Determining gene moderation of environmental risk factors for a mental disorder: a “perfect storm” of methodological problems

HELENA CHMURA KRAEMER^{1,2}

1 Stanford University, Stanford, CA USA

2 Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA USA

Key words

genes, environment, moderation, interaction, methods

Correspondence

Helena Chmura Kraemer, 1116 Forest Avenue, Palo Alto, CA 94301, USA.
Telephone: (+1) 650-328-7564
Email: hckhome@pacbell.net

Received 28 February 2011;
revised 10 January 2012;
accepted 17 February 2012

Abstract

For most of the twentieth century, the focus was on “nature” versus “nurture”, i.e. genetic versus environmental effects on disorders. Now it is increasingly recognized that a disorder may reflect genes and environments “working together”. A gene may moderate an environmental risk factor, it may be mediated by an environmental risk factor. The environmental risk factor may be proxy to the gene, and the two may be independent risk factors. Which of these situations pertain influences both subsequent research and clinical and policy decision-making. However, recent meta-analyses attempting to confirm the Caspi *et al.* (*Science*, 301, 386–389, 2003) hypothesis indicate that the methodological issues relating to establishing specifically a moderating effect of a gene on an environmental factor are not well understood. The discussion here concerns the definition of “moderator”, how it is distinct from other ways in which gene and environment can “work together”, the methods needed to establish such a moderator, and the public health significance of such efforts. Copyright © 2012 John Wiley & Sons, Ltd.

Introduction

It has long been thought that to understand the etiology of mental (and other) disorders, it will be necessary to understand how genetic and environmental risk factors “work together” or “interact”. In what follows, we will use the term “interaction” in the general sense that two factors somehow “work together” in their relationship to an outcome, and “statistical interaction” in the more narrow sense of a multiplicative term in a linear model (Greenland, 1993). Until recently, little research attention has been

paid to this issue, and recent research results have been confusing. To some extent, this situation results because such a research question falls into interdisciplinary “cracks”. Geneticists typically have little experience in the study of environmental risk factors, often assigning environmental influences to random error. Behavioral scientists interested in environmental risk factors such as stress, abuse, trauma, have put little emphasis on issues either related to genes or to interactions. Finally, the methodology to detect and identify how risk factors “work together” is not well understood. What results is a

“perfect storm” (a “perfect storm” is a critical situation created by a powerful concurrence of factors) of methodological problems with issues related to concept, design and analyses of studies investigating gene by environment interactions not clearly or consistently addressed. Then, the results in one study are contradicted in other studies; replication and confirmation cannot be achieved, and science progresses only very slowly.

A case in point: in 2003, Caspi *et al.* reported that, in exploring a community birth cohort, they found that the serotonin transporter polymorphism gene (5-HTT) moderated the effect of stressful life events on the occurrence of major depression (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition [DSM IV] criteria). In 2009, two meta analyses were published (Munafo *et al.*, 2009; Risch *et al.*, 2009) suggesting that this was a false positive finding. The two meta-analyses surveyed the same research literature, but not the same studies. Aside from the original Caspi *et al.* (2003) study, Risch *et al.* (2009) included 13 studies and Munafo *et al.* (2009) 14, but only seven in common, perhaps because different methodological criteria were used to identify valid studies (Cooper and Hedges, 1994). The issue here is not whether the Caspi *et al.* (2003) hypothesis is true or not, but the validity of methods used to address such a question. However, if true, this is a crucial result, identifying a genetic basis for a psychiatric disorder, and suggesting one reason why it has been so difficult to establish the genetic bases of psychiatric disorders in general.

To understand why this is so crucial an issue, consider the extreme hypothetical situation in Table 1, in which the combination of a certain genotype (G+ with probability P) plus the presence of a certain environmental risk factor (E+ with probability Q) is perfectly predictive of a disorder (D+). With unreliability of diagnosis, the risk difference

between the G + E + group and the others is reduced from 1.0 to $\rho = (Se + Sp - 1)$ (Se = sensitivity; Sp = specificity).

However, if the gene alone is considered, the risk difference would be further reduced to $Q\rho$, for

$$\begin{aligned} & \text{Prob}(DX + |G+) - \text{Prob}(DX + |G-) \\ &= [PQSe + P(1 - Q)(1 - Sp)]/P - (1 - Sp) = Q\rho, \end{aligned}$$

and, similarly, if the environmental factor alone is considered, to $P\rho$. If the proportion of those with E+ (Q) is low, the genetic association will appear weak, and if the proportion of those with G+ (P) is low, the environmental association will appear weak even here, when the underlying disorder is perfectly predictable from the combination of genes and environment. Thus the search for genes moderating environmental risk factors on subsequent onset of disorders is important from a scientific standpoint, but perhaps even more so, from a public health standpoint that, to date, has been largely ignored.

In the following we will first review what evidence is necessary to show that a gene moderates an environmental risk factor for a disorder, and how moderation is distinguished from other ways in which a gene and an environmental risk factor can “work together” (Kraemer *et al.*, 2008; Kraemer *et al.*, 2001). While these principles can be extended to more complex definitions of G, E, and DX, for clarity we here focus on binary G, E, and DX. We will discuss issues related to sampling, measurement and design necessary to generate valid and powerful results supporting a moderation effect, emphasizing public health as well as statistical significance. Finally we summarize the problems that contribute to the “perfect storm”, each individually well-recognized, but here occurring together.

