HHS Public Access

Author manuscript

*J Clin Epidemiol*. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

*J Clin Epidemiol*. 2015 September ; 68(9): 1046–1058. doi:10.1016/j.jclinepi.2015.05.029.

# Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations

**Chirag J. Patel**[1], **Belinda Burford**[2], and **John P.A Ioannidis**[1,3,4,5,6,*]

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, MA. 02115

[2]Melbourne School of Population and Global Health, The University of Melbourne, Victoria 3010, Australia

[3]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

[4]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA

[5]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA

[6]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA

## Abstract

**Objectives—**Model specification -- what adjusting variables are analytically modeled –may influence results of observational associations. We present a standardized approach to quantify the variability of results obtained with choices of adjustments called the “vibration of effects” (VoE).

**Study Design and Setting—**We estimated the VoE for 417 clinical, environmental, and physiological variables in association with all-cause mortality using National Health and Nutrition Examination Survey data. We selected 13 variables as adjustment co-variates and computed 8,192 Cox models for each of 417 variables’ associations with all-cause mortality.

**Results—**We present the VoE by assessing the variance of the effect size and in the −log10(p-value) obtained by different combinations of adjustments. We present whether there are multimodality patterns in effect sizes and p-values and the trajectory of results with increasing adjustments. For 31% of the 417 variables we observed a Janus effect, with the effect being in opposite direction in the 99$^{th}$ versus the 1$^{st}$ percentile of analyses. For example, the vitamin E variant α-tocopherol had a VoE that indicated higher *and* lower risk for mortality.

*To whom correspondence should be addressed: John PA Ioannidis, MD, DSc, Stanford Prevention Research Center, Medical School Office Building, Room X306, 1265 Welch Rd, Stanford, CA 94305, USA, +650-7255465, jioannid@stanford.edu.

**Author Contributions:** CJP and BB wrote the software code to conduct the VoE analysis. CJP, BB, and JPAI came up with the idea and wrote/edited the manuscript.

**Competing Interests:** The authors declare no competing interests.

**Conclusions—**Estimating VoE offers empirical estimates of associations are under different model specifications. When VoE is large, claims for observational associations should be very cautious.

### Keywords

Vibration of Effects (VoE); environment-wide association study; model specification; biostatistics

## Introduction

Observational associations between variables do not guarantee causality and they are often complex and influenced by other variables (confounders and effect modifiers). Accounting for covariates is typically achieved through statistical modeling, such as multivariable regression. However, what variables should one choose to account for in complex multivariate phenomena where many variables may be confounded or correlated [1]? *Model specification* can be a major issue in diverse fields, including epidemiology [2], economics [3–5], and psychological science and neurosciences [6]. Thousands of associations are published and many are often challenged and refuted by subsequent investigations [7–9]. Choices of models underlie our assumptions about association and about potential causes and effect [10]. Very often there is large uncertainty about what variables should be modeled and how they are related. Consequently, there is large heterogeneity in how investigators associate variables [2].

In discovery-based research in large datasets there is often no prior evidence or biological plausibility on what adjustment variables to include in statistical models. In other cases, unequivocal evidence and plausibility may exist to include some adjustment variables in the model, lack of consensus on some others, and no available guidance on yet another set of adjustment variables. Interpretation of effects may vary depending on the analytical choices made. A way to compute the extent of instability of the results due to model specification is needed to guide inference.

The “vibration of effects” (VoE)[2] describes the extent to which an estimated association changes under multiple distinct analytical modeling approaches. The VoE is related also to the previously described concept of “multiple modeling” [9] or statistical model induced variability (e.g., [11]). To estimate the VoE empirically, we can compute the distribution of the point estimates of measures of association (e.g. relative risks, odds ratios) and p-values that are possible under different analytical scenarios. The VoE measures how susceptible an association is under different modeling scenarios; the larger the VoE, the greater the instability of the results. One may also explore which specific scenarios most influence the estimated association. Here, we describe a framework to systematically evaluate the VoE for a set of adjustment covariates.

## Example of a controversial association

As an introductory example, we use the VoE framework to evaluate a contentious association between vitamin E (α-tocopherol) and mortality. Early publications of observational studies claimed large reductions in disease-related and mortality-related events

in association with vitamin E [30, 31]. However, clinical trials that followed were not able to support the early observational findings (e.g., [32–35]). Further still, meta-analyses of clinical trials have showed nearly the opposite of early observational studies including null [36] to even *increased* risk [37, 38] of vitamin E on adverse health-related outcomes including mortality. An important question is to understand the extent to which the results of observational studies on vitamin E may depend on how the observational data are analyzed, and in particular on the model specification, i.e. which other factors are taken into account in multivariable modeling. For a resolution, see the VoE analysis for vitamin E and mortality at the end of the Results.

