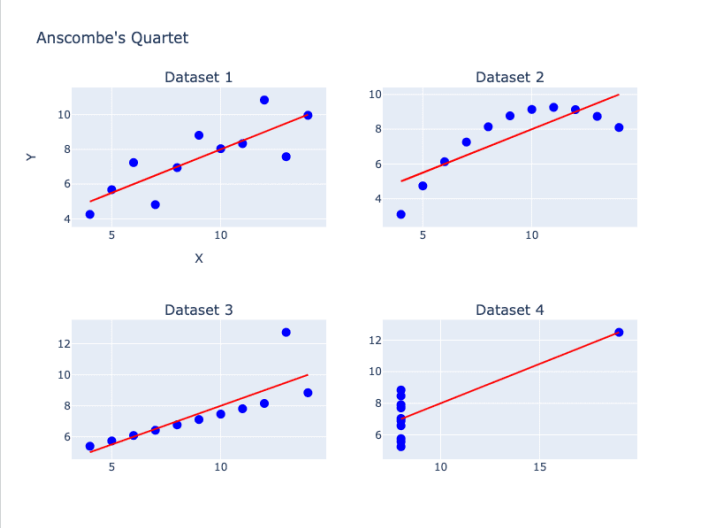
Références

- Claesen A, Gomes S, Tuerlinckx F, Vanpaemel W. Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. R Soc Open Sci. 2021 Oct 27;8(10):211037. doi: 10.1098/rsos.211037. PMID: 34729209; PMCID: PMC8548785.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. In Proceedings of the National Academy of Sciences (Vol. 115, Issue 11, pp. 2600–2606). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1708274114
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. In Psychological Science (Vol. 22, Issue 11, pp. 1359–1366). SAGE Publications. https://doi.org/10.1177/0956797611417632
- Syed, M. (2023, December 8). Some Data Indicating that Editors and Reviewers Do Not Check Preregistrations during the Review Process. https://doi.org/10.31234/osf.io/nh7qw

4 Il n'y a pas de compréhension de la distribution des données

À quoi faire attention ?

Trouver un tableau récapitulatif des données descriptives dans l'article, et/ou des tableaux récapitulant la distribution des données (histogrammes, graphiques à point) dans l'article ou les données supplémentaires.



Pourquoi faire attention ?

En 1973, Anscombe produit un quartet, c'est-à-dire une série de quatre graphiques, contenant 11 points. Dans les quatre graphiques, les moyennes sont identiques (9), la variance est identique (10), la corrélation est identique (0.81) et la régression linéaire est identique (la ligne rouge). Pourtant, la distribution des points est complètement différente, ce qui devrait mener à une interprétation différente des résultats.

Dans le cas où un article ne rapporte que les moyennes/écart-type, le résultat des analyses et ne rapportent pas de graphiques montrant la distribution des données, il est impossible de savoir d'où proviennent les résultats. Dans le quatrième graphique, il y a clairement une donnée aberrante (dans le troisième, il est possible qu'il y ait une donnée aberrante). Dans le deuxième graphique, les points sont distribués en cloche, il faut donc comprendre pourquoi elles ne sont pas distribuées uniformément autour de la droite de régression. Bref, sans visualisation des données, il est impossible de faire confiance seulement aux résultats statistiques. La visualisation des données descriptives est un processus obligatoire pour améliorer la transparence et la rigueur dans les sciences sociales (Tay et al. 2016).

Degré d'importance

Situationnelle.
Dans le cas où les données sont en libre accès, il est possible de créer ces visualisations par nous-même. Cependant, il est préférable de les avoir accessibles dans l'article ou en "supplementary material" afin de facilement pouvoir visualiser la distribution des données.

Références

- F. J. Anscombe, Graphs in Statistical Analysis, Am. Stat., vol. 27, n°1, 1973, pp. 17-21. DOI 10.1080/00031305.1973.10478966 JSTOR:2682899
- Tay, L., Parrigon, S., Huang, Q., & LeBreton, J. M. (2016). Graphical Descriptives. In Perspectives on Psychological Science (Vol. 11, Issue 5, pp. 692–701). SAGE Publications. https://doi.org/10.1177/1745691616663875

5 L'article est publié dans une revue prédatrice/peu filtrante

À quoi faire attention ?

il existe des signes que l'article est publié dans une revue qui ne possède aucun contrôle ou filtre de publication. Le TOP Factor publie une liste des revues selon leur degré de qualité, transparence et ouverture : https://www.topfactor.org/. Être dans la liste, voire y être bien classé est un signe de qualité. ScimagoJR.com publie un classement de journaux selon des quartiles. Être dans le quartile 1 ou 2 (en vert) n'est pas signe de qualité, mais être dans un quartile 3 ou 4 (orange ou rouge) est un signe qu'il faut se méfier de la revue. Enfin, il existe des éditeurs prédateurs qui possèdent de nombreux journaux problématiques, et de nombreux articles écrits dans des fermes à articles. C'est le cas de MDPI, Hindawi, ou Frontiers. Les articles publiés chez ces éditeurs doivent être lus avec un maximum de précautions.