Table 1. An extreme hypothetical situation in which disorder is completely determined by a combination of presence (G+) or absence (G-) of a certain gene and the presence (E+) and absence (E-) of a certain environmental risk factor (E+)

	Proportion in the population	Probability of disorder	Probability of positive diagnosis
G + E+	PQ	1.00	Se
G + E-	$P(1 - Q)$	0.00	$1 - Sp$
G-E+	$(1 - P)Q$	0.00	$1 - Sp$
G-E-	$(1 - P)(1 - Q)$	0.00	$1 - Sp$

Note: P , probability G+; Q , probability E+; Se = sensitivity of the diagnosis to the disorder (probability of a positive diagnosis when the disorder is present); Sp , specificity of the diagnosis to the disorder (probability of a negative diagnosis when the disorder is absent).

How may genes and environmental risk factors “work together”?

In what follows, *G* refers to a characteristic present at birth and constant over the lifetime of the subject (e.g. a genotype such as the 5-HTT in the Caspi *et al.* (2003) hypothesis). *E* refers to an event that happens, or a characteristic that develops, during the lifetime of the subject that can potentially be manipulated either to prevent its occurrence or to block its effect once it occurs (e.g. stress, trauma, abuse in the Caspi *et al.* (2003) hypothesis, but also, more generally, gene expressions). *DX* refers to the diagnosis of the presence/absence of a disorder (e.g. depression). Clearly, as we have defined them, *G* precedes both *E* and *DX* in time. However, to be a risk factor for *DX*, *E* must also precede *DX* in time (Kraemer *et al.*, 1997). Otherwise, *E* may be a concomitant (sign or symptom) or a consequence of *DX*, rather than a precursor, and play no role in the development or prevention of the disorder.

It is important to recognize that it is the timing of *G*, *E* and *DX* that is at issue, not the time at which their measurement is taken. Thus, for example, when *G* is a genotype, it can be measured at any time during the lifetime of the subject, but its timing is still at birth. However, if *E* is an environmental event or a gene expression, it may change over the lifetime of the subjects and thus when it is measured determines its timing. In rare cases, *E* too may be determined at a much later time from records uniformly compiled at the time of *E* (e.g. age at school entry), but not from retrospective recall after onset of *DX*, which might well be influenced by *DX*.

There are four distinct ways in which *G* and *E*, so defined, may “work together” to predict *DX* (Kraemer *et al.*, 2008; Kraemer *et al.*, 2001):

- *G moderates the effect of E on DX*: It must be shown that (1) *G* and *E* are uncorrelated, and (2) the effect of *E* on *DX* differs depending on what *G* is. As shown earlier, where moderation occurs, the effects of both *G* and *E* separately may be attenuated, perhaps even undetectable.
- *E mediates the effect of G on DX*: It must be shown that (1) *G* and *E* are correlated, and (2) the effect of *G* on *DX* can be shown to be completely or partially explained by *E*. Gene expression is typically the strongest mediator of the corresponding gene on *DX*, but environmental factors may mediate the gene as well. Mediators help to explain how and why *G* is associated with *DX* and thus also have both public health and scientific importance. Mediators provide clues as to what intervention might be necessary to prevent *DX*, whereas moderators provide clues for whom such intervention is necessary.
- *E is proxy to G in its effect on DX*: It must be shown that (1) *G* and *E* are correlated, and (2) when both *G* and *E* are considered in relation to *DX*, only *G* is related. In such cases, *E* is a “pseudo-correlate” of *DX*, and should be set aside from consideration for both public health and scientific considerations related to *DX*.
- *G and E are independent risk factors for DX*: Otherwise *G* and *E* likely represent two separate pathways to *DX*.

Precise definitions of scientific terms are crucial to scientific research (Finney, 1994). If the same term is used in multiple different ways, or multiple terms are used for the same concept, such imprecision fosters confusing and inconsistent research results. Part of the problem with the gene by environment interaction terminology lies in the lack of *precise* definitions that can be applied *consistently* in a research context.

In 1986, Baron and Kenny proposed conceptual definitions distinguishing moderators and mediators applicable in a research context that continue to be widely used. However, with their original definitions, the same demonstration that shows that *G* moderates *E* on *DX* also shows that *E* moderates *G* on *DX*. In many cases, if it can be shown that *E* mediates *G* on *DX*, it can also be shown that *G* mediates *E* on *DX*, and that *G* both moderates and is moderated by *E*. What is concluded is often guided by the assumptions that the researchers make. Not all researchers make the same assumptions, producing inconsistencies in their results.

The MacArthur revision of those definitions retained the same conceptual basis, but removed the ambiguities, by specifying temporal precedence and correlation criteria (Agras *et al.*, 2000; Kraemer *et al.*, 2008; Kraemer *et al.*, 2001). In the two millennia of discussion of the philosophical bases of causality, that the cause must precede its effect, and that the cause must be correlated with its effect, have remained the two constants, and these seem natural criteria to consider. With the revised definitions, the direction of moderation and mediation is determined by temporal precedence, and the distinction between moderation and mediation preserved by the criterion of correlation. Researchers examining the same evidence using these criteria then come to the same conclusions, and the conclusions are meaningful for public health applications. The orientation and goal of these revisions was to clarify the impact of risk factors in the public health area.