## Methods

### Data source: NHANES 1999–2000, 2001–2002 and 2003–2004

We downloaded NHANES examination, laboratory, questionnaire, and National Death Index (NDI) linked mortality data for 1999–2000, 2001–2002, and 2003–2004 surveys. Mortality information was collected from the date of the survey participation through December 31, 2006 and ascertained via a probabilistic match between NHANES and NDI death certificate information. The NDI matches individuals on personal and demographic criteria, such as social security number and date of birth and its performance has been described elsewhere (e.g., [12]). Overall, 9555, 11021, and 10100 participants were followed in the 1999–2000, 2001–2002, and 2003–2004 surveys respectively, with 611, 470, and 276 assumed death events.

### Variables of interest

A total of 993 self-reported, clinical, and molecular phenotypes were assayed on participants of NHANES between 1999 and 2004 and we associated 417 of these with all-cause mortality that had a minimum sample size of 1000 and a minimum of 100 deaths. These variables included 1) 179 serum or urine biomarkers of environmental exposures 2) 9 self-reported behavioral factors such as smoking, alcohol consumption and physical activity, 3) 85 self-reported nutritional intake information, 4) 27 self-reported health conditions, 5) 92 clinical factors such as body mass index (BMI), blood pressure, and cholesterol, and 6) 13 demographic variables.

Environmental exposure variables are physical/chemical biomarkers of external exposures measured in serum or urine, such as blood lead concentration or assays of infectious agents, such as HIV. Broadly, these included a serum marker of nicotine metabolism (cotinine), dioxins (n=7 markers), furans (n=10), heavy metals (n=15), hydrocarbons (n=21), nutrients (n=15), polychlorinated biphenyls (n=34), pesticides (n=22), phthalates (n=12), estrogenic compounds (n=6), bacterial (n=2), and viral organisms (n=6) (Table S2). With exception of assays detecting infectious agents (which were positive/negative assays), factors were continuous in scale. All continuous exposure markers were mean subtracted and standardized by the standard deviation (z-standardized) to facilitate comparisons.

Self-reported behavioral factors included current/past smoker (versus never smoker), alcohol consumption (e.g., “have you ever had five or more drinks per day?”, “how many drinks per

day in last month?"), and physical activity. Physical activity was estimated by deriving metabolic equivalents for self-reported leisure and normal-time activities [13] and treated as ordinal variables based on Health.gov physical activity guideline categories for no aerobic activity, low activity (medium intensity activity greater than baseline but fewer than 150 minutes per week), moderate activity (150 to 300 medium intensity minutes a week), and high activity (>300 minutes of medium intensive activity per week or >150 minutes of high intensity per week) as previously described [14, 15].

Self-reported food and nutrient consumption variables were determined from one in-person 24-hour interview (1999–2000, 2001–2002) or two 24-hour (2003–2004) in-person and on the phone interviews using the United States Department of Agriculture and Department of Health and Human Services food recall questionnaires [16–19]. These included daily estimated intake of 68 nutrients and vitamins, including vitamin A, B, C, E, carotenes, carbohydrates, sugars, proteins, fats, minerals (iron, sodium, potassium). All continuous food and nutrient consumption variables were z-standardized for comparison.

Clinical variables included body measurements assessed by a health professional, including height, weight, waist circumference, thigh circumference, triceps skinfold, subscapular skinfold, calf circumference. Laboratory values included serum measures of lipids (e.g., fasting LDL-cholesterol, HDL-cholesterol, triglycerides, glucose, insulin), creatinine (urine and serum), albumin (serum and urine), uric acid, total serum protein, C-reactive protein, thyroxine, and thyroid stimulating hormone, and others. Laboratory measures were log-transformed. All continuous clinical measures were z-standardized to facilitate comparison.

### Adjusting variables

We chose a set of fifteen variables as our set of possible adjustments. Because there is no perfect consensus on what variables should (or should not) be included as adjustments in association to all-cause mortality, we based our set on a large meta-analysis of 80 studies of physical activity on all-cause mortality [20]. Since age and sex are well-known factors related to mortality, we chose to keep these in all models ("baseline" variables) and then chose 13 other variables as the set of varying adjustments, the 12 that had the highest frequency of use among these 80 investigations (smoking, body mass index (BMI), hypertension, diabetes, cholesterol, alcohol consumption, education, income, family history of heart disease, heart disease, any cancer, physical activity) and race/ethnicity, because NHANES samples are racially heterogeneous previous studies examining mortality in NHANES adjusted for race/ethnicity (e.g., [21, 22]).

