Quantifying the impact of association of environmental mixture in a type 1 and type 2 error balanced framework

Vishal Midya*[1], Jiangang Liao[2], Chris Gennings[1], Elena Colicino[1], Susan L. Teitelbaum[1], Robert O. Wright[1] and Damaskini Valvi[1]

[1]Department of Environmental Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[2]Division of Biostatistics and Bioinformatics, Pennsylvania State College of Medicine, Hershey, PA, 17033

Abstract

In environmental epidemiology, analysis of environmental mixture in association to health effects is gaining popularity. Such models mostly focus on inferences of hypotheses or summarizing strength of association through regression coefficients and corresponding estimates of precision. Nonetheless, when a decision is made against alternative hypothesis, it becomes increasingly difficult to tease apart whether the decision is influenced by sample size or represents genuine absence of association and whether the result warrants further investigation. Similarly, in case of a decision made in favour of alternative hypothesis, a significant association may indicate influence of large sample and not a strong effect. Moreover, the disparate type 1 and type 2 errors, might render these inferences unreliable. Using Cohen's f^2 to evaluate the strength of explanatory associations in a more fundamental way, we herein propose a new concept, optimal impact, to quantify the maximum explanatory association solely contributed by an environmental-mixture after controlling for confounders and covariates such that the type 2 error remains at its minimum.

*vishal.midya@mssm.edu

optimal impact is built upon a novel hypothesis testing procedure in which the rejection region is determined in a way that type 1 and type 2 errors are balanced. Even when an association does not achieve statistical significance, its optimal impact might deem it meaningful and strong enough for further investigation. This idea was naturally extended to estimate sample size in designing studies by striking a balance between explanatory precision and utility. The properties of this framework are carefully studied and detailed results are established. A straightforward application of this procedure is illustrated using an exposure-mixture analysis of per-and-poly-fluoroalkyl substances and metals with serum cholesterols using data from 2017–2018 US National Health and Nutrition Examination Survey.

Keywords: Cohen's f^2, Exposure mixture models, Environmental Epidemiology

1 Introduction

There has been a welcome surge of interest in estimating health effects of exposure-mixture in environmental epidemiology (Bobb et al. (2014), Carrico et al. (2015), Colicino et al. (2020), Keil et al. (2020), Wheeler et al. (2021), Ferrari and Dunson (2021)). These developments are certainly promising but most of these methods use traditional null hypothesis significance testing (NHST) for exposure mixture-outcome associations. However, NHST has been severely criticized as a contributor to replication crisis in psychology, and biomedical sciences (Nakagawa and Cuthill (2007), Szucs and Ioannidis (2017)). NHST may contribute to selective reporting and subjectivity since it does not require us to designate what the data under alternative hypothesis should predict. Even for large sample sizes, it guarantees that any irrelevant and tiny effect sizes are detectable (Ioannidis et al. (2014), Wasserstein and Lazar (2016)). Additionally, the strength of association in these models is determined through regression coefficients and precision estimates like standard errors and p-values. Therefore as a direct consequence, when a decision is made against alternative hypothesis, it becomes increasingly difficult to tease apart whether the decision is influenced by sample size or it genuinely represents an absence of association. Similarly, in case of a decision made in favour of the alternative hypothesis, a significant association might indicate the influence of large sample and not a strong effect. In addition, the dependence on sample size and disparate type 1 and type 2 errors, further complicates reliability of any inference.

To circumvent such issues, researchers are starting to report in-sample scale-independent partial R^2 or F

type statistics along with regression estimates to indicate strength of explanatory association to quantify the effect of environmental mixture on health outcomes. But simply reporting these statistics does not alleviate the curses of NHST or imbalances of type 1 and type 2 errors. A long established index to report strength of explanatory association in a more fundamental way is Cohen's f^2 (Cohen (1988)), which evaluates the impact of additional variables in the context of multiple linear regression. Through the past three decades, Cohen's f^2 continues to be extensively used in behavioral sciences, sociology and biomedical sciences, due to its immense practical utility and ease of interpretation.

In this paper, we propose herein optimal impact using Cohen's f^2 to evaluate strength of explanatory association in a more fundamental and scale-independent way, by quantifying the maximum explanatory association solely contributed by an environmental-mixture on top of confounders and covariates such that the type 2 error remains at its minimum. optimal impact is built upon a novel hypothesis testing procedure in which the rejection region is determined in a way that type 1 and type 2 errors are balanced and both exponentially diminish to 0 as $n \to \infty$, under a meaningful deviation from null (and not just any deviation). This is similar to formulating a medical diagnostic test in which the threshold is adjusted to balance sensitivity and specificity (Zhou et al. (2011)). Utilizing the nuances in optimal impact, we also shed light on sample size estimation in designing time and cost effect studies from the perspective of explanatory power.

In subsections 2.1 and 2.2, we discuss Cohen's f^2 in linear and generalized linear models. Next, in 2.3, we develop the framework of type 1 and type 2 calibrated hypothesis testing. In subsection 2.4, we discuss theoretical implications of this hypothesis testing framework and consequently in subsection 2.5 we develop the concept of optimal impact. In section 3 we present a simulated example for illustration. In section 4, we estimate optimal impact of per-and-poly-fluoroalkyl substances and metals for serum cholesterols based on data from 2017–2018 US National Health and Nutrition Examination Survey (NHANES). Finally, we end this paper with a discussion.

2 Methods

Consider a common problem of testing if an exposure-mixture in a regression model is associated with outcome after adjusting for covariates or confounders. For example, consider linear model , $\boldsymbol{y} = \boldsymbol{X_0 b_0} +$

$\boldsymbol{X_1 b_1} + \varepsilon$, and we are interested to know the impact of the association of $\boldsymbol{X_1}$ after adjusting for $\boldsymbol{X_0}$ and formulate the hypothesis,

$$\begin{aligned} &H_0 : \text{Additional impact of association} = 0 \quad \text{vs.} \\ &H_1 : \text{Additional impact of association} = \delta \end{aligned} \tag{1}$$

where δ is a pre-defined meaningful quantity and $\delta > 0$. For example, Selya et al. (2012) reports that after controlling for gender and smoking quantity, the additional impact of the association between the outcome (nicotine dependence) and exposure (smoking frequency) is found to be 0.32.

Let, $S(\boldsymbol{y})$ be a test statistic based on observed data $\boldsymbol{y}$ and T be a type 1 and type 2 error calibrated cutoff which depends on sample size n and unknown parameters p_1 and effect size δ. Then one can define a testing procedure by its type 1 and type 2 errors as below

$$\begin{aligned} &\text{type 1 error} = P(S(\boldsymbol{y}) > \text{T}|\text{Additional impact} = 0) \\ &\text{type 2 error} = P(S(\boldsymbol{y}) < \text{T}|\text{Additional impact} = \delta). \end{aligned} \tag{2}$$

2.1 Cohen's f^2 in Linear Regression

Consider standard multiple linear regression model with error $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$, where $\boldsymbol{I_n}$ is an identity matrix of dimension $n \times n$. Let $\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_0}}$ be the maximum likelihood estimate (MLE) for model with only design matrix $\boldsymbol{X_0}$ whereas $\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_1}}$ and $\hat{\boldsymbol{b}}_{\boldsymbol{1},\boldsymbol{H_1}}$ be the MLEs for the model with design matrices $\boldsymbol{X_0}$ and $\boldsymbol{X_1}$. The standard test to compare a null and alternative is through F statistic, $F(\boldsymbol{y}) = \frac{(\text{SSR}_0 - \text{SSR}_1)/p_1}{\text{SSR}_1/(n-p_0-p_1)}$, where $\text{SSR}_0 = (\boldsymbol{y} - \boldsymbol{X_0}\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_0}})^t(\boldsymbol{y} - \boldsymbol{X_0}\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_0}})$ is the sum of squared errors under H_0 and $\text{SSR}_1 = (\boldsymbol{y} - \boldsymbol{X_0}\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_1}} - \boldsymbol{X_1}\hat{\boldsymbol{b}}_{\boldsymbol{1},\boldsymbol{H_1}})^t(\boldsymbol{y} - \boldsymbol{X_0}\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_1}} - \boldsymbol{X_1}\hat{\boldsymbol{b}}_{\boldsymbol{1},\boldsymbol{H_1}})$ is the sum of squared errors under H_1. Then $F(\boldsymbol{y}) \sim F_{p_1,n-p_0-p_1}(\gamma_n)$, where p_1 and $n - p_0 - p_1$ are the degrees of freedom and γ_n is the non-centrality parameter. As $n \to \infty$ while p_0, p_1 remain fixed, this F distribution can be approximated by chi-squared distribution, $\lim_{n\to\infty} p_1 F(\boldsymbol{y}) \sim \chi^2_{p_1}(\gamma_n)$. The non-centrality parameter γ_n equals 0 when $\boldsymbol{y}$ is generated under H_0. When $\boldsymbol{y}$ is generated under the alternative, γ_n has the form of $\gamma_n = \frac{\|(\boldsymbol{I_n} - \boldsymbol{P_{X_0}})\boldsymbol{X_1 b_1}\|^2}{\sigma^2}$, where $\boldsymbol{P_{X_0}} = \boldsymbol{X_0}\left(\boldsymbol{X_0^t X_0}\right)^{-1}\boldsymbol{X_0^t}$ is the projection matrix on to the linear space spanned by the column vectors

