PERSPECTIVE

Unconsidered issues of measurement noninvariance in biological psychiatry: A focus on biological phenotypes of psychopathology

Daniel P. Moriarity[1,2]✉, Keanan J. Joyner[3], George M. Slavich[4] and Lauren B. Alloy[1]

There is increasing appreciation that certain biological processes may not be equally related to all psychiatric symptoms in a given diagnostic category. Research on the biological phenotyping of psychopathology has begun examining the etiological and treatment implications of identified biotypes; however, little attention has been paid to a critical methodological implication of these results: measurement noninvariance. Measurement invariance is the ability of an instrument to measure the same construct, the same way, across different people, or across different time points for the same individual. If what a measure quantifies differs across different people (e.g., those with or without a particular biotype) or time points, then it is invalid to directly compare means on that measure. Using a running example of inflammatory phenotypes of depression, we first describe the biological phenotyping of psychopathology. Second, we discuss three types of measurement invariance. Third, we demonstrate how differential biology-symptom associations invariably creates measurement noninvariance using a theoretical example and simulated data (for which code is provided). We also show how this issue can lead to false conclusions about the broader diagnostic construct. Finally, we provide several suggestions for addressing these important issues to help advance the field of biological psychiatry.

Molecular Psychiatry; https://doi.org/10.1038/s41380-021-01414-5

INTRODUCTION

Many research questions in biological psychiatry use variables that index processes such as inflammatory activity, grey matter volume, and gene expression as predictors of an aggregate measure of psychopathology. An underlying assumption of these tests, as commonly performed, is that the psychopathology measure used assesses the same construct the same way each time it is administered, either across different people or across different time points for the same individual. However, this assumption might be untenable in light of growing evidence that some biological risk factors have differential associations with symptoms within a diagnostic construct (e.g., inflammatory proteins being most robustly associated with neurovegetative symptoms of depression [1]).

In this article, we first briefly describe the concept of biological phenotypes. Second, we discuss the concept of measurement invariance. Third, we use both a theoretical example and statistical simulation to illustrate how the presence of biological phenotypes of psychopathology induces measurement noninvariance. We also discuss how this issue can result in inappropriate conclusions about the relations between biology and behavior. Finally, we provide some recommendations for moving forward.

BIOLOGICAL PHENOTYPES OF PSYCHOPATHOLOGY

There is accumulating evidence that different psychiatric symptoms within some diagnostic categories (e.g., depression) may have different risk factors [2]. These findings have prompted interest in the symptom-level biological phenotyping of psychopathology (see Fig. 1 for an example of a nine-item measure of depression for which a risk factor is only related to three items). The thorough characterization of which specific symptoms of a disorder are associated with a given process may in turn help advance biological psychiatry and, in addition, precision medicine. For example, understanding that inflammation is associated primarily with neurovegetative symptoms of depression [1] can help clinicians identify patients who may possess an underlying atypical inflammatory phenotype, and this information can, in turn, guide decisions about who might benefit most from adjunctive anti-inflammatory treatments [3]. Further, this level of specificity will improve insight into whether biology—behavior associations are disorder specific or transdiagnostic in nature. For example, does irritability as an indicator of depression have the same biological correlates as irritability as an indicator of bipolar disorder or borderline personality disorder, and within non-clinical samples?

Studying biological phenotypes of psychopathology also has the potential to improve the replicability of psychiatric

[1]Department of Psychology, Temple University, Philadelphia, USA. [2]Department of Psychiatry, McLean Hospital/Harvard University Medical School, Boston, USA. [3]Department of Psychology, Florida State University, Tallahassee, USA. [4]Cousins Center for Psychoneuroimmunology and Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, USA. ✉email: Daniel.moriarity@temple.edu

Received: 3 June 2021 Revised: 19 November 2021 Accepted: 29 November 2021
Published online: 08 January 2022