Pourquoi faire attention ?

Certains journaux ne mettent pas en place correctement de processus de revue par les pairs. Les données issues de MDPI, Hindawi et Frontiers indiquent que ces éditeurs ne sont pas rigoureux (Hanson et al., 2023) : un taux exponentiel d'études publiées, un temps entre la soumission et la publication extrêmement court, de nombreuses éditions spéciales avec un processus de peer-reviewing allégé. Les rétractions d'articles sont exponentielles, dépassant en 2023 les 10 000 articles rétractés, majoritairement chez Hindawi (Van Noorden, 2023). C'est très loin d'être suffisant (Oransky, 2022), d'autant plus que le nombre d'articles publiés augmente exponentiellement. En plus de ces éditeurs que certains qualifient de prédateurs, les journaux des éditeurs moins prédateurs ne sont pas nécessairement rigoureux. Les faux articles issus de fermes à articles sont partout (Else, 2021), et il est extrêmement difficile de faire rétracter des articles (Allison et al., 2016). Les éditeurs cachent parfois la rétractation des articles en mettant un bandeau "rétracté" en bas des articles, peu lisible (Roche, 2023). Les peer-reviewers ne font pas un travail de filtre efficace (Marcoci et al., 2022). La transparence et l'automatisation de la détection d'erreur doit être possible pour s'assurer de la qualité des études publiées dans le journal.

Degré d'importance

Important.

La revue de Hanson et al. (2023) est très préoccupante. Elle indique que de nombreux éditeurs possédant des centaines de journaux sont prédateurs et n'ont pas un processus de filtre suffisant pour diminuer les probabilités de publier des articles de mauvaise qualité ou frauduleux. Cela représente plusieurs centaines de milliers d'articles faux et inventés publiés par an. S'assurer que les études que l'on lit ne proviennent pas de ces journaux semble obligatoire pour se prémunir de ces études inventées.

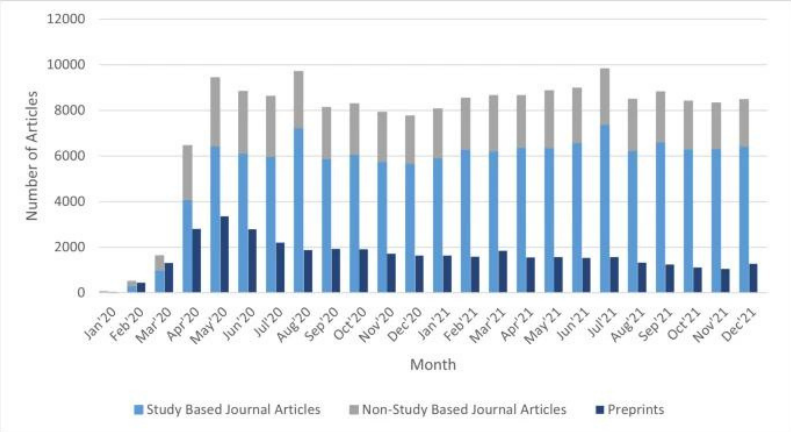
Références

- Allison, D., Brown, A., George, B. et al. Reproducibility: A tragedy of errors. Nature 530, 27–29 (2016).
- Else, H., & Van Noorden, R. (2021). The fight against fake-paper factories that churn out sham science. In Nature (Vol. 591, Issue 7851, pp. 516–519). Springer Science and Business Media LLC.
- Hanson, M. A., Barreiro, P. G., Crosetto, P., & Brockington, D. (2023). The strain on scientific publishing (Version 1). arXiv.
- Marcoci, A., Vercammen, A., Bush, M. et al. Reimagining peer review as an expert elicitation process. BMC Res Notes 15, 127 (2022).
- Oransky, I. (2022). Retractions are increasing, but not enough. In Nature (Vol. 608, Issue 7921, pp. 9–9). Springer Science and Business Media LLC.
- Roche, C. D. (2023). Visibility of retracted articles: journals struggle to make retractions in the first place. In BMJ (p. p1669). BMJ.
- Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023 — a new record. In Nature (Vol. 624, Issue 7992, pp. 479–481). Springer Science and Business Media LLC.

