

RESEARCH

Open Access



# The case-only design is a powerful approach to detect interactions but should be used with caution

Rui Dong<sup>1</sup>, Gao T. Wang<sup>1</sup>, Andrew T. DeWan<sup>2</sup> and Suzanne M. Leal<sup>1\*</sup>

## Abstract

**Background** The case-only design is a powerful approach to identify gene × gene and gene × environment interactions for complex traits. It has been demonstrated that for the case-only design to be valid the genetic and environmental factors must be independent in the population. Additionally, there is a rare disease assumption for the case-only design, but the impact of disease prevalence and other factors, e.g., size of main effects, on type I and II error rates has not been investigated.

**Methods** Through theoretical and extensive simulation studies, we investigated type I error, power, and bias of interaction term for a wide variety of disease prevalences, main and interaction effect sizes, sample sizes, and variant and environmental exposure frequencies.

**Results** For diseases with prevalence < 4%, the case-only design usually has well controlled type I error rates and is substantially more powerful to detect interactions than the case–control design, but for higher disease prevalences both type I and II error rates can be inflated and the estimate of interaction term biased. However, when one or both main effects are large there can be inflated type I error rate even for low disease prevalences, e.g., < 1%, but if there is no or only one main effect, type I error rate is controlled regardless of the disease prevalence. Additionally, type I error rate can increase with sample size.

**Conclusions** We determined the upper bound of the disease prevalence in order not to violate the rare disease assumption for the case-only design. To verify that a case-only design study does not have increased type I error rate, the bias of the interaction term should be estimated. Although the case-only design is a powerful method to detect interactions, prevalences for some complex traits are too high to implement this method without increasing type I error rates.

**Keywords** Case-only design, Complex traits, Gene × gene and gene × environment interactions, Rare disease

## Background

Genome-wide association studies (GWAS) have detected associations for thousands of complex traits with millions of single-nucleotide variants (SNVs). Most GWAS are designed to detect main effects, while interactions either between genetic variation and environmental factors (G × E) or genetic variations (G × G) may explain “missing heritability”. In the past due to insufficient sample sizes most studies were underpowered to detect interactions, however large biobanks that include genetic and

\*Correspondence:

Suzanne M. Leal  
sml3@cumc.columbia.edu

<sup>1</sup> Center for Statistical Genetics, Gertrude H. Sergievsky Center, and the Department of Neurology, Columbia University Medical Center, New York, NY 10032, USA

<sup>2</sup> Department of Chronic Disease Epidemiology and Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health, 1 Church Street, New Haven, CT 06510, USA



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

environmental data, e.g., UK Biobank, are making it possible to potentially detect interactions [1].

One method to detect  $G \times E$  interactions, the case-only design, has received considerable attention since Piegorsch et al. proposed it [2]. It can also be applied to study  $G \times G$  interactions, aka epistasis [3]. Piegorsch et al. stated that two assumptions must be met when applying the case-only design to estimate the interaction effect: (1) the disease should be rare (2)  $G$  and  $E$  or  $G$  and  $G$  are independent. The second assumption of independence is well studied [4–6] and has been evaluated using control data [2, 7]. For the case-only design the effect of deviation from Hardy–Weinberg equilibrium (HWE) in controls was also examined [8, 9]. Additionally, it was shown that population stratification can also introduce a bias [4]. However, the rare disease assumption has not been investigated. In practice the case-only design has been applied to complex traits with a wide range of disease prevalences, including breast, prostate and ovarian cancers, Crohn’s disease, and rheumatoid arthritis [10–14]. For these studies it is not clear whether the rare disease assumption was met and if it was appropriate to use the case-only design.

Using theoretical analysis and simulation studies, we evaluated the role of disease prevalence for the case-only design, assuming  $G$  and  $E$  are independent in the overall population. We determined that there is no set disease prevalence where type I error rates are always well controlled, since not only does disease prevalence impact type I error rate, but also main effects and sample sizes and to a lesser extent variant and environmental exposure frequencies. Generally, for diseases with a prevalence of  $< 4\%$  the case-only design has well controlled type I error rate. When main effects are large, e.g., odds ratio ( $OR$ )  $> 5.0$ , the disease prevalence must be  $< 1\%$  to control type I error rates. Contrarily, when there is no or only one main effect, the disease prevalence can be high, e.g.,  $20\%$  and type I error rates are controlled. The estimate of the interaction bias can be used to evaluate if type I error rate is controlled. To facilitate this evaluation, we provide the **CaseOnly** R code, which simulates and analyses data to assess type I error rates and statistical power in case-only designs, thereby offering researchers a robust tool for detecting interactions. The case-only design is a powerful method to detect interactions, and it is advantageous to apply it when type I error rate is controlled.

## Methods

### Simulation study

Data were generated for a genetic variant under a dominant and additive model and for a binary environmental

exposure. To evaluate type I error, the genetic variant and environmental exposure have no interaction effects. The main effects for the genetic variant ranged from  $\beta_G = \ln(1.050) \approx 0.049$  to  $\beta_G = \ln(3.846) \approx 1.347$  for the dominant model, and for the additive model  $\beta_G = \ln(1.2) \approx 0.182$  for the carrier of one risk allele. For the environmental exposure the main effects ranged from  $\beta_E = \ln(1.10) \approx 0.095$  to  $\beta_E = \ln(5.00) \approx 1.609$ . We also tested when both main effects were protective [ $\beta_G = \ln(0.952) \approx -0.049$  to  $\beta_G = \ln(0.26) \approx -1.347$ ,  $\beta_E = \ln(0.909) \approx -0.095$  to  $\beta_E = \ln(0.20) \approx -1.609$ ], as well as when one main effect was protective and the other increased risk. We also evaluated type I error when neither the genetic nor environmental factor had a main effect and when only one factor had a main effect. Data were also generated under the alternative where there was an interaction,  $\beta_{G \times E} = \ln(1.20) \approx 0.182$ . The minor allele frequency (MAF) of the genetic variant and the frequency of the environmental exposure was varied between 0.05 and 0.50. We also varied the disease prevalence between 1 and 20%. Using a uniform distribution and a random number generator, genetic variant and exposure data were generated for the  $i$ th sample and a logistic regression model,

$$\text{logit}(Y_i = 1) = \beta_0 + \beta_G G_i + \beta_E E_i + \beta_{G \times E} G_i E_i \quad (1)$$

was used to assign the  $i$ th sample as a case or control. For case–control design, we generated samples of 10,000 cases and 10,000 to 30,000 controls to obtain a ratio ( $R$ ) between controls and cases ( $R = 1, 2, \text{ and } 3$ ). We also generated samples of different number of cases (2,500, 5,000, 10,000, 20,000, and 50,000) to evaluate how sample size impacts type I error rate for the case-only design, with disease prevalence levels varying from 1% to 20%.