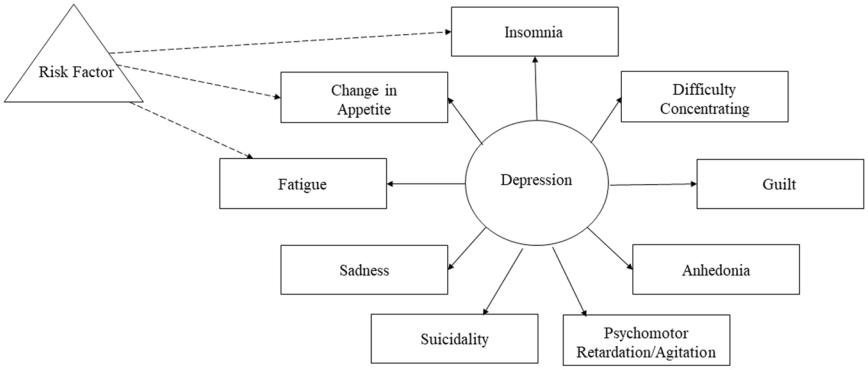

Fig. 1 Visual representation of a risk factor associated with a subset of symptoms. The triangle represents a risk factor, rectangles indicate individual depression symptoms, and the circle represents depression. Solid lines connect the individual items and depression. Dashed lines represent relations between a risk factor and a subset of items.

research [4]. Again, using an immunologic example, consider that the effect sizes between C-reactive protein (CRP) and depression symptoms across published studies is highly variable across studies [5]. Given evidence that CRP is not equally associated with all depression symptoms [6–8], inconsistent results between CRP and total depression symptoms are likely influenced by the sampling variability of which symptoms are endorsed across studies. Guided by phenotyping research, making psychiatric outcomes more nuanced or specific (i.e., specific symptoms or subtypes of depression) may increase replicability and shorten the research-to-practice timeline for syndromes that are characterized by high degrees of heterogeneity [9].

The implications of differential associations between a risk factor and the symptoms of a disorder extend beyond etiology, nosology, and treatment. Below, we examine an important methodological concern that has been largely ignored in extant discourse on phenotyping: measurement noninvariance. We will continue using the example of inflammation and depression to contextualize the issue of measurement noninvariance, discuss its consequences, and describe potential solutions. However, the issue of measurement invariance is universally applicable to all risk factors that are unequally associated with different symptoms on a measure.

MEASUREMENT INVARIANCE: A BRIEF OVERVIEW

In the context of psychological questionnaires, measurement invariance is the ability of a questionnaire to measure the same construct in the same way regardless of who takes it (e.g., people from two different groups) or when it is completed (e.g., same person at multiple points in a longitudinal study). Measurement invariance also can exist as a function of continuous variables (e.g., age). To keep the language consistent, we will focus on measurement invariance across groups.

Without measurement invariance, it is inappropriate to compare means—the most common level of analysis in biological psychiatry—because identical scores might not reflect the same level of a construct for both groups. As an example, consider if Person A steps on a scale on Earth and their weight is displayed in pounds and Person B steps on the same scale and the weight is displayed in pounds, but the scale is located on the moon. Despite using the same exact measurement instrument, these two numbers cannot be directly compared because the weight registered by the scale is influenced by a third factor—in this case, different levels of gravity.

The three most commonly discussed types of measurement invariance are configural, metric (sometimes referred to as "weak" invariance), and scalar invariance (sometimes referred to as "strong" invariance). We will briefly discuss configural and metric invariance, but focus on scalar invariance, for reasons described below. See Fig. 2 for visualization of the three kinds of measurement invariance and [10] for a more thorough review of measurement invariance and how to test it. Also, lest researchers assume that because they do not model their psychopathology variables in latent space and instead use sum scores, they are immune from the challenges raised herein, we want to highlight that sum scores actually are themselves latent variables (for an in-depth explainer, see [11]). Quoting from the authors of that article, "sum scoring corresponds to a statistical model and is not a model-free arithmetic calculation". Mathematically, a sum score from a set of items is a latent variable model that fixes all loadings and error variances to equivalence across items (among other assumptions). As such, all of the issues being discussed here are just as applicable to differences in sum scores as they are to latent variable means.

Configural invariance, the least strict form of measurement invariance, refers to equivalence of model form. That is, which variables (e.g., items) load onto which latent variables (e.g., depression) does not change as a function of a third variable (e.g., elevated inflammatory levels). An example of configural noninvariance is if all nine items on a depression questionnaire load onto the depression latent factor in a sample with normative inflammation, but only eight of the items load onto the depression factor in a group with elevated inflammation. If configural invariance is supported, the next form of invariance to check is metric. Metric invariance, in turn, refers to the equivalence of item loadings (how much an item is associated with a factor). For example, suppose item #9 had a loading of .3 on the depression factor in a sample with normative inflammation, but had a loading of .6 in a group with elevated inflammation.