In contrast, a separate area of research concerns the philosophical basis of inferences of causality and the development of complex mathematical models to implement those inferences (e.g. Pearl, 2000). The terms moderator/mediator are used in these discussions with definitions

not necessarily consistent with those of the MacArthur approach. In most cases, mathematical assumptions are made that guide the choice of the appropriate term, a problem that is avoided in the MacArthur approach. What is lost in the MacArthur approach is any inference of causality, but what is gained is the precision and consistency necessary for public health applications. Both approaches are clearly of importance.

Design issues

With the MacArthur approach, moderation can usually only be assessed in a prospective cohort study in order to establish the time precedence criterion. Since, for example, genotype can be measured at any time, at least two time points are needed to establish the time order of E and DX. Moderation can be assessed only in populations with variance in G, E and DX, not, for example, in a population, all of whom have E +.

In the Caspi *et al.* (2003) study, E specified life events in the 20–26 age period, with DX in the 26th year. Thus the time precedence criterion was generally satisfied. They found no indication of correlation between G and E, and a statistical interaction of G and E on DX that satisfied the definition that G moderated the effect of E on DX. However, in the Risch *et al.* (2009) meta-analysis, eight of the 13 studies, in the Munafo *et al.* (2009) meta-analysis, 11 of the 14 studies, had either cross-sectional or case-control designs, or were “case only” designs that cannot document the criteria for demonstration of moderation. Such studies are not designed to evaluate any moderator hypothesis or indeed any way in which genes and environment may “work together” to lead to a disorder. A common mistake is to interpret any statistical interaction effect in a linear model as “moderation” (Greenland, 1993). Not all statistical interactions signify moderation. Some signify mediation (Kraemer *et al.*, 2008), and some may signify neither.

Population: who and when?

It is known that gene frequencies, as well as distribution of environmental factors, differ from one population to another. Associations that exist among different variables may change from one population to another: A risk factor in one population could even be a protective factor in another. Whether or not G moderates the effect of E on DX may also differ from one population to another.

Caspi *et al.* (2003) sampled a birth cohort in the general population and followed subjects up to 26 years. In the Risch *et al.* (2009) meta-analysis, of the five prospective cohorts, two were twin samples, and of the other three,

one was seen in an age range of 44 to 52, the last two, later than 65 years of age. In the Munafo *et al.* (2009) meta-analysis, two of the prospective cohort studies were based on twin samples, the last seen at a mean age of 47. Thus, technically none of these cohort studies actually addressed the specific Caspi *et al.* (2003) moderator hypothesis, although some may have addressed other moderator hypotheses in other populations.

The importance of time

Time is here a vitally important factor. Consider the two hypothetical incidence curves from some entry age at $T = 0$, for those with G+ and G– shown in Figure 1. Obviously the association between G and DX, however that association is measured, changes as follow-up time (T) increases, even, as here, possibly becoming null at some time, or changing direction.

Moreover, as the follow-up time increases, who has E + may change depending on how E is defined (e.g. recent events versus cumulative events over time), and its relationship to G and DX may also change. For example, the death of a parent is not only less frequent but may be more traumatic at the age of 20 than it is at the age of 60 and have a different association with subsequent DX onset. Thus the Caspi *et al.* (2003) hypothesis related specifically to young adults (20–26 years), and would be neither confirmed nor refuted by showing presence/absence of moderation at other disparate ages.

Does that mean that every study seeking to identify genetic moderation must focus on a birth cohort and a long follow-up, as did Caspi *et al.* (2003), which would make discovery and confirmation difficult and costly,

Figure 1. Hypothetical survival curves with differential onset times that cross.

and independent replication/confirmation virtually impossible? No, because a genetic moderator effect is age specific, and genotype, G , can be measured at any time, it would actually be better to consider a more conservative formulation of the hypothesis to be tested, one that not only minimizes the cost and difficulties of such research but minimizes methodological problems (e.g. the biasing effects of accumulated non-random dropouts over long follow-up periods). For example, for a slightly modified hypothesis (cf. Caspi *et al.*, 2003, p. 388), one might sample the population of those aged 20 all with $DX-$, evaluate E at age 20, and follow all to age 26 monitoring both for onset of $DX+$ and for changes in E prior to such onset. This would then reduce a 26 year birth cohort follow-up to a six year follow-up, and would result in a more precise definition of the population to which the results pertain (those disorder-free at age 20). Such studies could then be repeated at various entry ages (10, 30, etc.) which might better inform public health considerations if and when intervention is necessary, on whom and of what kind.

Effect sizes and p values

In the last 20 years it has become increasingly apparent how misleading the “ p -value” has become (Cohen, 1995; Dar *et al.*, 1994; Hunter, 1997; Kline, 2005; Nickerson, 2000; Shrout, 1997; Thompson, 1999; Wilkinson and The_Task_Force_on_Statistical_Inference, 1999). Many mistakenly interpret the reported p -value as the probability that the null hypothesis is true, or the probability of non-replication, or using other such misinterpretations. The p -value is best interpreted as an indication that the sample size was large enough to detect a non-null effect size, not as an indicator of effect size.

