Traduction des articles de Jesse Singal concernant l'étude [Psychosocial Functioning in Transgender Youth after 2 Years of Hormones](), Chen et al., *NEJM*, janvier 2023 :

- [On Scientific Transparency, Researcher Degrees Of Freedom, And That NEJM Study On Youth Gender Medicine](),
- [The New, Highly Touted Study On Hormones For Transgender Teens Doesn't Really Tell Us Much Of Anything]()

Singal-Minded, Substack, 31 janvier et 7 février 2023



The NEW ENGLAND JOURNAL of MEDICINE

Partie 1 - Sur la transparence scientifique, les degrés de liberté des chercheurs

De nombreuses questions restent sans réponse, ce qui est regrettable.

Il existe un concept appelé "degrés de liberté du chercheur" qui est très important pour comprendre les formes de science bâclée qui sont moins salaces mais (beaucoup) plus courantes que, par exemple, la fraude pure et simple sur les données.

Comme l'[ont écrit]() Joseph P. Simmons, Leif D. Nelson et Uri Simonsohn [en 2011]() :

> Au cours de la collecte et de l'analyse des données, les chercheurs doivent prendre de nombreuses décisions : Faut-il collecter davantage de données ? Faut-il exclure certaines observations ? Quelles conditions doivent être combinées et lesquelles doivent être comparées ? Quelles variables de contrôle faut-il prendre en compte ? Faut-il combiner ou transformer des mesures spécifiques, ou les deux ?
>
> Il est rare, et parfois peu pratique, que les chercheurs prennent toutes ces décisions à l'avance. Au contraire, il est courant (et accepté) que les chercheurs explorent diverses alternatives analytiques, recherchent une combinaison qui produise une "signification statistique" et ne rapportent que ce qui a "fonctionné". Le problème, bien sûr, est que la probabilité qu'au moins une analyse (parmi de nombreuses autres) produise un résultat faussement positif au niveau de 5 % [niveau de confiance qui est un repère commun pour la signification statistique dans les sciences sociales] est nécessairement supérieure à 5 %.

Comme je le dis dans mon livre : "Si je vous vends une pilule sur la base de données montrant qu'elle réduit la tension artérielle par rapport à un groupe de contrôle recevant un placebo, mais que je ne vous dis pas que j'ai également testé son efficacité pour améliorer vingt-cinq autres paramètres de santé et que je n'ai rien trouvé pour chacun d'entre eux, **il s'agit d'un résultat très faible**. D'un point de

vue statistique, si vous disposez de suffisamment de données et que vous effectuez suffisamment de tests, vous pouvez toujours trouver quelque chose qui est, selon les normes des tests statistiques utilisés par les psychologues, "significatif".
Ou, formulé différemment, "Torturez les données suffisamment longtemps et elles avoueront n'importe quoi". Ensuite, une fois que vous avez obtenu l'aveu que vous voulez, vous pouvez HARK, ou émettre des hypothèses une fois que les résultats sont connus : *Ah, oui, c'est ce que nous nous attendions à trouver depuis le début*. Je savais que *la pilule aiderait à lutter contre l'hypertension.* Dans une telle situation, cependant, il existe un risque sérieux que ce que vous observez ne soit pas une pilule qui réduit l'hypertension artérielle, mais un bruit statistique.

La bonne nouvelle est que les chercheurs en psychologie et dans d'autres domaines touchés par les crises de réplication sont de plus en plus conscients de la manière dont ces pratiques peuvent générer des résultats faibles et non reproductibles. Les réformateurs scientifiques ont commencé à mettre en place des garde-fous qui restreignent efficacement les chercheurs et réduisent leurs degrés de liberté : Par exemple, il est possible d'inciter ou d'obliger les chercheurs à "préenregistrer" leurs hypothèses et à présenter exactement les tests statistiques qu'ils prévoient d'effectuer, de sorte que s'ils modifient leur plan d'analyse ou leur hypothèse en cours de route, cela sera visible pour tout le monde. Vous pouvez également les inciter ou les obliger à partager leurs données, ce qui permet aux autres chercheurs de vérifier plus facilement s'ils se sont livrés à des chicaneries statistiques.

Je pense que l'on peut raisonnablement affirmer que les résultats positifs rapportés dans "Psychosocial Functioning in Transgender Youth after 2 Years of Hormones", un article de recherche très attendu qui vient d'être publié dans le *New England Journal of Medicine*, peuvent être au moins partiellement expliqués par le *type de sélection statistique* qui a tendance à produire des résultats bancals.

L'article du *NEJM* s'inscrit dans le cadre de l'étude Trans Youth Care-United States (TYCUS), que les chercheurs décrivent comme "*une étude prospective d'observation évaluant les résultats physiques et psychosociaux du traitement médical de la dysphorie de genre dans deux cohortes distinctes de jeunes transgenres et non binaires*", l'une recevant des bloqueurs de puberté (l'équipe n'a pas encore rendu compte de ces résultats), et l'autre — celle-ci — recevant des hormones. L'étude TYCUS est menée dans quatre grandes cliniques spécialisées dans le traitement des jeunes transgenres […]. Les chercheurs cités comme coauteurs de cette étude comprennent certains des plus grands noms de la médecine et de la psychologie du genre chez les jeunes : Diane Chen, Johnny Berona, Yee-Ming Chan, Diane Ehrensaft, Robert Garofalo, Marco A. Hidalgo, Stephen M. Rosenthal, Amy C. Tishelman et Johanna Olson-Kennedy. Un certain nombre d'entre eux ont été des défenseurs déclarés des traitements de médecine de genre pour les jeunes.

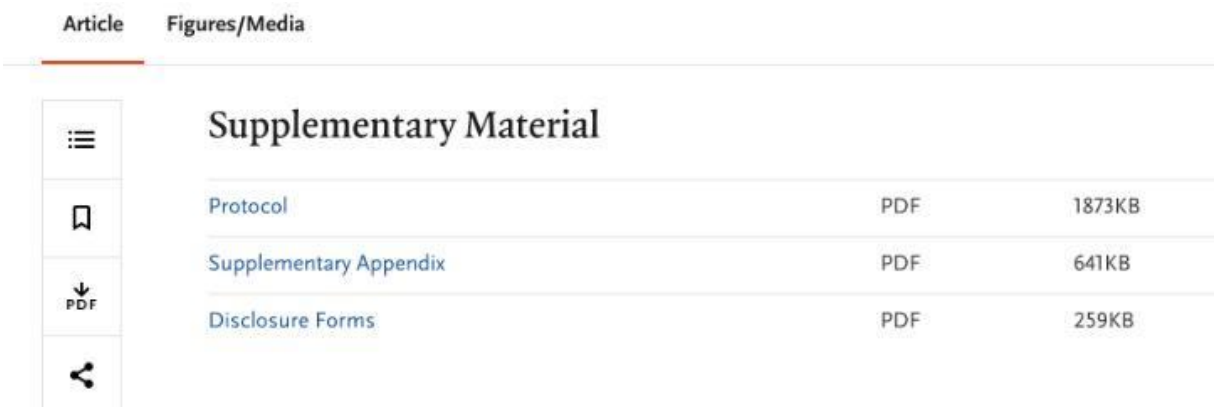
Cette équipe a reçu d'importantes subventions pour étudier cette population, et ce pour de bonnes raisons : Les cliniques américaines spécialisées dans les questions de genre sont très en retard dans la production de données utiles qui pourraient nous aider à mieux comprendre si et dans quelles circonstances la médecine de genre pour les jeunes est bénéfique pour les enfants souffrant de dysphorie de genre. Comme l'écrivent les auteurs dans leur protocole d'étude, leur objectif est de "*collecter des données critiques sur les modèles existants de soins pour les jeunes transgenres qui ont été couramment utilisés dans les milieux cliniques pendant près d'une décennie, bien qu'avec très peu de recherche empirique pour les soutenir*." Ils ont écrit cela en 2016, mais la situation n'a pas vraiment changé : il n'y a presque pas de bonnes données — ou même décentes — sur ces questions vitales. Sur la même page du protocole, ils écrivent : "*Cette recherche a une portée très importante car il s'agit de la première étude longitudinale recueillant des données — évaluant à la fois les résultats physiologiques et de santé mentale — pour évaluer les directives cliniques couramment utilisées pour les jeunes transgenres aux États-Unis*."

Dans leur étude publiée dans le *NEJM*, les chercheurs publient pour la première fois des données sur le suivi, au fil du temps, de la santé mentale de la cohorte de jeunes qui ont pris des hormones. Et les nouvelles semblent bonnes : L'équipe rapporte que deux ans après l'administration d'hormones, les enfants transgenres de leur étude ont connu une augmentation de leur "congruence d'apparence", c'est-à-dire le sentiment que leur apparence extérieure correspond à leur identité de genre, et de leur affect positif. Les garçons trans (filles natales) ont également connu une réduction de la dépression et de l'anxiété, ainsi qu'une augmentation de la satisfaction à l'égard de la vie, ce qui n'est pas le cas des filles trans (garçons natals).

Sur la base de ces résultats, l'étude est largement promue, à la fois par la plupart des médias grand public qui l'ont couverte et par les auteurs eux-mêmes, comme une preuve solide que les hormones améliorent le bien-être des jeunes transgenres. "*Nos résultats fournissent une base scientifique solide pour affirmer que les soins visant à affirmer le genre sont cruciaux pour le bien-être psychologique de nos patients*", a déclaré Garofalo, l'un des chercheurs principaux de l'étude, ainsi que le codirecteur de la clinique pour jeunes transgenres à l'hôpital pour enfants Lurie, dans un communiqué de presse de l'hôpital. "*Les résultats critiques que nous rapportons démontrent l'impact psychologique positif des hormones de confirmation du genre pour le traitement des jeunes souffrant de dysphorie de genre*", a ajouté Olson-Kennedy, directrice médical de la clinique de l'hôpital pour enfants de Los Angeles.

Je ne suis pas d'accord avec ces affirmations, mais j'y reviendrai dans la deuxième partie de ma critique de cette étude, qui se concentrera sur les résultats rapportés par les chercheurs. Cette partie porte principalement sur les **résultats qu'ils n'ont pas rapportés**, ce qui est un sujet important en soi. Mon argument en faveur de l'existence d'une situation douteuse est simple, et repose en grande partie sur le protocole d'étude que les auteurs ont rédigé dans le cadre de la procédure d'approbation

de leur Institutional Review Board (IRB). Ce document figure dans la section supplémentaire de la page *NEJM* de l'article et fait l'objet d'un lien :



Vous pouvez lire ou télécharger le protocole, l'article et une annexe supplémentaire que nous aborderons plus tard ici si vous le souhaitez.

Le protocole est un document long et riche, qui présente entre autres les procédures de l'étude pour les cohortes d'hormones et de bloqueurs. Une note dans la section supplémentaire du document explique qu'il comprend à la fois la version "originale" (2016) et la version "finale" du protocole. Je citerai la version finale, et donc la plus applicable, qui a été soumise le 11 mai 2021, bien qu'en ce qui concerne ce dont je vais parler, il n'y ait pas de différences substantielles entre les deux versions (à une exception près que j'évoquerai).
Le document de protocole fait office de pré-enregistrement *de facto* pour Chen et son équipe (ils ont également publié une version plus courte de ce document sous la forme d'un rapport enregistré, une forme de pré-enregistrement plus officielle), et il montre que dans l'étude du *NEJM*, les chercheurs ont simplement **exclu la plupart des variables clés dont ils avaient émis l'hypothèse qu'elles s'amélioreraient sous l'effet des hormones, et qu'ils ont modifié leur hypothèse de manière significative, de sorte que certaines de ces variables ont été mises à l'écart.**

Précisons ce que les auteurs ont fait et pourquoi cela soulève des questions. Ils énumèrent un certain nombre d'hypothèses dans leur protocole. L'une d'entre elles correspond à l'étude actuelle :
"**Hypothèse 2a** : *Les patients traités avec des hormones d'affirmation du genre présenteront une diminution des symptômes d'anxiété et de dépression, de dysphorie de genre, d'automutilation, de symptômes de traumatisme et de suicidalité, ainsi qu'une augmentation [sic] de l'estime corporelle et de la qualité de vie au fil du temps*."
La première sous-section de la section "Analyse statistique" est intitulée "*Objectif principal : Effets des interventions hormonales sur la santé mentale et le bien-être psychologique* ", et les auteurs y expliquent que leur analyse " *étudiera les changements au fil du temps dans la dysphorie de genre, la dépression, l'anxiété, les symptômes de traumatisme, l'automutilation, la suicidalité, l'estime corporelle et la qualité de vie* ". Il est donc assez clair, entre l'hypothèse et l'objectif principal, que ce qui les intéresse le plus est d'enquêter.