We coded the 15 adjustment variables as binary, categorical, or as continuous variables. Medical history items, including any hypertension, diabetes, heart disease, family history of heart disease, and cancer were treated as binary variables. We treated alcohol consumption as a binary variable (drinking five drinks per day: yes/no). Physical activity was estimated by deriving metabolic equivalents for self-reported leisure and normal-time activities[13] and treated as ordinal variables based on Health.gov physical activity guideline categories as previously described[14]. Serum total cholesterol in NHANES was assayed in participants who had fasted at least 8 hours prior to survey. Smoking was coded as a 3-category variable (self-reported current or past smoker vs. never smoker). BMI was coded as 5 categories

based on recent investigations between BMI and all-cause mortality [23], including underweight (BMI less than 18.5 kg/m$^2$), overweight (BMI greater than or equal to 25 and less than 30 kg/m2), obese (BMI greater than or equal to 30 kg/m$^2$ and less than 35 kg/m$^2$) or very obese (BMI greater than or equal to 35 kg/m$^2$) with the reference group being normal weight (BMI greater than or equal to 18.5 and less than 25 kg/m$^2$). Income was coded as tertile of income to poverty ratio (1$^{st}$ and 2$^{nd}$ tertile versus 3$^{rd}$ tertile). Education was coded as less than high school education or high school equivalent (vs greater than high school education). We added race/ethnicity as NHANES participants are representative of general population of the United States (e.g., Whites, Blacks, and Hispanics) and previous studies examining mortality in NHANES have routinely adjusted for race/ethnicity (e.g.,[21, 22]). Race/ethnicity was coded as Mexican American, Non-Hispanic Black, Other Hispanic, and Other (vs. Non-Hispanic Whites). We describe the adjustment variables for survivors and deceased individuals in Table S1 (see Appendix).

For analyses of the association of serum Vitamin D with all-cause mortality, we increased the number of rotating adjusting co-variates to 19 (from 13) to observe how the VoE changes under different sets of adjusting co-variates, as the options increase for adjustment further. In addition to the fifteen variables above, we selected urinary creatinine, urinary albumin, serum calcium as documented in a recent investigation vitamin D levels in mortality [24]. Further, we chose serum cadmium as we recently documented a potential association with mortality [25] along with 2 serum nutrient indicators, β-carotene and γ-tocopherol, which were associated with type 2 diabetes and cholesterol lipid levels [14, 26] and had consistent VoE patterns.

### Assessing the vibration of effects in all-cause mortality

We modeled 417 variables separately as a function of time to death with the Cox proportional hazards model. VoE may be utilized with any linear model, including linear or logistic regression. We used the R-project *survival and survey* library for all analyses and accounted for clusters pseudo-strata, pseudo-sampling units, and participant weights to accommodate the complex sampling of the data [27, 28].

We considered the VoE in the context of 13 adjustment variables not including age and sex. The total number of possible combinations of adjusting variables from the set of 13 total adjustments is 8,192. For all 8,192 models, sample sizes remained constant. Our algorithm for computing the VoE for a variable *x* (e.g., HDL-cholesterol) is the following:

1. Select a number *k* from 0 to *n* total adjustment variables (*n*=13). Categorical variables are considered as a single group (e.g., education or BMI is selected as a group of variables together)
2. For each combination of *k* number of adjustment variables (*k* number of distinct variables from the set of all variables $z_0$ through $z_n$), associate *x* with all-cause mortality with Cox-proportional hazards and assess the effect size (β) and p-value for *x*:

$$\text{Hazard Ratio} = \exp(\beta x + \gamma_1 age + \gamma_2 male + i_0 \delta_0 z_0 + \ldots + i_n \delta_n z_n)$$

Where $i_0 \ldots i_n$ are indicator variables (=1 or 0) whether the particular variable is in the model or not for a given *k*. Thus, the sum of indicator variables is equal to *k*. The number of iterations in this step is the total number of *k* number of combinations possible out of a set of *n* total variables.

3. If $k = n$, then stop. If $k < n$, then go back to step 1. for $k+1$

## Assessing summary statistics

We computed summary statistics for the VoE. These included the 1st, 50th (median), and 99th percentile of effect sizes and p-values. We created summary statistics derived from the distributions of VoE for a variable. The first was the "relative hazard ratio" (RHR), the ratio of the 99th percentile and 1st percentile HR (or the difference between the 99th and 1st percentile of effect size estimates). The RHR measures the difference between the larger and smaller effect sizes and this measure connotes a spread of effect sizes for different combinations of adjustments. The second was the "relative p-value" (RP), which is the difference between the 99th and 1st percentile of −log10(pvalue). We assessed the median effect size and p-value for a given variable for each *k*th level of adjustment.

## Systematic search for interactions between variables and adjustments

The VoE may vary depending on different strata of adjustment variables. Thus, we systematically searched for interactions between variables of interest and the 13 adjustment variables with the inclusion of a multiplicative interaction term. Interactions between variables are very often claimed in observational research, but the search for them is not systematic and the validity of the claimed interactions is often tenuous[29]. Specifically, our procedure to search for interactions between a variable *x* and the 13 possible adjusting variables was the following: For each adjusting variable *k* from 1–13:

$$\text{Hazard Ratio} = exp(\beta\, x + \gamma_1\, age + \gamma_2\, male + \delta_k\, z_k + \zeta_k\, x\, z_k)$$