For the case-only design, the following formula was applied to estimate the interaction effect  $\beta_{G \times E}$  where  $OR$  is odds ratio:

$$\begin{aligned} \hat{\beta}_{G \times E} &= \ln(OR_{G \times E|Y=1}) \\ &= \ln\left(\frac{P(G = 1, E = 1|Y = 1) \times P(G = 0, E = 0|Y = 1)}{P(G = 1, E = 0|Y = 1) \times P(G = 0, E = 1|Y = 1)}\right) \end{aligned} \quad (2)$$

For case–control design, a logistic regression model was used  $\text{logit}(Y_i = 1) = \beta_0 + \beta_G G_i + \beta_E E_i + \beta_{G \times E} G_i E_i$  to estimate the interaction effect  $\beta_{G \times E}$ . The p-value ( $P$ ) for both the case-only and case–control designs was computed using the Wald test for the coefficient of  $\beta_{G \times E}$  in the logistic regression model.

When estimating type I error under the null of no interaction ( $\beta_{G \times E} = 0$ ), 1,000,000 replicates were generated and analysed. Type I error was calculated for the different significance levels ( $\alpha$ ) varying from 0.001 to 0.05. Quantile–quantile (QQ) plots were also generated. For evaluating power under the alternative [ $\beta_{G \times E} = \ln(1.2)$ ],

100,000 replicates were generated and analysed. Power was estimated as the proportion of replicates with  $P < 0.05$ . For data generated under the null and alternative the distribution of  $\hat{\beta}_{G \times E}$  was obtained using all replicates.

**Theoretical and analytical studies**

We developed a theoretical framework to demonstrate the relationship between the interaction effect and the OR between G and E in the cases and controls. As stated in Piegorsch et al. [2], under the logistic model (1), the exponential value of interaction effect  $\beta_{G \times E}$  approximately equals to the OR between G and E among the cases when the disease prevalence is sufficiently low and G and E are independent in the population. Here we explicitly write the relationship as follows:

$$\exp(\beta_{G \times E}) = \frac{OR_{G,E|Y=1}}{OR_{G,E|Y=0}} \tag{3}$$

When the disease prevalence is sufficiently low,  $OR_{G,E|Y=0}$  is approximately the OR between G and E in the population (Supplementary Methods Sections 1, 2.1, and 2.2).

Besides, by comparing  $\hat{\beta}_{G \times E}$  in formula (2) and the true  $\beta_{G \times E}$  in formula (3), we show the analytical bias given the true  $\beta_G$ ,  $\beta_E$  and  $\beta_{G \times E}$  and the baseline prevalence of the disease (Supplementary Methods Sect. 2.3):

$$\begin{aligned} \text{bias} &= \hat{\beta}_{G \times E} - \beta_{G \times E} \\ &= \ln(OR_{G \times E|Y=0}) \\ &= \ln\left(\frac{P(G = 1, E = 1|Y = 0) \times P(G = 0, E = 0|Y = 0)}{P(G = 1, E = 0|Y = 0) \times P(G = 0, E = 1|Y = 0)}\right) \\ &= \ln\left(\frac{\exp(2\beta_0 + \beta_G + \beta_E) + \exp(\beta_0 + \beta_G) + \exp(\beta_0 + \beta_E) + 1}{\exp(2\beta_0 + \beta_G + \beta_E + \beta_{G \times E}) + \exp(\beta_0 + \beta_G + \beta_E + \beta_{G \times E}) + \exp(\beta_0) + 1}\right) \end{aligned} \tag{4}$$

**Results**

**Simulation studies**

**Type I error—case-only and case-control designs**

Simulation studies were used to evaluate type I error when testing for interactions ( $\beta_{G \times E}$ ) for the case-only and case-control designs. Tables 1a and Supplementary Table S1 show the type I error for the case-only and case-control designs for  $\alpha = 0.001, 0.01$  and  $0.05$  when there are main effects, i.e.,  $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$  under the dominant model. Similar results were observed for the additive model (Supplementary Table S2).

As the disease prevalence increases, type I error rate generally becomes higher in this setting and the type I error rate is inflated when disease prevalence is  $\geq 4\%$ . When disease prevalence is  $4\%$ , type I error for case-only design is  $0.001098$  and  $0.051387$  when  $\alpha = 0.001$  and  $\alpha = 0.05$ , respectively. When disease prevalence is  $20\%$ , type I

error rate is greatly inflated to  $0.002091$  and  $0.070049$ , for  $\alpha = 0.001$  and  $\alpha = 0.05$ , respectively. As observed in the QQ plot it is evident that type I error rate for the case-only design increases with increasing disease prevalence, but for the case-control design the type I error rate is well controlled even when the disease prevalence is  $20\%$  (Fig. 1 and Supplementary Fig. 1a). When one or both main effects are absent, the case-only design has well controlled type I error rate even when the disease prevalence is high, e.g.,  $20\%$  (Table 1b-d and Supplementary Figure S1b-d).