Dans l'annexe II du protocole, les chercheurs incluent un tableau utile des instruments qu'ils prévoient d'utiliser pour suivre ces variables et d'autres. Ce tableau est cependant un peu dépassé, car il utilise encore les "hormones du sexe opposé" plutôt que les "hormones d'affirmation du genre" et inclut quelques variables que les chercheurs ont par la suite abandonnées. Une version plus récente peut être téléchargée ici, et voici la partie du document qui énumère toutes les variables pour lesquelles des données ont été collectées à chaque vague de six mois de l'effort de recherche (ce sont les données sur lesquelles les chercheurs seraient le plus susceptibles de faire rapport dans une grande étude) :

Gender Affirming Hormone Cohort Survey Measures	
Construct	**Measure**
Time of Completion: Baseline, 6-month, 12-month, 18-month, & 24-month follow-up periods	
Demographics	Demographic questions for Cross-Sex Hormone Cohort
Religiosity & Spirituality	Modified Duke University Religion Index (DUREL)
Socio-Economic Status	Socioeconomic Status Questions (for Adolescents & Young Adults)
Gender Identity	Transgender Congruence Scale
	DSM 5 – Chicago adapted
Service Utilization	Dr. Olson's Service Utilization Questions
Depression	BDI-II
Anxiety	Revised Children's Manifest Anxiety Scale: Second Edition (RCMAS-2 – What I Think and Feel)
Quality of Life	Health-Related Quality of Life Scale (modified HIV QOL)
Suicidality	Suicidal Ideation Scale
Body Esteem	Body Esteem Scale
Body Image	Body Image Scale
Social Relationships	Emotional Support / Friendship / Loneliness / Perceived Hostility / Perceived Rejection – NIH Toolbox
Negative Affect	Anger / Fear / Sadness – NIH Toolbox
Psychological Well-being	General Life Satisfaction / Positive Affect – NIH Toolbox
Self-Efficacy	Self-Efficacy (CAT 13-17)– NIH Toolbox
Perceived Parent Support	Parent Support Scale – Youth Version
Resiliency	Gender Minority Stress and Resilience Scale
	Connor-Davidson Resilience Scale
Sexual Behavior	Sexual Risk Behavior Questions
STI History	STI Questions
Alcohol/Drug Use	Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST)
Autism	Autism-Spectrum Quotient (AQ-10) – Adult
History of Blocker Experience	Questions to obtain history of participant's blocker experience

D'accord, passons maintenant à la section "Mesures" de la présente étude dans le *NEJM*, où les chercheurs énumèrent leurs variables : "*En ce qui concerne les résultats longitudinaux, les participants ont rempli l' échelle de congruence transgenre (TCS Transgender Congruence Scale), l'inventaire de dépression de Beck-II, l'échelle d'anxiété manifeste révisée pour enfants (deuxième édition) et les mesures de l'affect positif et de la satisfaction de la vie de la batterie d'émotions de la boîte à outils des NIH (National Institutes of Health), à chaque visite durant l'étude.*"

Si vous comparez ce document au protocole, vous remarquerez que sur les huit variables clés qui intéressaient le plus les chercheurs — "**dysphorie de genre**, dépression, anxiété, **symptômes de traumatisme, automutilation, suicidalité, estime corporelle** et **qualité de vie**" — **celles que j'ai surlignées ne sont pas mentionnées dans l'article du *NEJM*.** Cela représente six variables sur huit, soit 75 % des variables couvertes par l'hypothèse des chercheurs dans leur document de protocole (y compris la version abrégée "officiellement" préenregistrée)[1].

En fait, la plupart de ces variables ne sont pas mentionnées du tout dans l'article du *NEJM* ou dans son annexe supplémentaire. La "*dysphorie de genre*" est mentionnée très tôt, car comment ne pas le faire dans un article sur la dysphorie de genre, **mais il n'y a aucune mention d'une quelconque échelle de dysphorie de genre** (c'est la seule variable manquante pour laquelle ils ont une explication partielle mais non compromettante, ce à quoi je vais arriver). Ni l'expression "*qualité de vie*" ni aucune mention de l'échelle de qualité de vie liée à la santé n'apparaissent dans l'article. Les auteurs font état du nombre de suicides accomplis et de cas d'idées suicidaires dans l'échantillon (plus d'informations à ce sujet dans la deuxième partie), mais il n'y a **aucune mention de l'échelle de suicidalité** — celle qu'ils ont utilisée dans l'article sur les caractéristiques de base — **qui leur permettrait d'analyser statistiquement le niveau de suicidalité** de l'échantillon au fil du temps, de la même manière qu'ils analysent les trajectoires longitudinales d'autres variables. (Ils **disposent en tout cas de données** sur l'estime corporelle et la suicidalité, puisqu'ils rapportent les chiffres de base de ces variables dans un article de 2021).

J'ai également consulté l'annexe supplémentaire de l'article du *NEJM* pour voir s'il y avait une explication à ces disparitions. Je suis tombé sur une courte section très prometteuse intitulée "*Rationale for Selecting Primary Mental Health Outcome Measures*", mais hélas, elle concerne un problème relativement mineur et sans rapport avec le sujet — elle n'explique pas où sont passées les variables clés. Ces variables ne sont pas non plus mentionnées dans le reste de l'annexe. (Si les auteurs avaient voulu expliquer l'absence de certaines variables sans occuper un espace potentiellement limité dans l'article lui-même, cela aurait été un bon endroit pour le faire).

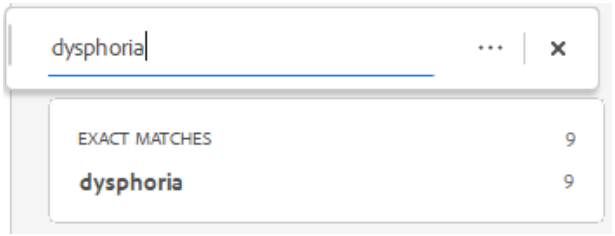
L'hypothèse des chercheurs change également dans l'étude du *NEJM*. Pour être honnête, lorsqu'ils font référence à leur hypothèse ici, c'est d'une manière moins formelle et plus familière — il n'y a pas de section officielle "Hypothèses" comme dans le document de protocole. Mais regardez quand même ce changement :

- **Hypothèse de la dernière version du protocole, publiée en 2021** : "*Les patients traités avec des hormones d'affirmation du genre présenteront une diminution des symptômes d'anxiété et de dépression, de dysphorie de genre, d'automutilation, de symptômes de traumatisme et de suicidalité, ainsi qu'une augmentation [sic] de l'estime corporelle et de la qualité de vie au fil du temps.*"
- **Hypothèse de l'étude *NEJM*, publiée en 2023** : "*Nous avons émis l'hypothèse que [après l'administration d'hormones aux enfants], la congruence de l'apparence, l'affect positif et la satisfaction de la vie augmenteraient et que les symptômes de dépression et d'anxiété diminueraient. Nous avons également émis l'hypothèse que les améliorations seraient secondaires au traitement de la dysphorie de genre, de sorte que l'augmentation de la congruence de l'apparence serait associée à des améliorations concomitantes des résultats psychosociaux.*"

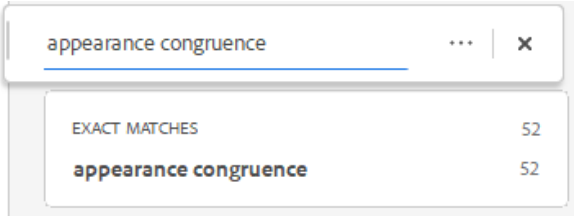
Il y a quelques similitudes, dans la mesure où la dépression et l'anxiété sont mentionnées dans les deux cas, mais les changements sont plutôt frappants. Un certain **nombre de variables dont l'hypothèse initiale était qu'elles seraient les plus importantes ont été supprimées**, y compris plusieurs – Dysphorie de genre, suicidalité et automutilation — universellement reconnues par les chercheurs sur le genre chez les jeunes comme étant d'une importance vitale. D'autres variables, comme "*la congruence de l'apparence, l'affect positif et la satisfaction à l'égard de la vie*", ont été incluses dans le protocole original**, mais n'ont pas été considérées comme particulièrement importantes** et n'ont pas été mentionnées dans les sections consacrées à l'hypothèse ou à l'objectif principal. (Et non, la qualité de vie et la satisfaction à l'égard de la vie ne sont pas le même concept — elles figurent sous deux variables différentes dans le protocole de l'étude, et il existe au moins une étude qui tente d'évaluer la force de la corrélation entre les deux).

Dans l'article du *NEJM*, les chercheurs semblent beaucoup plus intéressés par le concept de "congruence d'apparence" qu'ils ne l'étaient auparavant.

Alors que les termes vitaux de suicide (et ses variantes) et de dysphorie sont mentionnés respectivement huit et neuf fois dans l'article…



… la "congruence d'apparence" est mentionnée 52 fois



Elle apparaît même dans le tout premier paragraphe : "*Un objectif important de ce traitement est d'atténuer la dysphorie de genre en augmentant la congruence de l'apparence, c'est-à-dire le degré auquel les jeunes ressentent la concordance entre leur genre et leur apparence physique*".

Ce changement est remarquable. L'expression "*congruence d'apparence*" n'est pas mentionnée une seule fois dans le document de protocole, et la seule fois où le mot apparence apparaît dans ce contexte, c'est dans cette description de l'échelle de congruence transgenre (TCS), l'une des variables sur lesquelles les chercheurs ont recueilli des données : "*Une construction de la congruence pour conceptualiser le degré auquel les personnes transgenres se sentent authentiques et à l'aise avec leur identité de genre et leur apparence extérieure.*"

Même concernant ce test, il y a une sélection apparente. Dans le document de protocole et dans l'article du *NEJM*, les auteurs mentionnent l'administration du TCS. Mais **ils ne rapportent les résultats complets nulle part** dans l'article du *NEJM* — au lieu de cela, ils ne rapportent **que l'une des deux sous-échelles de l'échelle, Appearance Congruence** (encore une fois, nous savons qu'ils ont les données complètes parce qu'ils en ont fourni une partie dans leur article sur les mesures de base). Cela signifie que les chercheurs ont eu trois fois l'occasion d'analyser la situation : Ils pouvaient

analyser les changements au fil du temps sur l'échelle complète, puis sur chacune des deux sous-échelles. **Ils ne font état que d'un seul de ces trois résultats, ce qui leur permet d'obtenir ce qu'ils décrivent dans l'article comme leur plus forte conclusion** : Au cours des deux années de traitement aux hormones, l'enfant moyen de l'étude s'est amélioré d'environ un point sur cette sous-échelle de cinq points. Les chercheurs consacrent ensuite une grande partie de leur article à ce résultat, allant jusqu'à dire qu'ils ont émis l'hypothèse que la congruence de l'apparence serait importante — ce qui, à mon avis, donne l'impression qu'ils ont émis cette hypothèse depuis le début, alors que je ne vois aucune preuve qu'ils l'aient fait. Au contraire, **ils ont émis une hypothèse assez différente dans leur document de protocole, puis ils ont modifié cette hypothèse sans expliquer pourquoi**. (Je pense également que cette découverte sur la congruence de l'apparence est beaucoup moins impressionnante que les chercheurs ne le laissent entendre, mais je laisserai cela pour la partie 2).

L'approche des chercheurs est un peu similaire à la Boîte à outils de l'Institut national du vieillissement, "un ensemble multidimensionnel de mesures brèves évaluant les fonctions cognitives, émotionnelles, motrices et sensorielles des personnes âgées de 3 à 85 ans". Il s'agit essentiellement d'un mix de différents éléments d'enquête (plus d'informations et échelles de notation ici) — il y en a beaucoup. Si je remonte ce grand tableau et que j'indique les différents éléments que les chercheurs ont inclus dans cette batterie…

Gender Affirming Hormone Cohort Survey Measures	
Construct	**Measure**
Time of Completion: Baseline, 6-month, 12-month, 18-month, & 24-month follow-up periods	
Demographics	Demographic questions for Cross-Sex Hormone Cohort
Religiosity & Spirituality	Modified Duke University Religion Index (DUREL)
Socio-Economic Status	Socioeconomic Status Questions (for Adolescents & Young Adults)
Gender Identity	Transgender Congruence Scale
	DSM 5 – Chicago adapted
Service Utilization	Dr. Olson's Service Utilization Questions
Depression	BDI-II
Anxiety	Revised Children's Manifest Anxiety Scale: Second Edition (RCMAS-2 – What I Think and Feel)
Quality of Life	Health-Related Quality of Life Scale (modified HIV QOL)
Suicidality	Suicidal Ideation Scale
Body Esteem	Body Esteem Scale
Body Image	Body Image Scale
Social Relationships	Emotional Support / Friendship / Loneliness / Perceived Hostility / Perceived Rejection – NIH Toolbox
Negative Affect	Anger / Fear / Sadness – NIH Toolbox
Psychological Well-being	General Life Satisfaction / Positive Affect – NIH Toolbox
Self-Efficacy	Self-Efficacy (CAT 13-17)– NIH Toolbox
Perceived Parent Support	Parent Support Scale – Youth Version
Resiliency	Gender Minority Stress and Resilience Scale
	Connor-Davidson Resilience Scale
Sexual Behavior	Sexual Risk Behavior Questions
STI History	STI Questions
Alcohol/Drug Use	Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST)
Autism	Autism-Spectrum Quotient (AQ-10) – Adult
History of Blocker Experience	Questions to obtain history of participant's blocker experience

…vous verrez que les chercheurs ont demandé aux participants à l'étude de remplir des questions de la NIH Toolbox sur le soutien émotionnel, l'amitié, la solitude, l'hostilité perçue, le rejet perçu, la colère, la peur, la tristesse, la satisfaction générale de la vie, l'affect positif et l'auto-efficacité.
Aucun de ces éléments n'a été mis en avant par les chercheurs dans leur hypothèse initiale ou dans la section relative à l'objectif principal, de sorte que nous ne devrions probablement pas avoir d'idée préconçue sur ceux que nous devrions attendre d'eux dans le cadre d'une étude majeure, mais tout de même : **neuf des onze éléments sont introuvables, ce qui nous laisse avec les seules mesures de l'affect positif et de la satisfaction à l'égard de la vie**. Pourquoi ? Et pourquoi était-il plus important de rendre compte de la "*satisfaction à l'égard de la vie*", qui ne figure pas dans les sections relatives à l'hypothèse ou aux principaux résultats, que de la "*qualité de vie*", qui y figure ? De même, pourquoi rendre compte des éléments relatifs à l'affect positif, **mais pas de ceux relatifs à l'affect négatif ?** Si les hormones aident les enfants à se sentir mieux, ne devraient-ils pas éprouver moins de colère, moins de peur et moins de tristesse au fil du temps ?