The most common effect size used for association between binary measures, and the one used in both these meta-analyses, is the Odds Ratio ($OR = (p_1(1 - p_0) / ((1 - p_0)p_1)$ where p_1 and p_0 are the two probabilities to be compared), as in both meta-analyses, but there is accruing evidence that using the OR as an effect size is misleading (Greenland, 1987; Kraemer, 2004; Newcombe, 2006; Sackett, 1996) and that Risk Difference ($RD = p_1 - p_0$) is a better choice, with its reciprocal (Number Needed to Take, $NNT = 1/RD$) (Cook and Sackett, 1995; Kraemer and Kupfer, 2006) more likely to clearly communicate public health significance. OR and RD are measured on two different scales, OR with range from zero to infinity with null value at one, RD with range from -1 to $+1$ with null value at zero. However, it is known that when the two probabilities compared are complementary ($p_1 = 1 - p_0$),

$$RD = (OR^{1/2} - 1) / (OR^{1/2} + 1) = Y,$$

that, $RD = Y = 0$ if and only if $OR = 1$, and that RD and Y always have the same sign. Moreover, while the non-zero magnitude of Y always exceeds that of RD, the discrepancy is minor when both probabilities being compared are in the middle range (say, between 0.20 to 0.80).

Thus, for the survival curves in Figure 1, OR and RD are compared using the Y -rescaling of OR, to compare incidences by each time (Figure 2). Where the survival curves cross at about $T = 15$, $RD = Y = 0$ ($OR = 1$). Just before and after that cross-point, a relatively small discrepancy between Y and RD can be seen. Where the differences between Y and RD are major is when the denominator of OR is near zero or one. The mathematical problems with OR are all related to division by a number near zero, leading to an “explosion” of the magnitude of OR. When there is a discrepancy, OR may indicate strong association while RD may indicate trivial association.

To illustrate the problems with the clinical interpretation of OR, consider two points in Figure 2. For a follow-up time of $T = 0.05$, the probabilities being compared are 0.0050 for $G+$ and 0.0022 for $G-$, yielding $OR = 2.2$, and $NNT = 367$. For a follow-up time $T = 29$, the probabilities being compared are 0.689 for $G-$ and 0.498 for $G+$ (note the reversal), yielding the same $OR = 2.2$, but now $NNT = 5$. Preventing/blocking the effects of $G+$ in the first situation would make very little difference for individuals in that population, for most, whether $G+$ or $G-$, would have $DX-$ with or without intervention. However, in the second situation, preventing/blocking the effects of G could result in a 20% difference in the population prevalence/incidence of $DX+$.

The fact that when one of the incidences approaches zero or one (in which case the divisor of OR approaches zero), OR approaches infinity, regardless of what the other incidence is, and the consequent inability to clearly interpret OR in terms of public health significance, indicate strongly that RD should be preferred. In the following, RD will be used as the effect size.

Analysis issues

As illustration, in Table 2 are the results, as summarized by Munafo *et al.* (2009), from the Eley *et al.* (2004) study. A linear model relates some function of the probability in each cell (P_{ij} in Table 2) to $b_0 + b_1G + b_2E + b_3GE$, with the two values of G ($G+$ and $G-$) coded $\pm 1/2$ and the two values of E ($E+$ and $E-$) coded $\pm 1/2$ for clear interpretability of all the coefficients (Kraemer and Blasey, 2004). Which linear model is appropriate (if any), that is, the choice of the link function when selecting a linear

Figure 2. The relationship between Odds Ratio (OR) ($Y = (\sqrt{OR} - 1) / (\sqrt{OR} + 1)$) and Risk Difference (RD) for the survival curves shown in Figure 1.

Table 2. Descriptive statistics and computations of main effects and interactions in the Eley *et al.* (2004) study ($N=369$)

	E–	E+	Total
<i>Proportion of sample in each of the $G \times E$ cells (Q_{ij})</i>			
G–	0.114 = Q_{00}	0.057 = Q_{01}	0.171 = P'
G+	0.493 = Q_{10}	0.336 = Q_{11}	0.829 = P
Total	0.607 = Q'	0.393 = Q	1.000
<i>Proportion of sample in each of the $G \times E$ cells with D+</i>			
G–	0.595 = P_{00}	0.571 = P_{01}	0.587
G+	0.505 = P_{10}	0.621 = P_{11}	0.552
Total	0.522	0.614	0.558

Note: Overall Prevalence = 0.558; Overall Effect of G = $0.552 - 0.587 = -0.035$; Overall Effect of E = $0.614 - 0.552 = +0.092$; Conditional Effect of E: G+: $0.621 - 0.505 = +0.116$, G–: $0.571 - 0.595 = -0.024$; Conditional Effect of G: E+: $0.621 - 0.571 = +0.050$, E–: $0.505 - 0.595 = -0.090$; Main Effect of E: $(-0.24 + 0.116)/2 = +0.046$; Main Effect of G: $(-0.090 + 0.050)/2 = -0.020$; Interactive Effect: $0.116 - (-0.024) = 0.050 - (-0.090) = +0.139$.

model, is determined by the choice of the appropriate effect size. Since the effect size to be used here is the RD, the linear model applies to P_{ij} . If the effect size were the OR, it would apply to $\ln(P_{ij}/(1 - P_{ij}))$; if the effect size to be used was the Risk Ratio, it would apply to $\ln(P_{ij})$. In all these cases, the model will fit the data perfectly, but the answers will be different. It must be remembered that even a perfectly fitting mathematical model is not necessarily the *right* model, and the statistical interaction term in a specific linear model may not best indicate a moderator effect.