**Main effects and type I error for the case-only design**

We also evaluated type I error for a variety of main effects (Table 2a), the results suggest that if one or both main effects are strong, the disease prevalence should be  $< 4\%$  for well controlled type I error rate. If  $\beta_G$  is increased from  $\ln(1.2)$  to  $\ln(3.846)$  and  $\beta_E$  remains  $\ln(2)$ , and the disease prevalence is  $1\%$ , type I error rate increases from  $0.050132$  to  $0.056403$ . When both main effects are strong [ $\beta_G = \ln(3.846)$  and  $\beta_E = \ln(5)$ ], even when the disease prevalence is as low as  $1\%$ , type I error is very inflated, i.e.,  $0.141321$  for  $\alpha = 0.05$ . Interestingly, if the main effects are protective, the case-only design can be applied to diseases with higher prevalences without inflated type I error issues, e.g., when both main effects are strongly protective, e.g.,  $\beta_G = \ln(0.26)$  and  $\beta_E = \ln(0.2)$ , type I

error is  $0.048999$  when the disease prevalence is  $5\%$ . If the main effects are in the opposite directions but both strong, even for a disease prevalence of  $1\%$ , type I error is still inflated (Supplementary Table S3a). If both main effects are weak, then the disease prevalence can be  $> 4\%$  for well controlled type I error rate, e.g., if  $\beta_G = \ln(1.05)$  and  $\beta_E = \ln(1.1)$ , type I error is well controlled ( $0.049666$ ) even when the disease prevalence is  $20\%$ .

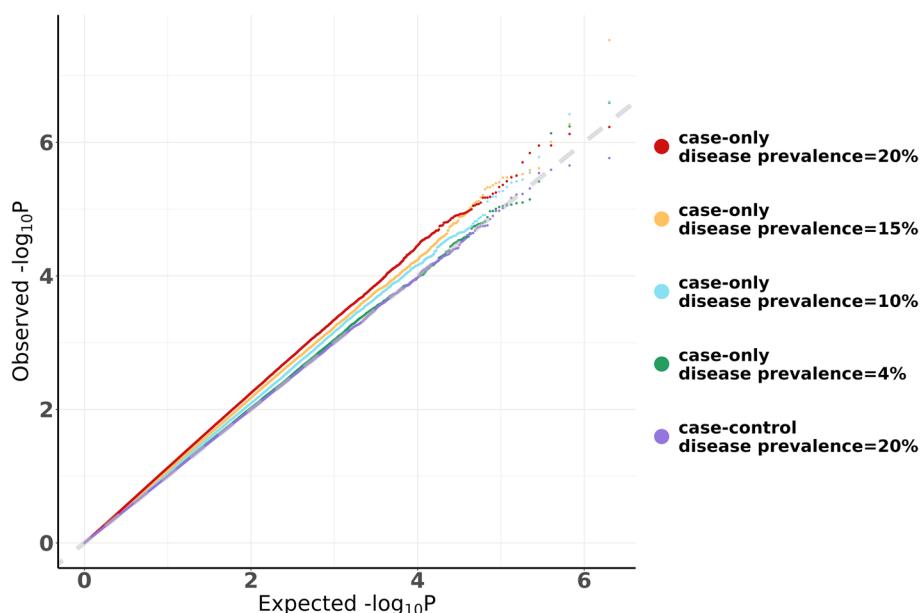
**Exposure frequencies and type I error for the case-only design**

Genetic variant and environmental exposure frequencies also affect type I error for the case-only design. Type I error rate first increases as variant and environmental exposure frequencies become greater, then decreases as the minor allele becomes the major allele or  $> 50\%$  of the population are exposed. For example, when the disease prevalence is  $5\%$  and frequency of environmental

**Table 1** Type I error rates for case-only and case-control designs

Disease prevalence	$\alpha$	1.0%	1.5%	2.0%	2.5%	3.0%	3.5%	4.0%	4.5%	5.0%	10.0%	15.0%	20.0%
<b>(a) Two main effects</b>													
case-control	0.001	0.000962	0.000970	0.001041	0.001009	0.000983	0.000958	0.000994	0.000983	0.000970	0.000976	0.000978	0.000997
case-only	0.001	0.001015	0.000942	0.000983	0.001038	0.000990	0.000995	<b>0.001098</b>	<b>0.001067</b>	<b>0.001083</b>	<b>0.001390</b>	<b>0.001667</b>	<b>0.002091</b>
case-control	0.05	0.050330	0.049824	0.049887	0.049905	0.050113	0.050161	0.049653	0.050160	0.049735	0.049845	0.049912	0.050371
case-only	0.05	0.050132	0.050603	0.049854	0.050842	0.050614	0.051195	0.051387	0.052100	0.052221	<b>0.057378</b>	<b>0.063536</b>	<b>0.070049</b>
<b>(b) Only a genetic main effect</b>													
case-control	0.001	0.001005	0.000929	0.001021	0.001002	0.000976	0.000999	0.001016	0.001030	0.000981	0.000958	0.000973	0.001068
case-only	0.001	0.001009	0.000969	0.000979	0.000990	0.000995	0.000940	0.001023	0.001030	0.001014	0.000966	0.000986	0.000984
case-control	0.05	0.049911	0.049822	0.050093	0.050012	0.049857	0.049930	0.049642	0.050276	0.050018	0.049894	0.049908	0.050146
case-only	0.05	0.050315	0.049506	0.050062	0.049971	0.049997	0.049994	0.049793	0.050205	0.049727	0.049807	0.049933	0.049700
<b>(c) Only an environment main effect</b>													
case-control	0.001	0.000998	0.000981	0.000997	0.000971	0.001026	0.000980	0.000973	0.000951	0.001022	0.001015	0.000960	0.000984
case-only	0.001	0.000952	0.000998	0.000936	0.000968	0.000975	0.001012	0.000962	0.001046	0.000995	0.000998	0.000961	0.000940
case-control	0.05	0.050296	0.050185	0.050248	0.049903	0.050320	0.049931	0.049868	0.049853	0.049658	0.049894	0.049753	0.049896
case-only	0.05	0.049769	0.049906	0.049880	0.050149	0.049980	0.050189	0.049935	0.049731	0.049396	0.049929	0.049621	0.050153
<b>(d) No main effects</b>													
case-control	0.001	0.001010	0.000981	0.000929	0.001040	0.000940	0.000984	0.000959	0.000999	0.001019	0.001012	0.000977	<b>0.001056</b>
case-only	0.001	0.000997	0.000904	0.000899	0.000930	0.000952	0.000959	0.000953	0.000959	0.001036	0.000958	0.000941	0.001027
case-control	0.05	0.049769	0.050132	0.049922	0.050361	0.050121	0.050052	0.049907	0.050104	0.050338	0.049919	0.049874	0.050470
case-only	0.05	0.049667	0.049608	0.049764	0.049975	0.050096	0.050202	0.049683	0.050152	0.050234	0.050031	0.049877	0.050120