Nous disposons de nombreuses informations utiles sur le protocole des chercheurs grâce à ce qu'ils ont mis en ligne. Ils devaient certainement **soumettre** leur document de protocole à l'approbation de

l'IRB (Institutional Review Board), et l'avoir dans leurs dossiers quelque part, mais je suis un peu ignorant de ces aspects bureaucratiques, et je ne sais donc pas s'ils étaient obligés de le publier en tant qu'exigence de la subvention ou du *NEJM*, ou s'ils l'ont fait par esprit d'ouverture. Quoi qu'il en soit, l'existence du document de protocole montre bien pourquoi ce type de transparence scientifique est utile : Dans ce cas, elle nous permet d'aller au-delà des résultats officiellement publiés, de comparer ces **résultats** au processus plus large qui les a produits et de poser des questions. Ce qui est moins clair, c'est l'absence de toute information sur les raisons pour lesquelles les auteurs ont fait les choix qu'ils ont faits pour l'article du *NEJM*.

Il va sans dire que je ne prétends pas que si les hormones ont un effet bénéfique sur les jeunes transgenres, **chaque** variable de cette étude devrait présenter un changement important et salutaire. Ce que je veux dire, c'est qu'en l'absence d'explication, nous ne pouvons que spéculer sur la manière dont les chercheurs ont pris toutes ces décisions subtiles — des décisions qui leur ont permis d'écrire, dans leur article final publié, qu'"*il y a eu des changements significatifs au sein des participants au fil du temps pour **tous** les résultats psychosociaux dans les directions hypothétiques*". Cette affirmation est à la limite de la tromperie.
Par "**tous**" les résultats psychosociaux, ils n'entendent pas : **tous ceux qu'ils ont mesurés et évalués** en termes de changement au fil du temps, mais : **ceux pour lesquels ils ont choisi de présenter des résultats**. Ce qui pourrait rendre leurs conclusions totalement insignifiantes. Pour prendre un exemple plus extrême, je ne peux pas tirer à pile ou face, encore et encore, jusqu'à ce que j'obtienne 10 têtes d'affilée quelques heures plus tard, et écrire ensuite "La présence de 10 têtes d'affilée suggère que la pièce n'est pas juste".

Pourquoi tant de variables ont-elles disparu ? Il y a quelques explications. L'une d'elles est que les chercheurs prévoient de rendre compte de ces résultats dans une prochaine étude, mais je ne vois pas très bien pourquoi ils ne le feraient pas dans le *NEJM,* et cela n'expliquerait toujours pas comment ils ont choisi les variables à prendre en compte. Une autre possibilité est que la revue elle-même ait demandé aux chercheurs de se concentrer sur des variables spécifiques. Le chemin vers la publication d'un article dans une revue aussi prestigieuse que le *NEJM* peut être semé d'embûches, et il se peut que vous soumettiez un premier projet qui vous enthousiasme, en raison de ce qu'il montre sur A, B et C, mais que les pairs évaluateurs massacrent votre beau bébé jusqu'à ce que, de nombreux mois, voire des années (sans parler des cheveux gris) plus tard, vous soupirez avec résignation et acceptiez de publier un article avec des résultats beaucoup moins excitants, sévèrement couverts, concernant les variables X, Y et Z, beaucoup moins séduisantes. Je suppose que cela aurait pu se produire dans le cas présent, mais cela n'aurait fait que renvoyer la question : "Pourquoi avez-vous choisi ces variables particulières ?" au *NEJM*, ce qui signifie que les questions méthodologiques sérieuses seraient restées sans réponse. En outre, comme nous le verrons dans la deuxième partie, le *NEJM* **n'a pas vraiment fait la pluie et le beau temps en ce qui concerne la rigueur**

méthodologique de cette étude, et je ne suis donc pas sûr de croire à cette version hypothétique des événements.

Globalement, si je devais deviner, je pense que l'explication la plus probable ici est que les chercheurs ont fait beaucoup d'"analyses exploratoires" jusqu'à ce qu'ils trouvent des résultats raisonnablement impressionnants, et qu'ils ont ensuite choisi de recentrer leurs efforts — et de remanier leur hypothèse — autour de ces résultats, en jetant certains résultats décevants dans un tiroir.
Si j'ai raison, il ne s'agissait pas nécessairement d'un processus intentionnel de leur part. Lorsqu'un grand nombre de personnes fouillent dans une grande quantité de données sans que des garde-fous soient mis en place, il est facile de perdre de vue tous les tests statistiques infructueux que vous avez effectués, tout en vous souvenant des résultats positifs qui soutiennent votre hypothèse préférée. Mais que j'aie raison ou non et que la sélection des données ait été intentionnelle ou non, les chercheurs auraient dû au moins se rendre compte de ce qui manquait dans leur article du *NEJM* et expliquer ce qui s'est passé quelque part.

Mais tout ce que je peux faire, c'est spéculer, **parce qu'ils ne veulent répondre à aucune question sur leur processus ou sur la possibilité de partager leurs données** pour que d'autres puissent étudier ces questions. J'ai envoyé des questions spécifiques au *NEJM*, aux contacts presse de deux des universités et à quatre membres de l'équipe (Chen, Hidalgo, Rosenthal et Tishelman), et à part une réponse du *NEJM* disant que je devais contacter directement les institutions des chercheurs, je n'ai eu de réponse que de la part d'un attaché de presse du Lurie Children's Hospital confirmant que les chercheurs n'accordaient pas d'interviews. Pour être honnête, la position de ne pas faire d'interviews semble être cohérente, quel que soit le journaliste qui pose la question. Dans mon dernier courriel, j'ai demandé à cette attachée de presse si l'équipe pouvait partager ses données — elle a dit qu'elle vérifierait, mais je n'ai pas eu de réponse. (En fait, je ne sais pas exactement comment fonctionne le partage des données — leur document de protocole indique que l'équipe les partagera éventuellement avec d'autres chercheurs des National Institutes of Health[2], mais il se pourrait qu'ils soient soumis à des restrictions lorsqu'il s'agit de journalistes choisis au hasard ou d'universitaires indépendants. **Le refus de partager les données n'est donc pas nécessairement un motif de suspicion dans ce cas).**

Ce refus de parler aux journalistes est une décision malheureuse de la part des chercheurs, surtout lorsqu'elle est associée à leurs citations élogieuses sur l'importance de leurs résultats — citations qui masquent beaucoup de nuances et de résultats manquants. En fin de compte, **cette équipe a prédit publiquement que 8 variables évolueraient dans une direction donnée. Puis, au moment de communiquer ses données, elle ne nous a dit ce qu'il advenait que de 2 de ces variables, et les 2 variables dont elle a fait état n'étaient même pas des résultats directs, étant donné que les filles transgenres n'ont pas connu de réduction de la dépression et de l'anxiété**. Si ces résultats sont si impressionnants, où sont toutes ces autres variables ?

Je crois que l'article fondateur sur le HARKing, c'est-à-dire l'émission d'hypothèses après que les résultats soient connus, est celui publié en 1998 par Norbert L. Kerr dans *Personality and Social Psychology Review*. À l'époque, ce phénomène n'était pas très connu, et certaines personnes l'ont même défendu ! L'idée était que si, en fouillant dans vos données, vous découvriez une nouvelle explication, pourquoi ne pas mettre à jour votre hypothèse pour en tenir compte ? À l'époque, de nombreux chercheurs talentueux et bienveillants ne comprenaient pas vraiment les inconvénients statistiques et autres de cette méthode, si bien que Kerr a dû affirmer que les inconvénients du HARKing l'emportaient sur les avantages. En ce sens, il s'agit d'une lecture étrange par rapport aux normes actuelles — de nos jours, la plupart des chercheurs comprennent pourquoi ces pratiques conduisent à une science bancale.

Selon Kerr :

> Tout le monde serait probablement d'accord pour dire que, toutes choses étant égales par ailleurs, un "bon" écrit scientifique (c'est-à-dire clair, cohérent, engageant, passionnant) est meilleur qu'un "mauvais" écrit scientifique (c'est-à-dire incohérent, peu clair, turgescent, peu engageant). Mais tout le monde reconnaîtrait probablement aussi que l'auteur d'un rapport scientifique a des contraintes sur ce qu'il ou elle peut écrire sous prétexte de raconter une bonne histoire. Les rapports scientifiques ne sont pas de la fiction, et un scientifique est soumis à des contraintes différentes de celles d'un auteur de fiction. Même si les ajouts peuvent améliorer l'histoire, le scientifique ne peut pas fabriquer ou déformer les résultats empiriques. La question ultime est de savoir si de telles contraintes devraient s'appliquer aux aspects fictifs de HARKing (par exemple, la représentation inexacte de certaines hypothèses comme étant celles qui ont guidé la conception de l'étude).

Les auteurs de l'article du *NEJM* ont-ils "présenté de manière inexacte certaines hypothèses comme étant celles qui ont guidé la conception de l'étude" ? Peut-être est-ce une affirmation trop forte, mais je n'en suis pas sûr. Les chercheurs sont très clairs sur les variables qui les intéressent le plus dans le document de protocole qui est censé sous-tendre cette étude — ils émettent l'hypothèse que "*les patients traités avec des hormones du sexe opposé présenteront moins de symptômes d'anxiété et de dépression, de dysphorie de genre, d'automutilation, de symptômes de traumatisme et de suicidalité et augmenteront [sic] l'estime corporelle et la qualité de vie au fil du temps*". Puis, dans l'étude qui est l'une des principales raisons pour lesquelles ils ont collecté toutes ces données — une étude qui inclut la ligne "*Les auteurs se portent garants de l'exactitude et de l'exhaustivité des données et de la fidélité de l'étude au protocole*" — leur hypothèse est substantiellement différente, et ils présentent **leur intérêt pour la congruence de l'apparence comme une hypothèse qu'ils avaient depuis le début**, alors qu'il n'y a aucune preuve que c'était le cas. Ce changement et la disparition de toutes ces variables restent presque entièrement inexpliqués.

Comme je l'ai déjà mentionné, les auteurs proposent une explication partielle à l'absence de variables relatives à la dysphorie de genre. À l'origine, ils ont recueilli des données sur la dysphorie de genre à l'aide de deux instruments, l'échelle de dysphorie de genre d'Utrecht (UGDS) et le questionnaire sur l'identité de genre et la dysphorie de genre pour les adolescents et les adultes (GIDYQ-AA). Dans un article paru en 2019 dans Transgender Health, ils décrivent certains des défauts apparents de ces mesures et expliquent qu'ils ont cessé de collecter des données à leur sujet :

> La nécessité d'une mesure améliorée pour saisir les éléments nuancés de la dysphorie de genre et son potentiel d'intensification ou d'atténuation au fil du temps a été soulignée par les membres de notre équipe transgenre qui ont été en première ligne avec les jeunes participant à l'étude. Après mûre réflexion, nous avons choisi d'inclure l'UGDS dans cette étude, dans l'espoir de démontrer ses limites à saisir l'expérience dynamique des jeunes atteints de dysphorie de genre. Pour des raisons similaires, nous avons également inclus le questionnaire sur l'identité de genre et la dysphorie de genre pour les adolescents et les adultes (GIGDQAA) [GIDYQ-AA].
>
> Au cours des deux dernières années, des participants ont fait part de leurs préoccupations concernant la détresse qu'ils ressentent lorsqu'ils sont confrontés à certains éléments de ces deux instruments. Il a été décidé de supprimer l'UGDS et le GIGDQAA [GIDYQ-AA] de l'évaluation des participants, sauf pour les participants ayant participé à une sous-étude visant à recueillir leurs commentaires sur les éléments inclus dans ces deux échelles. Notre équipe a estimé que l'échelle de congruence transgenre (TCS) et l'échelle de stress et de résilience des minorités de genre (Gender Minority Stress and Resilience Scale) sont probablement les meilleures mesures existantes pour recueillir des informations sur les facteurs distaux et proximaux de la dysphorie de genre. [notes de bas de page omises].