of $\boldsymbol{X_0}$ (Wilks (1938), Brown et al. (1999)) (Section S.1 of the supplementary material). γ_n quantifies the additional impact in y due to X_1 relative to the error variance σ^2. For the common regression design in which the predictor vector of each subject is drawn from a common population, γ_n grows linearly on n. Note that γ_n does not depend on $\boldsymbol{y}$ but depends on the design matrix $\boldsymbol{X}$ and underlying parameter $\boldsymbol{b_1}$ and σ^2. A long established index of quantifying additional impact in linear regression is Cohen's f^2,

$$f^2 = \frac{R^2_{\boldsymbol{y},\boldsymbol{X_0},\boldsymbol{X_1}} - R^2_{\boldsymbol{y},\boldsymbol{X_0}}}{1 - R^2_{\boldsymbol{y},\boldsymbol{X_0},\boldsymbol{X_1}}},$$

where $R^2_{\boldsymbol{y},\boldsymbol{X}_0,\boldsymbol{X}_1}$ and $R^2_{\boldsymbol{y},\boldsymbol{X}_0}$ are the squared multiple correlation for $\boldsymbol{X_0}, \boldsymbol{X_1}$ under H_1 and $\boldsymbol{X_0}$ under H_0 respectively. The f^2 quantifies the proportion of impact in $\boldsymbol{y}$ accounted by $\boldsymbol{X_1}$ on top of the impact accounted by $\boldsymbol{X_0}$, a concept that most researcher can relate to intuitively (Selya et al. (2012)).We then establish the following Lemma to connect Cohen's f^2 and non-centrality parameter γ_n in linear regression.

Lemma 1.

$$f^2 \xrightarrow[n\to\infty]{P} \frac{\gamma_n}{n}.$$

The proof is presented in Section S.2 of the supplementary material. We can borrow the common convention for f^2 (Cohen (1988)) and call $f^2 \geq 0.02$, $f^2 \geq 0.15$ and $f^2 \geq 0.35$ as representing small, moderate and large effect size respectively. This can serve as the guidance in understating the effect size obtained from the data.

2.2 Cohen's f^2 in Generalized Linear Models.

Now consider a generalized linear model (McCullagh and Nelder (1989)). Let, the outcome variable $\boldsymbol{y}$ follows the exponential family, $\exp[\{\boldsymbol{y}\theta - b(\theta)\}/a(\phi) + c(\boldsymbol{y}, \boldsymbol{\phi})]$. Let $\mu = E(\boldsymbol{y})$ and $g(.)$ be the link function. μ is related to the canonical parameter θ through the function $\mu = b'(\theta)$, where b' denotes the first derivative of b. Then the model is completed by $\eta = g(\mu) = \boldsymbol{X_0 b_0} + \boldsymbol{X_1 b_1}$. Here the standard test is likelihood ratio test for testing of hypothesis, $\Lambda(\boldsymbol{y}) = 2\left\{\ell(\hat{\boldsymbol{b}}_{\boldsymbol{1},\boldsymbol{H_1}}, \hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_1}} | \boldsymbol{X_0}, \boldsymbol{X_1}) - \ell(\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_0}} | \boldsymbol{X_0})\right\}$. As the sample size $n \to \infty$, the likelihood ratio statistic $\Lambda(\boldsymbol{y})$ follows a central chi-squared distribution $\chi^2_{p_1}$ with p_1 degrees of freedom, when $\boldsymbol{y}$ is generated under the model in H_0. $\Lambda(\boldsymbol{y})$ follows a non-central chi-squared distribution $\chi^2_{p_1}(\gamma_n)$

with degrees of freedom p_1 and non-centrality parameter γ_n, when $\boldsymbol{y}$ is generated under H_1. However, there is no simple and explicit form for γ_n. Self et al. (1992) and Shieh (2000) defined γ_n in likelihood ratio test for generalized linear models as $\gamma_n := E_{\boldsymbol{y}\sim H_1}\{\Lambda(\boldsymbol{y})\}$. Similar to the non-centrality parameter in linear regression, γ_n grows linearly with n since $E_{\boldsymbol{y}\sim H_1}\{\Lambda(\boldsymbol{y})\} = n\Big\{2a^{-1}(\phi)\,\{b'(\theta)[\theta-\theta^*]-[b(\theta)-b(\theta^*)]\}\Big\}+o(1)$ (where θ and θ^* denote the canonical parameter values evaluated at $(\boldsymbol{b_0},\boldsymbol{b_1})$ and $(\boldsymbol{b_0^*},\boldsymbol{b_1}=\boldsymbol{0})$ and $\boldsymbol{b_0^*}$ is the limiting value of $\hat{\boldsymbol{b}}_{\boldsymbol{0},\boldsymbol{H_0}}$ as described in equation (2.2) of Self and Mauritsen (1988) (see Cordeiro (1983) and section 2 of Shieh (2000) for a detailed derivation)).

Consider the adjusted coefficient of determination for generalized linear models as shown below (note that the definition of squared multiple correlation is generally accepted for linear regression but it is not directly applied to a generalized linear models (for example a logistic regression) and there remains a lack of general consensus discussed in literature (see Liao and McGee (2003) for more details)), $R_l^2 = 1 - \frac{\ell(\boldsymbol{y}|\text{ Any predictor }\boldsymbol{X})}{\ell(\boldsymbol{y}|\text{ Intercept Only Model})}$. Then Cohen's f^2 in generalized linear models can be written as,

$$f^2 = 2\left\{\frac{R^2_{l,H_1}-R^2_{l,H_0}}{1-R^2_{l,H_1}}\right\} := 2\left\{\frac{\ell(\boldsymbol{b_{1,H_1}},\boldsymbol{b_{0,H_1}}|\boldsymbol{X_0},\boldsymbol{X_1})-\ell(\boldsymbol{b_{0,H_0}}|\boldsymbol{X_0})}{\ell(\boldsymbol{b_{1,H_1}},\boldsymbol{b_{0,H_1}}|\boldsymbol{X_0},\boldsymbol{X_1})}\right\}, \tag{3}$$

where, R^2_{l,H_0} and R^2_{l,H_1} are the adjusted coefficient of determinations under the null and the alternative respectively. Similar to Lemma 1, we connect f^2 and γ_n in generalized linear models as below,

Lemma 2.

$$f^2 \xrightarrow[n\to\infty]{P} \frac{\gamma_n}{O(n)}.$$

For simplicity and keeping similarity with Lemma 1 the $O(n)$ is not expanded further. The proof and more details are presented in Section S.3 of the supplementary material. Both the Lemma (1) and (2) help connect the non-centrality parameters and Cohen's f^2, which will be utilized later in the following sections.