If both configural and metric invariance are supported, the next step is to test for scalar invariance for the items with metric invariance. Scalar invariance refers to equality of item intercepts/thresholds (i.e., what level of endorsement of an item to expect if the latent variable associated with the item is 0). For example, consider a sample in which, when the true latent score of depression is 0, none of the items are endorsed. An example of scalar noninvariance would be if, in a different sample (e.g., one with elevated inflammation), when the true latent score of depression was 0, there would still be a couple of items likely to be endorsed (because they are attributable to inflammation instead of depression). If item intercepts differ between groups, then observed mean differences in the construct (e.g., depression) do not accurately capture true mean differences in the latent variable (see below for an illustration). Therefore, if scalar invariance is not met, any statistical test comparing mean differences on the total number of depression symptoms would

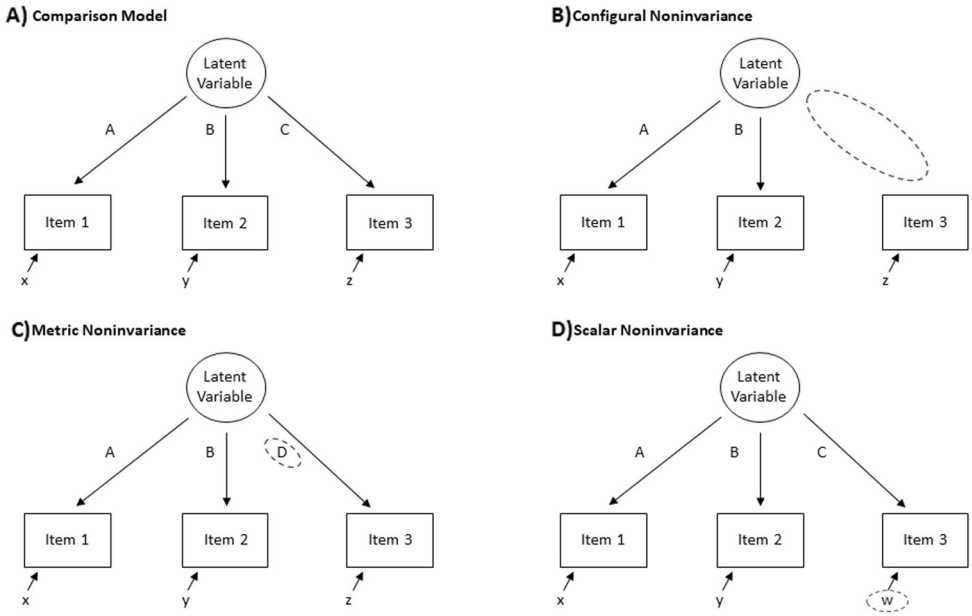

Fig. 2 Visual representations of measurement noninvariance. a The comparison model, **b** configural noninvariance, **c** metric noninvariance, and **d** scalar noninvariance. Focal differences associated with the specified type of noninvariance are highlighted by a dashed circle. Uppercase letters = factor loadings; lowercase letters = intercepts.

be confounded by the lack of scalar invariance, precluding interpretable group-difference analyses. As illustrated below, *unequal associations between a variable and individual items on a measure will always induce scalar noninvariance*. In fact, it is analogous to the definition of scalar noninvariance, highlighting a potential limitation of much extant research in biological psychiatry.