J'ai trouvé cela un peu étrange — ils ont utilisé ces deux échelles dysphorie de genre (DG) non pas pour mesurer la DG, comme l'explique leur document de protocole, mais parce qu'ils pensaient qu'elles étaient mauvaises et qu'ils voulaient le démontrer ? Ceci mis à part, cela se vérifie, car le document de protocole comprend une "*lettre de modification*" de 2019 qui supprime ces deux instruments. (Ils sont toujours administrés lors de la visite à 24 mois).

Mais si la Transgender Congruence Scale et la Gender Minority Stress and Resilience Scale sont, en fait, "*probablement les meilleures mesures existantes*" pour évaluer la dysphorie de genre, **pourquoi sont-elles toutes deux absentes de l'article du *NEJM*, à l'exception de cette sous-échelle de la TCS ?** Et selon le protocole, les enfants de cette cohorte ont également été interrogés sur leurs symptômes de dysphorie de genre selon le DSM-5 jusqu'à ce qu'une lettre d'amendement distincte de 2021 mette fin à cette question. **Mais ces données ne figurent pas non plus dans l'article du *NEJM*.** Pourquoi ces données sont-elles absentes de l'article du *NEJM* ?

En bref, que l'abandon de l'UGDS et du GIDYQ-AA soit justifié ou non, les chercheurs ont recueilli des données sur trois autres mesures qui, selon eux, ont un objectif similaire, **mais ils ne les ont pas publiées**. Il est décevant de constater que cette étude ne fournit aucune donnée sur la DG, étant donné que l'atténuation de la DG est **la** justification médicale ostensible pour mettre les enfants sous bloqueurs et/ou hormones en premier lieu, malgré le manque de preuves solides publiées sur ces traitements[1]. Il serait intéressant de savoir si des pairs ou des rédacteurs du *NEJM* ont demandé aux auteurs pourquoi leur article ne contenait pas de données longitudinales sur la dysphorie de genre, la suicidalité et l'automutilation, compte tenu de l'importance de ces variables et de l'intérêt préalable démontré par l'équipe de recherche pour le suivi de ces résultats.

Rien de ce que je dis au sujet des degrés de liberté des chercheurs ou de la méthode HARKing n'est nouveau ou controversé. Encore une fois, les chercheurs savent depuis des années qu'il ne faut pas faire cela ; si vous ne tenez pas compte du fait que **vous avez fait un tas d'autres comparaisons statistiques dont vous n'avez pas fait état,** cela peut remettre en question toute votre analyse, parce que des résultats qui semblent statistiquement significatifs peuvent passer sous ce seuil une fois que vous avez fait les corrections appropriées.

Voici un tableau pratique, tiré de l'annexe supplémentaire, qui présente les scores moyens obtenus pour les cinq variables que les chercheurs ont relevées au départ et à 24 mois (la dernière vague de collecte de données pour cette étude) :

Table S5. Paired Samples *t*-tests Comparing Scores at Baseline and 24 Months

	n	baseline	24 Months	*p*-value	effect size
Appearance congruence	213	2.86 (0.74)	3.86 (0.76)	<0.001	-1.12
Depression	211	16.39 (11.88)	13.95 (12.76)	<0.001	0.20
Anxiety	208	60.25 (11.18)	57.38 12.00)	<0.001	0.25
Positive affect	215	42.90 (10.05)	43.72 (12.03)	0.37	-0.05
Life satisfaction	217	39.92 (10.55)	44.61 (12.29)	<0.001	-0.39

Note. Variables are presented as mean (SD). Results are based on *t*-tests (baseline minus 24-months). Negative *t*-test values indicate increases in appearance congruence, positive affect, and life satisfaction. Effect sizes are Cohen's *d* (ranges: 0.20, small; 0.50, medium; 0.80, large).

Si l'on met de côté certaines questions que j'aborderai dans la deuxième partie, comme le fait que les chercheurs vantent une augmentation de 0,82 point sur deux ans sur une échelle d'affect positif de 100 points comme preuve de l'efficacité des hormones, il manque **beaucoup de choses** dans ce tableau.
En réalité, le graphique devrait ressembler à ceci :

[1] Je suppose que quelqu'un pourrait dire "Ce n'est pas juste — ils ont au moins publié les résultats de la sous-échelle de congruence de l'apparence". Mais il ne s'agit pas d'une mesure validée de la dysphorie de genre, un phénomène qui va bien au-delà de l'inconfort lié à l'apparence.

Table S5. Paired Samples *t*-tests Comparing Scores at Baseline and 24 Months

	n	baseline	24 Months	*p*-value	effect size
Appearance congruence	213	2.86 (0.74)	3.86 (0.76)	<0.001	-1.12
Depression	211	16.39 (11.88)	13.95 (12.76)	<0.001	0.20
Anxiety	208	60.25 (11.18)	57.38 12.00)	<0.001	0.25
Positive affect	215	42.90 (10.05)	43.72 (12.03)	0.37	-0.05
Life satisfaction	217	39.92 (10.55)	44.61 (12.29)	<0.001	-0.39
Gender dysphoria	?	?	?	?	?
Trauma symptoms	?	?	?	?	?
Suicidality	?	?	?	?	?
Self-injury	?	?	?	?	?
Body esteem	?	?	?	?	?
Quality of life	?	?	?	?	?

Note. Variables are presented as mean (SD). Results are based on *t*-tests (baseline minus 24-months). Negative *t*-test values indicate increases in appearance congruence, positive affect, and life satisfaction. Effect sizes are Cohen's *d* (ranges: 0.20, small; 0.50, medium; 0.80, large).

Et il ne s'agit là que **des variables que les chercheurs ont mentionnées dans leur hypothèse** ; ils ont jugé bon de choisir parmi toutes les **autres** variables, de sorte que nous pourrions vraiment faire une version beaucoup plus longue de ce tableau, avec beaucoup plus de points d'interrogation.

Si vous remplissez ces points d'interrogation, les chercheurs peuvent-ils encore présenter leur étude comme une découverte impressionnante ? La raison pour laquelle vous êtes censé communiquer de manière claire et transparente sur vos choix méthodologiques est d'empêcher les critiques de poser de telles questions en premier lieu. Le fait que tout cela ne soit pas clair devrait être considéré comme une lacune de la part des auteurs, de la revue, ou des deux.

Ou, pour le dire autrement : Visualisons les résultats qu'ils **ont** rapportés (encadrés en vert), parmi tous les résultats qu'ils **auraient pu** rapporter :

Gender Affirming Hormone Cohort Survey Measures	
Construct	**Measure**
Time of Completion: Baseline, 6-month, 12-month, 18-month, & 24-month follow-up periods	
Demographics	Demographic questions for Cross-Sex Hormone Cohort
Religiosity & Spirituality	Modified Duke University Religion Index (DUREL)
Socio-Economic Status	Socioeconomic Status Questions (for Adolescents & Young Adults)
Gender Identity	Transgender Congruence Scale **(Incomplete data -- one subscale)**
	DSM 5 – Chicago adapted
Service Utilization	Dr. Olson's Service Utilization Questions
Depression	BDI-II
Anxiety	Revised Children's Manifest Anxiety Scale: Second Edition (RCMAS-2 – What I Think and Feel)
Quality of Life	Health-Related Quality of Life Scale (modified HIV QOL)
Suicidality	Suicidal Ideation Scale
Body Esteem	Body Esteem Scale
Body Image	Body Image Scale
Social Relationships	Emotional Support / Friendship / Loneliness / Perceived Hostility / Perceived Rejection – NIH Toolbox
Negative Affect	Anger / Fear / Sadness – NIH Toolbox
Psychological Well-being	General Life Satisfaction / Positive Affect – NIH Toolbox
Self-Efficacy	Self-Efficacy (CAT 13-17)– NIH Toolbox
Perceived Parent Support	Parent Support Scale – Youth Version
Resiliency	Gender Minority Stress and Resilience Scale
	Connor-Davidson Resilience Scale
Sexual Behavior	Sexual Risk Behavior Questions
STI History	STI Questions
Alcohol/Drug Use	Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST)
Autism	Autism-Spectrum Quotient (AQ-10) – Adult
History of Blocker Experience	Questions to obtain history of participant's blocker experience

Le fait est que, quelle que soit la manière dont on le formule ou dont on le visualise, les chercheurs expliquent si peu leur processus, le chemin parcouru depuis leur protocole d'étude initial jusqu'au produit fini — un article publié dans l'une des meilleures revues de recherche au monde — que, tant qu'ils n'auront pas comblé certaines de ces lacunes, je ne pourrai m'empêcher d'être sceptique. Et je pense que vous devriez l'être aussi.

Dans leur protocole d'étude, y compris dans une version qu'ils ont soumise à une base de données de pré-enregistrement, les chercheurs ont émis l'hypothèse que les membres de cette cohorte connaîtraient une amélioration sur huit mesures, y compris celles qui sont presque universellement reconnues par les chercheurs sur le genre chez les jeunes comme des résultats importants, tels que la dysphorie de genre, la suicidalité et l'automutilation. Puis, dans l'article publié par le *NEJM,* les chercheurs ont modifié leur hypothèse et six de ces variables n'apparaissaient nulle part. Les deux variables restantes — l'anxiété et la dépression — ont évolué positivement chez les garçons transgenres (femmes natales), mais pas chez les filles transgenres (hommes natals). Les chercheurs ont également fait état de trois autres variables, sans expliquer comment ils les avaient choisies (deux se sont améliorées pour les filles et les garçons trans, et une uniquement pour les garçons trans).
À mon avis, cette question des variables manquantes remet en question tout l'effort, simplement parce que si de nombreuses variables suivies par les chercheurs ne se sont **pas** améliorées, voire ont empiré, le fait qu'ils aient pu sélectionner cinq variables qui ont montré une certaine amélioration pourrait ne rien signifier du tout. Il se peut que nous n'ayons affaire qu'à du **bruit statistique**, mais nous ne pouvons pas en être sûrs puisque les chercheurs dissimulent un grand nombre de leurs résultats.

Partie 2 - Cette nouvelle étude dont on fait grand cas, ne nous apprend pas grand-chose

Imaginons que les auteurs aient manifesté un intérêt préalable pour les cinq variables qu'ils ont rapportées depuis le début, et procédons en conséquence dans notre évaluation de leur étude. Pour les besoins de cet article, je ne vais pas non plus contester, ni même évaluer en profondeur, les techniques statistiques spécifiques employées par les auteurs : Comme je le montrerai, même si nous admettons qu'ils ont pris les bonnes décisions (ce qui n'est pas forcément le cas) et que nous prenons leurs conclusions pour argent comptant, les résultats restent au mieux ambigus.

Un petit éclaircissement politique avant de commencer : Si vous écrivez des articles critiques sur la médecine de genre pour les jeunes, vous entendrez beaucoup de gens qui sont horrifiés que vous puissiez le faire étant donné les menaces auxquelles sont confrontées les personnes transgenres aux États-Unis (adultes et enfants confondus). Et comme l'a noté Dave Weigel dans un article paru vendredi dans *Semafor*, Donald Trump vient de rendre publique une nouvelle proposition très folle visant à restreindre sévèrement les droits des transgenres pour les adultes et l'accès à la médecine de genre pour les enfants et les adolescents.

Au risque de me répéter, je suis opposé au type de politiques que Trump propose — à la fois des restrictions directes sur la médecine de genre pour les jeunes et sa proposition encore plus radicale de codifier dans la politique fédérale une interdiction pour les adultes de changer leur sexe légal. Cette dernière partie, en particulier, est carrément cruelle et inutile, si ce n'est qu'elle constitue de la viande rouge pour les électeurs évangéliques qu'il espère courtiser en 2024. […]

Mais comme Weigel l'a noté dans son article, il y a une différence notable entre ce que Trump propose et les débats plus substantiels et plus courants qui font actuellement rage sur la transition médicale des jeunes. En fin de compte, étant donné la popularité croissante de ces traitements et les déclarations parfois trop confiantes de leurs défenseurs, les questions entourant la médecine de genre pour les jeunes ont un besoin urgent de réponses, indépendamment de l'identité du président ou des menaces qui pèsent sur la communauté LGBT. L'argument "Ce n'est pas le moment d'en discuter" une "tactique de déraillement". Pour être honnête, j'ai rencontré cette tactique bien avant que les républicains ne s'emparent de cette question. Si nous attendons qu'il n'y ait plus de réactionnaires essayant de tirer profit de la peur des personnes transgenres avant de déterminer exactement si et dans quelle mesure la médecine du jeune genre fonctionne — des questions qui restent, selon toutes les normes préexistantes et largement acceptées de la preuve médicale, non résolues — nous continuerons à voler à l'aveuglette.

Et c'est tout ce que je vais dire à ce sujet — je suis un écrivain scientifique plus qu'un expert en opinions publiques- et si je pimente chaque paragraphe avec des parenthèses réitérant que j'ai les "bonnes" croyances sur les politiques Trumpistes, cela deviendra rapidement illisible. De plus, cela n'a pas vraiment d'importance : Ce qui suit est correct ou incorrect en soi, indépendamment des convictions de l'auteur.

