

Analyse de l'« [Utah Review](#) » sur les traitements hormonaux chez les mineurs avec dysphorie de genre

Magali Pignard, [Transition-mineurs](#), juillet 2026

Concernant l'évaluation de la qualité des études individuelles

Une des forces de l'Utah Review est que les auteurs ont **évalué la qualité des études** selon des outils standards comme la Newcastle-Ottawa Scale ([NOS](#)) pour les études observationnelles et l'outil [NIH](#) pour les études avant-après (explication des items : tableaux I.3 et I.4 ; résultats globaux des évaluations : tableaux I.12 à I.17 ; évaluations détaillées : Appendix I.I).

Cependant, selon [SEGM](#), les évaluations de qualité apparaissent trop indulgentes.

En exemple, voici une comparaison que j'ai faite, concernant l'étude d'Achille et al. [2020](#) : évaluée par l'Utah Review et par Taylor et al. [2024](#) (hormones) :

Achille et al., [2020](#)

Selon l'Utah Review (NOS, p. 577/1 051) : 7/9 → « élevée »	Selon Taylor et al., 2024 (NOS modifiée) 4/8 → « faible »
--	---

Cette évaluation illustre un exemple flagrant où une **notation (1/1)** est un peu **déconnectée de l'analyse que l'évaluateur formule lui-même dans son propre commentaire** :

- Sur la **durée de suivi** :
 - Taylor et al. **refusent le point** (0/1) car le suivi n'est pas ancré à l'initiation du traitement et la variable d'exposition utilisée est la simple réception d'un traitement à un moment quelconque, et non sa durée.
 - L'Utah Review **accorde le point**, tout en reconnaissant dans son commentaire qu'« *on ne sait pas quand les traitements ont été initiés. tous les participants n'avaient pas débuté leur traitement à l'évaluation initiale* ».
- Le domaine de l'**attrition est le plus révélateur** :
 - Taylor et al. **refusent le point** : 116 participants sont entrés dans l'étude, seuls 50 sont rapportés, sans aucune information sur les 66 perdus de vue.
 - L'Utah Review **accorde le point** en considérant que « *tous les sujets pris en compte* », tout en reconnaissant dans son propre commentaire que « *le fait de restreindre l'analyse aux participants ayant complété les trois questionnaires signifie qu'il n'y avait plus d'attrition dans l'échantillon analysé, mais cela introduit un biais de sélection* ». C'est donc paradoxal car l'Utah Review **identifie le problème mais accorde quand même le point**.

À noter : Achille et al. a été évaluée également avec l'échelle du NIH, en tant qu'étude pré/post, alors **qu'elle n'est pas une étude Pré/Post**.

Dans son analyse, SEGM donne deux exemples d'études largement citées dans la littérature :

Chen et al., 2023

Selon l'Utah Review NOS, p. 581/1 051 : 8/9 → « élevée »	Selon Cheung et al., 2024 (NOS modifiée) : 5/7 → « modérée »
---	---

La différence la plus frappante concerne le critère de « cohorte non exposée », autrement dit, un groupe de comparaison. Ce groupe est tout simplement **absent de l'étude**, comme Chen et al. l'admettent eux-mêmes. Malgré cela, l'Utah Review **leur accorde le point**, en estimant que les participants venaient « *de la même communauté que la cohorte exposée* ». Cheung et al., eux, **classent logiquement ce critère comme non applicable**.

Les deux évaluations divergent aussi concernant l'attrition (le fait que certains participants n'ont pas été suivis jusqu'au bout) et le contrôle des facteurs de confusion. Cheung et al. sanctionnent l'absence de contrôle de certaines co-interventions importantes. L'Utah Review se montre plus indulgente sur ces mêmes points.

Les deux scores sont néanmoins plutôt élevés, ce qui montre les limites de la Newcastle – Ottawa Scale (NOS). Cet outil est utile pour repérer les problèmes classiques des études observationnelles (attrition, comparabilité des groupes, durée du suivi) mais **il passe à côté de problèmes plus subtils de cette étude**, comme la sélection des résultats rapportés, les changements d'hypothèses en cours de route, comme le révèle l'[analyse détaillée](#) de Jesse Singal (en français [ici](#)) : sur les **huit variables** principales que les chercheurs avaient annoncé vouloir suivre (notamment la dysphorie de genre, la suicidalité, l'automutilation ou la qualité de vie) **six ne sont finalement pas rapportées dans l'article principal**. Singal relève aussi que les auteurs ont progressivement déplacé le « centre de gravité » de l'étude vers une variable appelée *appearance congruence*, très peu mise en avant au départ dans le protocole de l'étude, qui devient ensuite le principal résultat positif mis en avant.

Au final, un score NOS élevé **ne garantit donc pas qu'une étude soit exempte de problèmes sérieux**.

Tordoff et al., 2022

Selon l'Utah Review : 6/9 (NOS, p. 595/1 051) → « modérée »	Selon Taylor et al., 2024 (NOS modifiée) 3,5 /8 → « faible »
--	---

Ici, les divergences portent sur quatre points : la représentativité de l'échantillon, la mesure de l'exposition, la durée et la structure du suivi, ainsi que l'attrition.

Sur la représentativité : l'Utah Review juge acceptable un échantillon recruté dans une seule clinique multidisciplinaire, malgré l'absence d'environ 30 % des patients éligibles. Taylor et al. sont beaucoup plus sévères, et estiment que cette sélection compromet la représentativité de l'étude.

Sur la mesure de l'exposition : l'Utah Review considère que l'auto-déclaration des traitements est suffisante. Pour Taylor et al., cela ne permet pas de mesurer de façon fiable quand et comment le traitement (bloqueurs ou hormones) a été initié.

Mais le problème le plus important est la structure même du « groupe comparateur ». L'étude est souvent présentée comme comparant des jeunes traités à un **groupe stable de jeunes non traités, ce qui est inexact**. En réalité, il s'agit **d'une seule cohorte de jeunes, dont le statut de traitement varie durant l'étude** : comme l'analyse [Singal](#) et Abbruzzese et al. [2023](#), la majorité des participants initialement classés « non traités » ont soit commencé un traitement en cours de suivi, soit quitté l'étude. À 12 mois, il ne restait plus que **7 participants non traités avec des données de santé mentale exploitables**.

À cela s'ajoute une **perte de données fortement reliée au statut de traitement** : à 12 mois, les données de santé mentale manquaient pour 80 % des jeunes non traités, contre seulement 17 % des jeunes traités. Ce déséquilibre rend toute comparaison entre les deux catégories très difficiles à interpréter, les caractéristiques des jeunes traités et non traités étant probablement trop différentes ; sans compter la possibilité que les jeunes initiant le traitement en cours d'étude soient ceux dont la santé mentale est la plus stable, la santé mentale étant précisément ce que l'étude cherche à évaluer (possible « confusion par indication »).

À noter : Tordoff et al. a été évaluée également avec l'échelle du NIH, en tant qu'étude pré/post, alors **qu'elle n'est pas une étude Pré/Post**.