Where $\zeta_k$ corresponds to the interaction term. We collected the test-statistic and p-value for the interaction term to determine presence of interactions. For categorical adjustment variables, we assessed presence of interaction each category of the variable separately. Thus, we tested for interactions between variables of interest and: 2 categorical variables corresponding to education (< high school and high school education vs. > high school education), 2 variables corresponding to the largest race groups (Non-Hispanic Whites and Mexican Americans vs. Whites), 4 BMI variables (vs. normal weight), 2 variables corresponding to current and past smoking, and 1 variable each for coronary disease, family history for coronary disease, hypertension, diabetes, any cancer, total cholesterol (mean subtracted standardized by the SD), physical activity, drink five drinks per day, age, and male sex resulting in 20 effective tests of interaction for a variable *x*. To maximize power of detection of interaction, we restricted the search to 107 variables with the least missing data and sample sizes greater than 6,500 (range of sample size = 6,500–8,607). Thus, this resulted in 2,140 total of interactions tested (107*22=2,140). We considered the strongest interactions as those that surpassed a Bonferroni level of statistical significance ($p < 0.05/2{,}140 = 2\text{x}10^{-5}$).

For interactions that surpassed Bonferroni level of significance, we computed the VoE for each strata of the adjustment variable separately. For example, if we found a significant interaction between a variable *x* and any family history of heart disease was found we assessed the VoE for *x* for samples who answered without family history of heart disease and with family of heart disease separately. If an interaction was found between a variable *x* and total cholesterol, we categorized samples as having less than 200 mg/dL, greater or equal than 200 and less than 239 mg/dL, and greater than 239 mg/dL total cholesterol. We show the VoE for the strongest interaction between platelet count and total cholesterol (interaction $p=1x10^{-10}$, total cholesterol main effect p=0.01, and platelet count main effect $p=1x10^{-9}$) in Fig. S4.

## Results

### Estimating the VoE

VoE is estimated by computing the hazard ratio (HR) and p-value for a variable of interest while adjusting for all possible combinations of adjustments from a finite set of adjustment variables. Our algorithm for computing the VoE for a variable *x* (e.g., serum Vitamin D) is shown in Fig. 1.

First, we downloaded 417 self-reported, clinical, and molecular measures with linked all-cause mortality information in participants from NHANES 1999–2004 (Fig. 1A). Mortality information was collected from the date of the survey participation through December 31, 2006 and ascertained via a probabilistic match between NHANES and National Death Index (NDI) death certificate information (e.g.,[12]). We chose variables of interest that had data on at least 1000 participants and at least 100 death events during follow-up.

Next, we describe the VoE methodology for the association between serum vitamin D and all-cause mortality (Fig. 1B). The total number of combinations of adjusting variables from the set of $n$=13 total adjustments is 8,192 (or, in general, $2^n$ models, Fig. 1C). We chose a set of thirteen variables as the set of possible adjustments (Fig. 1BC, Table S1). Because there is no consensus on what variables should (or should not) be included as adjustments in association with all-cause mortality, we based the selection of these 13 variables on a large meta-analysis of 80 studies of physical activity on all-cause mortality[20]. The most common adjustment variables in these 80 investigations included (in decreasing order of frequency) age, smoking, BMI, hypertension, diabetes, cholesterol, alcohol consumption, education, income, sex, family history of heart disease, heart disease, and any cancer. Since age and sex are well-known factors related to mortality, we chose to keep these in all models ("baseline" variables). The HR and the respective p-value for the association of that variable with all-cause mortality are estimated for all 8,193 models with different combinations of 13 adjusting variables using Cox proportional hazards time-to-event regression (Fig. 1D). We visualized the VoE for a given variable by plotting the HR versus −log10(pvalue) as 2-dimensional histogram and a contour plot (Fig. 1E).

We created metrics to express the distributions of VoE for a variable (Fig. 1F). The first was the "relative hazard ratio" (RHR), the ratio of the $99^{th}$ percentile and $1^{st}$ percentile HR. The RHR connotes the spread of hazard ratios for different combinations of adjustments. The

second was the "range of P-value" (RP), which is the difference between the 99$^{th}$ and 1$^{st}$ percentile of −log10(pvalue). The RP measures the range of p-values over all estimates. We also assessed whether associations appeared on both sides of the null (HR<1 and HR>1): depending on what adjustments are chosen the results may suggest that the variable of interest is associated with either increased or decreased mortality. We also visualized trends corresponding to the number of adjusting variables (*k*), plotting the median effect size and p-value for each *k* from 0 to 13. We recorded the proportion of estimates that achieved different levels of nominal statistical significance (p<0.05, 0.0001).

The 417 variables included 179 serum or urine biomarkers of environmental exposures (e.g., serum cadmium, mercury, or pesticide level), 9 self-reported behavioral factors such as smoking and physical activity, 84 self-reported nutritional intake information (from a food frequency questionnaire), 27 self-reported health conditions (e.g., diabetes), 92 clinical factors (e.g., body mass index [BMI] and cholesterol), and 13 sociodemographic variables (e.g., income). All continuous variables were log-transformed and z-standardized for comparison. Table S2 describes these 417 variables.