Pour être clair, cette étude du *New England Journal of Medicine* représente une amélioration significative par rapport à ce qui est considéré comme de la recherche dans le domaine de la médecine du genre chez les jeunes (bien que ce ne soit pas une mince affaire). Il est excellent que les chercheurs suivent de près des cohortes d'enfants qui prennent des bloqueurs de puberté et des hormones, et qu'ils recueillent de nombreuses données sur leurs trajectoires de santé mentale et physique. Il est également utile que cette équipe ait préenregistré son protocole. Mais le fait est que cette étude particulière ne fournit pas de preuves substantielles que les hormones améliorent la santé mentale des enfants transgenres.

1. Les enfants de cette étude avaient un taux de suicide alarmant.

Bien que les auteurs aient eu d'autres problèmes de transparence dans leur étude, ils notent, dès le résumé, que deux participants sont morts par suicide. Dans le corps de l'article, ils écrivent que *"l'un [suicide] s'est produit après 6 mois de suivi et l'autre après 12 mois de suivi*". Ainsi, environ un an après avoir commencé à prendre des hormones, deux des 315 enfants de cette étude sont morts. Ils notent également qu'il y a eu 11 cas d'"*idées suicidaires pendant la visite de l'étude*".

Table 2. Adverse Events.

Event	No. of Events in Sample
Any event	15
Death by suicide	2
Suicidal ideation reported during study visit	11
Severe anxiety triggered by study visit	2

Laissons de côté la question de l'idéation, car les chercheurs ne nous ont pas fourni les informations nécessaires pour l'évaluer. L'"idéation suicidaire" peut signifier des choses très différentes, allant de pensées occasionnelles et fugaces de suicide à, beaucoup plus sérieusement, l'existence d'un plan et la possession de tous les outils nécessaires pour le mettre en œuvre. *"Je pense qu'il est juste de dire que leur utilisation de l'expression "idées suicidaires" est ambiguë*", a déclaré un chercheur sur le suicide que j'ai parfois sollicité sur cette question, mais à qui j'ai toujours garanti l'anonymat parce qu'il n'est absolument pas impliqué dans la lutte contre la médecine du genre chez les jeunes. Les chercheurs du *NEJM* ont bien administré une **échelle** d'idéation suicidaire afin d'obtenir plus de détails, mais comme nous l'avons vu dans la partie 1, ils n'ont tout simplement pas rapporté ces données. Il n'y a donc aucun moyen de savoir si les 11 incidents d'idées suicidaires signalés lors des visites avec les chercheurs sont élevés, faibles ou entre les deux.

En ce qui concerne le taux de suicides accomplis, une manière courante de mesurer et de comparer les taux de suicide et d'autres résultats de ce type est le nombre de suicides pour 100 000 personnes, par an. Aux États-Unis, ce chiffre est d'environ 13,9, bien qu'il puisse être recalculé chaque année et qu'il varie considérablement selon les sous-groupes. L'estimation annuelle la plus proche que nous puissions obtenir pour la population générale dans la tranche d'âge de l'étude (12-20 ans) est de 14,2 suicides pour 100 000 membres de la tranche d'âge 15-24 ans.

Comme me l'a fait remarquer Michael Biggs, professeur de sociologie à l'Université d'Oxford et critique fréquent de la recherche sur le genre chez les jeunes, ce chiffre était d'environ 317 décès par suicide pour 100 000 patients-années dans l'étude du *NEJM.* C'est un chiffre assez élevé. Il convient d'être prudent, car les taux pour 100 000 sont étranges lorsque le nombre brut d'événements est aussi faible : un suicide de moins aurait réduit le taux de moitié, et un suicide de plus l'aurait augmenté de 50 %. Pour ce que cela vaut, lorsque j'ai posé la question aux chercheurs spécialisés dans les suicides, ils ont répondu *: "Je dirais oui. Je suis d'accord pour dire que deux décès par suicide dans ce groupe d'âge pour cette taille d'échantillon est élevé par rapport à la population générale."*

Mais il serait incorrect de dire : "*Aha, les enfants de votre groupe avaient un taux élevé de suicide — votre traitement ne fonctionne pas"*. Il s'agissait d'enfants qui avaient déjà des problèmes de santé mentale ; la dysphorie de genre elle-même peut être très pénible et la population LGBT est connue, de manière plus générale, pour avoir des taux élevés de problèmes de santé mentale. Il ne faut donc probablement pas s'attendre à ce que les enfants participant à cette étude aient le même taux de suicide que leurs pairs de la population générale. Lorsque j'ai soulevé ce point dans un courriel adressé au chercheur chargé de l'étude sur le suicide, il m'a répondu qu'il était d'accord pour dire qu'"*une comparaison avec les taux de suicide dans un autre groupe ayant des problèmes de santé mentale serait probablement plus appropriée qu'un taux dans la population générale*".

Malheureusement, nous ne disposons pas de beaucoup de données à ce sujet. Dans la lettre Suicide d'adolescents transgenres référés par les cliniques au Royaume-Uni (janvier 2022, *Archives of Sexual Behavior*), Biggs avait précédemment calculé que le taux de suicide accompli à la clinique Tavistock en Angleterre était de 13 pour 100 000 patients-années, ce qui est bien inférieur à ce qui a été observé dans l'étude du *NEJM.* La seule autre comparaison convenable entre des pommes et des pommes dont nous disposons ici est également mentionnée dans sa lettre :

> Une seule étude publiée fait état de suicides mortels chez les adolescents transgenres. La clinique pédiatrique belge spécialisée dans les questions de genre a conseillé 177 jeunes âgés de 12 à 18 ans, qui lui avaient été adressés entre 2007 et 2016 : cinq d'entre eux (2,8 %) se sont suicidés (Van Cauwenberg et al., 2021). L'âge moyen de référence était de 15 ans, ce qui implique une durée moyenne de 3 ans avant la transition vers une clinique pour adultes, ce qui se traduit par un taux de suicide annuel de 942 pour 100 000. Il s'agit du taux de mortalité par suicide le plus élevé jamais enregistré pour une population transgenre.

L'échantillon du *NEJM* n'avait donc pas le taux de suicide élevé de la cohorte belge, mais il est indéniable qu'il est élevé.

Cela ne signifie pas que les suicides de la cohorte du *NEJM* **ont été causés** par les hormones. *"Bien sûr, nous ne pouvons pas attribuer ces suicides aux hormones du sexe opposé, car nous n'avons pas de groupe de contrôle"*, a déclaré Biggs dans un courriel. "*De même, nous ne pouvons pas attribuer l'amélioration aux hormones du sexe opposé !*" Nous reviendrons sur ce dernier point, mais il n'est pas nécessaire d'avancer un argument de cause à effet pour s'inquiéter. L'une des justifications les plus courantes pour expliquer pourquoi la médecine du genre chez les jeunes en vaut la peine, malgré la myriade d'inconnues qui subsistent, est que les enfants se tueront s'ils ne suivent pas ce traitement. Eh bien, voici un échantillon d'enfants qui y ont eu accès dans des environnements supposés de haute qualité, avec beaucoup de soutien et de suivi, et qui ont tout de même eu un taux de suicide très élevé. Comment cela peut-il ne pas soulever de questions ? Les chercheurs n'ont rien à dire à ce sujet, si ce n'est qu'ils notent le nombre de suicides accomplis et les cas d'"*idéation*".

Le problème devient encore plus inquiétant lorsque l'on examine le protocole de l'étude que nous avons passé tant de temps à étudier dans la 1re partie et que l'on constate que **les enfants souffrant de graves problèmes psychiatriques, y compris de suicidalité, ont été exclus de l'étude dès le départ** : "*La présence de symptômes psychiatriques graves (par ex, hallucinations actives, troubles de la pensée) qui nuiraient à la capacité de l'individu à fournir un véritable consentement éclairé ou à participer à l'ACASI [auto-entretien assisté par ordinateur] de base*" était un critère d'exclusion, tout comme le fait d'être "*visiblement désemparé (par exemple, suicidaire, homicide, comportement violent) au moment du consentement ou de l'ACASI de base […]*".

Les enfants pouvaient donc présenter un certain degré de suicidalité et participer à l'étude — les chercheurs **ne** considèrent pas la suicidalité comme une variable binaire — mais les enfants **très** suicidaires ou **très** mal en point ont été exclus, ce qui signifie que nous ne nous attendions pas à ce que cette cohorte soit particulièrement suicidaire à l'arrivée. Et pourtant, il y a eu deux suicides. Ce n'est pas bon, et cela devrait être considéré comme un signal d'alarme qui mérite une explication. Nous n'en avons pas, car les chercheurs **ne font pas état du niveau global de suicidalité** de la cohorte au fil du temps, **bien que cela fasse partie de leur hypothèse de base**. *Le New England Journal of Medicine*, qui publiait ce qu'il savait être une étude susceptible d'attirer l'attention sur un sujet très controversé, constamment associé au suicide dans les conversations publiques, ne leur a pas demandé de le faire.

2. La plupart des améliorations constatées dans la cohorte ont été modestes.

Telles sont les améliorations observées au fil du temps, toutes statistiquement significatives, dans les variables rapportées par les chercheurs, selon leur modèle statistique. Rappelons que l'étude porte sur une période de deux ans :

- Congruence de l'apparence : augmentation de 0,96 sur 5 points
- Affect positif : augmentation de 1,6 sur 100 points
- Satisfaction à l'égard de la vie : augmentation de 4,64 points sur 100

- Dépression : diminution de 2,54 points sur 63
- Anxiété : diminution de 2,92 points sur 100

Ces chiffres correspondent aux changements moyens pour l'ensemble du groupe.
Les changements statistiquement significatifs ont été observés pour les deux sexes pour les variables en rose. Pour celles en orange : les garçons trans, mais pas les filles, ont vu des avantages.

La congruence de l'apparence et l'affect positif sont les deux seules variables pour lesquelles les chercheurs ont pu constater des augmentations salutaires chez les deux sexes au cours des deux années de l'étude. (La congruence de l'apparence fera bientôt l'objet d'une section distincte, je vais donc l'ignorer ici).
Compte tenu de ces différences, il aurait été utile que les chercheurs indiquent clairement quels étaient les changements moyens pour les garçons transgenres par rapport aux filles transgenres, qui, après tout, ont reçu des hormones totalement différentes. Ils ne le font pas (pas plus qu'ils n'expliquent pourquoi les deux tiers de leur échantillon étaient des filles natales). Si vous savez lire le tableau 3 [page 7 de l'étude], vous pouvez en quelque sorte inverser la logique de certaines de ces informations, mais elles devraient vraiment être plus claires.
La première question à se poser dans une situation comme celle-ci, où l'on constate des améliorations statistiquement significatives mais de faible ampleur, est de savoir si elles ont de l'importance. Un effet statistiquement significatif peut ne pas être *cliniquement* significatif, ce qui signifie qu'il ne représenterait pas une amélioration ou une aggravation notable de l'état mesuré. Les statisticiens discutent sans cesse de taille d'effet, mais il n'est souvent pas facile de déterminer si un petit effet est **trop** petit. Je pense que parfois, pour les effets très faibles, il est bon de faire appel au bon sens.
L'affect positif, mesuré par un instrument de la NIH Toolbox, a augmenté de 1,6 point en deux ans sur une échelle de 100 points. Si un chercheur vantant les mérites d'un traitement pour votre enfant vous dit que celui-ci améliorera son score sur cette échelle de 1,6 % en deux ans, vous avez le droit d'être sceptique. Je ne crois vraiment pas qu'il faille s'en préoccuper ou y voir une preuve que les hormones aident les enfants (j'ai vérifié si ces documents sur la boîte à outils des NIH contenaient des informations sur l'interprétation des changements dans les scores d'affect positif ou de satisfaction de la vie, mais il ne semble pas que ce soit le cas).
Qu'en est-il des autres résultats qui paraissent faibles, mais qui ne sont pas si faibles que cela ? Devons-nous nous en préoccuper ?
Tout ce que nous pouvons faire, c'est essayer de chercher dans la littérature et de trouver d'autres comparaisons. L'un des documents les plus intéressants que j'ai trouvés sur la question de la signification clinique par rapport à la signification statistique est une méta-analyse de la base de données Cochrane de revues systématiques sur les "antidépresseurs de nouvelle génération pour la dépression chez les enfants et les adolescents" (Cochrane est considéré comme l'un des meilleurs pour ce type d'évaluation minutieuse de la recherche). Les auteurs y résument les études qui ont comparé ces antidépresseurs à un placebo, et qui ont utilisé le Children's Depression Rating Scale-Revised (CDRS-R) pour évaluer les symptômes. L'échelle CDRS-R va de 17 à 113, ce qui

signifie qu'elle comporte 97 points. Les auteurs décrivent des différences aussi élevées que 3,51 comme "*petites et sans importance*". Encore une fois, une différence qui ne semble pas minime en soi peut ne pas avoir d'importance sur le plan clinique.