In this sample, the overall probability of DX+ was 0.558 (see Table 2 for all calculations). The effect size of G in the total population (ignoring E) was -0.035 , and the effect size of E in the total population (ignoring G) was $+0.092$. Among those with G+, the effect size of E was $+0.050$, and among those with G– was -0.090 . The “main effect” of E (b_2 in the linear model) is then the average of these two conditional effect sizes: -0.020 , and the GXE statistical interactive effect (b_3 in the linear model) is the difference between these two conditional effect sizes: $+0.139$.

It should be noted that the overall effect of E (+0.092), the conditional effects of E in the G+ (+0.050) and G− (−0.090) subgroups and the main effect of E (−0.020) are all different, not merely because of estimation error, but because they refer to the effects of E in different subpopulations. “Adjusting for” any variable in a linear model, does not give a “more correct” answer than in absence of such adjustment, it simply changes the population to which the answer applies, and is correct only if the assumptions of the linear model used hold.

This specific example was chosen because the moderation was “qualitative”. A *quantitative moderator* is one in which E+ is either a risk factor or a protective factor in both the G+ and G− subpopulations, while a *qualitative moderator* is one in which E+ is a risk factor in one subpopulation and a protective factor in the other. When a moderator is quantitative, a single intervention appropriately addressing E would benefit both groups. When a moderator is qualitative, a public health effort addressing E might benefit those in one group and harm those in the other.

Many genetics researchers believe that in absence of an overall effect of G, or a main effect of G, on DX, there is no possibility of an interaction effect. To behavioral researchers or statisticians this is a puzzling claim. As Caspi *et al.* (2010) comment: “this claim is statistically unwarranted ... Waiting for genome-wide association studies (GWAS) to throw up candidate genes may be ill-advised because GXE interactions may conceal good candidates from GWAS” (p. 10). One troubling result of this ill-advised claim is that when a small genetic effect is detected, genetic researchers appeal for larger sample sizes, as if increasing the sample size will increase the size of the effect. A larger sample size would undoubtedly decrease the *p*-value, but will only stabilize the estimate of a small, perhaps trivial, genetic effect. However, that small genetic effect may be (as in Table 1) important if it conceals an ignored gene by environment interaction.

With binary G, E, and DX, the sample estimate of the statistical interaction effect is the difference between the two RDs (thus indicating moderation) and is:

$$P_{11} - P_{10} - P_{01} + P_{00},$$

where the sample proportions in each cell in Table 2 can be used to estimate the population probabilities. The sample estimate is then an unbiased estimate of the population interactive effect. In a sample of size *N* with sample sizes in each cell and with the probability of DX+ in each cell as shown in Table 2, the squared standard error of this estimator is:

$$\sigma^2 = P_{11}(1 - P_{11})/N_{11} + P_{10}(1 - P_{10})/N_{10} + P_{01}(1 - P_{01})/N_{01} + P_{00}(1 - P_{00})/N_{00},$$

which can be estimated as s^2 , by substituting the observed proportions for P_{ij} in each cell and the actual sample sizes (N_{ij}) in each cell. Then the approximate 95% two-tailed confidence interval is given by $(P_{11} - P_{10} - P_{01} + P_{00}) \pm 1.96 s$. If the value zero is not included in the 95% two-tailed confidence interval, the result is “statistically significant” using a 5% two-tailed test.

Power: The issue of power for a statistical test relates closely to the issue of length of a confidence interval. Low power of the 5% significance test means a wide 95% confidence interval and thus a lack of precision in estimation of the true effect.

To have approximately 80% power with a 5% two-tailed test, to detect any moderator effect with magnitude greater than RD* (the critical value or selected threshold of clinical significance) the total sample size, *N*, must exceed

$$N \geq (1.96 + 0.28)^2 / (RD^2 H),$$

where 1.96 reflects the choice of a 5% two-tailed test, 0.28 the selection of 80% power, RD* the critical effect size, or threshold of clinical significance, and *H* is the harmonic mean of the proportion of subjects in the sample in each cell:

$$1/H = 0.25(1/Q_{11} + 1/Q_{10} + 1/Q_{01} + 1/Q_{00}).$$

This sample size is also the minimal sample size to have 80% chance of having the 95% two-tailed confidence interval that excludes the null effect size.

If one chose to use a representative sample from a population to test a moderator hypothesis, where the probability of G+ is *P* (*P'* = 1 − *P*), and that of E+ is *Q* (*Q'* = 1 − *Q*), the expected proportions in the four cells are *PQ*, *PQ'*, *P'Q*, *P'Q'*, (since G and E are uncorrelated) and $H = 4PP'QQ'$. Thus if either G+ or E+ is very rare or very common (*P* or *Q* approaching zero or one), the necessary sample size might have to be very large to ensure adequate power.