Data were generated under the null with no interaction for four scenarios: (a) two main effects [ $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$ ]; (b) only a genetic main effect [ $\beta_G = \ln(1.2)$  and  $\beta_E = 0$ ]; (c) only an environmental main effect [ $\beta_G = 0$  and  $\beta_E = \ln(2)$ ]; and (d) no main effects [ $\beta_G = 0$  and  $\beta_E = 0$ ]. The MAF for the genetic variant is 0.2 and the frequency of the environmental exposure is 0.1. The sample sizes are  $N = 10,000$  cases for both the case-only and case-control designs and additionally for the case-control design there are  $N = 10,000$  controls. Type I error is calculated as the proportion of instances that the null hypothesis is rejected among all 1,000,000 replicates. Type I errors exceeding 5% of the alpha level are highlighted in bold in the table



**Fig. 1** Quantile–Quantile plot for the case-only and case–control designs. Data were generated under the null with no interaction when there are two main effects [ $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$ ]. The MAF for the genetic variant is 0.2 and the frequency of the environmental exposure is 0.1. A total of 1,000,000 replicates were generated. The sample sizes are  $N = 10,000$  cases for both the case-only and case–control designs and additionally for the case–control design there are  $N = 10,000$  controls

exposure is 10%, type I error is 0.049830 (MAF=0.05), 0.052221 (MAF=20%), and 0.051684 (MAF=50%), respectively. However, the impact of MAF and the environmental exposure frequency on type I error rate is limited compared to the influence of main effects and disease prevalence (Table 3a).

#### Sample size and type I error for the case-only design

We also evaluated the role of sample size plays on type I error for the case-only design. When the disease prevalence is 4% and there are two main effects [ $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$ ], type I error is 0.050325, 0.051155, and 0.058031 for 2 500, 10,000, and 50,000 cases for  $\alpha = 0.05$ . If the sample size is 20,000 the disease prevalence should be  $\leq 2\%$  for well controlled type I error rate. For a disease prevalence of 2% and a sample size of 50,000 cases, type I error is inflated (0.051977). On the other hand, if the sample size is 2 500, type I error is still well controlled (0.050102) when the disease prevalence is 5% (Table 4).

#### Statistical power—case-only and case–control designs

When there are two main effects [ $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$ ] under the dominant model with disease prevalence  $\leq 5\%$ , the case-only design has higher power than case–control designs even when  $R = 3$ . Similar

results were observed for the additive model (Supplementary Figure S2). When disease prevalence is 1%, the power for case-only and case–control ( $R = 1$ ) designs are 0.93 and 0.56, respectively, and even when disease prevalence increases to 4%, the power of case-only design (0.87) is  $1.58\times$  greater than case–control design (0.55). The power of case-only design drops substantially when disease prevalence increases to 10%, but it is not until the disease prevalence reaches 20% that case–control design ( $R = 1$ ) has greater power (Fig. 2a).

When one or both main effects are absent, the power for the case-only design is also significantly higher than for the case–control design, even though the former has a much smaller sample size, e.g.,  $N = 10,000$  for case-only and  $N = 20,000$  for case–control ( $R = 1$ ) designs. For a disease prevalence of 4%, when there is only a main genetic effect [ $\beta_G = \ln(1.2)$  and  $\beta_E = 0$ ], the power is 0.76 and 0.48 for the case-only and case–control designs, respectively. As the disease prevalence increases to 20% the power for the case-only design is still greater than for the case–control design but the gain in power is not as great, i.e., 0.56 vs. 0.47. The result is similar when there is only one main environmental effect [ $\beta_G = 0$  and  $\beta_E = \ln(2)$ ] or no main effects [ $\beta_G = 0$  and  $\beta_E = 0$ ] (Fig. 2b–d).

Higher MAF and frequency of environmental exposure may also have an impact on power. For example, when MAF remains 0.2 and disease prevalence is 4%,

**Table 2** Impact of main effects on type I error and analytical bias for the case-only design

$\beta_G$	$\beta_E$	Disease prevalence											
		1.0%	1.5%	2.0%	2.5%	3.0%	3.5%	4.0%	4.5%	5.0%	10.0%	15.0%	20.0%
<b>(a) Type I error</b>													
$\ln(3.846)$	$\ln(5)$	0.141321	0.252332	0.396712	0.551859	0.695653	0.810552	0.891836	0.942837	0.972389	0.999998	1.000000	1.000000
$\ln(3.846)$	$\ln(2)$	0.056403	0.063479	0.072751	0.085891	0.101210	0.119065	0.138098	0.160377	0.183840	0.480285	0.735506	0.873881
	$\ln(1.2)$	0.050132	0.050603	0.049854	0.050842	0.050614	0.051195	0.051387	0.052100	0.052221	0.050016	0.063536	0.070049
	$\ln(1.05)$	0.049890	0.049729	0.049665	0.049657	0.049644	0.049992	0.049896	0.049923	0.050019	0.049651	0.049992	0.049666
	$\ln(0.952)$	0.049816	0.049991	0.049856	0.049738	0.049618	0.050020	0.050054	0.049950	0.049860	0.049651	0.049931	0.050423
	$\ln(0.833)$	0.050030	0.049658	0.049459	0.049875	0.049607	0.049515	0.049339	0.049723	0.049573	0.049940	0.051270	0.052579
	$\ln(0.26)$	0.048962	0.048521	0.048793	0.048931	0.049388	0.049655	0.050008	0.050319	0.051327	0.063964	0.088547	0.125843
	$\ln(0.26)$	0.047237	0.047101	0.046880	0.047278	0.047025	0.047459	0.047611	0.048081	0.048999	0.063687	0.096078	0.152817
<b>(b) Analytical bias</b>													
$\ln(3.846)$	$\ln(5)$	-0.039119	-0.057894	-0.076150	-0.093890	-0.111119	-0.127843	-0.144066	-0.159797	-0.175042	-0.302652	-0.391307	-0.449137
$\ln(3.846)$	$\ln(2)$	-0.012563	-0.018684	-0.024697	-0.030605	-0.036407	-0.042105	-0.047698	-0.053188	-0.058576	-0.106964	-0.145864	-0.175993
	$\ln(1.2)$	-0.001667	-0.002479	-0.003277	-0.004061	-0.004831	-0.005588	-0.006330	-0.007060	-0.007775	-0.014227	-0.019476	-0.023631
	$\ln(1.1)$	-0.000048	-0.000072	-0.000095	-0.000118	-0.000141	-0.000164	-0.000186	-0.000208	-0.000230	-0.000434	-0.000612	-0.000764
	$\ln(0.952)$	-0.000044	-0.000066	-0.000087	-0.000109	-0.000130	-0.000151	-0.000171	-0.000192	-0.000212	-0.000404	-0.000574	-0.000724
	$\ln(0.833)$	-0.000928	-0.001389	-0.001847	-0.002303	-0.002756	-0.003207	-0.003655	-0.004101	-0.004544	-0.008831	-0.012836	-0.016536
	$\ln(0.26)$	-0.005304	-0.007952	-0.010597	-0.013240	-0.015878	-0.018514	-0.021146	-0.023774	-0.026399	-0.052388	-0.077767	-0.102296
	$\ln(0.26)$	-0.008789	-0.013196	-0.017611	-0.022035	-0.026467	-0.030907	-0.035354	-0.039810	-0.044273	-0.089276	-0.134820	-0.180639