Plus directement applicable à la présente discussion, voici un article publié en 2015 par des chercheurs examinant le concept de différence clinique minimale importante, ou "*la plus petite différence de score considérée comme cliniquement valable par le patient : MICD (Minimal Clinically Important Difference)*", en ce qui concerne le Beck Depression Inventory 2 (BDI-II), qui est l'outil utilisé par les chercheurs du *NEJM* pour mesurer la dépression dans leur étude. Les auteurs de l'étude de 2015 *"ont estimé une MCID de 17,5 % de réduction des scores par rapport à la ligne de base.… L'estimation correspondante pour les personnes souffrant de dépression depuis plus longtemps et n'ayant pas répondu aux antidépresseurs était plus élevée, à savoir 32 %*." Dans le modèle statistique du chercheur du *NEJM*, les patients avaient un score de base moyen de 15,46 et une réduction moyenne de 2,54. Étant donné que 17,5 % de 15,46 correspondent à environ 2,51, si l'on se fie à ces estimations, l'enfant moyen de l'étude du *NEJM a* tout juste remarqué une amélioration de ses symptômes dépressifs au cours des deux années où il a pris des hormones. (Je n'ai pu trouver aucune recherche sur ce qui constitue une amélioration cliniquement significative sur l'instrument d'anxiété utilisé par les chercheurs, le Revised Children's Manifest Anxiety Scale, pour lequel les chercheurs ont observé une amélioration moyenne de 2,92 points sur 100).

Bien entendu, l'enfant moyen ne présentait que des symptômes dépressifs légers au départ, ce qui complique l'analyse. D'autres patients avaient des scores BDI-II plus élevés au départ, ce qui signifie au moins qu'ils ne considéreraient les réductions comme intéressantes que si elles étaient significativement plus importantes.

Il s'agit d'un tableau utile tiré de l'annexe supplémentaire :

Table S6. Proportions of Youth Scoring in the Clinical Range for Depression and Anxiety at Each Timepoint

	Baseline	**6-month**	**12-month**	**18-month**	**24-month**
Beck Depression Inventory-II n (%)	*n*=307	*n*=281	*n*=248	*n*=210	*n*=219
Minimal Depression	149 (48.5)	152 (54.1)	143 (57.7)	125 (59.5)	126 (57.5)
Mild Depression	53 (17.3)	46 (16.4)	41 (16.5)	25 (11.9)	41 (18.7)
Moderate Depression	57 (18.6)	43 (15.3)	24 (9.7)	30 (14.3)	22 (10)
Severe Depression	48 (15.6)	40 (14.2)	40 (16.1)	30 (14.3)	30 (13.7)
Revised Children's Manifest Anxiety Scale 2	*n*=308	*n*=282	*n*=248	*n*=209	*n*=216
M (*SD*)	60.0 (11.5)	58.6 (11.6)	58.6 (11.3)	56.8 (11.4)	57.4 (12.1)
n (%) in Clinical range (*T*>60)	181 (58.8)	145 (51.4)	115 (46.4)	90 (43.1)	103 (47.7)

Note. % calculated as valid percent using the n for each timepoint as the denominator.

Je ne sais pas ce qu'il faut en penser. Les chercheurs soulignent le fait que de nombreux enfants sont passés à des niveaux inférieurs de dépression et d'anxiété au cours de l'étude, mais ils reconnaissent également qu'un bon nombre d'entre eux sont restés dans la fourchette clinique au moment du suivi.

C'est tout à fait vrai. Il s'agit d'une situation véritablement compliquée à interpréter, en partie parce que les chiffres de base sont très hétérogènes : Il n'est pas juste d'insister sur l'absence d'amélioration chez un enfant qui n'allait pas si mal au départ. C'est un problème qui se pose dans certaines recherches cliniques sur la médecine du genre chez les jeunes, qui impliquent généralement des cohortes ayant fait l'objet d'un dépistage préalable de graves problèmes de santé mentale : Ils n'ont pas beaucoup de marge de manœuvre pour s'améliorer, de sorte que les statistiques sont quelque peu défavorables aux chercheurs qui cherchent à démontrer des améliorations en matière de santé mentale. (D'un autre côté, il est utile de savoir que la situation des enfants de cette cohorte ne s'est pas détériorée de manière significative, pour la plupart d'entre eux).

Il aurait été utile que les chercheurs fournissent des informations plus précises sur, par exemple, l'amélioration numérique moyenne parmi les enfants se situant dans les catégories modérée ou sévère de la mesure de la dépression. Comme nous l'avons vu précédemment, le fait d'indiquer plutôt le pourcentage de participants dans différentes catégories cliniques peut masquer de nombreuses informations utiles : Dans les cas les plus extrêmes, une baisse d'un point que le patient ne remarque même pas peut le faire passer de la catégorie "modérée" à la catégorie "légère". **Il serait également utile de savoir si la proportion relativement importante d'enfants qui n'ont pas fourni de données à 24 mois, soit environ un tiers d'entre eux, différait à d'autres moments du reste du groupe, car si les enfants dont les données à 24 mois sont manquantes se portaient en moyenne mieux ou moins bien que leurs pairs dans l'étude, cela pourrait sérieusement fausser les résultats**. (J'ai peut-être oublié quelque chose, mais je pense que les chercheurs ne font cette comparaison que pour le très petit nombre de vrais abandons, plutôt que pour les enfants qui sont techniquement restés dans l'étude mais qui n'ont pas fourni de données sur certains points lors de l'observation finale).

En fin de compte, il semble difficile de nier que beaucoup d'enfants sont restés avec un mal-être malgré deux années d'accès régulier à une clinique spécialisée dans les questions de genre et à un médicament conçu pour améliorer leur état de santé mentale :

- **au départ, 18,6 %** des enfants souffraient de **dépression modérée,** et deux ans plus tard**, 10 %** d'entre eux en souffraient encore**.** En ce qui concerne la **dépression sévère**, l'amélioration est encore moindre : **15,6 % à 13,7 %** ;
- même constat pour l'**anxiété : 58,8 %** des enfants avaient un score clinique au départ et **47,7 %** l'avaient au moment du suivi.

D'un autre côté, il est difficile d'ignorer que certains enfants **ont connu** des **réductions significatives de la dépression et/ou de l'anxiété** et/ou d'autres symptômes. Cela ne constitue-t-il pas au moins une preuve de l'efficacité des hormones ?

Malheureusement…

3. Il est impossible d'attribuer les améliorations observées dans cette étude aux hormones plutôt qu'à d'autres formes de traitement pratiquées dans ces cliniques.

Il s'agit d'une étude longitudinale **sans groupe de comparaison**. Comme n'importe quel diplômé d'un cours de statistiques universitaire vous le dira, il est beaucoup plus difficile de prétendre qu'une influence particulière est responsable des changements observés au fil du temps.

Si vous suivez **deux** groupes par ailleurs similaires pendant deux ans, que vous donnez à l'un d'eux un médicament et à l'autre un placebo, et que vous observez des différences entre les groupes, vous pourriez alors commencer à faire des déductions causales raisonnablement sûres sur les effets du médicament, bien que le **degré de confiance** dépende d'un grand nombre de facteurs. Mais avec un seul groupe, il s'agit d'un problème statistique notoire, même dans des situations relativement simples. Si je vous donne un médicament contre la grippe et que vos symptômes se sont considérablement améliorés 5 jours plus tard, cela signifie-t-il que le médicament a agi ? Peut-être. Mais les personnes ont aussi tendance à aller mieux avec le temps. Les scores sur certaines échelles ont tendance à revenir à la moyenne si la première observation est très élevée ou très basse. Et ainsi de suite. Encore une fois, il s'agit de principes statistiques de base — il ne s'agit pas de pinaillage.

Cette étude du *NEJM* est toutefois beaucoup plus complexe qu'une étude sur la grippe de deux semaines. Et nous avons de bonnes raisons de penser que d'autres facteurs ont pu contribuer aux améliorations (pour la plupart minimes) observées par les chercheurs.

Comme ils le notent au début de leur article, "*toutes les cliniques participantes ont une équipe multidisciplinaire qui comprend des prestataires de soins médicaux et de santé mentale et qui détermine en collaboration s'il y a dysphorie de genre et si des soins médicaux d'affirmation du genre sont appropriés. Pour les mineurs, le consentement des parents est nécessaire pour entamer un traitement médical. Les publications des équipes d'étude individuelles fournissent des détails sur les approches de soins spécifiques à chaque site".*

Cette dernière phrase cite quatre articles, et si vous les lisez, vous constaterez qu'il est fait mention de **thérapies et de médicaments** pour les patients dont la santé mentale est mise à rude épreuve, au-delà de leurs problèmes de genre. Pour prendre un exemple, l'équipe du programme d'identité de genre de l'hôpital pour enfants Lurie de Chicago écrit que "*dans les cas où une thérapie fondée sur des preuves et une psychopharmacothérapie sont indiquées, un psychologue et un psychiatre peuvent constituer l'équipe de traitement du patient — le psychologue servant de thérapeute principal et le psychiatre assurant la gestion des médicaments psychotropes".*

Tous les enfants participant à cette étude ont été examinés dans l'une de ces cliniques multidisciplinaires. Il est donc logique que les enfants souffrant de graves problèmes d'anxiété et de dépression aient eu accès à une psychothérapie, à des médicaments ou aux deux. Il est également logique que plus les symptômes de santé mentale d'un enfant étaient graves au départ, plus il était probable qu'il bénéficie de l'une de ces interventions, **et plus il** avait de chances de s'améliorer. Cela

signifie que même lorsqu'il s'agit du sous-ensemble d'enfants dont les améliorations sont notables, nous devons nous poser la question :

Leurs symptômes ont-ils diminué à cause des hormones, des médicaments ou de la thérapie ? Ou s'agit-il d'une combinaison des trois ?

Il n'y a aucun moyen de le savoir. Il est donc impossible d'interpréter pleinement ces résultats. C'est regrettable, car je pense que les chercheurs auraient pu tenir compte de ce facteur dans leurs modèles statistiques, étant donné qu'ils avaient certainement accès aux dossiers des patients. L'une des questions que je leur ai envoyées portait précisément sur ce sujet, mais comme je l'ai mentionné dans mon dernier article, ils n'accordent aucune interview et ne répondent pas aux questions. Malgré tout, les chercheurs proclament avec assurance dans leur résumé que *"[les hormones d'affirmation du genre] améliorent la congruence de l'apparence et le fonctionnement psychosocial".* Il s'agit là d'un langage causal direct, mais leur méthodologie est loin de le justifier. De plus, ils ne **mentionnent même pas cette confusion potentielle**, qui, une fois encore, est vraiment le genre de chose que l'on apprend au cours de la première année d'un cours d'introduction aux statistiques au niveau de l'université.

Imaginez que vous soyez confronté à cette question lors de l'examen de mi-parcours d'un tel cours : "Un groupe d'enfants souffrant de problèmes de santé mentale a accès à des conseils psychologiques, à des médicaments psychotropes et au traitement X pendant deux ans. À la fin de cette période, leur état s'est considérablement amélioré."
• Est-ce dû au suivi psychologique ?
• Est-ce dû au médicament psychotrope ?
• Est-ce dû au traitement X ?
☑ Sans plus d'informations, il est impossible de savoir.

100 % des professeurs de statistiques vous diront que la bonne réponse est la dernière proposition (Un commentaire d'Annelou de Vries et Sabine Hannema accompagnant l'étude mentionne la question de la thérapie, mais pas celle des médicaments).

Il existe un autre problème potentiel lorsqu'il s'agit de mettre les améliorations constatées dans cette étude sur le compte des hormones d'affirmation du genre, en tant que telles. Ce problème m'a été signalé par un chercheur spécialisé dans la dépression avec lequel j'ai échangé quelques courriels. Il a souligné que les chercheurs *"ignorent totalement le fait évident que la testostérone a des effets importants sur l'humeur, l'anxiété et les antidépresseurs".* J'ai entendu d'autres personnes soulever ce point dans le passé, et il pose un autre véritable défi aux chercheurs en médecine du genre pour les jeunes. Il existe en effet **des preuves** [Association entre le traitement à la testostérone et l'atténuation des symptômes dépressifs chez les hommes ; revue systématique et méta-analyse, Walter et al. 2019] **que la testostérone a des effets bénéfiques sur l'humeur**, ce qui soulève la possibilité qu'elle puisse entraîner des améliorations des symptômes de santé mentale qui n'ont rien à voir avec le traitement de la dysphorie de genre en tant que telle. (Je tiens à préciser que les études auxquelles je fais référence et que je cite en lien concernent généralement des hommes natals plutôt que des femmes natales, des adultes plutôt que des jeunes, et couvrent parfois des périodes nettement plus courtes que deux ans. Nous ne pouvons donc pas affirmer avec certitude que la testostérone a les

mêmes effets sur les jeunes femmes natales et sur les hommes natals plus âgés, mais c'est certainement une possibilité, et il existe des preuves anecdotiques de ce phénomène chez les garçons et les hommes transgenres).

