The Complexities of Evaluating the Exposome in Psychiatry: A Data-Driven Illustration of Challenges and Some Propositions for Amendments

Sinan Guloksuz[1,2], Bart P. F. Rutten[1], Lotta-Katrin Pries[1], Margreet ten Have[3], Ron de Graaf[3], Saskia van Dorsselaer[3], Boris Klingenberg[1], Jim van Os[1,4,5], John P. A. Ioannidis[*,6–8]; for the European Network of National Schizophrenia Networks Studying Gene-Environment Interactions Work Package 6 (EU-GEI WP6) Group

[1]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Centre, Maastricht, the Netherlands; [2]Department of Psychiatry, Yale University School of Medicine, New Haven, CT; [3]Department of Epidemiology, Netherlands Institute of Mental Health and Addiction; Utrecht, the Netherlands; [4]Department of Psychiatry, Brain Centre Rudolf Magnus, University Medical Centre Utrecht, Utrecht, the Netherlands; [5]Department of Psychosis Studies, King's College London, King's Health Partners, Institute of Psychiatry, London, UK; [6]Stanford Prevention Research Center, Department of Medicine, Department of Health Research and Policy and Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA; [7]Department of Statistics, Stanford University School of Humanities and Science, Stanford, CA; [8]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA

*To whom correspondence should be addressed; Stanford Prevention Research Center, Stanford University, Medical School Office Bldg, 1265 Welch Rd, Stanford, CA 94305, US; tel: (650) 725-5465, fax: (650) 725-6247, e-mail: jioannid@stanford.edu

Identifying modifiable factors through environmental research may improve mental health outcomes. However, several challenges need to be addressed to optimize the chances of success. By analyzing the Netherlands Mental Health Survey and Incidence Study-2 data, we provide a data-driven illustration of how closely connected the exposures and the mental health outcomes are and how model and variable specifications produce "vibration of effects" (variation of results under multiple different model specifications). Interdependence of exposures is the rule rather than the exception. Therefore, exposure-wide systematic approaches are needed to separate genuine strong signals from selective reporting and dissect sources of heterogeneity. Pre-registration of protocols and analytical plans is still uncommon in environmental research. Different studies often present very different models, including different variables, despite examining the same outcome, even if consistent sets of variables and definitions are available. For datasets that are already collected (and often already analyzed), the exploratory nature of the work should be disclosed. Exploratory analysis should be separated from prospective confirmatory research with truly pre-specified analysis plans. In the era of big-data, where very low P values for trivial effects are detected, several safeguards may be considered to improve inferences, eg, lowering P-value thresholds, prioritizing effect sizes over significance, analyzing pre-specified falsification endpoints, and embracing alternative approaches like false discovery rates and Bayesian methods. Any claims for causality should be cautious and preferably avoided, until intervention effects have been validated. We hope the propositions for amendments presented here may help with meeting these pressing challenges.

> "In all things which have a plurality of parts, and which are not a total aggregate but a whole of some sort distinct from the parts, there is some cause"
>
> Aristotle

Introduction

Observational epidemiological studies have identified various environmental exposures associated with psychosis spectrum disorder (PSD). We have recently discussed the challenges of environmental research in psychiatry and how the exposome framework, an agnostic exposure-wide analytic approach akin to genome-wide analysis, might help us with embracing the complexity of the environment.[1] In this article, we aim to extend this discussion. Specifically, we provide a data-driven illustration of the tangled environment in relation to mental health outcomes, address various challenges, and suggest some strategies that may reduce bias and increase reproducibility.

Analytical Challenges

Dense Correlation Patterns

Exposures come in bundles of closely connected, correlated items.[2] This creates a challenge to understand which of these multiple items are essential or most important

for the association with mental health outcomes, such as PSD, and if particular combinations need to be present. Further, mental health outcomes exhibit a dynamic network structure, with an interplay between exposures and symptom domains.[3–5] Figure 1A, which shows the correlation globe for various demographic characteristics, exposures, and clinical outcomes derived from the baseline assessment of the Netherlands Mental Health Survey and Incidence Study-2 (NEMESIS-2),[6,7] also confirms this. Given the complexity (often to the degree that it is even difficult to differentiate exposure from outcome) an assumption of independent and specific effects for exposures is not realistic.