Data ( $N = 10,000$  cases) were generated under the null with no interaction ( $\beta_{G \times E} = 0$ ). The MAF for the genetic variant is 0.2 and the frequency of the environmental exposure is 0.1. Type I error is calculated as the proportion of instances that the null hypothesis is rejected among all 1,000,000 replicates ( $\alpha = 0.05$ ). Analytical bias was calculated using formula (4). Type I errors exceeding 5% of the alpha level are highlighted in bold in the table



**Table 3** Impact of variant and environmental exposure frequencies on type I error and analytical bias for the case-only design

MAF	Frequency of environmental exposure	Disease prevalence											
		1.0%	1.5%	2.0%	2.5%	3.0%	3.5%	4.0%	4.5%	5.0%	10.0%	15.0%	20.0%
<b>(a) Type I error</b>													
0.05	0.10	0.049544	0.049293	0.049588	0.049806	0.049691	0.049811	0.050086	0.049976	0.049830	0.051979	<b>0.054398</b>	<b>0.056482</b>
0.20	0.05	0.049945	0.050021	0.049959	0.049814	0.050021	0.050543	0.050531	0.050543	0.051266	<b>0.054281</b>	<b>0.057564</b>	<b>0.061556</b>
0.20	0.20	0.049905	0.050096	0.050687	0.051004	0.051012	0.051657	0.051967	0.052483	<b>0.052905</b>	<b>0.060343</b>	<b>0.069782</b>	<b>0.079699</b>
0.50	0.10	0.049891	0.050474	0.050255	0.050288	0.050658	0.050898	0.050997	0.051125	0.051684	<b>0.055462</b>	<b>0.060344</b>	<b>0.065009</b>
0.50	0.50	0.049928	0.050109	0.050259	0.050434	0.049975	0.050435	0.050898	0.050810	0.050750	<b>0.055206</b>	<b>0.060188</b>	<b>0.066779</b>
<b>(b) Analytical bias</b>													
0.05	0.10	-0.00175	-0.002601	-0.003436	-0.004255	-0.005058	-0.005845	-0.006618	-0.007375	-0.008117	-0.014751	-0.020063	-0.024195
0.20	0.05	-0.001744	-0.002591	-0.003422	-0.004238	-0.005037	-0.005821	-0.006590	-0.007344	-0.008082	-0.014685	-0.019975	-0.024098
0.20	0.20	-0.001532	-0.002282	-0.003020	-0.003748	-0.004464	-0.005169	-0.005864	-0.006548	-0.007221	-0.013374	-0.018518	-0.022706
0.50	0.10	-0.001557	-0.002318	-0.003066	-0.003803	-0.004529	-0.005242	-0.005945	-0.006636	-0.007315	-0.013503	-0.018644	-0.022809
0.50	0.50	-0.00115	-0.001718	-0.002282	-0.002840	-0.003394	-0.003944	-0.004488	-0.005028	-0.005563	-0.010640	-0.015208	-0.019246

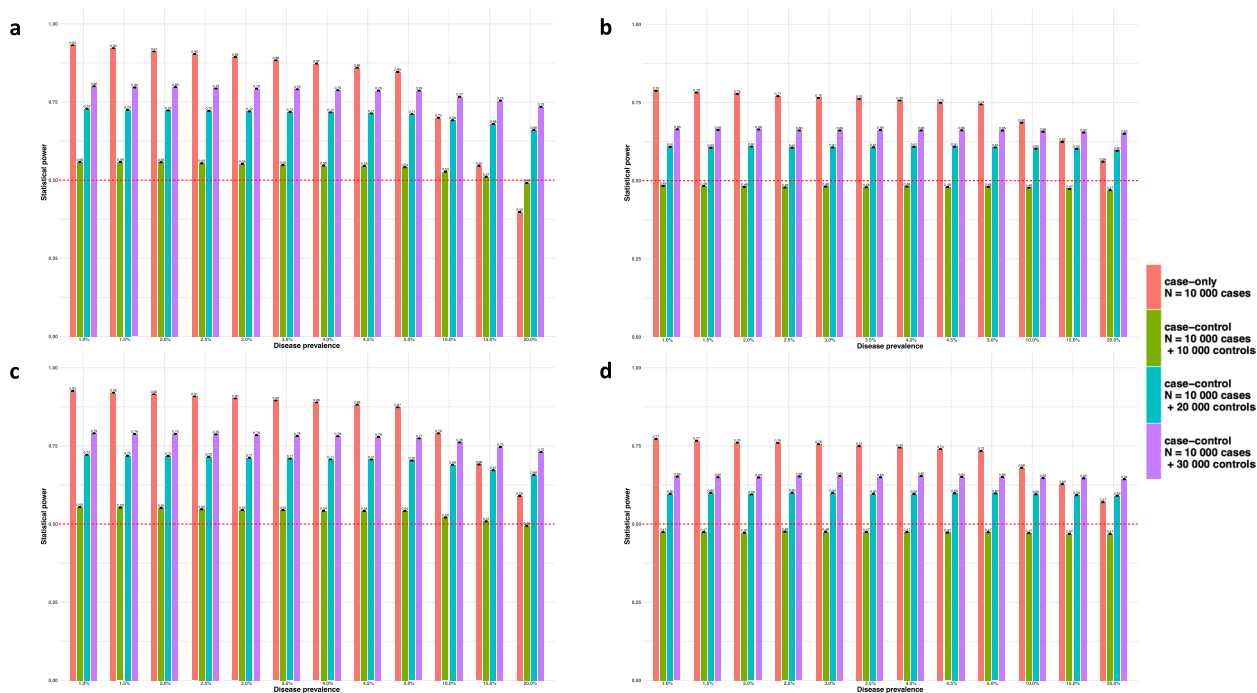
Data (N = 10,000 cases) were generated under the null with no interaction ( $\beta_G = \ln(1.2)$ ,  $\beta_E = \ln(2)$ , and  $\beta_{G \times E} = 0$ ). Type I error is calculated as the proportion of instances that the null hypothesis is rejected among all 1,000,000 replicates ( $\alpha = 0.05$ ). Analytical bias was calculated using formula (4). Type I errors exceeding 5% of the alpha level are highlighted in bold in the table