A THEORETICAL EXAMPLE AND SIMULATION OF MEASUREMENT NONINVARIANCE

Imagine a scenario in which a researcher tests whether individuals with atypically elevated CRP report more depression symptoms on the Patient Health Questionnaire (PHQ)-9 [12] as compared to individuals with normative levels of CRP. The findings suggest that CRP levels are specifically related to changes in appetite and increased fatigue and no other depression symptoms on the PHQ-9 [7]. If the researcher simply summed the items on the PHQ-9 (or used them to load onto a single, latent variable of depression) and then tested group differences, it is possible that they would find a statistically significant mean difference that could, at least in part, be driven by actual differences in these two specific symptoms. Further, because we would expect that the items measuring changes in appetite and fatigue would have a systematically higher rate of endorsement (i.e., higher intercepts/scalar non-invariance) in the elevated CRP group relative to the non-elevated CRP group, identical scores across these groups likely reflect different symptom profiles. Consequently, although there might be a statistically significant difference between the group means, these means are reflective of different depression constructs (e.g., one where endorsement of all nine symptoms is approximately equal, and one where changes in appetite and fatigue are featured proportionally more than the other symptoms), confounding inferences about differences in total depression scores between groups. It is important to reiterate that, although a group-differences design is used in this example, measurement noninvariance can exist as a function of a continuous variable (for a description of moderated nonlinear factor analyses, see [13]). We recommend using such approaches when there are not clinically/theoretically meaningful cut-offs for biological variables of interest. Furthermore, although we have focused on scalar noninvariance because it is invariably induced by unequal associations between a risk factor of interest and mean levels of individual symptoms on a measure, it is possible that certain biological processes also are associated with other types of noninvariance (e.g., configural or metric).

As a didactic resource, annotated R code that can be used to simulate 100 versions each of two different datasets, each consisting of two groups (representative of the theoretical elevated and non-elevated CRP groups above) with 250 participants each, is provided in the Supplemental Materials. The first dataset has group differences for only a subset of three variables (henceforth referred to as "symptoms"); the second dataset has group differences for all of the symptoms measured (i.e., the high-risk group only increased risk for 3/9 symptoms in the first dataset, but equally increased risk for 9/9 symptoms in the second dataset). Tests of the three types of measurement invariance described above also are provided. Only the dataset that yields group differences in a subset of symptoms consistently has scalar noninvariance (i.e., in 100% of the simulations conducted, compared to only 2% of the simulations when there was an equal group difference across all symptoms). Notably, this is the only type of noninvariance that systematically differs between the datasets.

As a follow-up to illustrate how scalar noninvariance can lead to false conclusions about the broader construct that items measure, group differences in the latent symptom total score were tested in the datasets available in the Supplemental Materials with the systematic group difference present for just a subset of symptoms. Even though the simulated datasets were not simulated to have differences at the latent factor level—and, therefore, we would expect a false-positive group-difference for approximately 5% of the samples given a conventional alpha of .05—a significant group-difference in the latent factor was observed in 63% of simulations. Therefore, there was a greatly inflated risk of falsely concluding group-differences in the latent factor when scalar noninvariance was present. In addition to illustrating the issues considered in this article, the code can be adapted to test for measurement invariance in readers' own data.

MOVING FORWARD

We have used the example of inflammatory phenotypes of depression [1, 14] to illustrate how unequal associations between a given biological process and different symptoms on a measure induces scalar noninvariance; however, this is a relevant concern for several subfields in psychiatry. For example, polygenetic risk scores for schizophrenia are primarily associated with positive psychotic symptoms [15]. Additionally, symptom-level endorsement of depression in women varies as a function of early vs. late onset Major Depressive Disorder (MDD), presence/absence of a family history of MDD, and exposure to adversity [16]. Several reproductive biomarkers also have shown unequal associations with perinatal depression symptoms [17]. Further, differences in grey matter volume have domain-specific associations with obsessive-compulsive traits (e.g., less right insula volume associated with higher "contamination/washing"; [18]), and symptom-specific associations with depression (e.g., hippocampal volume is positively associated with loss of interest and irritability, but negatively associated with changes in appetite and sadness; [19]). Therefore, research on all of these topics might be affected by unconsidered issues of measurement noninvariance.

Because there are many biological processes that have not yet been investigated using symptom-specific approaches, the true breadth of this problem is unknown. However, given increasing evidence across psychopathologies and biological processes that not all symptoms of a disorder have the same risk factors, it is plausible that measurement noninvariance is pervasive in biological psychiatry. Testing measurement invariance can provide insight into which specific subfields of psychiatry—or areas in psychiatry and psychology more generally—may be missing differential associations between biological processes and symptoms of specific disorders. To this end, it is imperative that biological psychiatry tests units of measurement smaller than diagnoses and total symptom scores [9]. By diversifying the level of psychopathological measurement explored, it will be possible to determine at what level biology-psychopathology associations most consistently exist (i.e., diagnosis vs. subscale vs. symptom).