Soyons encore une fois extrêmement généreux et laissons de côté cette possibilité. Pour résumer le reste de cette section, même si nous zoomons **uniquement sur** les enfants qui se sont beaucoup mieux portés dans cette étude, en ignorant les petites tailles d'effet moyennes, les suicides, les résultats vraiment décevants pour les filles trans, et ce qui semble être un véritable carnaval de sélection de variables, **nous n'avons tout simplement aucun moyen de savoir — point final — si leur état s'est amélioré grâce aux hormones, à la thérapie, aux médicaments, ou à une combinaison des trois.**

Je ne suis pas doué pour les statistiques, mais je n'ai pas besoin de l'être pour faire valoir mon point de vue. Ce n'est pas ésotérique. Il s'agit d'une question fondamentale, bien connue des sciences sociales. Il compromet sérieusement notre capacité à déterminer si les enfants de cette étude ont bénéficié de la prise d'hormones, et il donne une mauvaise image de la décision du *New England Journal of Medicine* d'autoriser les chercheurs à utiliser un langage causal aussi fort et direct pour décrire leurs résultats.

4. La seule amélioration importante concerne une variable qui n'a peut-être pas beaucoup d'importance.

Jusqu'à présent, on peut dire qu'il s'agit de *peanuts*, en ce qui concerne les tailles d'effet moyennes. Les auteurs mettent toutefois en évidence un résultat plus impressionnant :

> L'amélioration de la congruence de l'apparence est l'un des principaux objectifs des hormones d'affirmation de genre (GAH), et nous avons observé une amélioration de la congruence de l'apparence au cours des deux années de traitement. Il s'agit d'un effet modéré et de l'effet le plus important observé sur l'ensemble de nos résultats, ce qui est cohérent avec l'effet observé dans les recherches portant sur d'autres échantillons, qui ont noté des effets importants des GAH sur l'image corporelle et des effets faibles à modérés sur la santé mentale. La congruence de l'apparence a également été associée à chaque résultat psychosocial évalué au départ et pendant la période de suivi, de sorte que l'augmentation de la congruence de l'apparence a été associée à une diminution des symptômes de dépression et d'anxiété et à une augmentation de l'affect positif et de la satisfaction à l'égard de la vie. Ces résultats suggèrent que la congruence de l'apparence est un mécanisme par lequel [les hormones d'affirmation de genre] influencent le fonctionnement psychosocial.

Plus précisément, sur une période de deux ans, les enfants participant à l'étude ont connu une amélioration **d'environ 1 point sur l'échelle de congruence d'apparence** sur les 5 points de l'échelle de congruence transgenre (Transgender Congruence Scale, TCS).

La TCS est composée de 10 ou 12 questions comprenant les éléments suivants, résumés dans un tableau tiré d'une étude de 2021 [Évaluation psychométrique de l'échelle de congruence transgenre] parue dans *Sexuality Research and Social Policy*, qui était la plus récente que j'ai pu trouver sur ses propriétés psychométriques :

Table 3 Factor loadings for original TCS and TCS-10 reduced models

Items by original factors	Item #	Original λ	TCS-10 λ	Factor in TCS-10 model/reason removed
Appearance Congruence (Kozee et al., 2012)				
My outward appearance represented my gender identity	1	.493	.481	Appearance Congruence
I experienced a sense of unity between my gender identity and my body	2	.498	.447	Appearance Congruence
My physical appearance adequately expressed my gender identity	3	.561	.604	Appearance Congruence
I was generally comfortable with how others perceived my gender identity when they look at me	4	.624	.643	Appearance Congruence
My physical body represented my gender identity	5	.559	–	Removed due to high covariance
The way my body currently looks did not represent my gender identity (Reversed)	6	.276	.269	Appearance Congruence
I was happy with the way my appearance expressed my gender identity	7	.457	.491	Appearance Congruence
I did not feel that my appearance reflects my gender identity (Reversed)	8	.280	.316	Appearance Congruence
I felt that my mind and body were consistent with one another	9	.436	.398	Appearance Congruence
Gender Identity Acceptance (Kozee et al., 2012)				
I was not proud of my gender identity (Reversed)	10	.391	–	Removed due to high covariance
I was happy that I have the gender identity that I do	11	.555	.494	Gender Identity Acceptance
I had accepted my gender identity	12	.537	.561	Gender Identity Acceptance

L'automne dernier, j'ai écrit un article sur le problème des chercheurs dans ce domaine qui vantent des résultats apparemment impressionnants en s'appuyant sur des instruments qui ne signifient pas grand-chose. Dans ce cas, je parlais d'une recherche sur la "chirurgie du haut" chez les adolescents, dont la principale conclusion était que les scores des enfants s'étaient "*améliorés*", après avoir subi une double mastectomie, sur des échelles qui semblaient consister principalement en des questions leur demandant s'ils avaient actuellement des seins. En d'autres termes, si l'on demande à une personne ayant une dysphorie de genre d'indiquer si elle est d'accord avec l'affirmation *"Je m'inquiète que les gens regardent ma poitrine"* — un élément réel de l'échelle en question — avant et après l'intervention chirurgicale, il serait choquant **que son score ne s'améliore pas**. Mais cela ne nous dit évidemment pas grand-chose sur le fait que les doubles mastectomies *"fonctionnent"* à long terme, dans le sens où nous voudrions qu'une intervention chirurgicale majeure fonctionne. L'article en question n'incluait pas d'éléments plus substantiels sur la santé mentale des enfants.

Je pense que c'est un peu ce qui se passe ici. Pour être juste, les auteurs de l'article du *NEJM* font également état de mesures plus courantes et mieux validées, y compris celles portant sur l'anxiété et la dépression, mais comme nous l'avons vu, ces changements étaient faibles, d'une importance clinique discutable, et ne s'appliquaient pas aux personnes en transition homme vers femme.

L'amélioration de la congruence de l'apparence a été ressentie par les deux sexes, et c'est le **seul** effet de taille décente que les chercheurs ont découvert, comme ils le notent eux-mêmes.

Mais on peut affirmer que le jeu est quelque peu truqué ici. Prenons des questions comme : *"Mon apparence extérieure représentait mon identité de genre"* et "*Mon apparence physique exprimait*

adéquatement mon identité de genre". D'une part, ces éléments sont tellement similaires que je suis surpris qu'ils figurent tous les deux sur l'échelle — ils semblent vraiment redondants, et il semble presque impossible que si l'un augmente ou diminue, l'autre ne suive pas directement, ce qui pourrait gonfler artificiellement les changements observés dans les scores des personnes interrogées. (En théorie, lorsqu'une échelle est validée pour la première fois, quelqu'un vérifie ce genre de problèmes, mais je ne prétendrai pas avoir examiné ce processus en profondeur ici. Il convient toutefois de noter que dans l'article de 2021, les auteurs indiquent qu'un élément de la sous-échelle de congruence de l'apparence **a été** supprimé en raison de sa forte covariance avec d'autres éléments).

Mais surtout, il semble presque impossible d'imaginer que les résultats d'une personne sur un point comme celui-ci ne se soient **pas "améliorés"** à mesure que les changements physiques des hormones s'installaient et que leur corps correspondait à ce qu'ils ressentaient en eux-mêmes. Je veux dire que je ne veux pas ignorer complètement ce résultat — ce serait certainement une mauvaise nouvelle s'il n'y avait pas d'amélioration, parce que cela pourrait suggérer que leur identité de genre ou leurs objectifs de transition ont changé au milieu du traitement (ce qui ne serait pas idéal) — mais je ne suis pas sûr que nous devrions être impressionnés par ce résultat.

C'est d'autant plus vrai que **l'échelle de congruence de l'apparence ne semble pas vraiment corrélée à d'autres mesures plus solides du bien-être**. C'est du moins ce qu'ont constaté les auteurs de l'article de 2021 :

Table 4 Correlations between *TCS-10 Total* and subscales with gender and well-being constructs including means, standard deviations, and range

	Total	*AC*	*GIA*	*M*	*SD*	*Range*
TCS-10 Total	–	.96***	.50***	3.24	.60	1.70–5.00
- Appearance Congruence (AC)	.96***	–	.22**	3.18	.67	1.00–5.00
- Gender Identity Acceptance (GIA)	.50***	.22**	–	2.34	.61	1.00–3.33
TC[3]	.58***	.54***	.33***	56.28	9.18	28.00–89.00
GPSQ	−.23**	−.18*	−.26***	41.83	8.20	15.00–62.00
GMSR						
- Discrimination	−.19**	−.11	−.29***	3.63	1.69	0.00–5.00
- Rejection	−.12	−.06	−.21**	4.15	2.06	0.00–6.00
- Victimization	−.04	.05	−.26***	3.99	2.26	0.00–6.00
- Non-affirmation	−.25***	−.28***	.00	14.14	4.69	0.00–24.00
- Internalized transphobia	−.18**	−.09	−.34***	16.77	6.75	0.00–30.00
- Pride	.36***	.26***	.44***	18.66	5.84	4.00–32.00
- Negative expectations	−.02	−.03	.02	20.25	6.52	0.00–36.00
- Nondisclosure of gender identity	.09	.13	−.07	10.98	4.06	0.00–20.00
PHQ-9	−.11	−.05	−.22**	13.01	5.81	0.00–27.00
GAD-7	−.02	.02	−.11	10.74	4.60	0.00–21.00
PANAS						
- Positive	.31***	.26***	.25***	28.90	6.23	11.00–45.00
- Negative	−.03	.03	−.19**	26.17	7.42	9.00–40.00
SWLS	.43***	.43***	.17*	21.59	5.77	6.00–35.00

Note: *TCS*, Transgender Congruence Scale; *TC*[3], Trans Collaborations Clinical Check-in; *GPSQ*, Gender Preoccupation and Stability Questionnaire; *GMSR*, Gender Minority Stress and Resilience; *PHQ-9*, Patient Health Questionnaire (9-item); *GAD-7*, Generalized Anxiety Disorder (7-item); *PANAS*, Positive and Negative Affect Scale; *SWLS*, Satisfaction with Life Scale. *N*'s range from 202 to 208 due to missing data

***$p < .001$, **$p < .01$, *$p < .05$

Ne vous inquiétez pas si vous ne pouvez pas l'interpréter. Le fait est que cette mesure ne présente qu'une corrélation incohérente avec d'autres mesures mieux établies. Sans entrer dans les détails, les auteurs notent que cette constatation est en contradiction avec des recherches antérieures sur le TCS

et ses sous-échelles, qui ont révélé **une** corrélation assez forte entre le TCS et d'autres éléments. Mais la question de savoir si les changements dans le TCS ou ses sous-échelles ont une grande importance clinique reste ouverte — une question qui mérite d'être approfondie, mais qui n'a pas encore trouvé de réponse. Si c'est la plus grande amélioration que vous constatez dans votre grande étude du *NEJM*, vous devriez vous poser quelques questions.

Les chercheurs affirment qu'ils ont effectué d'autres travaux statistiques pour étayer l'idée que la congruence de l'apparence pourrait être particulièrement importante. Ils incluent ce visuel sophistiqué :