To gain power in testing an “a priori” moderator hypothesis, one might design a study prospectively stratifying the sample on G, taking half the sample with G+ and half with G−. Then $H = QQ'$. The sample size necessary to establish moderation will then be much less. Moreover, if E is established at the time of entry, one could also stratify the sample into two groups, one at relatively high risk of E+, and the other at lower risk of E+, and sample about equal numbers into each of the four cells, one might even further decrease the necessary sample size.

Thus in a sample of military servicemen at the end of their enlistment ($T=0$), one might stratify on G taking half the sample with $G+$ and half with $G-$. Then one might stratify each G stratum, taking half from those reporting combat stressful events ($E+$) and half from those not ($E-$). The sample to be followed, say for 5 to 10 years post end of service to observe possible onsets of PTSD, would include only those so selected. Then $H=0.25$ which is its maximal value, thus requiring the minimal value of N for adequate power. Since follow-up is the most costly and difficult aspect of a prospective study, minimizing the total number to be followed up, is a cost-effective approach to such studies.

In *testing* an “a priori” moderator hypotheses, samples stratified by G and/or E , in order to bring the cell sizes closer to equal, are recommended. However, for *exploration* to detect possible moderators, representative samples are preferred, since in such cases, many different genes and/or environmental risk factors are examined relative to the outcome, as was probably true in the Caspi *et al.* (2003) study. Then because of multiple testing, the probability of false positives increases and because of under-powered designs, the probability of false negatives increases. What results from such exploration are not conclusions, but hypotheses to be evaluated against the existing research theory and results, and if viable, to be formally tested in a future study specifically designed for the purpose.

Neither the original Caspi *et al.* (2003) study, nor any of the studies in the meta-analyses, appears to have been specifically designed to test an “a priori” moderator hypothesis. Consequently, as both meta-analyses emphasize, the power to detect moderation was very low. In general, inclusion of under-powered studies in any meta-analysis, although often done, is at least questionable (Kraemer *et al.*, 1998), for it exacerbates the “file drawer” problem (Rosenthal, 1979) in meta-analysis. For example, with the distribution in the Eley *et al.* (2004) study, the sample size necessary to have 80% power to detect any interactive effect greater than 0.10 with a 5% two-tailed test, is 5803. The actual sample size was 369. The interactive effect was not statistically significant at the two-tailed 5% level, but since the point estimate of the interactive effect was 0.139, greater than the critical effect size of 0.10 here, if this were a valid hypothesis-testing study of a moderator effect, the hypothesis of a clinically significant moderator effect would remain viable. Often such non-significant results are mistakenly interpreted as “proving” the null hypothesis, but such a result more likely indicates inadequate design (power).

What is a clinically significant moderator effect?

How large an effect in a moderator study is large enough for public health significance? The effect size above ignores the distribution of subjects in the population in the four cells (Q_{ij} in Table 2). That distribution affects only the precision of its estimation. However, the public health impact depends strongly on that distribution. Thus this effect is a poor indicator of the clinical importance of a moderator in any population.

In the Eley *et al.* (2004) study, in the absence of knowledge of G and E , we estimate that 55.8% of the population would have $DX+$ (see Table 2). If G were recognized as a risk factor for DX , this could not change the situation, since G itself cannot be changed by intervention. One can only change that which G moderates, or that which mediates G on DX . If E were here recognized as a risk factor, with successful universal efforts to prevent or block the high risk form of E , the incidence might be reduced to 52.2%. Since $E+$ is a risk factor for those with $G+$ but a protective factor for those with $G-$ (a qualitative moderator), such universal prevention efforts might *increase* the risk of $DX+$ in the $G-$ population. However, if G were recognized as a moderator of E , then efforts might be made to prevent/block $E+$ in the $G+$ subpopulation, and to prevent/block $E-$ in the $G-$ subpopulation, which, if successful, might further reduce incidence to 51.6%. The public health question is whether reducing incidence from 55.8% to 52.2% with recognition of E , then to 51.6% with recognition of G moderating E , is worth the cost that might be entailed in such prevention/blocking efforts. That is a health economics question, not a statistical question.

What is important is that the moderator effect size important to public health considerations is *not* the interaction effect in a study designed to test whether G moderates the effect of E on DX , but the reduction in incidence that could potentially be achieved by giving the appropriate interventions to the appropriate subjects to prevent the disorder. To evaluate this, estimates of the Q_{ij} would need to be obtained, from the total representative sample, or from the screening sample if a stratified sample is used.

The importance of moderator analysis is that it helps to identify the appropriate subjects for the appropriate intervention. One might well show a very large interaction effect that has no potential public health impact, as well as a relatively small interaction effect with a large potential public health impact.

Conclusion

None of the individual methodological problems here discussed is new or unknown. What is unusual is the

convergence of all these problems in one area of research, creating a “perfect storm” afflicting a research area vitally important to understanding the etiology of many disorders. There are several components to this “storm”.