Vibration of Effects

Different studies often present very different models, including different variables, even though they might be examining the same outcome. Differences in data availability and definitions in different datasets and studies are partly responsible for this lack of analytical standardization. However, there is little agreement between investigators about which variables (eg, confounds, moderators,

and modifiers) should be included in a model, even if all variables and definitions were available. Besides limited agreement between different consortia, work packages or writing groups within the same consortium working on parallel research objectives may also use divergent analytical strategies that are not methodologically linked and eventually lead to heterogeneity and inability to combine the results in an over-arching model, as has been commendably described in the case of a cohort investigating the association between cortisol and mental disorders.[8]

To provide an example of how model and variable selections may dramatically influence the results, we tested the "vibration of effects" (VoE)[9]—the degree of variation of results under multiple different model specifications—by assessing the variability of odds ratios (OR) and *P*-values under logistic regression models using different combinations of adjustments in the NEMESIS-2 dataset. To increase reliability and consistency, we analyzed data using Stata 14.2[10] in the "long format," corrected for clustering of multiple observations at 3 time points (T0 at baseline, $n = 6646$; T1 at year 3, $n = 5303$; T2 at year 6, $n = 4618$) within subjects. Conforming to previous analyses,[5] presence of any psychotic experience was our
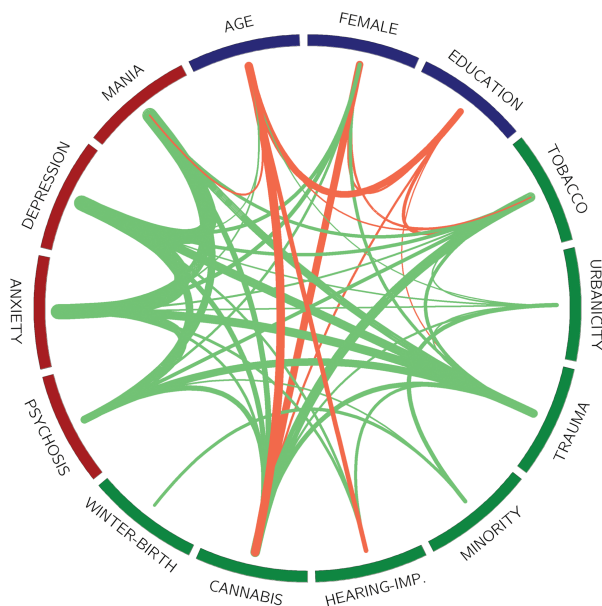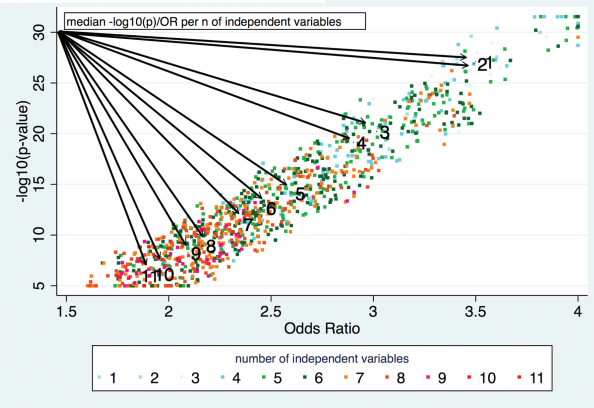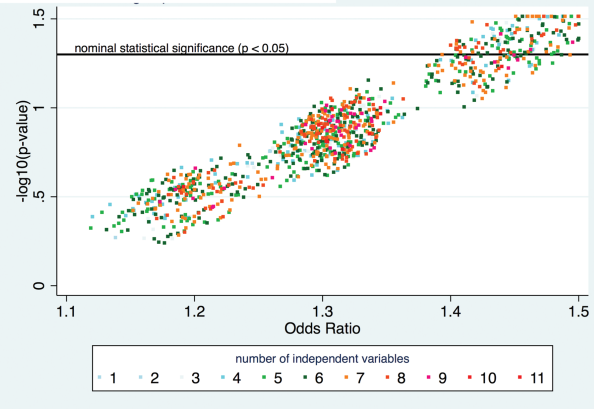