## Prototypical patterns of the VoE

We describe four prototypical patterns from the set of 417 variables (Fig. 2, Fig. S1 for all 417 variables). The first pattern is exemplified by the association between serum levels of vitamin D and mortality (Fig. 2A). All of the HR estimates are less than 1.00, indicating that higher levels of vitamin D tend to be associated with longer survival (all HR<0.76); however the magnitude of the estimated effect is dependent on the number of adjustment variables and the association is attenuated when adjusting for more variables, from HR=0.64 with no adjustment (*k*=0) to 0.75 with all thirteen adjustment variables included (*k*=13). On the other hand, the p-values are less than the nominal level of statistical significance (p=0.05, black line). Most of the results are centered on HR~0.72 and p~$10^{-4}$ (2-dimensional mode). In this first pattern, one concludes that while adjustment weakens the magnitude relationship between vitamin D levels and mortality, inferences regarding the relationship are similar throughout all scenarios of adjustment. Of the 417 variables, 53 (13%) exhibited similar behavior to vitamin D, where all associations were beyond the level of nominal statistical significance but the association was attenuated with a greater number of adjustment variables (Fig. S1).

The second pattern is exemplified by the relationship between thyroxine and mortality, displays how increasing adjustment might change inference (Fig. 2B). Higher thyroxine levels tend to be associated with longer survival, but p-values become greater than the nominal level of statistical significance (p=0.05) with ≥9 adjustment variables on average. Of the 417 variables, 91 (22%) variables had similar behavior to thyroxine in which HR were attenuated and the p-values rose above the nominal level of significance (p>0.05) as the number of adjusting variables, *k*, increased (Fig. S1 and Table S3).

The third pattern, as exemplified by an indicator of kidney function, urinary creatinine, and mortality shows an opposite trend (Fig. 2C). For *k*=5 to *k*=13 number of adjustment variables, the association tends to become stronger in HR and statistical significance; however, the trend is less clear for *k*=0 to *k*=4, where the median p-values increase. Twenty-

six (6%) of the 417 variables exhibited similar behavior to urinary creatinine where the effect sizes increased and p-values decreased for larger *k*.

In the last pattern, as exemplified by α-tocopherol (Fig. 2D), the estimated HRs can be both greater and less than the null value (HR > 1 and HR ≤ 1) depending on what adjustments were made. We call this the *Janus effect* after the two-headed representation of the ancient Roman god. For α-tocopherol, most of the HR and p-values were concentrated around 1 and non-significance, respectively. However, 1% of the models had a HR<0.875 (12.5% decreased risk of death for 1SD increase in exposure) with a nominally significant p-value ($p \leq 0.05$); while another 1% of the models had HR>1.05 (5% increased risk for death for 1SD increase of exposure), albeit without reaching nominal significance. The *Janus effect* is common: 131 (31%) of the 417 variables had their $99^{th}$ percentile HR>1 and their $1^{st}$ percentile HR<1.

Examples like those in Figs. 2A–D represented the VoE patterns for 72% of the 417 associations. Other patterns included VoE where all p-values were >0.05 and the strength of the association decreased (n=50, 12%), increased (n=27, 6%), or showed no dependence (n=15, 4%) with increasing number of adjustment variables *k* (Table S3). Rarer patterns included variables where all p-values were <0.05 and there was an increasing strength of association (n=5, 1%) or no clear relationship with increasing *k* (n=4, 1%), and those having p-values with a range less than and greater than 0.05 with no clear relationship with *k* (n=15, 4%).

## Identifying “multimodality of effects” with VoE

By empirically estimating the VoE it is also possible to detect whether one or more adjustment variables makes a marked difference in the results, leading to multiple modes (Fig. 3) which we call *multimodality of effects*. Multimodality of effects was clearly seen in 71 of the 417 (17%) assessed variables. For example, the overall VoE for serum cadmium on mortality indicates strong association with mortality (Fig. 3A); all of the HRs are >1.2 per 1 SD change in serum cadmium levels and p-values in all analytical scenarios are <0.05. However, two modes are visually evident (Fig. 3A). To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the thirteen adjustment variables in the model, leading to thirteen separate visualizations. In serum cadmium, we observed the two distinct modes were indicative of models that contained or did not contain current or past smoking (Fig. 3B). Specifically, models that contained the smoking adjustment variable (Fig. 3B, yellow points) had HR lower than the models without the smoking adjustment and lower −log10(p-values) (Fig. 3B, black points). One source of cadmium exposure includes smoking and we concluded that the correlation between smoking and exposure to cadmium might be driving the multimodal behavior of VoE. Further, we observed that models that included (or did not include) alcohol drinking also resulted in separate modes in p-values (Fig. 3C).

We observed three modes in the association between triglyceride levels and mortality (Fig. 3DEF). The multimodal plots indicated that total cholesterol and diabetes were driving these modes. For example, in models that did not contain these two adjustments, the associations

had smaller p-values and a smaller range of HR. Further, in models containing diabetes, HR were attenuated. The multimodal pattern seems reasonable in light of the high correlation between triglyceride levels and total cholesterol levels/risk for diabetes. We observed a similar pattern for other cardiometabolic indicators, including fasting blood glucose and insulin (Fig. S1).