With these points in mind, we conclude with some recommendations to facilitate the exploration of measurement noninvariance as a function of biological measures and strategies to navigate this issue should it be found: First, *test for measurement noninvariance of symptom measures as a function of biological processes* to identify subfields for which this is a concern that needs to be addressed. Second, *when measurement noninvariance is found, modify analytic strategy as appropriate*. It is important to emphasize that ideal choice of analytic adjustment is influenced by a few considerations including: the type(s) of noninvariance observed, sample size, diagnostic philosophy, and the specific nuances of one's research question. For example, it is possible to adjust model constraints to create latent models with measurement invariance (for more details, see [10]). However, especially in the case of scalar noninvariance, noninvariance is an indication to select a more detailed analytic approach. Options include: hierarchical models (e.g., within the HiTOP framework [20]) where biology predicts multiple levels of measurement (e.g., total score, subscales, specific symptoms), analyses taking a differential item functioning approach (i.e., tests how the probability of endorsing an item might change as a function of a biological variable), or symptom-level analyses. Third, *when analyzing heterogenous psychopathological constructs, consider exploring multiple levels of measurement* (e.g., total score vs. subscale vs. specific items of a symptom measure, as described above) *as an a priori analytic strategy* [9]. This will help isolate at which level of measurement a biological process is associated with a behavioral phenotype and at what level it might be appropriate to aggregate similarly associated components. Further, this level of inquiry protects against problems with diagnostic heterogeneity [21] and facilitates dimensional conceptualizations of mental illness consistent with RDoC [22].

At the same time, it is important to note that many extant measures were explicitly created with unidimensional sum scores/latent variables in mind and the psychometrics of the constituent parts of the measures (subscales, individual items) must also be considered. Additionally, many biological psychiatry studies have small sample sizes that may preclude tests of measurement invariance. Approximating the necessary sample size for tests involves understanding of the underlying factor structure and strength of the relation between the biological and psychological variables of interest, and thus, is outside the scope of this paper. It is also important to note that sample size influences ideal fit indices, a topic that is described in more detail in [10]. Further, should a sample be underpowered for testing measurement invariance, it is advisable to consider preliminary analyses in a larger dataset, perhaps one with openly-available data (e.g., Midlife in the United States [23], UK Biobank [24], Adolescent Brain Cognitive Development Study [25]).

CONCLUSION

In conclusion, growing evidence suggests that many biological processes are unequally associated with symptoms in a given diagnostic category. As demonstrated above, these biological phenotypes of psychopathology can induce measurement noninvariance, which precludes valid comparison of unidimensional sum scores/latent variables on a measure as a function of the associated biological construct being assessed. Looking forward, researchers should explicitly test for measurement noninvariance before analyzing aggregate symptom measures and continue investigating biological phenotypes of psychopathology using more detailed analytic techniques.

REFERENCES

1. Majd M, Saunders EFH, Engeland CG. Inflammation and the dimensions of depression: a review. Front Neuroendocrinol. 2020;56.
2. Fried EI, Nesse RM, Zivin K, Guille C, Sen S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. Psychol Med. 2014;44:2067–76.
3. Slavich GM, Irwin MR. From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. Psychol Bull. 2014;140:774–815.
4. Moriarity DP. Building a replicable and clinically-impactful immunopsychiatry: methods, phenotyping, and theory integration. Brain Behav Immun-Heal. 2021;16.
5. Mac Giollabhui N, Ng TH, Ellman LM, Alloy LB. The longitudinal associations of inflammatory biomarkers and depression revisited: systematic review, meta-analysis, and meta-regression. Mol Psychiatry. 2020:1–13.
6. Fried EI, von Stockert S, Haslbeck JMB, Lamers F, Schoevers RA, Penninx BWJH. Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. Psychol Med. 2019. 2019. https://doi.org/10.31234/osf.io/84ske.
7. Moriarity DP, Horn SR, Kautz MM, Haslbeck JM, Alloy LB. How handling extreme C-reactive protein (CRP) values and regularization influences CRP and depression criteria associations in network analyses. Brain Behav Immun. 2021;91:393–403.
8. Milaneschi Y, Kappelmann N, Ye Z, Lamers F, Moser S, Jones PB, et al. Association of inflammation with depression and anxiety: evidence for symptom-specificity and potential causality from UK Biobank and NESDA Cohorts. Mol Psychiatry. https://doi.org/10.1101/2021.01.08.20248710.
9. Moriarity DP, Alloy LB. Beyond diagnoses and total symptom scores: diversifying the level of analysis in psychoneuroimmunology research. Brain Behav Immun. 2020;89:1–2.
10. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Dev Rev. 2016;41:71–90.
11. McNeish D, Wolf MG. Thinking twice about sum scores. Behav Res Methods. 2020;52:2287–305.
12. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606–13.