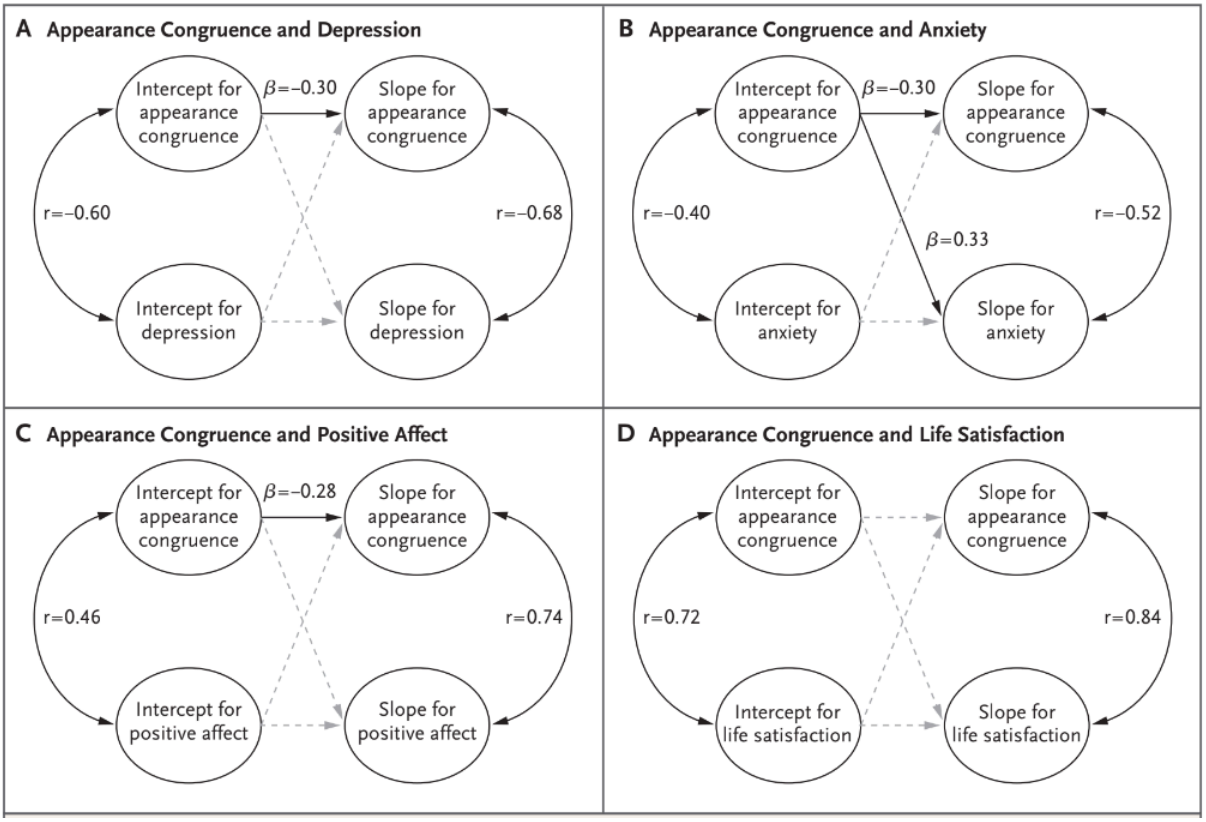


Figure 1. Congruence de l'apparence et dépression, anxiété, affect positif et satisfaction de la vie.
Les modèles de courbe de croissance latente à processus parallèle sont représentés. Un modèle de courbe de croissance latente linéaire a été ajusté pour chaque résultat, avec des estimations basées sur le modèle des scores de base (*intercept*) et des taux de changement linéaire au fil du temps (*pente*). Les modèles de processus parallèles permettent de tester la manière dont les aspects des trajectoires sont liés les uns aux autres. Chaque partie de la figure fournit des estimations des corrélations entre les scores de base de la congruence de l'apparence et chaque résultat (*corrélations d'interception*, arcs affichés sur le côté gauche de chaque partie), des corrélations entre le taux de changement de la congruence de l'apparence et le taux de changement de chaque résultat (*corrélations de pente*, arcs affichés sur le côté droit de chaque panneau), et des effets des scores de base sur les pentes (lignes droites au milieu de chaque panneau). Les lignes noires pleines et les arcs indiquent des *effets significatifs* (les intervalles de confiance pour les estimations des variables ne contiennent pas 0) ; les *effets non significatifs* sont indiqués par des lignes grises en pointillé. Tous les modèles ont été contrôlés pour l'âge, le sexe désigné à la naissance, l'identité raciale et ethnique, et les soins précoces d'affirmation du genre (non illustrés pour faciliter l'interprétation).

L'un de mes interlocuteurs privilégiés pour les questions de statistiques qui me dépassent est Stuart Ritchie, psychologue chercheur et auteur de l'excellent livre *Science Fictions : Comment la fraude, les préjugés, la négligence et le battage médiatique minent la recherche de la vérité* ; il tient également un compte Substack, et il est rédacteur scientifique au journal britannique *i*.

Je lui ai envoyé ce tableau et les affirmations des auteurs concernant leur modélisation de la courbe de croissance latente et ce qu'elle montrait sur l'importance de la congruence de l'apparence. Par chance, Ritchie avait publié des recherches utilisant des techniques statistiques similaires.
"Pour moi, il est logique que la congruence de l'apparence change en même temps que les autres variables de santé mentale, mais cela ne dit rien sur la question de savoir si le changement de congruence de l'apparence changerait la santé mentale dans un sens causal", a écrit Ritchie dans un courriel. "*La causalité aurait très bien pu aller dans l'autre sens (les personnes qui se sentent mieux, mentalement parlant, ont tendance à moins se préoccuper de leur apparence).* Ritchie résume cette partie de l'article du NEJM comme *"une façon fantaisiste de jouer avec les corrélations, rien de causal, plutôt intéressant mais pas un argument décisif".*
Le chercheur anonyme spécialisé dans la dépression a fait exactement la même remarque, de manière indépendante, dans un courriel. *"Les modèles de courbe de croissance latente ne sont pas très convaincants"*, écrit-il. *"Bien sûr, ils suggèrent que les changements dans la congruence de l'apparence sont corrélés avec les changements dans d'autres variables psychologiques, mais il n'y a aucune preuve de* ***causalité****, et nous savons certainement que les mesures d'apparence auto-déclarées dépendent fortement de l'état d'esprit — de sorte qu'une augmentation de la dépression, par exemple, entraînerait presque certainement une diminution de la congruence de l'apparence".*
Ritchie et le chercheur sur la dépression ont également tous deux **noté indépendamment le nombre élevé de liens statistiquement non significatifs** (les lignes en pointillé) dans le diagramme ci-dessus.

5. Les chercheurs n'envisagent même pas la possibilité que ces traitements soient moins efficaces qu'ils ne le pensaient — leur seule réponse est : "plus d'hormones".

Dans une partie subtilement révélatrice de cet article, les chercheurs tentent d'expliquer que les hommes natals n'ont pratiquement pas montré d'amélioration au cours de leurs deux premières années de traitement aux hormones :

> Étant donné que certains changements phénotypiques clés médiés par les œstrogènes peuvent prendre entre 2 et 5 ans pour atteindre leur effet maximal (par exemple, la croissance des seins), nous supposons qu'une période de suivi plus longue pourrait être nécessaire pour observer un effet sur la dépression, l'anxiété et la satisfaction de la vie. En outre, les changements associés à une puberté médiée par la testostérone endogène (par exemple, une voix plus grave) peuvent être plus prononcés et observables que ceux associés à une puberté médiée par les œstrogènes endogènes. Nous émettons donc l'hypothèse que les différences observées en matière de dépression, d'anxiété et de satisfaction à l'égard de la vie chez les jeunes désignés comme femmes à la naissance par rapport à ceux désignés comme hommes à la naissance peuvent être liées à des expériences différentielles de stress lié à l'appartenance à une minorité de genre, qui pourraient résulter de

> différences dans l'acceptation sociétale des femmes trans par rapport aux hommes trans. En effet, le stress lié à la minorité de genre est systématiquement associé à des résultats plus négatifs en matière de santé mentale, et la recherche suggère que les jeunes femmes trans peuvent subir plus de stress lié à la minorité que les jeunes hommes trans.

Deux choses à ce sujet : Premièrement, je ne pense pas que cela corresponde vraiment au fait que les améliorations de la congruence de l'apparence étaient statistiquement égales entre les hommes et les femmes nés. Bien sûr, la congruence de l'apparence n'est qu'un aspect d'une transition réussie, mais on pourrait penser que si tous ces autres obstacles empêchaient les filles trans de se sentir mieux après deux ans d'hormones, cela se verrait dans la variable qui suit le plus étroitement les effets physiques de ces hormones.

Plus important encore, la science est censée être ouverte d'esprit. Si vous évaluez un nouveau traitement, vous êtes censé envisager la possibilité qu'il ne fonctionne pas comme prévu. Il se peut que tout ce que les auteurs disent dans cet extrait soit vrai, **mais il est impossible d'ignorer la chaîne des événements : Ils ont mis une cohorte d'enfants sous hormones, deux d'entre eux se sont suicidés, et les hommes nataux ne semblent avoir connu aucune amélioration mesurable, si ce n'est une toute petite amélioration sur une échelle d'affect positif et une autre d'une importance discutable sur une échelle de congruence d'apparence plutôt tautologique**. Leur réponse est de dire… peut-être que les enfants ont juste besoin d'être sous hormones plus longtemps. **Il n'y a même pas un moment de pause, de réflexion ou d'incertitude**. C'est la fuite en avant.

Le fait est que certains membres de cette équipe ne sont pas seulement des cliniciens et des chercheurs — ils sont aussi **des défenseurs inébranlables de ces traitements**. Ils croient **fermement** que ces traitements aident les enfants transgenres, et ils bénéficient matériellement de leur administration et de leur participation au débat public à leur sujet. Cette situation, dans laquelle de fervents défenseurs aux opinions préexistantes clairement affirmées produisent ce qui est censé être une preuve de premier ordre (il s'agit du *New England Journal of Medicine,* après tout), ne devrait-elle pas nous préoccuper **un peu ?**

Il est probablement inévitable que, dans certains cas, les défenseurs d'un traitement fassent également de la recherche sur ce traitement. Mais nous devrions au moins reconnaître les pièges potentiels dans ce cas. Lorsque Cochrane ou le National Institute for Health and Care Excellence (NICE) du Royaume-Uni publient des données scientifiques, on s'attend à ce que les bureaucrates chargés de les produire entrent dans le processus sans avoir d'a priori et s'investissent raisonnablement dans l'examen des données **d'une manière juste et impartiale**. En fait, si l'examen accablant par le NICE des preuves concernant l'administration de bloqueurs de puberté aux enfants et d'hormones aux adolescents a marqué un tournant majeur dans cette discussion, c'est en partie **parce qu'il** *s*'agit d'une institution digne de confiance.

Si Cochrane publiait une étude montrant qu'un antidépresseur particulier l'emporte largement sur les autres, et qu'il était révélé par la suite qu'un coauteur de cette étude avait un mari qui travaillait pour l'entreprise pharmaceutique qui produisait l'antidépresseur en question — et que cette relation n'avait

pas été divulguée — cela jetterait immédiatement une ombre sur l'article. Il s'agirait d'un conflit d'intérêts.

Pourquoi cette logique **ne** s'appliquerait-elle **pas** ici ? Un seul article du *NEJM n'*est pas la même chose qu'un examen complet des preuves, bien sûr, mais pourquoi devrions-nous complètement ignorer la réalité, à savoir que les humains étant des humains, une équipe de chercheurs profondément investie dans un traitement pourrait être moins capable d'évaluer soigneusement, sans passion, les preuves qui le soutiennent ?

Il y a peut-être une raison pour que cette logique ne s'applique pas du tout ici. Mais s'il y en a une, je ne la trouve pas. En particulier, en l'absence persistante de tout type d'examen systématique de la médecine du genre chez les jeunes aux États-Unis — un point important que Moti Gorin a soulevé [Le remède aux soins de genre pédiatriques politisés, novembre 2022], chaque étude sur ce sujet va revêtir une importance démesurée. C'est une raison de plus pour exiger que les chercheurs à l'origine de ces études respectent les normes les plus strictes en matière de rigueur et de transparence.

Je ne pense pas que ce soit le cas ici, et ce n'est pas la première fois que de fervents défenseurs de la transition de genre chez les jeunes produisent des recherches douteuses à ce sujet. Comme je l'ai noté dans la première partie, dans ce cas, il n'est peut-être pas juste de mettre tous les points d'interrogation sur le dos des chercheurs eux-mêmes ; certains d'entre eux peuvent être le résultat de la contribution éditoriale du *NEJM.* Il n'en reste pas moins que cet effort de recherche dure depuis plusieurs années et qu'il a coûté plusieurs centaines de milliers de dollars. Les chercheurs ont publié ce qui était censé être l'une de leurs études phares, et il manque des preuves absolument cruciales dont nous avons besoin pour évaluer ces traitements. Toutes ces années plus tard, **nous ne savons même pas si la cohorte a connu une réduction de la suicidalité ou de la dysphorie de genre.** Ce que nous savons n'est pas encourageant : un taux de suicide élevé, des améliorations infimes, voire minimes, sauf sur une mesure discutable, **aucune** amélioration sur la plupart des mesures pour les filles transgenres.

Encore une fois : Il se pourrait qu'en ajustant correctement toutes les variables manquantes et non déclarées, les chercheurs n'aient rien trouvé du tout. Il est frustrant de constater que nous sommes encore dans une situation où cette possibilité existe — un plus grand nombre de ces questions auraient déjà dû trouver une réponse.

1. Quelques remarques à propos de ces variables : Le protocole n'indique pas clairement quel instrument les chercheurs ont utilisé pour mesurer les symptômes de traumatisme, mais je ne vois aucun signe qu'il soit couvert par les variables incluses dans l'étude du *NEJM*. Quoi qu'il en soit, le terme "traumatisme" n'apparaît pas dans l'article, ce qui signifie qu'il n'est pas signalé. Par ailleurs, l'"automutilation" et la "suicidalité" sont effectivement des variables distinctes. Dans leur protocole, les chercheurs mentionnent l'"automutilation — Questions sur le fait de savoir si et où le participant s'est volontairement blessé" séparément de l'échelle d'idéation suicidaire. C'est ainsi que j'en suis arrivé à mon décompte de huit variables.

2. Extrait du protocole : "Les données seront mises à la disposition d'autres chercheurs des NIH dans le cadre de l'accord de partage des données avec [l'Eunice Kennedy Shriver National Institute of Child Health and Human Development] après un délai raisonnable qui comprend suffisamment de temps pour préparer et soumettre à la publication quatre manuscrits présentant les résultats fondamentaux du projet. À partir de l'année 3, des publications évaluées par des pairs seront élaborées en rapport avec les hypothèses transversales et les questions de recherche figurant dans les objectifs primaires et secondaires, bien que la plupart des publications soient de nature longitudinale et seront élaborées au cours de l'année 5. La diffusion des résultats auprès des fonctionnaires de l'État et du comté, des décideurs politiques et des organisations commencera au cours de l'année 3, lorsque les données préliminaires seront disponibles. Dans l'étude elle-même, les auteurs notent également : "Il n'y a eu aucun accord concernant la confidentialité des données entre le commanditaire (Eunice Kennedy Shriver National Institute of Child Health and Human Development), les auteurs et les institutions participantes".