2.3 Constructing a type 1 and type 2 error balanced hypothesis testing framework

2.3.1 Formulation of Error Calibrated Cutoff.

Let the test statistic be $S(\boldsymbol{y}) = p_1 F(\boldsymbol{y})$ for a linear regression and $S(\boldsymbol{y}) = \Lambda(\boldsymbol{y})$ for other generalized linear models as in Section 2.1. For a given cutoff T, the type 1 error and type 2 error are given by

$$\begin{aligned} \alpha(\mathrm{T}) &= P\left(S(\boldsymbol{y}) > \mathrm{T} \mid H_0\right) = P\left(\chi^2_{p_1} > \mathrm{T}\right)(1 + o(1)) \\ \beta(\mathrm{T}) &= P\left(S(\boldsymbol{y}) < \mathrm{T} \mid H_1\right) = P\left(\chi^2_{p_1}(\gamma_n) < \mathrm{T} | \gamma_n = n\delta\right)(1 + o(1)), \end{aligned} \tag{4}$$

where $\chi^2_{p_1}$ denotes central chi-squared random variable with p_1 degrees of freedom and $\chi^2_{p_1}(\gamma_n)$ denotes non-central chi-squared random variable with p_1 degrees of freedom and γ_n as the non-centrality parameter. Our central idea is to choose T so that type 1 error α and the type 2 error β satisfy the relationship, $\alpha(\mathrm{T}) = \beta(\mathrm{T})$. Using the chi-square approximation to test statistic $S(\boldsymbol{y})$, we can solve for the calibrated cutoff T by equation

$$P\left(\chi^2_{p_1} > \mathrm{T} | \gamma_n = 0\right) = P\left(\chi^2_{p_1}(\gamma_n) < \mathrm{T} | \gamma_n = n\delta\right). \tag{5}$$

When T is fixed, the left size of equation (5) remains constant as $n \to \infty$ while the right side diminishes to 0 rapidly under non-centrality parameter $n\delta$. Therefore, equation (5) implies $\mathrm{T} \to \infty$ as $n \to \infty$. In the Theorem stated below we elaborate more on T. The results in theorem 1 depend on the normality approximation of the non-central chi-square distribution, i.e. for large n, equation (5) was rewritten as,

$$P\left(\chi^2_{p_1} > \mathrm{T} | \gamma_n = 0\right) = \Phi\left(\frac{\mathrm{T} - p_1 - n\delta}{\sqrt{2(p_1 + 2n\delta)}}\right) + o(1). \tag{6}$$

Theorem 1. *Consider the hypothesis of interest to be $H_0 : f^2 = 0$ vs. $H_1 : f^2 \geq \delta$, where f^2 denotes Cohen's f^2. Assume data $\boldsymbol{y}$ is generated under the alternative with $f^2 = \delta$. Then following the constraint $\alpha = \beta$ as in (5) and for large n, the error calibrated cutoff T has the following expression,*

$$T = \left(\frac{\delta n}{2K - 1} + c_1 n^{\frac{1}{2k}}\right)(1 + o(1)) \tag{7}$$

Further, the type 1 error (α) *or the type 2 error* (β) *rates can be expressed as,*

$$\frac{d}{dn}\log\{\alpha(T)\} = \frac{d}{dn}\log\{\beta(T)\} = -\frac{\delta(K-1)^2}{2(2K-1)^2} + o(1) \tag{8}$$

where, $K \to (2+\sqrt{2})(1+o(1))$ *and* c_1 *is a constant of integration.*

The proof is presented in Section S.4 of the supplementary material. Theorem 1 sheds light on the structure of the cutoff T and the rates of the corresponding type 1 or type 2 errors when the sample size n is large. Since both the errors go to 0 as $n \to \infty$, this procedure for testing of hypothesis is consistent while keeping the error rates equal. It should be noted that both the errors decay at an exponential rate and therefore deems useful even at moderate sample sizes. In order to convince the accuracy of Theorem 1, we presented the type 1 and type 2 error rates as well as the rate of change of T with respect to n using the results from theorem 1 and corresponding numerical results from equation (6). As seen from Table 1, irrespective of the Cohen's f^2, as n increases, the rate of change of T, $\log$ (type 1) and $\log$ (type 2) converge to the corresponding theoretical rates specified in Theorem 1. Further, the error rates only depend f^2 but not on p_1 (the number of the exposures on top of the baseline covariates). In addition, through Monte Carlo simulation we showed the calibrated type 1 and type 2 errors remain approximately same under T in a linear regression framework (Section S.5 and Table S.1 of the supplementary material).

2.4 Type 2 error function

Theorem 1 suffices when the data is generated under the Cohen's $f^2 = \delta$ and the alternative is set at $H_1 : f^2 = \delta$. But what happens when the true f^2, under which the data is generated, is not δ ? To highlight and investigate these subtleties, let us assume that the true Cohen's $f^2 = \epsilon$, under which the data is generated. Then given the hypothesis in (1), the type 2 error function is, $\beta(\epsilon) = P\left(\chi^2_{p_1}(\gamma_n) < \mathrm{T}(\delta)|\gamma_n = n\epsilon\right)$.

2.4.1 Case 1: when $0 < \epsilon < \delta$.

Note that, $\beta(0) \to 1$ as $n \to \infty$. Since $1 - \beta(0) = \beta(\delta) \to 0$ when $n \to \infty$, there exists a point $\epsilon^* \in (0, \delta)$ such that $\beta(\epsilon) > 0.5$ for $\epsilon \in (0, \epsilon^*)$ and $\beta(\epsilon) \leq 0.5$ for $\epsilon \in [\epsilon^*, \delta)$. The ϵ^* is the root to

$P\left(\chi^2_{p_1}(\gamma_n) < \mathrm{T}(\delta)|\gamma_n = n\epsilon^*\right) = \frac{1}{2}$. The following corollary introduces an asymptotic expression for ϵ^* under the hypothesis (1).

Corollary 1. *Under the hypothesis (1) and as* $n \to \infty$, $\epsilon^* = \frac{\delta}{2K-1} + o(1)$ *with* $K = (2+\sqrt{2})(1+o(1))$

2.4.2 Case 2: when $\epsilon \geq \delta$.

First note that $\beta(\epsilon)$ is monotonically decreasing in ϵ (see Section S.6 of the supplementary material). Therefore the type 2 error is even smaller when the true Cohen's f^2 is larger than δ. Hence asymptotically the $\beta(\epsilon) \to 0$ (see supplementary material). Now we investigate what happens to the type 2 error when $\boldsymbol{y}$ is generated while ϵ lies between $(0, \delta)$.

2.4.3 Neutral effect size, null and alternative neighborhood.

When true Cohen's f^2, $\epsilon \in [0, \epsilon^*)$, T prefers the H_0 with probability greater than $\frac{1}{2}$. Therefore, we can think of the interval $[0, \epsilon^*)$ as an expanded null hypothesis which nicely connects to the interval null hypothesis discussed in literature (Morey and Rouder (2011), Kruschke (2013) and Liao et al. (2020), Midya and Liao (2021)). We name ϵ^* as the "neutral effect size". Note that, ϵ^* decreases as n increases although very slowly due to a term $O\left(\frac{1}{n^{1-\frac{1}{2K}}}\right)$ and eventually converges to $\frac{\delta}{2K-1}$. The interpretation of "neutral effect size" is that, as long as the true Cohen's f^2, denoted by ϵ, originates below ϵ^*, the probability of rejecting the H_0 gets smaller than $\frac{1}{2}$ even when the true effect size ϵ is far from zero. Whereas, if true effect size ϵ originates above ϵ^*, the probability of rejecting the null becomes greater than $\frac{1}{2}$. Similarly, an interval of the form $\{x|x > \epsilon^*\}$ can be conceived such that, for any $\epsilon > \epsilon^*$, the probability of rejecting the null remains greater than $\frac{1}{2}$. The interval denoted is named as the "alternative neighborhood".

To demonstrate the concepts discussed above, $\beta(\epsilon)$, i.e. the type 2 error function is plotted for $n = 250$ and $n = 1000$ at $p_1 = 5$ and $\delta = 10\%$ in Figure 1 with ϵ from 0 to 10% on x-axis. $\beta(\epsilon)$ decreases smoothly starting from 0 as ϵ increases. The neutral effect sizes, ϵ^* are 2.6% and 2.2% for $n = 250$ and $n = 1000$ respectively. The shaded red region denotes the "alternative neighborhood" with type 2 error below $\frac{1}{2}$, whereas the shaded blue region denotes the "null neighborhood" with type 2 error above $\frac{1}{2}$. As long as the true effect size (ϵ) of the underlying data, is greater than the neutral effect size, the error calibrated cutoff T will favour the alternative. Whereas the null will be favoured only when the true effect sizes are less than

the neutral effect sizes. Note that, neutral effect size, ϵ^* decreases from 2.6% to 2.2% as the sample size n increases but the rate of decrease is very slow. This plot therefore reinforces the idea that no matter how large the sample size is, this error calibrated cutoff will only reject null in the favour of the alternative if and only if the true effect size originates from a neighbourhood $\{f^2 | f^2 > \epsilon^*\}$.

2.5 Putting estimation in the type 1 and type 2 error balanced hypothesis testing framework

2.5.1 Notion of optimal impact .

What should be a prudent choice of δ ? The aim should be to choose larger δ with minimum type 2 error. Since for any data, $f^2 \geq 0$, choosing a slightly smaller deviation from zero, minimizes the type 2 error but rapidly increases the type 1 error.