6 L'article n'a pas été publié - le cas des préprints*

À quoi faire attention ?

L'article est disponible sur une archive - PsyArxiv, BioArxiv, SSRN, mais n'a pas été publié dans un journal scientifique. Il n'a pas encore été mis à l'épreuve du peer-reviewing et n'a donc pas eu au moins un premier filtre pour s'assurer de sa qualité.



Drzymalla et al. (2022). Le cas des préprints (en bleu foncé) sur le COVID-19. On observe une augmentation des préprints autour de mai 2020. Sur 20698 préprints, seulement 36% étaient publiés un an après.

Pourquoi faire attention ?

Les préprints augmentent les chances qu'un article soit cité dans d'autres articles, sur les réseaux sociaux et dans les médias (Bagchi et al., 2024; Conroy, 2019; Xu et al., 2021). Une revue des préprints publiés pendant le COVID-19 indique qu'une partie importante des préprints n'ont pas été publiés en tant qu'article par la suite (Añazco et al. 2021). De plus, il y a des changements qui amènent des différences importantes entre les préprints et les études publiées (Miller, 2022). Ainsi, les préprints ne semblent pas être des sources fiables d'information en tant que tels.

Degré d'importance

Modéré.
Les préprints possèdent les mêmes risques de biais des études publiées, et les autres risques s'y appliquent. Ils ajoutent également un manque de filtre par la revue par les pairs, dans les revues de qualité suffisante.

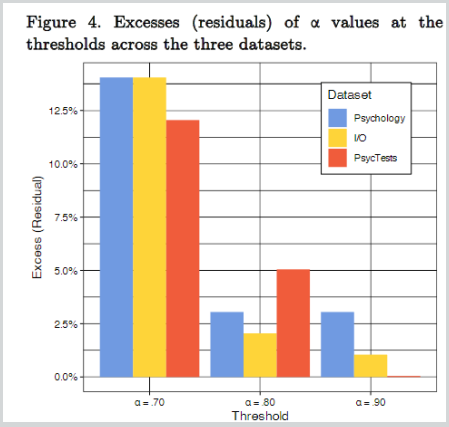
Références

- Añazco, D., Nicolalde, B., Espinosa, I., Camacho, J., Mushtaq, M., Gimenez, J., & Teran, E. (2021). Publication rate and citation counts for preprints released during the COVID-19 pandemic: the good, the bad and the ugly. In PeerJ (Vol. 9, p. e10927). PeerJ. https://doi.org/10.7717/peerj.10927
- Bagchi, C., Malmi, E., & Grabowicz, P. (2024). Promotion of Scientific Publications on ArXiv and X Is on the Rise and Impacts Citations (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2401.11116
- Conroy, G. (2019). Preprints boost article citations and mentions. Nature, https://www.nature.com/nature-index/news/preprints-boost-article-citations-and-mentions
- Drzymalla, E., Yu, W., Khoury, M.J. et al. COVID-19-Related manuscripts: lag from preprint to publication. BMC Res Notes 15, 340 (2022). https://doi.org/10.1186/s13104-022-06231-9
- Miller, N. (2022). How different are preprints from their published versions? 2 studies have some answers. https://journalistsresource.org/media/two-studies-examine-preprints/
- Xu, F., Ou, G., Ma, T., & Wang, X. (2021). The consistency of impact of preprints and their journal publications. In Journal of Informetrics (Vol. 15, Issue 2, p. 101153). Elsevier BV. https://doi.org/10.1016/j.joi.2021.101153

7 Il est impossible d'accéder à la procédure exacte

À quoi faire attention ?

Dans la partie méthode, il est possible de vérifier si les auteurs ont correctement indiqué toute la procédure à laquelle les participants ont été exposés. Dans les articles de meilleure qualité, la procédure est totalement disponible dans un serveur en ligne ou dans un matériel supplémentaire, parfois à travers un flux de travail (par exemple, Qualtrics propose de télécharger la procédure), parfois avec des captures d'écran des fenêtres avec les questions posées aux participants. Dans un certain nombre de cas, les articles rapportent les questions posés, mais parfois, ils ne les rapportent pas.