### Summary of common patterns of the VoE

Fig. 4 shows the distribution of the fold-deviation of HR from the null (HR=1.00), the −log10(pvalue), RHR, and RP for all 417 variables considered. The "fold deviation" is the difference of the median VoE-estimated HR from 1 (the null value). The median fold deviation was 1.13-fold (25% percentile: 1.05 fold, 75% percentile: 1.24 fold, Fig. 4A). Moreover 50% of the variables had a median p-value less than or greater than 0.27 (25% percentile: 0.04, 75% percentile: 0.59, Fig. 4B). The median RHR was 1.15 ($5^{th}$ percentile: 1.07, 25% percentile: 1.11, 75% percentile: 1.22, $95^{th}$ percentile: 1.70, Fig. 4C). The median RP was 1.07 ($5^{th}$ percentile: 0.31, 25% percentile: 0.589, 75% percentile: 2.03, $95^{th}$ percentile: 5.09). We observed that most associations could vary by at least 1.15-fold in the magnitude of the HR and by one order of magnitude (log10(p-value)) in the level of statistical significance and much larger changes were not uncommon. We observed a weak correlation between RHR and RP (Fig. S2, $\rho$ =0.09, p=0.06).

Returning to the prototypical examples that we discussed above, the RHR for vitamin D and thyroxine was moderate 1.14 ($44^{th}$ percentile) and 1.15 ($51^{st}$ percentile) (Fig. 4C, Fig. 2AB). However, their RPs were among the largest and equal to 4.7 ($93^{rd}$ percentile) and 2.90 ($84^{th}$ percentile), respectively (Fig. 4D, Fig. 2AB). For urinary creatinine, the scenarios of adjustment had less prominent VoE. The RHR and RP for urinary creatinine was 1.07 ($5^{th}$ percentile) and 0.98 ($47^{th}$ percentile) (Fig. 4CD). The RHR for $\alpha$-tocopherol (with the *Janus* effect) was higher (1.21, $71^{st}$ percentile, Fig. 4C). Variables that demonstrated multimodality, such as serum cadmium and triglycerides, tended to have larger VoE. For example, serum cadmium had a RHR of 1.29 ($82^{nd}$ percentile) and one of the highest RPs, 8.29 ($99^{th}$ percentile). Serum triglycerides had a RHR of 1.18 ($64^{th}$ percentile) and an RP of 1.93 ($73^{rd}$ percentile).

### Increasing the set of adjusting variables for serum Vitamin D

We increased the set of adjusting variables (n=19) to observe differences in the VoE distribution under new assumptions for the association between Vitamin D and mortality (Fig. S3). We observed that as the number of possible models expanded from 8,192 to 524,288, there was not a dramatic shift in the RHR/RP (Fig. S3 and Fig. 2A) of 1.14/4.7 versus 1.19/4.22 for the original VoE and the VoE under new assumptions (n=19 variables), respectively. However, we did observe further attenuation of the effects as a function of number of adjusting variables and some models now indicated a null association between vitamin D and mortality (Fig. S3).

### Investigation of interactions with the VoE

Effects may vary depending on different levels/categories of adjustment variables, known as "interactions". We extended the VoE methodology to systematically consider interactions

between variables of interest and adjustment variables (see Methods). The strongest interaction included that between platelet count and total cholesterol. Specifically, we observed that the effects for platelet count were systematically opposite depending on cholesterol levels and most of the effects were most significant at higher levels of cholesterol (Fig. S4). Notably, we found that VoE will increase in the presence of an interaction.

## Resolution of the example on vitamin E and mortality

As shown in figures 2D and 4D, the association between higher levels of α-tocopherol points to both higher *and* lower risk for mortality depending on the choice of the adjustment variables in the model in a pattern that is characteristic of what we will call the *Janus effect*, named after the two-faced Roman god. In our primer, we show that a majority (4821 of 8192 or 59%) of models show HR less than 1, indicating decreased risk for death for α-tocopherol. The remainder of models indicate increased risk for death (HR > 1, 41%). A number of models are even significant at a nominal p-value threshold of 0.05 (81 of 8192, or 1%). These nominally significant models ranged in their complexity, from univariate to having up to 8 other adjustment variables. The association between α-tocopherol and mortality is one that is highly sensitive to model choice. This extreme VoE may explain, at least in part, the inability to replicate the observational associations in randomized trials and the hotly debated discrepancies.