For any given δ, we reject the null if and only if $\Lambda(\boldsymbol{y}) \geq \mathrm{T}(\delta)$. Given this simplicity of the decision rule, one can choose the maximum value of δ such that the alternative $H_1 : f^2 = \delta$ will always be preferred against the null; but when that maximum value of f^2 is crossed, the null can no longer be rejected. Denote this particular choice of Cohen's f^2 by δ^*. Since $\mathrm{T}(\delta)$ is an increasing function of δ, one can obtain an unique δ^* by solving,

$$\delta^* = \mathrm{argmax}_{\delta > 0}\{\mathrm{T}(\delta) \leq \Lambda(\boldsymbol{y}) | \Lambda(\boldsymbol{y}) > \mathrm{T}(\delta = 0)\}.$$

Note that, while considering δ^*, we exclude those scenarios where $\Lambda(\boldsymbol{y}) < \mathrm{T}(\delta)$ for all δ, implying that the null $H_0 : f^2 = 0$ will be accepted, no matter what δ is chosen.

Corollary 2. *Under the hypothesis in (1), the maximum value of Cohen's f^2 such that the asymptotic type 2 error is at its minimum, is given by:*

$$\delta^* = \frac{\Lambda(\boldsymbol{y}) - c_1 n^{\frac{1}{2K}}}{n}(2K - 1) + o(1)$$

To contextualize and interpret δ^*, consider the following hypothesis and the null and alternative neighborhood it induces.

$$H_0 : f^2 = 0 \quad \text{vs} \quad H_1 : f^2 = \delta^*. \tag{9}$$

Note that an asymptotic estimate of the true Cohen's f^2, is given by $\frac{\Lambda(\boldsymbol{y})}{n}$ (Lemma (1) and (2)). The neutral effect size for the hypothesis in (9) is $\frac{\Lambda(\boldsymbol{y})-c_1 n^{\frac{1}{2K}}}{n} = \epsilon^*$ (say) (from Corollary (1) and (2)). Therefore asymptotically, the null neighborhood is $[0, \epsilon^*)$ and the alternative neighborhood is $\{f^2 | f^2 \geq \epsilon^*\}$. As long as the true Cohen's $f^2 \geq \epsilon^*$, the null will be rejected in support of the alternative. But if one chooses a larger $\delta = \delta^* + h$, for any $h > 0$, in hypothesis (9) the alternative neighborhood will be squeezed to $\{f^2 | f^2 \geq \epsilon^* + \frac{h}{2K-1}\}$. Hence, even if the true Cohen's f^2 is larger than ϵ^* and lies within $[\epsilon^*, \epsilon^* + \frac{h}{2K-1})$, the null will no longer be rejected. Further, note that the type 2 error function $\beta(h) = P\left(\chi^2_{p_1}(\gamma_n) < \mathrm{T}(\delta^* + h) | \gamma_n = n\hat{\delta}\right)$ can be expressed as, $\beta(h) = \Phi\left(\frac{\frac{nh}{2K-1} - p_1}{\sqrt{2(p_1 + 2\Lambda(\boldsymbol{y}))}}\right) + o(1)$ and it attains minimum when $h = 0$, i.e. when δ attains its maximum at δ^*. Any $h > 0$ therefore rapidly increases the type 2 error as the sample size increases. In summary, for a given data, δ^* quantifies the maximum "impact" by any exposure-mixture in a larger model on top of the smaller baseline model, such that the type 2 error is at its minimum.

Definition 1. *Considering the stochasticity in $\boldsymbol{y}$ and given sample size n, we define, impact* $:= E_{\boldsymbol{y}}\{\delta^*\}$ *and optimal impact* $:= E_{\boldsymbol{y}}\{\delta^*\}$ *with* $n \to \infty$.

impact depends on the sample size n. But it is a function of $\Lambda(\boldsymbol{y})$ and undertakes asymptotic convergence based on weak law of large numbers (Lemma 3.1 of Vuong (1989)). Under this framework, therefore, the expected impact converges to an optimal quantity, $E_{\boldsymbol{y}}\left\{\frac{\Lambda(\boldsymbol{y})}{n}(2K-1)\right\}$ as $n \to \infty$. The implication of optimal impact is very vital. Under the hypothesis in (9), it captures the optimal "sample size stabilized" impact in outcome impacted solely by the larger exposure-mixture model on top of baseline covariate only model, such that the asymptotic type 2 error is at its minimum.

2.5.2 Data driven estimation of optimal impact and sufficient sample size.

optimal impact can be estimated by bootstrapping a large size N (say $N = 5000$ or 10000) with replacement from the original sample of size n, with $n < N$. Moreover, because of its convergence, one can find a sample size and a corresponding impact such that it will be in a "practically close neighbourhood" of the optimal impact.

Consider the equivalence tests for the ratio of two means with prespecified equivalence bounds (Schuirmann (1987) and Phillips (1990)). N and n_s be the sample sizes under which we estimate optimal impact and im-

pact respectively. Let δ^s and δ^{opt} be the underlying random variables for the impact and optimal impact respectively. We are interested in the distribution of the log transformed ratios of δ^s and δ^{opt}, i.e. $\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}$. Consider the hypothesis of non-equivalence as below,

$$\begin{aligned} H_0 &: \mu\left(\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right) < l_R \text{ or } \mu\left(\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right) > l_U \\ H_1 &: \quad l_R \leq \mu\left(\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right) \leq l_U \end{aligned} \tag{10}$$

where, l_R and l_U are the lower and upper equivalence bounds with $l_R < 0$ and $l_U > 0$. The null hypothesis will be rejected to favour the alternative if a two-sided $100(1-2\alpha)\%$ CI is completely included within l_R and l_U. We will assume $l_R = \log(0.8)$ and $l_U = \log(1.25)$ following typical practice (Phillips (2009)) but less stricter values can be chosen for practical purposes. We approximate $\mu\left(\frac{\delta^s}{\delta^{opt}}\right)$ and $\sigma\left(\frac{\delta^s}{\delta^{opt}}\right)$ by using Taylor series expansions (detailed in Section S.2 of supplementary material). The mean and variance after logarithmic transformation are found using direct application of delta theorem on $\frac{\delta^s}{\delta^{opt}}$. Finally, we declare alternative hypothesis if the 2α level CI on $\mu\left(\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right)$ is within the equivalence limits, i.e.,

$$l_R \leq \log\left(\hat{\mu}\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right) - \frac{t_{1-\alpha,M-1}}{\sqrt{M}}\frac{\hat{\sigma}\left(\frac{\delta^s}{\delta^{opt}}\right)}{\hat{\mu}\left(\frac{\delta^s}{\delta^{opt}}\right)} \quad \text{and} \quad \log\left(\hat{\mu}\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right) + \frac{t_{1-\alpha,M-1}}{\sqrt{M}}\frac{\hat{\sigma}\left(\frac{\delta^s}{\delta^{opt}}\right)}{\hat{\mu}\left(\frac{\delta^s}{\delta^{opt}}\right)} \leq l_U$$

where, $t_{1-\alpha,M-1}$ is the $100(1-\alpha)\%^{\text{th}}$ quantile in a standard t-distribution. As long as the hypothesis of non-equivalence in (10) is rejected in favour of the alternative, n_s can be regarded as a "sufficient sample size" at equivalence bounds of $[\log(\frac{8}{10}), \log(\frac{10}{8})]$ with a corresponding impact of $\hat{\mu}(\delta^s)$.

3 Simulated examples

Consider a normally distributed outcome and one single exposure with five baseline covariates based on sample size of 300. Further assume, the R^2 for the baseline covariate only model is 20%, and the true and unknown impact due to the exposure is 5.8%. Therefore, the R^2 for the larger model with a single exposure and five covariates is 20.8% (the mean correlation between the covariates is set at 0.3 and the error variance is assumed to be 5). See Section S.7 of the supplementary material for the data generating process.

Assume a researcher collected this data and intends to find the association between the outcome and the

exposure after controlling for the five baseline covariates. As a first step, the optimal impact is estimated by bootstrapping a size $N = 5000$ based on the original sample of $n = 300$. We obtain an estimate of optimal impact : 6.1% (which is very close to the true impact of 5.8%). Similarly, impacts are estimated at bootstrapped samples of sizes $N = 200, 300, 400, 500, 600$ and 2500 to illustrate the gradual convergence of optimal impact as the bootstrap size increases (Fig. 2-A).

Further note that, as precision increases with sample size, the absolute value of the regression coefficient remain stable (Fig. 2-B) while the p-values keep getting smaller (Fig. 2-D). For the original sample size of $n = 300$, the corresponding p-value of the regression estimate of the exposure, is not significant. The researcher therefore might want to collect more data and increase the original sample size based on statistical power calculation and sample size determination - which estimates that a total sample size of around 1000 is required assuming 80% power and type 1 error fixed at 5%.