Hussley et al. (2023) - % d'excès d'alpha juste au dessus des seuils 0.70, 0.80 et 0.90

On observe également un manque de standardisation des rapports des qualités psychométriques des mesures, et il est donc difficile de savoir si ces mesures sont fiables (Parsons et al., 2019).

Pourquoi faire attention ?

Un article expérimental ne peut être considéré comme scientifique que s'il est reproductible : nous avons les informations nécessaires pour le reproduire et vérifier que ces résultats sont concluants avec des recherches indépendantes. Sans accès à la procédure exacte, il n'est pas possible de faire des réplications identiques, et il est souvent nécessaire de faire des réplications éloignées. En cas d'échec, on ne peut pas attribuer l'échec à un manque d'effet ou à une mauvaise opérationnalisation de la procédure, ce qui augmente les problèmes de reproductibilité. De nombreux articles développent leurs propres échelles, qui ne sont jamais validées rigoureusement (Elson et al.,2023). Celles-ci sont couplées avec des possibilités d'alpha-hacking, où les distributions d'indicateurs de fiabilités comme l'alpha de Cronbach se trouvent juste au-dessus du seuil acceptable (.70, Hussey et al., 2023).
Pour ces raisons, il est indispensable de connaître les questions exactes posées aux participants, ainsi que l'ordre dans laquelle elles ont été posées.

Degré d'importance

Modéré.
Beaucoup de chercheurs ne rendent pas leurs procédures disponibles alors qu'elles sont correctes, principalement à cause d'un manque de formation et de sensibilisation à ces problèmes. D'un autre côté, il existe des études avec des procédures disponibles dont il manque des preuve d'efficacité dans la mesure de l'effet étudié.

Références

- Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2023, February 3). An aberrant abundance of Cronbach's alpha values at .70. https://doi.org/10.31234/osf.io/dm8xn
- Elson, M., Hussey, I., Alsalti, T. et al. Psychological measures aren't toothbrushes. Commun Psychol 1, 25 (2023). https://doi.org/10.1038/s44271-023-00026-9
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. In Advances in Methods and Practices in Psychological Science (Vol. 2, Issue 4, pp. 378–395). SAGE Publications. https://doi.org/10.1177/2515245919879695

8

Il n'y a pas de filtre ou de vérification des données

À quoi faire attention ?

Dans la partie résultat, au début, il est possible de voir combien de participants ont été filtrés sur des bases de qualité. Les participants peuvent être filtrés en fonction de questions de compréhension, mais aussi des questions sur leur fluence dans la langue dans laquelle a été conduite l'étude, leur sérieux dans la passation du questionnaire etc.
Cette procédure est essentielle quand les questionnaires sont passés dans les laboratoires auprès d'étudiants ayant une motivation faible à répondre au questionnaire et qui peuvent répondre aléatoirement, mais aussi dans les questionnaires en ligne pour lesquels il existe des risques de réponses par des robots ou des individus inattentifs.

Outcome Measure	MTurk (N = 500)	CloudResearch (N = 505)	Prolific (N = 496)
Select Strongly Agree*	93.80%$^{a}_{vw}$	97.03%$^{b}_{vx}$	98.39%$^{b}_{x}$
Pass Arithmetic Check*	98.40%$^{a}_{v}$	99.41%$^{a}_{v}$	99.19%$^{a}_{v}$
Pass Color Recall Check*	75.76%$^{a}_{v}$	95.84%$^{b}_{w}$	98.58%$^{c}_{x}$
Leave Textbox Blank*	98.80%$^{a}_{v}$	100.00%$^{b}_{w}$	100.00%$^{b}_{w}$
Pass Colorblindness Check*	88.60%$^{a}_{v}$	98.02%$^{b}_{w}$	98.59%$^{b}_{w}$
Unique Worker ID*	99.60%$^{a}_{v}$	100.00%$^{a}_{v}$	100.00%$^{a}_{v}$
Unique IP Address*	97.20%$^{a}_{v}$	99.21%$^{b}_{w}$	98.79%$^{a}_{vw}$
Unique Geolocation*	53.40%$^{a}_{v}$	90.10%$^{b}_{w}$	89.92%$^{b}_{w}$
Time > 3 Minutes*	82.60%$^{a}_{v}$	88.91%$^{b}_{w}$	90.12%$^{b}_{w}$
Meaningful or Blank Open Response*	82.20%$^{a}_{v}$	99.01%$^{b}_{w}$	99.19%$^{b}_{w}$
Self-Reported High Data Quality*	54.40%$^{a}_{v}$	79.80%$^{b}_{w}$	85.89%$^{c}_{x}$
High-Quality Respondents*	26.40%$^{a}_{v}$	61.98%$^{b}_{wx}$	67.94%$^{c}_{x}$
Total Cost	$575	$625	$640
Cost We Paid per High-Quality Respondent	$4.36	$2.00	$1.90

Douglas et al. (2023). Selon les critères, entre 26% et 70% des répondants produisent des réponses de qualité en moyenne, nécessitant obligatoirement plusieurs filtres

Pourquoi faire attention ?