13. Bauer D. A more general model for testing measurement invariance and differential item functioning. Psychol Methods. 2017;22:507–26.
14. Dooley LN, Kuhlman KR, Robles TF, Eisenberger NI, Craske MG, Bower JE. The role of inflammation in core features of depression: insights from paradigms using exogenously-induced inflammation. Neurosci Biobehav Rev. 2018;94:219–37.
15. Isvoranu AM, Guloksuz S, Epskamp S, van Os J, Borsboom D. Toward incorporating genetic risk scores into symptom networks of psychosis. Psychol Med. 2020;50:636–43.
16. van Loo HM, Van Borkulo CD, Peterson RE, Fried EI, Aggen SH, Borsboom D, et al. Robust symptom networks in recurrent major depression across different levels of genetic and environmental risk. J Affect Disord. 2018;227:313–22.
17. Santos H, Fried EI, Asafu-Adjei J, Jeanne, Ruiz R. Network structure of perinatal depressive symptoms in Latinas: relationship to stress and reproductive biomarkers. Res Nurs Heal. 2017;40:218–28.
18. Okada K, Nakao T, Sanematsu H, Murayama K, Honda S, Tomita M, et al. Biological heterogeneity of obsessive-compulsive disorder: a voxel-based morphometric study based on dimensional assessment. Psychiatry Clin Neurosci. 2015;69:411–21.
19. Hilland E, Landrø NI, Kraft B, Tamnes CK, Fried EI, Maglanoc LA, et al. Exploring the links between specific depression symptoms and brain structure: a network study. Psychiatry Clin Neurosci. 2020;74:220–1.
20. Kotov R, Krueger RF, Watson D, Bagby M, Carpenter WT, Caspi A. The hierarchical taxonomy Of Psychopathology (HiTOP). J Abnorm Psychol. 2017:1–83.
21. Feczko E, Miranda-dominguez O, Marr M, Graham AM, Nigg JT, Fair DA. The heterogeneity problem: approaches to identify psychiatric subtypes. Trends Cogn Sci. 2019:1–18.
22. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine D, Quinn K, et al. Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry Online. 2010;167:748–51.
23. Ryff CD, Seeman T, Weinstein M. Midlife in the United States (MIDUS 2): Biomarker Project, 2004-2009. Ann Arbor, MI Inter-University Consort Polit Soc Res. [Distributor]. 2017:10.
24. Allen NE, Sudlow C, Peakman T, Collins RUK. Biobank data: come and get it. Sci Transl Med. 2014;6:4–7.
25. Volkow ND, Koob GF, Croyle RT, Bianchi DW, Gordon JA, Koroshetz WJ, et al. The conception of the ABCD study: from substance use to a broad NIH collaboration. Dev Cogn Neurosci. 2018;32:4–7.

ACKNOWLEDGEMENTS

DPM was supported by National Research Service Award F31 MH122116 and an APF Visionary Grant. KJJ was supported by a Ford Foundation Predoctoral Fellowship administered by the National Academy of Sciences, Engineering, and Medicine and National Institute of Drug Abuse R36 DA050049. GMS was supported by National Institutes of Health grant K08 MH103443 and by grant OPR21101 from the California Initiative to Advance Precision Medicine. LBA was supported by National Institute of Mental Health R01 MH101168.

AUTHOR CONTRIBUTIONS

DPM generated the idea for the manuscript, wrote the manuscript, and ran analyses. KJJ helped refine theoretical underpinning of the manuscript, consulted on code, and provided feedback on the manuscript. GMS and LBA provided feedback on the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41380-021-01414-5.

Correspondence and requests for materials should be addressed to Daniel P. Moriarity.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.