- *Imprecise Use of Scientific Language:* moderator, mediator, interaction, significance.
- *Over-generalization.*

Not only are the inferences from a study in one population applied to other populations, but here the inferences from a study of one construct are applied to other constructs. The one factor common to all these studies was the HTTP gene. Not only did the studies used different diagnoses (Munafo *et al.*, 2009) at different ages, but used many different measures of “stress”. A count of life events, episodes of maltreatment, a stress index, unemployment, number of chronic diseases, etc., do not all measure the same underlying construct.

- *Mathematical Modeling Issues*

Many geneticists have expressed the mistaken belief that in absence of a genetic effect on the disorder, gene by environment interaction cannot occur. This is untrue. Many behavioral scientists seem to think that one can freely choose whether or not to include a statistical interaction term in a linear model (Aiken and West, 1991; Cohen *et al.*, 2003; Kromrey and Foster-Johnson, 1998). This is untrue. If one omits the statistical interaction term in the model when there is interaction in the population, not only is there a loss of precision, but the estimates of the other effects in the model are likely to be biased. Testing the statistical interaction term and finding it not statistically significant at the 5% significance level is not justification for omitting a statistical interaction term in the model. Such lack of statistical significance may be due to inadequate sample size rather than absence of interaction, and the biasing effects remain.

Similarly, showing a statistically significant goodness-of-fit test indicates a poor-fitting, thus non-viable model. However, not showing a statistically significant result on a goodness-of-fit test does not indicate a viable model, since that may be the result of insufficient power to detect important deviations from the assumptions of the model. As noted earlier, the linear model here could be applied to $\ln(P_{ij}/(1 - P_{ij}))$ or to $\ln(P_{ij})$, rather than to P_{ij} , achieving

perfect fit to the data in each case, but obtaining different results. The choice of model must be based on what is appropriate in the specific context of the research, here on the choice of an interpretable effect size for the interaction.

Box and Draper (1986) said: “Essentially, all models are wrong, but some are useful”(p. 424) and “. . . the practical question is how wrong do they have to be to not be useful” (p. 74). Here we avoided most problems by focusing on binary G, E and DX, with a model that fits the data perfectly, and choosing an effect size that could be clinically interpreted. Moreover, no assumptions about causality were made, nor were any inferences about causality drawn.

- *Continued Problems with Meta-analysis:*

Meta-analysis is a valuable tool to summarize what is known about a certain population effect size from multiple valid studies estimating that effect size (Cooper and Hedges, 1994; Hedges and Olkin, 1985; Lau *et al.*, 1992; Light and Pillimer, 1984). Yet meta-analysis has long been criticized (Egger and Smith, 1995; Eysenck, 1992; Lewin, 1996; Meinert, 1989; Shadish and Sweeney, 1991; Shapiro, 1994; Thompson and Pocock, 1991) because users of the method are not comprehensive in their review of the literature, not careful in setting aside studies that are not valid studies of the specific phenomenon of interest, as appears true with the two meta-analyses here discussed. Underpowered studies (Kraemer *et al.*, 2006; Maxwell, 2004), too continue to be problematic in meta-analysis.

All these problems affected these reviews. When one or two such problems occur in a study, the harm may be minimal. However, when, as here, there is a confluence of such errors, it is possible to adversely affect subsequent research and clinical and policy decision-making. Whether the Caspi *et al.* (2003) hypothesis is true or not, we still do not know. However, finding that type of gene by environment association may be vitally important to determining the aetiology of psychiatric disorders. It is important that the methodology to detect such associations be accurate and precise.

Declaration of interest statement

The author has no competing interests.

References

Agras W.S., Walsh B.T., Fairburn C.G., Wilson G. T., Kraemer H.C. (2000) A multicenter comparison of cognitive-behavioral therapy and	interpersonal psychotherapy for bulimia nervosa. <i>Archives of General Psychiatry</i>, 57(5), 4599–4466.	Aiken L.S., West S.G. (1991) <i>Multiple Regression: Testing and Interpreting Interactions</i>, Newbury Park, CA, Sage Publications.