De nombreuses études ont été faites sur la qualité des données récoltées par des questionnaires en ligne. Brühlmann et al (2020) indiquent que 46% de leurs participants semblent avoir au moins une fois fait preuve d'un manque de rigueur dans leurs réponses, ce taux n'était que de 7% dans l'étude de Schneider et al. (2017). Il existe une énorme variabilité de ce taux en fonction du fournisseur de données, mais aussi des paiements aux participants et dans les filtres (les participants européens étant généralement plus sérieux que les Américains ou que les Indiens, CloudResearch et Prolific étant généralement plus qualitatifs que Mturk).

Degré d'importance

Modéré
L'important n'est pas de vérifier le taux de filtre des participants, mais de vérifier s'il y a eu des filtres pour éliminer les participants non sérieux dans leurs réponses. À nouveau, cela est plus facile quand les données sont ouvertes et la procédure entièrement disponible afin de s'assurer des questions exactes posées et des réponses des participants.

Références

- Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. In Methods in Psychology (Vol. 2, p. 100022). Elsevier BV. https://doi.org/10.1016/j.metip.2020.100022
- Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. In J. S. Hallam (Ed.), PLOS ONE (Vol. 18, Issue 3, p. e0279720). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0279720
- Schneider, S., May, M., & Stone, A. A. (2017). Careless responding in internet-based quality of life assessments. In Quality of Life Research (Vol. 27, Issue 4, pp. 1077–1088). Springer Science and Business Media LLC. https://doi.org/10.1007/s11136-017-1767-2

B

Bonus : Tableau des risques dans les études cliniques

Critère	Description
Validité des inférences	Y a-t-il suffisamment de transparence en ce qui concerne la collecte de données et les analyses statistiques, et les conclusions sont-elles étayées par des données probantes ? Y a-t-il des preuves que le traitement et non d'autres facteurs (ex. Le manque de double aveugle) explique la différence entre le groupe d'intervention et le groupe témoin ? Y-a-t-il eu un peer-reviewing indépendant ? Les rapports des reviewers sont-ils accessibles au public ?
Conflits d'intérêts	Les conflits d'intérêts potentiels sont-ils signalés de manière transparente dans le document ? Quelle est la nature de ces conflits d'intérêts et, en présence de conflits d'intérêts graves, y a-t-il suffisamment de mesures de protection (ex. préenregistrement) ? Toutes les mesures incluses sont-elles accessibles et déclarées ?
Sécurité et effets indésirables	Est-il facile de trouver toutes les informations pertinentes concernant les événements indésirables dans l'étude ? Existe-t-il un critère clair pour décider si un événement indésirable est lié au traitement ? La composante psychothérapeutique de l'étude est-elle standardisée et entièrement décrite ? A-t-on fait appel à un thérapeute qualifié pour effectuer les traitements ?
Groupe témoin	Un groupe témoin est-il inclus pour tester les effets placebo, les effets d'attente et la régression à la moyenne ? Si aucun groupe témoin n'est inclus, les interprétations sont-elles suffisamment prudentes ?
Taille de l'échantillon	Une analyse de puissance ou de sensibilité est-elle fournie, et comprend-elle une justification de l'ampleur minimale de l'effet d'intérêt ? L'étude est-elle suffisamment puissante pour détecter une différence entre le groupe d'intervention et le groupe témoin (et non par rapport à pas d'effet).
Biais de sélection	L'échantillon étudié diffère-t-il de la population d'intérêt ? Y a-t-il une déclaration sur les contraintes de généralisabilité ? Des renseignements démographiques (p. ex., sexe, âge, milieu socioéconomique) et cliniques (p. ex., gravité, comorbidités) sont-ils fournis ?
Durée de l'étude	Les scientifiques suivent-ils les patients pendant une période de temps suffisante pour justifier la conclusion que le traitement a été réussi, c'est-à-dire que les gens sont revenus à un niveau normal de charge symptomatique, de bien-être et de fonctionnement ?
Problème de rupture de l'aveugle	Des efforts ont-ils été faits pour réduire au minimum le risque de non-aveugle (ex. en utilisant des placebos actifs) ? L'efficacité de l'aveuglement a-t-elle été évaluée et rapportée ?
Effet Placebo	La conception de l'étude tient-elle compte des effets placebo, par exemple en comparant le groupe d'intervention à un groupe témoin ? L'étude comprend-elle des mesures permettant d'évaluer les attentes du patient et du thérapeute à l'égard des résultats du traitement, tant au début que pendant le traitement ?
Mécanismes d'action	Les inférences concernant les mécanismes d'action potentiels sont-elles étayées par des preuves? Les données et les documents sont-ils en libre accès pour permettre une réplication et des analyses secondaires ?