**Table 4** Impact of sample size on type I error rates for the case-only design

Number of cases	$\alpha$	Disease prevalence											
		1.0%	1.5%	2.0%	2.5%	3.0%	3.5%	4.0%	4.5%	5.0%	10.0%	15.0%	20.0%
2,500	0.001	0.000983	0.000958	0.000942	0.001019	0.000959	0.000911	0.000962	0.000956	0.001024	<b>0.001057</b>	<b>0.001097</b>	<b>0.001216</b>
	0.05	0.049557	0.049826	0.049762	0.050058	0.050021	0.049818	0.050325	0.050030	0.050102	0.051262	<b>0.053111</b>	<b>0.054364</b>
5,000	0.001	0.000996	0.001050	0.000976	0.000976	0.000972	0.001000	0.000949	0.001010	<b>0.001060</b>	<b>0.001153</b>	<b>0.001275</b>	<b>0.001413</b>
	0.05	0.049919	0.050110	0.049596	0.050317	0.050346	0.050453	0.050565	0.050731	0.050874	<b>0.053727</b>	<b>0.056344</b>	<b>0.059515</b>
10,000	0.001	0.000994	0.001004	0.001058	0.001037	<b>0.001057</b>	<b>0.001106</b>	<b>0.001115</b>	<b>0.001068</b>	<b>0.001097</b>	<b>0.001357</b>	<b>0.001731</b>	<b>0.001966</b>
	0.05	0.050155	0.050387	0.050449	0.050325	0.051138	0.050999	0.051155	0.051520	0.052183	<b>0.057479</b>	<b>0.063428</b>	<b>0.070083</b>
20,000	0.001	0.000980	0.000998	0.001012	<b>0.001052</b>	0.000992	<b>0.001114</b>	<b>0.001162</b>	<b>0.001186</b>	<b>0.001267</b>	<b>0.001838</b>	<b>0.002487</b>	<b>0.003475</b>
	0.05	0.049976	0.050759	0.050546	0.051070	0.051843	<b>0.052863</b>	<b>0.053234</b>	<b>0.053472</b>	<b>0.054632</b>	<b>0.065503</b>	<b>0.078536</b>	<b>0.090791</b>
50,000	0.001	0.001034	0.000988	<b>0.001076</b>	<b>0.001160</b>	<b>0.001208</b>	<b>0.001294</b>	<b>0.001373</b>	<b>0.001544</b>	<b>0.001647</b>	<b>0.003449</b>	<b>0.005998</b>	<b>0.008934</b>
	0.05	0.050299	0.051092	0.051977	<b>0.053040</b>	<b>0.054681</b>	<b>0.055883</b>	<b>0.058031</b>	<b>0.059773</b>	<b>0.061797</b>	<b>0.089890</b>	<b>0.122479</b>	<b>0.154596</b>

Data were generated under the null with no interaction but two main effects [ $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$ ]. The MAF for the genetic variant is 0.2 and the frequency of the environmental exposure is 0.1. Type I error is calculated as the proportion of instances that the null hypothesis is rejected among all 1,000,000 replicates. Type I errors exceeding 5% of the alpha level are highlighted in bold in the table





**Fig. 2** Statistical power of the case-only and case-control designs Data were generated under the alternative with  $\beta_{G \times E} = \ln(1.2)$  for (a) two main effects [ $\beta_G = \ln(1.2)$  and  $\beta_E = \ln(2)$ ]; (b) only a genetic main effect [ $\beta_G = \ln(1.2)$  and  $\beta_E = 0$ ]; (c) only an environmental main effect [ $\beta_G = 0$  and  $\beta_E = \ln(2)$ ]; and (d) no main effects ( $\beta_G = 0$  and  $\beta_E = 0$ ). The MAF for the genetic variant is 0.2 and the frequency of the environmental exposure is 0.1. Power is calculated as the proportion of instances that the null hypothesis is rejected among all 100,000 replicates ( $\alpha = 0.05$ )

the power for the case-only design is 0.65, 0.87, 0.97 for frequency of environmental exposure being 0.05, 0.10, 0.20, respectively. However, when MAF increases from 0.05 to 0.20 to 0.50 with frequency of environmental exposure fixed as 10%, under the dominant model, the power will first increase from 0.53 to 0.87, then decrease to 0.73 (Supplementary Figure S3).