## Discussion

Almost all reported findings in observational quantitative research to-date in fields like epidemiology consider only a single or a few modeling scenarios. It is often not clear whether this/these model(s) was/were selected a priori. It is often suspected that selective reporting abounds, i.e. several models are tested and only those with the most impressive results are presented with particular attraction for nominally significant results[39]. There are ongoing efforts to enhance transparency, improve reporting, and reveal whether models are pre-specified, and if so, how[40]. We suggest that instead of presenting a single effect and p-value for an association of interest, one can present the median effect, the median p-value, the RHR and the RP across all possible analyses using different adjustments in addition to the pattern of the VoE, whether a Janus effect exists, and whether there are clear multimodality (and if so, due to what). While we used here $99^{th}$ and $1^{st}$ percentiles to report the RHR and RP, we acknowledge that these percentiles are arbitrary and are used for illustrative purposes. It is possible that selective reporting may cherry pick even more extreme values than these percentiles. Plus, the full number of model specifications may approach infinity. For example, if we consider not only 13 but 417 covariates, this would be $2^{417}$ choices. As we show here, for most associations the magnitude of the effect and even the presence or not of nominal significance may depend on the analytical model used. Presenting the VoE metrics and plots can offer a proper picture of the stability of postulated associations. This may help avoiding surprises when some unstable associations are not replicated in subsequent studies, a common feature in observational research [9, 41] that can lead to contradictory statements and endless debates.

Some of the patterns described here have been documented, albeit in different contexts. For example, Altham has described the attenuation of effect size as a function of increasing the number of covariates [42]. The opposite effect, where effect sizes increase on addition of covariates is a pattern attributed to “suppressor variables” in the psychometric literature [43]. In situations where effect sizes connote no major difference in risk for mortality, we can expect patterns such as “Janus effect”. The VoE framework allows investigators to assess empirically the relative frequency of these patterns.

Model selection may be dependent on the dataset and overfitting causes poorer performance in independent datasets [43–45]. Methods such as Bayesian Model Averaging (BMA) [46, 47], relative importance based on variance decomposition [48], and bootstrap-based procedures aim to avoid selection of overfit models. Specifically, BMA does “all-subsets regression” to infer the best model of those tested while taking into account posterior probability of models and it has been applied to ascertain model uncertainty (e.g., [49]). Relative importance estimates the increase of variance explained when individual variables are added to a linear regression model. Stability of models, such as those ascertained via relative importance measures, can also be assessed through other simulation methods, such as the “bootstrap” [44]. In bootstrap-based procedures, an automated model selection procedure is performed on many random samples from the dataset. Variables that are selected in multiple random samplings of the data are hypothesized to be robust candidates to describe associations in an independent dataset. While related to our method, these methods aim to select the best possible model(s). The VoE, on the other hand, focuses on how much the effect sizes change due to model selection. Even though a specific model may have better statistical support over other models, it is often impossible to isolate a single model that conceptually is definitely the best versus thousands of competitors.

Here, we described the VoE empirically as a function of estimates from 8,192 specifications. However, these different specifications may not define the entire scope of model flexibility that can be encountered in large discovery-based observational studies with many variables. Visible and invisible “multiplicities” can be introduced anywhere in the design and analysis process[50, 51]. For example, eligibility criteria can be altered to include fewer or more participants in the analysis; in our study, we computed the VoE over a fixed sample size. Further still, we do not explore flexibility in outcome definition (e.g., cardiovascular-related mortality vs. all-cause mortality). Flexibility in defining outcomes may result in outcome reporting bias with further instability in the results [52, 53]. The VoE quantifies how the single or few of “main effect[s]” reported in biomedical investigations are but a handful of findings in a vast space of possible findings and, perhaps, it may behoove of investigators to report all possible “main effects” for a modeling scenario.

We emphasize there is no consensus for what variables should be included in a model or a model selection procedure even for high priority risk factors and outcomes, such as physical activity and mortality. For example, in a comprehensive meta-analysis of association between physical activity and mortality, Samitz and colleagues observed different combinations of 26 variables were utilized in individual investigations [20]. Out of the 80 independent studies, the most common variables in the individual analyses included age (in 86.3% of studies), smoking (78%), and body mass index (55%). But, some studies included

mental disorder (5%), hormone replacement therapy (4%), liver disease (4%), and renal disease (4%). Because of lack of consensus, we selected for exploration the most common adjusting variables from these independent studies.

Nevertheless, additional and/or other adjustment variables can be chosen. In a supplementary analysis for Vitamin D, we selected 19 variables resulting in exponentially more models to compute (524,288) and impractical as codified here. Estimating the VoE for 20 or more variables may be possible with algorithms such as “leaps and bounds” [54].

The fundamental analytical methods used to compute effects, such as regression, are also subject to choice. For example, we conducted survival analyses with the Cox semi-parametric tool, but one could have chosen other modeling methods. An empirical study testing 8 different statistical methods for adverse effect identification found that results were often discrepant [55].

The way datasets are sampled and dataset heterogeneity may also lead to different results[56]. If several of these additional options are considered as mainstream options, the VoE is expected to increase further. Thus, VoE should be presented with specification of the type and range of choices that it has encompassed. Finally, in any observational research, there may be many unaccounted, unmeasured, or unknown confounders. These may invalidate a claimed association, even when it shows limited VoE and consistently strong and statistically significant results based on the models that can be estimated based from available data and measurements. For example, while serum vitamin D was found to have consistently significant results in the VoE analyses with 15 considered variables, other studies do not show such consistency[57] and this inconsistency became apparent only when 19 variables were considered. Therefore, the VoE estimates should be seen as a lower limit of the uncertainty that is inherent in observational associations. Routine adoption of the VoE concept and analyses may place observational associations in a more proper context, acknowledging their potential instability.