Sample size estimation using optimal impact, strikes a balance between precision and utility of explanatory power. We estimated $\mu\left(\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right)$ based on 2000 iterations and used the hypothesis of non-equivalence in (10) to compare the impacts in $N = 200, 300, 400, 500, 600$ and 2500 with respect to the estimated optimal impact at $N = 5000$ (Fig. 2-C). At $N = 600$, the $\mu\left(\log\left\{\frac{\delta^s}{\delta^{opt}}\right\}\right)$ and its 95% CI lies within the bounds of $l_R = \log(8/10)$ and $l_U = \log(10/8)$, whereas at $N = 500, 400$ and 300, it breaches the upper bound of $l_U = \log(10/8)$ but stays within the bounds of $l_R = \log(7/10)$ and $l_U = \log(10/7)$. Accordingly, the researcher can choose a sufficient sample size of $N = 600$ or $N = 300$ at equivalence bounds of $[\log(8/10), \log(10/8)]$ or $[\log(7/10), \log(10/7)]$ respectively with corresponding impacts 7% and 7.9%. These impacts are within a close neighbourhood of the estimated optimal impact of 6%. The optimal R^2 of the larger exposure mixture models at sample sizes 600 and 300, will be 20.9% and 21.1% - making them equivalent in most practical purposes.

3.1 optimal impact and p-value

Assume a population of size 5000. A researcher is interested to find the association between an exposure and health-outcome after controlling for some baseline covariates. They plan to conduct a preliminary study with a sample size of 300 and then eventually increase the sample to 1000. For the true population, the optimal impact is a set at 2.9% with β(p-value): $0.06(1.41 \times 10^{-5})$. In Figure S.1 of supplementary material, we

show that as sample size increases, the p-value corresponding to the regression coefficient of the exposure decreases and eventually crosses the canonical cutoff 0.05; whereas the mean of optimal impacts converge to the true value. Even at original sample size 300, the mean of optimal impacts remain very close to the true value. Thus, irrespective of the size of p-values, scale-independent optimal impacts remain practically unaltered with the change of sample size.

4 Application in exposure-mixture association of Per-and poly-fluoroalkyl substances (PFAS) and metals with serum lipids among US adults

Endocrine-disrupting chemicals (EDCs) are a diverse class of environmental pollutants with "emerging concern" that interfere with multiple metabolic and hormonal systems in human (Futran Fuhrman et al. (2015)). EDCs include pesticides, pharmaceutical agents, toxic metals, plasticizers which are used in many commercial products (Buhari et al. (2020)). PFAS are exclusively man-made EDCs and environmentally persistent chemicals which are used to manufacture a wide variety of consumer and industrial products, non-stick, stain and water resistant coatings, fire suppression foams, and cleaning products (Liu et al. (2018) and Jain and Ducatman (2018)). Both PFAS and metals have been associated with increase in cardiovascular disease (CVD) or death using many cross-sectional and longitudinal observational studies and experimental animal models (Meneguzzi et al. (2021)). Hypercholesterolemia is one of the significant risk factors for CVD and it is characterized by the presence of high levels of cholesterol in the blood. High serum low-density lipoprotein (LDL), total serum cholesterol levels, and low levels of high-density lipoprotein (HDL) in the blood are one of the incriminating factors for the pathogenesis of this disorder (Buhari et al. (2020)).

Using the theory discussed in the sections above, we aim to quantify and contrast the optimal impacts by PFAS and metal mixture on serum lipoprotein-cholesterols after adjusting for baseline covariates. The goal therefore is to estimate optimal impacts and corresponding sufficient sample sizes for both the PFAS and metal mixture.

4.0.1 Study Population

We have used a cross-sectional data from the 2017–2018 US NHANES (CDC and NCHS (2018)). The present study has data on 683 adults. Data on baseline covariates (age (in years), gender, ethnicity, body mass index (bmi) (in kg/m^2), smoking status, ratio to family income to poverty) were downloaded and matched by IDs of the NHANES participants. See Table 2 for details on characteristics of the study population. To adjust for oversampling of non-Hispanic black, non-Hispanic Asian, and Hispanic in NHANES 2017-2018, a weight variable was added in the regression models. List of individual PFAS, metals and their lower limit of detection can be found in Section S.8 in the supplementary material.

4.0.2 Methods for exposure-mixture analysis

We will use WQS regression as our explanatory model but other exposure-mixture models such as Bayesian weighted quantile sum regression and Bayesian kernel machine regression can also be utilized, so long as the likelihood ratio test statistic can be estimated. All the PFAS and metals were converted to deciles. Further intentionally, no validation set was used to keep the problem of estimating optimal impact and finding best partitioning the dataset, separate from one another. As an additional analysis, both the serum cholesterols were also dichotomized using their 90th percentile, to demonstrate the utility of optimal impact on binary outcomes. The optimal impacts were estimated using bootstrapped samples of 5000 from the original sample of size 683 and Monte Carlo simulations were repeated 100 times.

4.1 Results

4.1.1 PFAS and Metal mixture have higher optimal impacts on LDL-C than HDL-C

For metals and PFAS, the optimal impacts of continuous HDL-C were $9.6\%[95\%$ CI: $(9.1\%, 10.0\%)]$ and $10.7\%[95\%$ CI: $(10.2\%, 11.1\%)]$ respectively, whereas for continuous LDL-C, those were $14.7\%[95\%$ CI: $(14.2\%, 15.2\%)]$ and $16.2\%[95\%$ CI: $(15.6\%, 16.7\%)]$ respectively. Both the EDC mixture have relatively higher impact on LDL-C than HDL-C. Further, for both the cholesterols, metal-mixture has slightly higher impact than the PFAS-mixture (Fig.3 - A and B). After dichotomizing both the cholesterols at their 90th percentile, the optimal impacts for metal-mixture remained similar to the continuous choles-

terol outcome (HDL-C: $9.8\%[95\%$ CI: $(9.4\%, 10.2\%)]$ and LDL-C: $17.2\%[95\%$ CI: $(16.6\%, 17.8\%)]$), but slightly decreased for PFAS-mixture (HDL-C: $6.9\%[95\%$ CI: $(6.5\%, 7.2\%)]$ and LDL-C: $11.5\%[95\%$ CI: $(11.0\%, 12.0\%)]$). The decrease might have been due to some loss of information while dichotomizing the outcome (Fig.3 - C and D).

As a post-hoc analysis and simple demonstration, we also calculated the sufficient sample sizes for this data-set at the equivalence bounds of $[\log(\frac{75}{100}), \log(\frac{100}{75})]$. For both metal and PFAS-mixture, the mean of log ratio estimates and their corresponding 95% CI for the original sample size at 683, lie well within the equivalence bounds. Further even at a decreased sample size of 483, means of log ratio estimates and their 95% CIs , still remain with the equivalence bounds. Therefore, N = 483, is a sufficient sample size at equivalence bounds $[\log(\frac{75}{100}), \log(\frac{100}{75})]$ for both metal and PFAS-mixture (Fig. 4). But further decrease in the sample size, would not be sufficient, at this pre-fixed equivalence bounds.

5 Discussion

In this paper, we presented the idea of optimal impact of exposure mixture in association to health outcomes within a type 1 and type 2 error balanced hypothesis testing framework to evaluate the strength of explanatory associations. The utility of an explanatory model is evaluated through its strength of association (Shmueli (2010)) and therefore this framework provide a systematic way for theory building in environmental epidemiology. optimal impact does not get perturbed by study sample sizes as long as the studies are well representative of the target population. For an exposure mixture to have large optimal impact but statistically non-significant regression coefficient might be a result of smaller sample size but not a genuine absence of association. Further, this idea was naturally extended to estimate sample size in designing studies by striking a balance between explanatory precision and utility of association estimates.

This framework has its limitation, the bootstrapped estimation of optimal impact assumes the original sample is well representative of the true target population. Any estimation of optimal impact, therefore carries this implicit assumption. But such an assumption is at the core of many statistical analyses and a well designed study can ideally alleviate such issues or could be corrected to be well representative. In addition, this current theory is based on likelihood ratio test of nested models but future work can extend this

framework to strictly non-nested or overlapping models. Progress can be made to estimate optimal impact on high dimensional setting where the number of parameters is strictly lesser than the sample size but $p \to \infty$.

optimal impact might be of practical importance when one finds null associations at the time of data analysis with prefixed sample sizes. Further sample size determination based on preliminary data might utilize optimal impact in designing more cost-efficient human studies. Because of its connect to Cohen's f^2, optimal impact remains a standardized effect size, which is unitless and allows for direct comparisons between several outcomes measured on different scales or separate studies or in meta-analysis. Thereafter, optimal impact can be potentially used to compare and choose between multiple outcomes with varying units and scales. Additionally, by connecting the error balanced testing of hypothesis framework to Cohen's f^2, we circumvented the issues of NHST. In the end, quantifying the impact of exposure-mixture on several health-outcomes might have direct implication for policy decisions and when used together with regression estimates, might prove to be very informative.