**Bias—analytical results**

In formula (4), we derived the analytical bias of  $\hat{\beta}_{G \times E}$  given the baseline risk ( $\frac{1}{1+e^{-\beta_0}}$ ), main effects ( $\beta_G$  and  $\beta_E$ ), and true interaction effect ( $\beta_{G \times E}$ ). Under the null, when one or both main effects are absent, the bias of  $\hat{\beta}_{G \times E}$  equals 0. However, when the main effects either both increase or decrease the risk of developing disease,  $\hat{\beta}_{G \times E}$  underestimates the true interaction effect  $\beta_{G \times E}$ , leading to a negative bias of  $\hat{\beta}_{G \times E}$ , and the bias is greater with higher disease prevalence. For example, when  $\beta_G = \ln(1.2)$ ,  $\beta_E = \ln(2)$  for the following disease prevalences 1%, 4%, and 20% the biases are  $-0.001667$ ,  $-0.006330$ , and  $-0.023631$ , respectively. Stronger main effects also lead to greater bias of  $\hat{\beta}_{G \times E}$ , e.g., when both main effects are strong [ $\beta_G = \ln(3.846)$  and  $\beta_E = \ln(5)$ ], the bias of  $\hat{\beta}_{G \times E}$  reaches  $-0.144066$  even when the disease prevalence is only 4%. When main effects ( $\beta_G$  and  $\beta_E$ ) are in opposite directions,  $\hat{\beta}_{G \times E}$  for the case-only

design overestimates  $\beta_{G \times E}$  (a positive bias), with the bias increasing as the disease prevalence increases. As long as there are two non-zero main effects,  $\hat{\beta}_{G \times E}$  is biased (Table 3b, Supplementary Table S3, Supplementary Figure S4 and Supplementary Sect. 2.3).

Under the alternative [ $\beta_{G \times E} = \ln(1.2)$ ], even when there are no main effects, the bias of  $\hat{\beta}_{G \times E}$  is negative, indicating that the case-only design underestimates the interaction effect. When  $\beta_G = \ln(1.2)$ ,  $\beta_E = \ln(2)$  and  $\beta_{G \times E} = \ln(1.2)$ , the bias is  $-0.005613$ ,  $-0.021628$ , and  $-0.089011$  for disease prevalences of 1%, 4%, and 20%, respectively. The bias of  $\hat{\beta}_{G \times E}$  is smaller when  $\beta_G$  and  $\beta_E$  are in opposite directions compared to when  $\beta_G$  and  $\beta_E$  are either both positive or negative, e.g., when  $\beta_G = \ln(1.2)$ ,  $\beta_{G \times E} = \ln(1.2)$ , and the disease prevalence is 20%, the bias is  $-0.089011$  when  $\beta_E = \ln(2)$ , but only  $-0.008341$  when  $\beta_E = \ln(0.5)$ . However, the bias still increases as the disease becomes more prevalent regardless of the main effects (Supplementary Table S4 and Supplementary Figure S5).

The bias of  $\hat{\beta}_{G \times E}$  is closely related to type I and II error rates for the case-only design. Under the null, when there are two main effects the bias of  $\hat{\beta}_{G \times E}$  always leads to an increase in type I error rate, and  $\hat{\beta}_{G \times E}$  for the case-only design either underestimates or overestimates  $\beta_{G \times E}$ . For the case-only design, as the disease

prevalence increases, the bias of  $\hat{\beta}_{G \times E}$  is greater, leading to higher type I error rates. The case-only design may suffer from inflated type I error rate if the absolute value of the analytical bias of  $\hat{\beta}_{G \times E} > 0.006$  (both main effects increase risk),  $> 0.01$  (one main effect increases and the other decreases risk), and  $> 0.025$  (both main effects are protective) (Tables 3 and 4). Under the alternative, if the bias of  $\hat{\beta}_{G \times E}$  is in the same direction as  $\beta_{G \times E}$ , this will boost the power of the case-only design. However, it should be noted even when  $\beta_{G \times E}$  and the bias of  $\hat{\beta}_{G \times E}$  are in opposite directions, e.g.,  $\beta_{G \times E}$  is positive and the bias of  $\hat{\beta}_{G \times E}$  is negative, the power for the case-only design is greater than when a case-control sample is analysed for disease prevalences  $< 15\%$  (Fig. 2 and Supplementary Table S4).

## Discussion

We evaluated the case-only design through theoretical and simulation analysis, and showed that several factors including disease prevalence, main effects, variant and environmental exposure frequencies, and sample size may impact bias of the interaction term and type I and II error rates. While previous studies stated that for the case-only design a rare disease assumption is necessary, they often lack clarity on what constitutes a "rare" disease. Our simulations investigated various disease prevalence thresholds, from 1% to 20%, to assess their effects on type I error rates and bias in estimating interaction terms. Compared with the conventional case-control design, the power of the case-only design can be a magnitude greater, but it should only be applied when type I error rate is controlled. Generally, under the assumption of independence between G and E, for disease prevalences  $< 4\%$  type I error rate of the case-only design is not inflated, but for higher disease prevalences type I error rate can be high and the estimation of interaction effect biased. When disease prevalence is  $> 20\%$  the power for the case-only design can be lower than analysing a case-control sample. When one or both main effects are absent, the disease prevalence does not impact type I error rate. However, caution is required since there may be a failure to detect main effects even when they exist. The analytical bias can be calculated using formula (4), to aid in evaluating if the case-only design is appropriate to use. Besides the analytical bias, when the sample size is large ( $> 10,000$  cases), the case-only design requires lower disease prevalence to avoid inflation in type I error rate.

Stronger main effects may lead to a greater bias of the interaction estimate and higher type I error rate. Though it is uncommon for complex traits to have a genetic risk factor with  $OR > 1.5$  there are several environmental exposures that have large main effects [15–17]. We

recommend applying the case-only design to test for interactions for complex traits with low prevalences (e.g., ovarian cancer and celiac disease), and limit testing of interactions to variants and environmental exposures that do not have strong main effects.

Because of the bias of  $\hat{\beta}_{G \times E}$  may be in the same or opposite direction of  $\hat{\beta}_{G \times E}$  itself, type I and II error rates can both increase with increasing disease prevalence. Therefore, for the case-only design unlike most statistical tests as type I error rate increases power decreases.