- Baron R.M., Kenny D.A. (1986) The Moderator-Mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**(6), 1173–1182.
- Box G.E.P., Draper N.R. (1986) *Empirical Model-building and Response Surfaces*, New York, John Wiley & Sons.
- Caspi A., Hariri A.R., Holmes A., Uher R., Moffitt T.E. (2010) Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. *American Journal of Psychiatry*, **167**(5), 509–527.
- Caspi A., Sugden K., Moffitt T.E., Taylor A., Craig I.W., Harrington H.L., McClay J., Mill J., Martin J., Braithwaite A. (2003) Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, **301**(5639), 386–389.
- Cohen J. (1995) The Earth is round ($p < .05$). *American Psychologist*, **49**(12), 997–1003.
- Cohen J., Cohen P., West S., Aiken L. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Hillsdale, NJ, Lawrence Erlbaum Associates.
- Cook R.J., Sackett D.L. (1995) The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal*, **310**(6977), 452–454.
- Cooper H., Hedges L.V. (1994) Research synthesis as a scientific enterprise. In Cooper H., Hedges L.V. (eds) *The Handbook of Research Synthesis*, pp. 4–14, New York, Russell Sage Foundation.
- Dar R., Serlin R.C., Omer H. (1994) Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Research*, **62**(1), 75–82.
- Egger M., Smith D.G. (1995) Misleading meta-analysis: lessons from “an effective, safe, simple” intervention that wasn’t. *British Medical Journal*, **310**(6982), 752–754.
- Eley T.C., Sugden K., Corsico A., Gregory A.M., Sham P., McGuffin P., Plomin R., Craig I.W. (2004) Gene-environment interaction analysis of serotonin system markers with adolescent depression. *Molecular Psychiatry*, **9**(10), 908–915.
- Eysenck H.J. (1992) Meta-analysis: sense or nonsense? *Pharmaceutical Medicine*, **6**, 113–119.
- Finney D.J. (1994) On biometric language and its abuses. *Biometric Bulletin*, **11**(4), 2–4.
- Greenland S. (1987) Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, **125**(5), 761–768.
- Greenland S. (1993) Basic problems in interaction assessment. *Environmental Health Perspectives Supplements*, **101**(Supplement 4), 59–66.
- Hedges L.V., Olkin I. (1985) *Statistical Methods for Meta-analysis*, Orlando, FL, Academic Press.
- Hunter J.E. (1997) Needed: a ban on the significance test. *Psychological Science*, **8**(1), 3–7.
- Kline R.B. (2005) *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, Washington, DC, American Psychological Association.
- Kraemer H.C. (2004) Reconsidering the Odds Ratio as a measure of 2×2 association in a population. *Statistics in Medicine*, **23**(2), 257–270.
- Kraemer H.C., Blasey C. (2004) Centring in regression analysis: a strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, **13**(3), 141–151.
- Kraemer H.C., Gardner C., Brooks J.O., Yesavage J.A. (1998) The advantages of excluding underpowered studies in meta-analysis: inclusionist versus exclusionist viewpoints. *Psychological Methods*, **3**(1), 23–31.
- Kraemer H.C., Kazdin A.E., Offord D.R., Kessler R.C., Jensen P.S., Kupfer D.J. (1997) Coming to terms with the terms of risk. *Archives of General Psychiatry*, **54**(4), 337–343.
- Kraemer H.C., Kiernan M., Essex M.J., Kupfer D.J. (2008) How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, **27**(2), S101–S108.
- Kraemer H.C., Kupfer D.J. (2006) Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, **59**(11), 990–996.
- Kraemer H.C., Mintz J., Noda A., Tinklenberg J., Yesavage J.A. (2006) Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, **63**(5), 484–489.
- Kraemer H.C., Stice E., Kazdin A., Kupfer D. (2001) How do risk factors work together to produce an outcome? Mediators, moderators, independent, overlapping and proxy risk factors. *The American Journal of Psychiatry*, **158**(6), 848–856.
- Kromrey J.D., Foster-Johnson L. (1998) Mean centering in moderated multiple regression: much ado about nothing. *Educational and Psychological Measurement*, **58**(1), 42–68.
- Lau J., Elliott M.A., Jimenez-Silva J., Kupelnick B., Mosteller F., Chalmers T.C. (1992) Cumulative meta-analysis of therapeutic trials for myocardial infarction. *The New England Journal of Medicine*, **327**(4), 248–254.
- Lewin D.I. (1996) Meta-analysis: a new standard or clinical fool’s gold? *Journal of NIH Research*, **8**, 30–31.
- Light R.J., Pillimer D.B. (1984) *Summing Up: The Science of Reviewing Research*, Cambridge, MA, Harvard University Press.
- Maxwell S. (2004) The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, **9**(2), 147–163.
- Meinert C.L. (1989) Meta-analysis: science or religion? *Controlled Clinical Trials*, **10**(4), 257–263.
- Munafo M.M., Durrant C., Lewis G., Flint J. (2009) Gene X environment interactions at the serotonin transporter locus. *Biological Psychiatry*, **65**(3), 211–219.
- Newcombe R.G. (2006) A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, **25**(24), 4235–4240.
- Nickerson R.S. (2000) Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, **5**(2), 241–301.
- Pearl J. (2000) *Causality: Models, Reasoning, and Inference*, Cambridge, Cambridge University Press.
- Risch N., Herrell R., Lehner T., Liang K.Y., Eaves L., Hoh J., Griem A., Kovacs M., Ott J., Merikangas K.R. (2009) Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *Journal of the American Medical Association*, **301**(23), 2462–2471.
- Rosenthal R. (1979) The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, **86**(3), 638–641.
- Sackett D.L. (1996) Down with odds ratios! *Evidence-Based Medicine*, **1**(6), 164–166.
- Shadish W.R., Sweeney R.B. (1991) Mediators and moderators in meta-analysis: there’s a reason we don’t let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology*, **59**(6), 883–893.
- Shapiro S. (1994) Meta-analysis/Shmeta analysis. *American Journal of Epidemiology*, **140**(9), 771–772.
- Shrout P.E. (1997) Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, **8**(1), 1–2.
- Thompson B. (1999) Journal editorial policies regarding statistical significance tests: heat is to fire as p is to importance. *Educational Psychology Review*, **11**(2), 157–169.
- Thompson S.G., Pocock S.J. (1991) Can meta-analysis be trusted? *Lancet*, **338**(8775), 1127–1130.
- Wilkinson L., The Task Force on Statistical Inference. (1999) Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, **54**(8), 594–604.