6 Software

The codes used in the article is available on GitHub (vishalmidya/Quantification-of-variation-in-environmental-mixtures)

7 Supplementary Material

Supplementary material is provided in a separate file.

Acknowledgments

Conflict of Interest: None declared.

References

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2014). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508.

Brown, B. W., Lovato, J., and Russell, K. (1999). Asymptotic power calculations: description, examples, computer code. *Statistics in Medicine*, 18(22):3137–3151.

Buhari, O., Dayyab, F., Igbinoba, O., Atanda, A., Medhane, F., and Faillace, R. (2020). The association between heavy metal and serum cholesterol levels in the us population: National health and nutrition examination survey 2009–2012. *Human & Experimental Toxicology*, 39(3):355–364. PMID: 31797685.

Carrico, C., Gennings, C., Wheeler, D. C., and Factor-Litvak, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *Journal of Agricultural, Biological, and Environmental Statistics*, 20:100–120.

CDC and NCHS (2017-2018). Us national health and nutrition examination survey data.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Colicino, E., Pedretti, N. F., Busgang, S. A., and Gennings, C. (2020). Per- and poly-fluoroalkyl substances and bone mineral density. *Environmental Epidemiology*, 4.

Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):404–413.

Ferrari, F. and Dunson, D. B. (2021). Bayesian factor analysis for inference on interactions. *Journal of the American Statistical Association*, 116(535):1521–1532.

Futran Fuhrman, V., Tal, A., and Arnon, S. (2015). Why endocrine disrupting chemicals (edcs) challenge traditional risk assessment and how to respond. *Journal of Hazardous Materials*, 286:589–611.

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, 383(9912):166–175.

Jain, R. B. and Ducatman, A. (2018). Associations between lipid/lipoprotein levels and perfluoroalkyl substances among us children aged 6–11 years. *Environmental Pollution*, 243:1–8.

Keil, A. P., Buckley, J. P., O'Brien, K. M., Ferguson, K. K., Zhao, S., and White, A. J. (2020). A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environmental Health Perspectives*, 128(4).

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573.

Liao, J. G. and McGee, D. (2003). Adjusted coefficients of determination for logistic regression. *The American Statistician*, 57(3):161–165.

Liao, J. G., Midya, V., and Berg, A. (2020). Connecting and contrasting the bayes factor and a modified rope procedure for testing interval null hypotheses. *The American Statistician*, 0(0):1–19.

Liu, H.-S., Wen, L.-L., Chu, P.-L., and Lin, C.-Y. (2018). Association among total serum isomers of perfluorinated chemicals, glucose homeostasis, lipid profiles, serum protein and metabolic syndrome in adults: Nhanes, 2013–2014. *Environmental Pollution*, 232:73–79.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Meneguzzi, A., Fava, C., Castelli, M., and Minuz, P. (2021). Exposure to perfluoroalkyl chemicals and cardiovascular disease: Experimental and epidemiological evidence. *Frontiers in Endocrinology*, 12:850.

Midya, V. and Liao, J. (2021). Systematic deviation in mean of log bayes factor: Implication and application. *Communications in Statistics-Theory and Methods*, pages 1–10.

Morey, R. D. and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4):406.

Nakagawa, S. and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605.

Phillips, K. F. (1990). Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, 18:137–144.

Phillips, K. F. (2009). Power for testing multiple instances of the two one-sided tests procedure. *The International Journal of Biostatistics*, 5(1).

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680.

Self, S. G. and Mauritsen, R. H. (1988). Power/sample size calculations for generalized linear models. *Biometrics*, 44(1):79–86.

Self, S. G., Mauritsen, R. H., and Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48(1):31–39.

Selya, A., Rose, J., Dierker, L., Hedeker, D., and Mermelstein, R. (2012). A practical guide to calculating cohen's f2, a measure of local effect size, from proc mixed. *Frontiers in Psychology*, 3:111.

Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 56(4):1192–1196.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3):289 – 310.

Szucs, D. and Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, 11:390.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Wasserstein, R. L. and Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Wheeler, D. C., Rustom, S., Carli, M., Whitehead, T. P., Ward, M. H., and Metayer, C. (2021). Assessment of grouped weighted quantile sum regression for modeling chemical mixtures and cancer risk. *International Journal of Environmental Research and Public Health*, 18(2).

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Zhou, X., Obuchowski, N. A., and McClish, D. K. (2011). *Measures of Diagnostic Accuracy*, chapter 2, pages 13–55. John Wiley & Sons, Ltd.

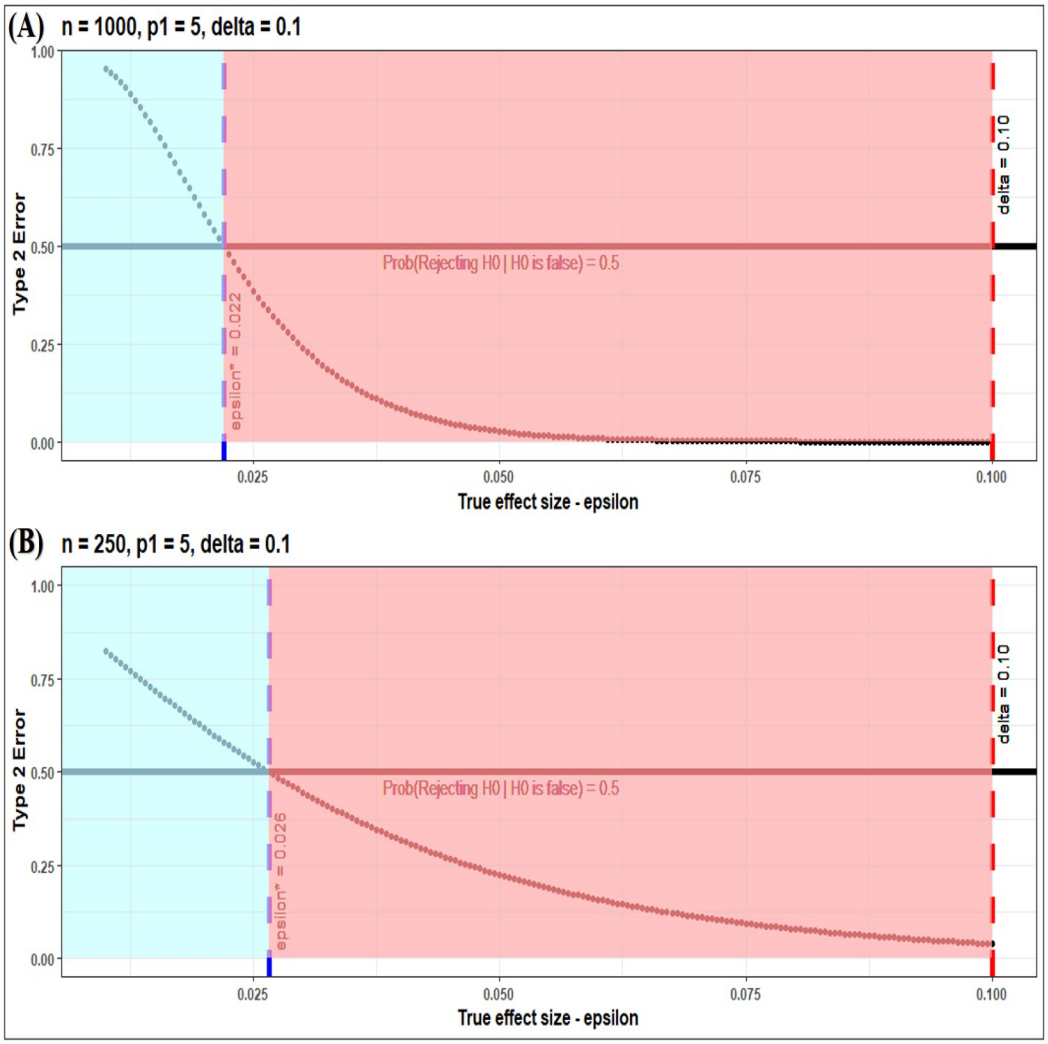

Figure 1: Null and alternative neighborhoods induced through neutral effect size and corresponding type 2 error function for sample sizes $n = 250$ and $n = 1000$