Fig. 1. (A) The correlation globe (created using Circos[25] and ClicO FS[26]). Only nominally statistically significant (at the *P* < .05 level) correlations are shown. The thickness of each line indicates the degree of the correlation, while the color represents direction (green = positive; red = negative). (B) and (C) demonstrate the vibration of effects for cannabis use and hearing impairment on psychosis expression, respectively.

binary outcome, while independent variables included demographic features (age, sex, and educational level) and risk factors previously associated with PSD (ethnic minority status, cannabis use, tobacco use, urbanicity at childhood, childhood-trauma, hearing impairment, winter-birth, and family history of mental disorder). Only participants with no missing values on each of the variables (97% of the total) were included in the analysis. The ORs and $-\log10(P$-value)s were derived from each of the 1024 potential models per exposure with different sets of adjusting variables. As previously proposed,[9] we present the relative OR (ROR = the ratio of the 99th and 1st percentiles) and relative P-value (RP = the difference between the 99th and 1st percentiles of $-\log10[P$-value]). Figures 1B and 1C present the VoE for cannabis use (a widely evaluated, seemingly strongly associated factor) and hearing impairment (a less-studied factor). Cannabis use was associated with psychosis in all models, but the magnitude of the OR estimates was attenuated largely as a function of the number of adjusting variables (ROR = 2.23, RP = 26.15). For hearing impairment, all the OR estimates were above 1, but the level of statistical significance was heavily dependent on the model specification (ROR = 1.28, RP = 1.13), with only 152 models (15%) having $P < .05$. These findings demonstrate the importance of acknowledging the VoE stemming from model choice in environmental research. Moreover, additional VoE may stem from other analytical decisions, eg, how to handle missing data, whether to apply more or less stringent exclusion criteria, or how exposures and outcome should be operationalized.

Challenges in Measuring Exposure

Dynamics

The dynamic nature of environment makes measuring the exposures a tremendously challenging task, particularly when trying to account for the effects of the timing, duration, severity, and extent of repeated exposures over time.[11]

High-level vs Granular Exposures

A major dilemma is whether one should aim to identify associations at a macro-level or at increasing levels of fine granularity—the genetic equivalent: chromosome, chromosomal area, gene, gene region, genetic variant, or gene-environment interaction levels. Insight from genetics suggests that it rarely suffices to know the high-level chromosome level defect (eg, trisomy 21), but most often one needs to proceed to more detailed granularity. Macro-level measurements may or may not capture the effect of specific more granular factors that they encompass. A similar situation may exist for environmental exposures. One needs to decide whether the problem is best studied at a macro-level (the equivalent level of a full

chromosome), eg, "war" or "divorce" or "sexual abuse"; or a substantial gain in information would be obtained by dissecting granular components of the exposure. A related, but separate decision is whether interventions should aim at the high or more granular levels. High-level interventions may not be easy (eg, stopping or preventing war), while granular level interventions pose challenges to understand as they may become more personalized and circumstantial, as granularity increases.

Subjectivity and Bias in Measurements

Objective measurement may be largely unattainable for subjective experiences (eg, childhood adversities), where differences in the sociocultural background also influence appraising, reporting, and disclosing. Further, the retrospective collection of most exposures, as is most often the case, are subject to recall and response biases. Therefore, over- or under-estimation may be driven by several factors: misattribution of current mental condition to early life events[12] that might be motivated by reaching a closure; consequences of particular personality traits and dispositions[13]; and underreporting of illegal activity or socially undesirable behavior.[14]

Promoting Better Research Practice

Exposure-Wide Assessments and Systematic Comparison Across Studies

The interdependence of exposures is the rule rather than the exception (figure 1A). Results for a specific exposure need to be seen systematically in the context of and compared to the results of all other exposures in the same dataset and other similar analyses in other datasets. Exposure-wide systematic approaches are needed to separate genuine strong signals and genuine strong heterogeneity from selective reporting.