Yang et al. claimed that no assumption about disease prevalence is required stating that the cross-product term of the  $2 \times 2$  table (presence or absence of the risk allele for genes 1 and 2 among cases) measures the departure from the multiplicative joint effects of relative risk aka the risk ratio ( $RR$ ) (not  $OR$ ) [3]. The  $RR$  is used for observational cohort studies that have incidence cases and only approximate the  $OR$  for diseases with low prevalence. Therefore, using  $RR$  to define interaction is inaccurate and both  $\hat{\beta}_{G \times E}$  estimated by the case-only design and the "interaction" defined by  $RR$  have high type I error rates when the disease is prevalent. In fact, the case-only design measures the "interaction" effect defined by  $RR$  (Supplementary Methods Sect. 2.4) and they are both biased estimates of  $\beta_{G \times E}$  (Supplementary Figures S4 and S5).

In addition to multiplicative interaction, there are other types of interactions such as additive [18] and sufficient-cause interactions [9]. Interaction on an additive scale, often uses the index of relative excess risk due to interaction, which should be used with caution due to use of the relative risk as discussed in the paragraph above. Sufficient-cause interaction comprises a set of conditions or events that lead to a specific outcome, that is equivalent to the scenario when the main effects for G and E are both absent.

## Conclusions

Although the case-only design is a powerful method to detect interaction due to increased type I error rates for a variety of scenarios, e.g., main effects, allele and environmental exposure frequencies are not sufficiently low, type I error rate should be evaluated. Our research contributes to the existing literatures by establishing clear guidelines on the acceptable thresholds for disease prevalence that maintain type I error rates under control. This can be done by analytically examining the bias [formula (4)] and performing simulation studies by implementing the R code, CaseOnly, that simulates and analyses data to evaluate type I error rate. CaseOnly can also be used to evaluate power. Although the case-only design is a powerful method to detect interactions it should be used with caution.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-025-11318-1>.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

### Acknowledgements

Not applicable.

### Authors' contributions

RD, GTW and SML designed the theoretical and simulation analysis, performed data interpretation, and RD and SML drafted the manuscript. SML directed the study's implementation. ATD and SML obtained the funding. All authors provided critical review for important intellectual content of the manuscript and approved the final manuscript for submission.

### Funding

This work was supported by grant R01DC017712 to SML and ATD from National Institute and other Communications Disorders (NIDCD).

### Data availability

The source code used to simulate and analyze data in this study is available in the CaseOnly repository (<https://github.com/RuiDongDR/CaseOnly>), and the simulations can be replicated through running the Rscript CaseOnly.R.

### Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare no competing interests.

Received: 8 May 2024 Accepted: 3 February 2025

Published online: 06 March 2025

## References

- Wang H, Zhang F, Zeng J, Wu Y, Kemper KE, Xue A, et al. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci Adv*. 2019;
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13:153–62.
- Yang Q, Khoury MJ, Sun F, Flanders WD. Case-Only Design to Measure Gene-Gene Interaction: *Epidemiology*. 1999;10:167–70.
- Wang L-Y, Lee W-C. Population Stratification Bias in the Case-Only Study for Gene-Environment Interactions. *Am J Epidemiol*. 2008;168:197–201.
- Gatto NM. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. *Int J Epidemiol*. 2004;33:1014–24.
- Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, et al. Using Principal Components of Genetic Variation for Robust and Powerful Detection of Gene-Gene Interactions in Case-Control and Case-Only Studies. *Am J Hum Genet*. 2010;86:331–42.
- Khoury MJ, Flanders WD. Nontraditional Epidemiologic Approaches in the Analysis of Gene Environment Interaction: Case-Control Studies with No Controls! *Am J Epidemiol*. 1996;144:207–13.
- Lee W-C, Wang L-Y, Cheng KF. An easy-to-implement approach for analyzing case-control and case-only studies assuming gene-environment independence and hardy-weinberg equilibrium. *Stat Med*. 2010;29:2557–67.
- Lee W-C. Testing for sufficient-cause gene-environment interactions under the assumptions of independence and hardy-weinberg equilibrium. *Am J Epidemiol*. 2015;182:9–16.
- Lash TL, Bradbury BD, Wilk JB, Aschengrau A. A case-only analysis of the interaction between N-acetyltransferase 2 haplotypes and tobacco smoke in breast cancer etiology. *Breast Cancer Res*. 2005;7:R385.
- Neslund-Dudas C, Levin AM, Rundle A, Beebe-Dimmer J, Bock CH, Nock NL, et al. Case-only gene-environment interaction between ALAD tagSNPs and occupational lead exposure in prostate cancer. *Prostate*. 2014;74:637–46.
- Liu G, Mukherjee B, Lee S, Lee AW, Wu AH, Bandera EV, et al. Robust Tests for Additive Gene-Environment Interaction in Case-Control Studies Using Gene-Environment Independence. *Am J Epidemiol*. 2018;187:366–77.
- Helbig KL, Nothnagel M, Hampe J, Balschun T, Nikolaus S, Schreiber S, et al. A case-only study of gene-environment interaction between genetic susceptibility variants in NOD2 and cigarette smoking in Crohn's disease aetiology. *BMC Med Genet*. 2012;13:14.
- Clarke GM, Pettersson FH, Morris AP. A comparison of case-only designs for detecting gene  $\times$  gene interaction in rheumatoid arthritis using genome-wide case-control data in Genetic Analysis Workshop 16. *BMC Proc*. 2009;3:S73.
- Tuomi T, Huuskonen MS, Virtamo M, Tossavainen A, Tammilehto L, Mattson K, et al. Relative risk of mesothelioma associated with different levels of exposure to asbestos. *Scand J Work Environ Health*. 1991;17:404–8.
- Van Hylckama VA, Helmerhorst FM, Vandembroucke JP, Doggen CJM, Rosendaal FR. The venous thrombotic risk of oral contraceptives, effects of oestrogen dose and progestogen type: results of the MEGA case-control study. *BMJ*. 2009;339:b2921–b2921.
- Bloemenkamp KWM, Rosendaal FR, Büller HR, Helmerhorst FM, Colly LP, Vandembroucke JP. Risk of Venous Thrombosis With Use of Current Low-Dose Oral Contraceptives Is Not Explained by Diagnostic Suspicion and Referral Bias. *Arch Intern Med*. 1999;159:65.
- Lee W-C. Sample size calculations for additive interactions. *Epidemiology*. 2013;24:774.

## Publisher's Note

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