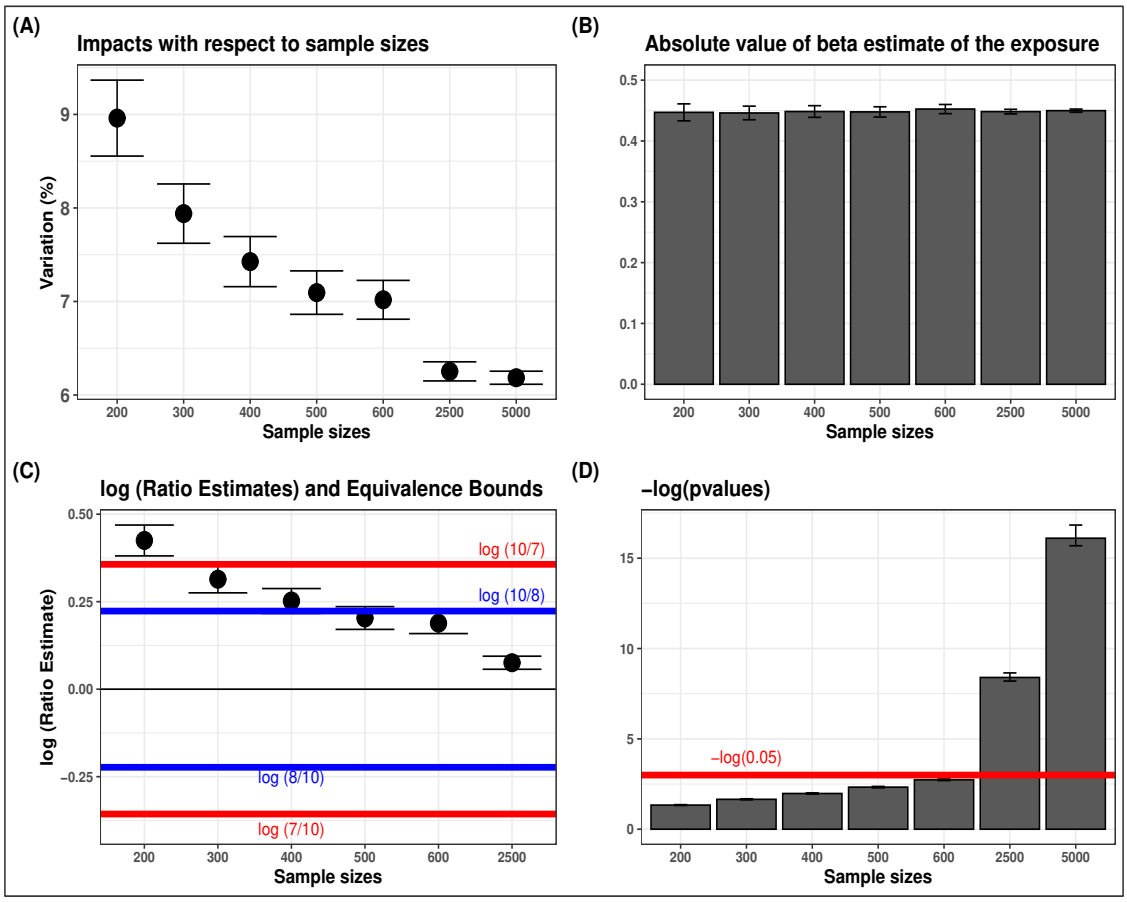

Figure 2: Results from simulated example in linear regression. (A) Illustration of impacts for different bootstrapped sample sizes and its convergence to optimal impact , (B) Absolute values of the β estimates of the exposure, (C) Sufficient sample sizes with respect to choices of equivalence bounds (D) P-values in log (base = e) scale as bootstrapped sample size increases

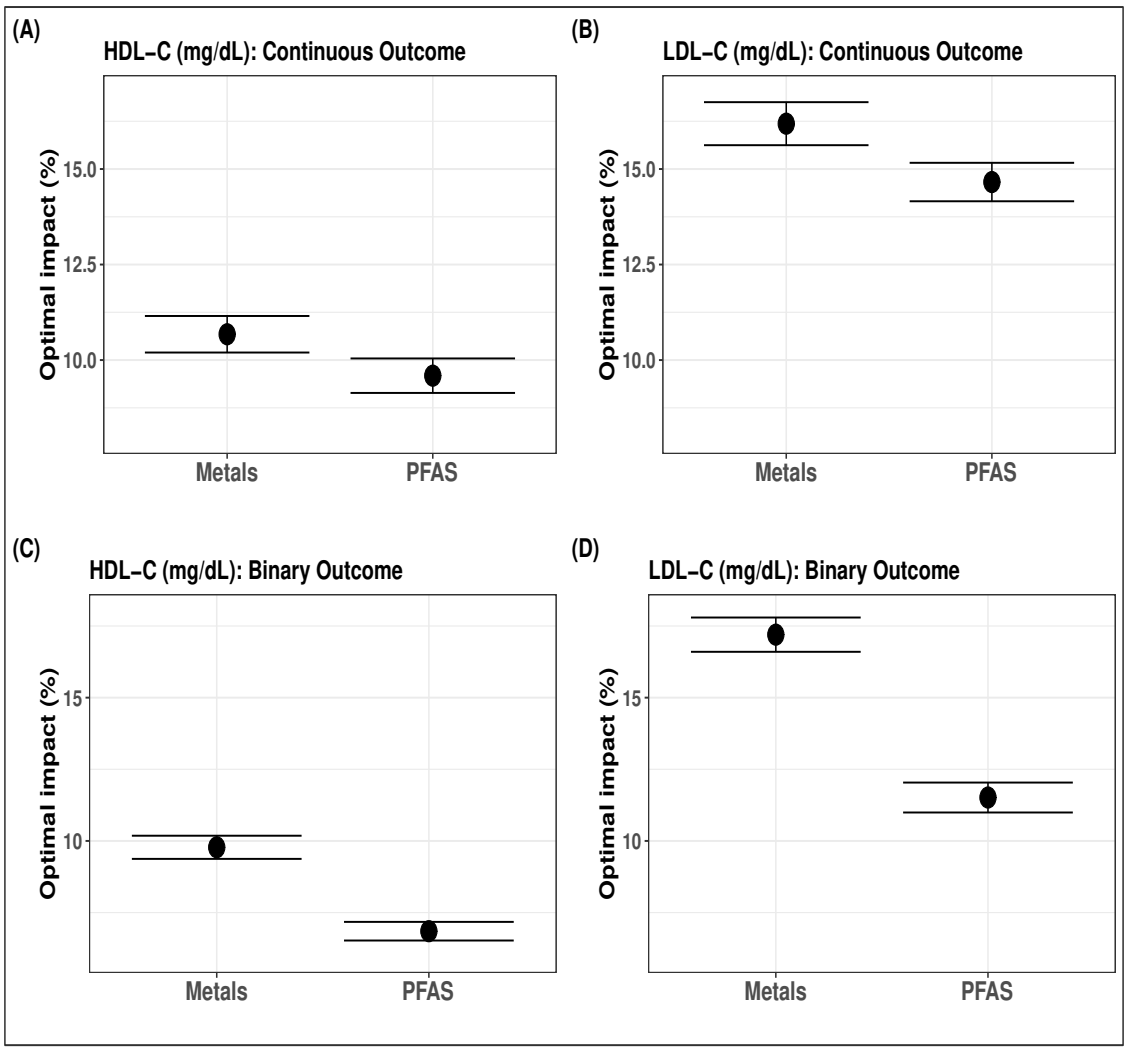

Figure 3: optimal impacts of EDC exposure-mixture to quantify the variability in serum lipoprotein-cholesterols