For example, the association between urbanicity with psychosis outcomes may fluctuate over time and across populations.[15] Urbanicity may be associated with psychosis in high-income countries but not in developing countries.[16] This inconsistency may be explained primarily by the variation of other factors linked to urbanicity across different settings, such as ethnic density, substance use practices, patterns of pollution, and infectious agents,[17] which are elements of the wider totality of exposures, the "exposome." To validate both the association signal and the heterogeneity, the entirety of the available exposures ("exposome") need to be evaluated agnostically in each study/dataset, instead of taking the easy (profitable for "salami slicing" publications) way out of 1-exposure to 1-outcome at a time analysis using the same single dataset and selective reporting.

Eventually, the specificity of the reported associations should be interpreted in the light of the density, magnitude, and heterogeneity of correlations in the dataset and across the field.[18]

Pre-registration and Transparency

Pre-registration of protocols and analytical plans is still uncommon in environmental research. Even in otherwise very carefully designed collaborative cohorts, a common practice is to share the data with each group of investigators focusing on a particular exposure/outcome to test a series of hypotheses and report one finding at a time in isolation. As there is neither a registered protocol nor a record of conducted analyses, it is often impossible to track down how many different ways or times the dataset was analyzed. In the absence of transparency and protocol registration, there is minimal control over entertaining the option to analyze (data dredging) and publish (selective reporting) based on personal agendas. Therefore, if the potential for VoE is large (a common situation) measures of effect risk being merely measures of expert opinion. In the extreme, the literature may be a vote count of investigators/reviewers/editors/grant committees, who are often the same people in different roles each time.

An essential question to raise is whether one can genuinely pre-register analysis plans for datasets that are already collected. In this common scenario, the presence of a pre-specified analytical approach is doubtful unless there is proof by a time-stamped protocol that it preceded data collection. Therefore, it is vital to separate exploratory analysis from prospective confirmatory research and communicate findings with appropriate caution. In an explorative analysis, there is nothing wrong with following an intuitive approach and analyzing data from multiple perspectives that have not been fully anticipated upfront. Simply, this flexibility should be disclosed and findings should then be validated in other studies using a pre-specified plan.

The Importance of Power and Effect Size

As effect sizes become smaller with the perusal of more subtle associations, underpowered research not only increases the risk for yielding false-negative findings but also—even more so—false-positives and exaggerated effects because of various biases including, but not limited to, flexibility in analytical strategy, selective reporting, publication bias, and the winner's curse.[19] However, the new wave of big data and overpowered studies (eg, nation- or worldwide samples) give rise to an important, yet less familiar, complication: reaching a nominal statistical significance (often way beyond the traditional threshold of 0.05) for trivial effects that are possibly erroneous and almost certainly useless, even if they were causal (something that is almost always unknown). Several safeguards may help with ameliorating these understated yet highly counter-productive effects of overpowered analyses: (1) lowering the threshold of statistical significance,[20] (2) prioritizing effect size over significance, (3) exploring pre-specified falsification endpoints, where completely unrelated and improbable hypotheses that are forecasted to generate null effects are tested first to calibrate statistics that may represent just null findings,[21] (4) embracing alternative approaches over classical null hypothesis testing, such as false discovery rates and Bayesian methods.[11] The first solution is a simple yet temporary fix owing to exponentially growing data volumes and our verified dexterity in chasing significance.[22] The second solution is as easy to implement as the first one and may help differentiate potentially meaningful effects from trivial ones. It should be acknowledged that with very large sample sizes, almost everything will seem nominally significantly correlated with everything. The third solution requires pre-registration of falsification points. The fourth solution requires a reconstruction of our way of thinking and therefore more time and long-term recasting of training curricula.

Refraining From Hasty Causality Claims

Observational studies offer hints but may not necessarily prove causality. Multiple criteria have been proposed for assessing causality, the most famous being the ones proposed by Bradford Hill,[23] but they need to be reevaluated in the current era of big data and new challenges.[24] Experimental studies with randomization will be needed, whenever feasible, to pursue the most promising observational leads and translate them into equivalents of helpful interventions.

Conclusion

Identifying modifiable factors is the key to improve outcomes in mental health, and environmental research helps us in reaching this goal. However, we need to overcome several challenges. Some of these challenges are inherent and more difficult to resolve, but many are related to research practices and should be addressed to optimize our chances of success.