Référence

- Le tableau provient de van Elk, M., & Fried, E. I. (2023). History repeating: guidelines to address common problems in psychedelic science. In Therapeutic Advances in Psychopharmacology (Vol. 13). SAGE Publications. https://doi.org/10.1177/20451253231198466

Glossaire

Biais de publication

L'omission de publier les résultats en fonction de «l'orientation ou de la force des résultats de l'étude» (Dickersin et Min, 1993, p. 135). Le biais de publication est la tendance à ce que des résultats nouveaux et significatifs sont plus publiés que des réplications et des résultats nuls. Ce biais se matérialise généralement par un nombre disproportionné de résultats significatifs et des tailles d'effet exagérées. Ce processus conduit à ce que la littérature scientifique publiée ne soit pas représentative de l'étendue complète de toutes les recherches, et sous-représente spécifiquement les résultats nuls.

Puissance Statistique

La puissance statistique est la probabilité à long terme qu'un test statistique rejette correctement l'hypothèse nulle si l'hypothèse alternative est vraie. Puissance a priori : le chercheur pose la question «compte tenu d'une taille d'effet, de combien de participants aurais-je besoin pour une puissance de X %?». La puissance de sensibilité : «étant donné une taille d'échantillon connue, quelle taille d'effet pourrais-je détecter avec une puissance de X %?».

P-valeur

Valeur statistique utilisée pour évaluer le résultat d'un test d'hypothèse dans le cadre d'un test de signification d'hypothèse nulle (NHST). Il s'agit de la probabilité d'observer un effet, ou un effet plus extrême, en supposant que l'hypothèse nulle soit vraie (Lakens, 2021).
Selon l'APA : «une valeur de p est la probabilité selon un modèle statistique spécifié qu'un résumé statistique des données (par exemple, la différence moyenne de l'échantillon entre deux groupes comparés) soit égal ou plus extrême que sa valeur observée» (Wasserstein et Lazar, 2016, p.131).

P-Hacking

Exploiter des techniques qui peuvent augmenter artificiellement la probabilité d'obtenir un résultat statistiquement significatif en répondant au critère standard de signification statistique (généralement $\alpha = 0,05$). Par exemple, effectuer plusieurs analyses et ne déclarer que celles dont $p<.05$, supprimant des données jusqu'à ce que $p < .05$, selectionner des variables à analyser en se basant sur les résultats significatifs.

Préprint

Une version accessible au public de tout type de manuscrit scientifique ou de résultat de recherche précédant une publication officielle. Les préprints sont généralement hébergées sur un dépôt (par exemple arXiv) qui facilite la diffusion en partageant les résultats de la recherche plus rapidement que par le biais d'une publication traditionnelle. Les préprints peuvent être publiées à n'importe quel moment du cycle de recherche, mais ils sont le plus souvent publiées au moment de la soumission (c'est-à-dire avant l'évaluation par les pairs).

Pré-enregistrement

Pratique consistant à publier le plan d'une étude, y compris les questions et hypothèses de recherche, la conception de la recherche et l'analyse des données avant que les données n'aient été recueillies ou examinées. Un document de préenregistrement est horodaté et généralement enregistré auprès d'une partie indépendante (par exemple, un dépôt) afin qu'il puisse être partagé publiquement avec d'autres (éventuellement après une période d'embargo). Le pré-enregistrement fournit une documentation transparente de ce qui était prévu à un moment donné et permet à des tiers d'évaluer les changements qui ont pu se produire par la suite.

Références

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., ... & Aczel, B. (2022). A Community-Sourced Glossary of Open Scholarship Terms. Nature human behaviour, 6(3), 312-318.

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., ... & Aczel, B. (2022). A Community-Sourced Glossary of Open Scholarship Terms. Nature human behaviour, 6(3), 312-318.