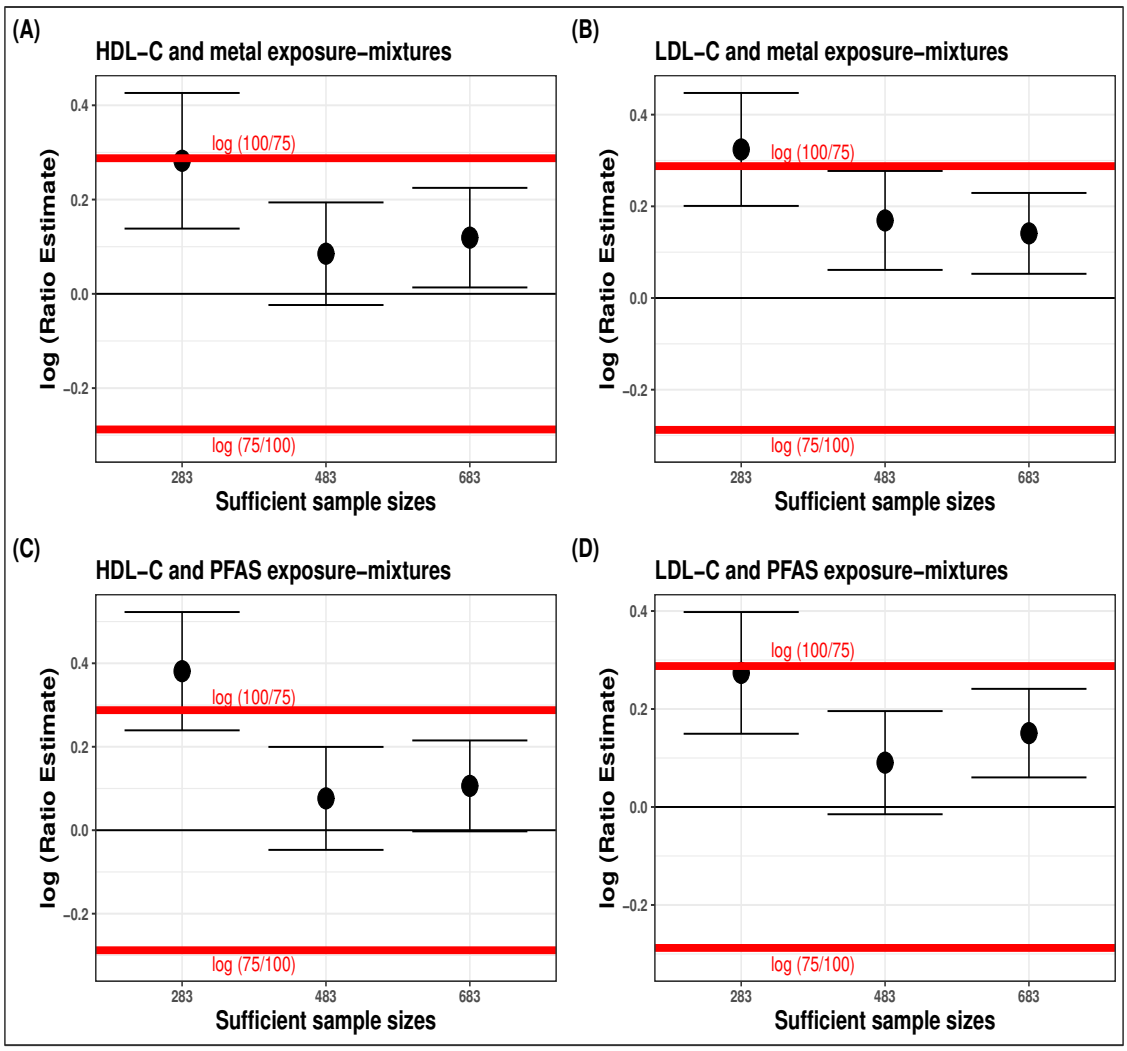

Figure 4: Sufficient sample sizes to estimate the optimal impacts in serum lipoprotein-cholesterols explained solely by EDC exposure-mixture at equivalence bounds of $[\frac{75}{100}, \frac{100}{75}]$

Table 1: Rates of error calibrated cutoff T, $\log$ (type 1) or $\log$ (type 2) with respect to sample size n, based on Theorem 1 and equation (6)

			Numerical approximation using Equation (6)		Using Theorem (1)	
p_1	f^2	n	$\frac{d}{dn}$T	$\frac{d}{dn}\log(\alpha)$	$\frac{d}{dn}$T	$\frac{d}{dn}\log(\alpha)$
1	2.5%	250	0.0042895	-0.0031	0.0042893	-0.0021
1	10%	250	0.0171573	-0.0100	0.0171573	-0.0086
5	2.5%	250	0.0043764	-0.0025	0.0042893	-0.0021
5	10%	250	0.0172491	-0.0090	0.0171573	-0.0086
1	2.5%	500	0.0042896	-0.0028	0.0042893	-0.0021
1	10%	500	0.0171573	-0.0094	0.0171573	-0.0086
5	2.5%	500	0.0043764	-0.0024	0.0042893	-0.0021
5	10%	500	0.0172491	-0.0087	0.0171573	-0.0086

Table 2: Study characteristics of the population under investigation - data from National Health and Nutrition Examination Survey 2017–2018

	Total	Male	Female	% observations ≥ LLOD
Sample size (n)	683	339	344	
Baseline Covariates				
Age (years)	49.51 (18.77)	50.38 (18.81)	48.65 (18.73)	
Ethnicity				
Mexican American	88	43 (49%)	45 (51%)	
Other Hispanic	58	23 (40%)	35 (60%)	
Non-Hispanic White	260	135 (52%)	125 (48%)	
Non-Hispanic Black	155	79 (51%)	76 (49%)	
Other Race - Including Multi-Racial	122	59 (48%)	63 (52%)	
Body mass index (kg/m^2)	29.59 (7.90)	28.67 (6.36)	30.49 (9.09)	
Smoking Status				
Never	402	170 (42%)	232 (58%)	
Smoked at least 100 cigarettes in life but don't smoke now	163	100 (61%)	63 (39%)	
Smoked at least 100 cigarettes in life and smoke now	118	69 (58%)	49 (42%)	
Ratio of family income to poverty	2.56 (1.61)	2.64 (1.63)	2.48 (1.59)	
Outcomes				
HDL-C (mg/dL)	53.91 (15.53)	49.19 (13.10)	58.56 (16.33)	
LDL-C (mg/dL)	109.35 (37.11)	108.99 (35.35)	109.71 (38.83)	
PFAS exposures (Unadjusted geometric means with 95% confidence intervals)				
PFDeA (ng/mL)	0.20 (0.19, 0.21)	0.21 (0.19, 0.22)	0.20 (0.18, 0.22)	68.73 %
PFHxS (ng/mL)	1.10 (1.03, 1.17)	1.49 (1.38, 1.61)	0.81 (0.74, 0.89)	99.12%
Me-PFOSA-AcOH (ng/mL)	0.13 (0.12, 0.14)	0.14 (0.13, 0.15)	0.12 (0.11, 0.13)	38.64%
PFNA (ng/mL)	0.42 (0.39, 0.44)	0.46 (0.42, 0.5)	0.38 (0.34, 0.42)	91.74%
PFUA (ng/mL)	0.14 (0.13, 0.15)	0.14 (0.13, 0.15)	0.14 (0.13, 0.15)	41.59%
n-PFOA (ng/mL)	1.28 (1.22, 1.35)	1.52 (1.42, 1.64)	1.08 (1, 1.17)	99.41%
n-PFOS (ng/mL)	3.26 (3.04, 3.5)	4.11 (3.74, 4.51)	2.59 (2.35, 2.86)	99.41%
Sm-PFOS (ng/mL)	1.28 (1.19, 1.37)	1.73 (1.58, 1.89)	0.95 (0.86, 1.04)	98.82 %
Lead, Cadmium, Total Mercury, Selenium, & Manganese exposures (Unadjusted geometric means with 95% confidence intervals)				
Cd (μg/L)	0.32 (0.3, 0.34)	0.29 (0.27, 0.32)	0.35 (0.32, 0.38)	91.36%
Pb (μg/dL)	0.91 (0.86, 0.96)	1.09 (1, 1.18)	0.76 (0.7, 0.82)	100%
Mn (μg/L)	9.45 (9.21, 9.7)	8.91 (8.62, 9.22)	10.01 (9.64, 10.41)	100%
THg (μg/L)	0.78 (0.72, 0.84)	0.81 (0.73, 0.9)	0.75 (0.67, 0.83)	84.77%
Se (μg/L)	188.62 (186.75, 190.52)	189.28 (186.57, 192.04)	187.97 (185.38, 190.6)	100%

Data presented as mean(SD) or $n(\%)$; LLOD: lower limit of detection (in ng/mL); LDL-C: low-density lipoprotein-cholesterol (mg/dL) ; HDL-C: high-density lipoprotein-cholesterol (mg/dL); PFDeA: Perfluorodecanoic acid; PFHxS: Perfluorohexane sulfonic acid; Me-PFOSA-AcOH: 2-(N-methylperfluoroctanesulfonamido)acetic acid; PFNA: Perfluorononanoic acid; PFUA: Perfluoroundecanoic acid; PFDoA: Perfluorododecanoic acid; n-PFOA: n-perfluorooctanoic acid; Sb-PFOA: Branch perfluorooctanoic acid isomers; n-PFOS: n-perfluorooctane sulfonic acid; Sm-PFOS: Perfluoromethylheptane sulfonic acid isomers; Pb: Lead; Cd: Cadmium; THg: Total Mercury; Se: Selenium; Mn: Manganese