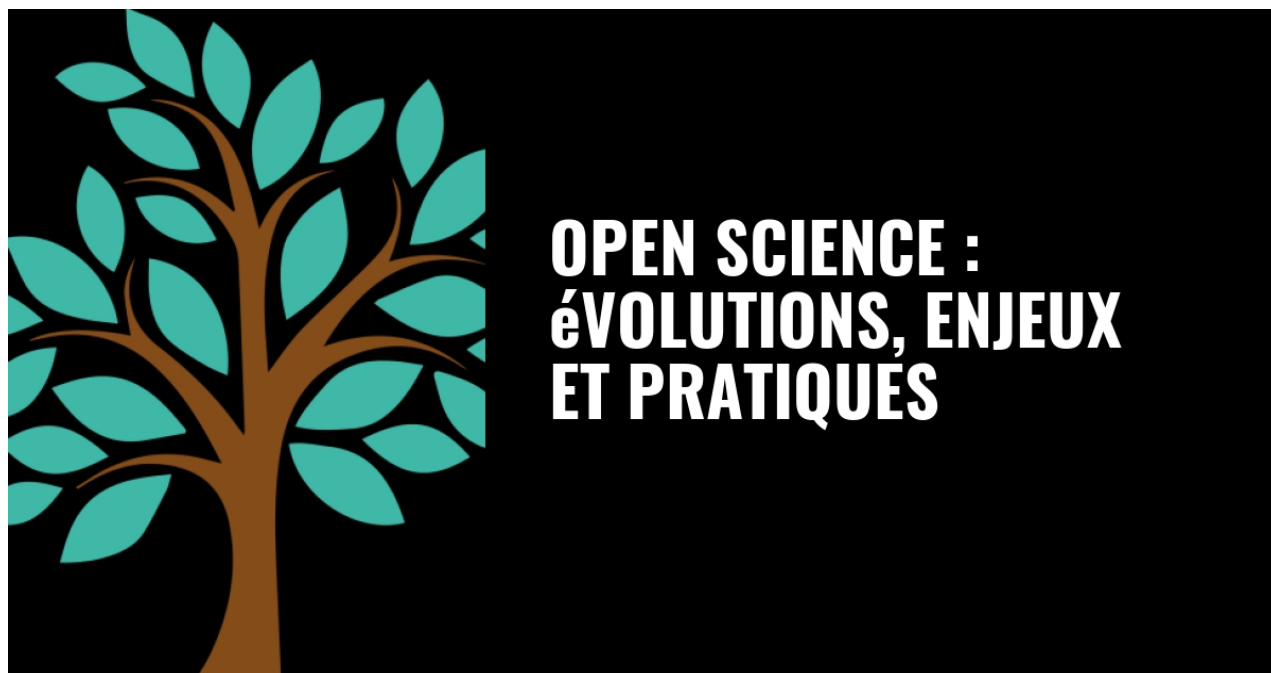


Les éditeurs sévissent face aux réutilisations inappropriées de bases de données publiques en santé

 openscience.pasteur.fr/2025/12/02/les-editeurs-sevissent-face-aux-reutilisations-inappropriees-de-bases-de-donnees-publiques-en-sante/

CeRIS - Institut Pasteur

2 décembre 2025



En septembre dernier, les éditeurs [PLOS](#) et [Frontiers](#) ont annoncé qu'ils modifiaient leur politique d'évaluation des **recherches s'appuyant sur des bases de données publiques** en sciences sociales et en santé, ce qui a conduit au **rejet automatique** de la grande majorité des articles utilisant ces bases de données.

À l'origine de cette situation : l'**explosion du nombre d'études de mauvaise qualité basées sur ces données publiques**, dans le seul objectif d'augmenter le nombre de publications, probablement sous l'impulsion de [paper mills](#) et facilitées par l'utilisation de l'IA générative.

Ces bases de données permettent en effet d'innombrables combinaisons de variables, de sorte que des corrélations peuvent apparaître « statistiquement significatives » par hasard, sans qu'**aucune question de recherche légitime** ne soit formulée : on analyse de manière sélective certaines parties de la base de données (en choisissant un état de santé, un facteur environnemental ou physiologique pouvant y être associé, ainsi qu'un groupe de population), puis on recherche un résultat statistiquement significatif, sans hypothèse directrice. On peut par exemple choisir d'analyser le lien entre les niveaux de vitamine D et la dépression chez les hommes de plus de 65 ans, ou entre une mauvaise santé dentaire et le diabète chez les femmes âgées de 18 à 45 ans.

Des chercheurs ont notamment [étudié](#) les articles s'appuyant sur les données de la base américaine NHANES (*National Health and Nutrition Examination Survey*). Au cours des trois dernières années, ils ont constaté une augmentation rapide du nombre d'études

« stéréotypées » analysant le lien entre un facteur unique et un état de santé. En moyenne, 4 articles par an étaient publiés entre 2014 et 2021. Ce nombre est passé à 33, 82 et 190 respectivement en 2022, 2023 et 2024. Ils considèrent que cette augmentation accroît le **risque d'introduire des conclusions trompeuses dans la littérature scientifique**, notamment parce que des questions de santé multifactorielles sont analysées comme des études à facteur unique, en sélectionnant les données, et sans analyse statistique pour exclure les faux-positifs.

C'est donc ce constat qui a conduit les éditeurs PLOS et Frontiers à faire évoluer leur politique, bien qu'ils soient conscients de l'importance des études menées sur ces bases de données, lorsqu'elles sont menées de manière rigoureuse. Une bonne pratique recommandée par PLOS pour démontrer la rigueur de son étude observationnelle est de joindre une checklist [STROBE](#) à son manuscrit.

Sources :

- [Journals and publishers crack down on research from open health data sets](#), par Cathleen O'Grady, *Science Insider*, 8 octobre 2025.
- [Low-quality papers are surging by exploiting public data sets and AI](#), par Cathleen O'Grady, *Science Insider*, 14 mai 2025.

Sur un sujet connexe : le serveur de préprint arXiv a [fait évoluer sa politique de modération](#) et n'acceptera plus les articles de synthèse (*reviews*) ou de position dans le domaine de l'informatique, sauf s'ils ont été *reviewés* et acceptés pour publication dans une revue ou une conférence. Cette mesure fait suite à une **augmentation du nombre d'articles de mauvaise qualité**, dont beaucoup semblent avoir été rédigés à l'aide de l'IA générative. Ces articles prennent souvent la forme de « bibliographies annotées » comprenant des citations d'articles n'ayant parfois aucun rapport avec le sujet, ce qui laisse soupçonner des liens avec des [citation mills](#).

Source : [Preprint site arXiv is banning computer-science reviews: here's why](#), par Davide Castelvecchi, *Nature news*, 7 novembre 2025.