Funding

Acknowledgments

Utrecht. The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Guloksuz S, van Os J, Rutten BPF. The exposome paradigm and the complexities of environmental research in psychiatry. *JAMA Psychiatry*. 2018. doi: 10.1001/jamapsychiatry.2018.1211. [Epub ahead of print].

2. Patel CJ, Ioannidis JP. Studying the elusive environment in large scale. *JAMA*. 2014;311:2173–2174.

3. Isvoranu AM, Borsboom D, van Os J, Guloksuz S. A network approach to environmental impact in psychotic disorder: brief theoretical framework. *Schizophr Bull*. 2016;42:870–873.

4. Guloksuz S, van Nierop M, Lieb R, van Winkel R, Wittchen HU, van Os J. Evidence that the presence of psychosis in non-psychotic disorder is environment-dependent and mediated by severity of non-psychotic psychopathology. *Psychol Med*. 2015;45:2389–2401.

5. Pries LK, Guloksuz S, Ten Have M, et al. Evidence that environmental and familial risks for psychosis additively impact a multidimensional subthreshold psychosis syndrome. *Schizophr Bull*. 2018;44:710–719.

6. de Graaf R, Ten Have M, van Dorsselaer S. The Netherlands mental health survey and incidence study-2 (NEMESIS-2): design and methods. *Int J Methods Psychiatr Res*. 2010;19:125–141.

7. de Graaf R, ten Have M, van Gool C, van Dorsselaer S. Prevalence of mental disorders and trends from 1996 to 2009. Results from the Netherlands Mental health survey and incidence study-2. *Soc Psychiatry Psychiatr Epidemiol*. 2012;47(2):203–213.

8. Rosmalen JG, Oldehinkel AJ. The role of group dynamics in scientific inconsistencies: a case study of a research consortium. *PLoS Med*. 2011;8:e1001143.

9. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68:1046–1058.

10. *Stata Statistical Software* [computer program]. Version 14. College Station, TX: StataCorp LP; 2015.

11. Ioannidis JP, Loy EY, Poulton R, Chia KS. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci Transl Med*. 2009;1:7ps8.

12. Bendall S, Jackson HJ, Hulbert CA, McGorry PD. Childhood trauma and psychotic disorders: a systematic, critical review of the evidence. *Schizophr Bull*. 2008;34:568–579.

13. Reuben A, Moffitt TE, Caspi A, et al. Lest we forget: comparing retrospective and prospective assessments of adverse childhood experiences in the prediction of adult health. *J Child Psychol Psychiatry*. 2016;57:1103–1112.

14. Buchan BJ, M LD, Tims FM, Diamond GS. Cannabis use: consistency and validity of self-report, on-site urine testing and laboratory testing. *Addiction*. 2002;97(suppl 1):98–108.

15. Jongsma HE, Gayer-Anderson C, Lasalvia A, et al.; European Network of National Schizophrenia Networks Studying Gene-Environment Interactions Work Package 2 (EU-GEI WP2) Group. Treated Incidence of psychotic disorders in the multinational EU-GEI study. *JAMA Psychiatry*. 2018;75:36–46.

16. DeVylder JE, Kelleher I, Lalane M, Oh H, Link BG, Koyanagi A. Association of urbanicity with psychosis in low- and middle-income countries. *JAMA Psychiatry*. 2018;75:678–686.

17. Plana-Ripoll O, Pedersen CB, McGrath JJ. Urbanicity and risk of schizophrenia-new studies and old hypotheses. *JAMA Psychiatry*. 2018;75:687–688.

18. Patel CJ, Ioannidis JP. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Community Health*. 2014;68:1096–1100.

19. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.

20. Ioannidis JPA. The proposal to lower p value thresholds to.005. *JAMA*. 2018;319:1429–1430.

21. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? *JAMA*. 2013;309:241–242.

22. Bruns SB, Ioannidis JP. p-Curve and p-Hacking in observational research. *PLoS One*. 2016;11:e0149144.

23. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295–300.

24. Ioannidis JP. Exposure-wide epidemiology: revisiting Bradford Hill. *Stat Med*. 2016;35:1749–1762.

25. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19:1639–1645.

26. Cheong WH, Tan YC, Yap SJ, Ng KP. ClicO FS: an interactive web-based service of Circos. *Bioinformatics*. 2015;31:3685–3687.