The research process, including model selection, may be driven by incentives for the investigator themselves [58]. To avoid an excess of cherry picking of model specifications with most favorable/significant results, incentives for researchers must change. Journal editors, funding agencies, and referees must coordinate efforts to ensure researchers are reporting transparent and unbiased results. Furthermore, analytical software and computational infrastructure must be broadly accessible. Toward this goal, the data, code, and figures for this Primer are available in a repository located here: http://chiragjpgroup.org/voe.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The authors thank Profs. Andrew Gelman and Bin Yu for their comments. All data, software code, and additional figures can be found at the following website: http://chiragjpgroup.org/voe

**Funding:** This work was supported by a National Institute of Environmental Health Sciences K99 ES023504, R21 ES0250252 and a PhRMA foundation award to CJP.

## Abbreviations

| | |
|---|---|
| **VoE** | Vibration of Effects |
| **NHANES** | National Health and Nutrition Examination Survey |
| **NDI** | National Death Index |
| **BMI** | Body mass index |
| **HR** | Hazard Ratio |
| **RHR** | Relative Hazard Ratio |
| **RP** | Relative p-value |

## References

1. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med. 2007; 4:e352. [PubMed: 18076282]
2. Ioannidis JPA. Why most discovered true associations are inflated. Epidemiology. 2008; 19:640–8. [PubMed: 18633328]
3. Sala-I-Martin X. I Just Ran Two Million Regressions. Am Econ Rev. 1997; 87:178–83.
4. Leamer E. Sensitivity Analyses Would Help. Am Econ Rev. 1985; 57:308–13.
5. Leamer E, Leonard H. Reporting the Fragility of Regression Estimates. Rev Econ Stat. 1983; 65:306–17.
6. Akil H, Martone ME, Van Essen DC. Challenges and opportunities in mining neuroscience data. Science. 2011; 331:708–12. [PubMed: 21311009]
7. Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. False-positive results in cancer epidemiology: a plea for epistemological modesty. J Natl Cancer Inst. 2008; 100:988–95. [PubMed: 18612135]
8. Fiedler K. Voodoo Correlations Are Everywhere -- Not Only in Neuroscience. Perspect Psychol Sci. 2011; 6:163–71. [PubMed: 26162135]
9. Young SS, Karr A. Deming, data and observational studies. Significance. 2011; 8:116–20.
10. Pearl, J. Causality: Models, Reasoning and Inference. Cambridge: Cambridge University Press; 2009.
11. Young SS, Yu M. Association of bisphenol A with diabetes and other abnormalities. J Am Med Assoc. 2009; 301:720–1. author reply 1–2.
12. Fillenbaum GG, Burchett BM, Blazer DG. Identifying a national death index match. Am J Epidemiol. 2009; 170:515–8. [PubMed: 19567777]
13. Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ, et al. Compendium of physical activities: an update of activity codes and MET intensities. Med Sci Sports Exerc. 2000; 32:S498–504. [PubMed: 10993420]
14. Patel CJ, Cullen MR, Ioannidis JP, Butte AJ. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. Int J Epidemiol. 2012; 41:828–43. [PubMed: 22421054]
15. US Department of Health and Human Services. 2008 Physical Activity Guidelines for Americans. Washington DC: US Department of Health and Human Services; 2008.
16. Blanton CA, Moshfegh AJ, Baer DJ, Kretsch MJ. The USDA Automated Multiple-Pass Method accurately estimates group total energy and nutrient intake. J Nutr. 2006; 136:2594–9. [PubMed: 16988132]

17. U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center, Food Surveys Research Group, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, et al. What We Eat in America, NHANES 2003–2004. 2003.
18. U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center, Group FSR, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, et al. What We Eat in America, NHANES 2001–2002. 2001.
19. U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center, Food Surveys Research Group, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, et al. What We Eat in America, NHANES 1999–2000. 1999.
20. Samitz G, Egger M, Zwahlen M. Domains of physical activity and all-cause mortality: systematic review and dose-response meta-analysis of cohort studies. Int J Epidemiol. 2011; 40:1382–400. [PubMed: 22039197]
21. Arrieta A, Russell LB. Effects of leisure and non-leisure physical activity on mortality in U.S. adults over two decades. Ann Epidemiol. 2008; 18:889–95. [PubMed: 19041587]
22. Tellez-Plaza M, Navas-Acien A, Menke A, Crainiceanu CM, Pastor-Barriuso R, Guallar E. Cadmium exposure and all-cause and cardiovascular mortality in the U.S. general population. Environ Health Perspect. 2012; 120:1017–22. [PubMed: 22472185]
23. Flegal KM, Graubard BI, Williamson DF, Gail MH. Cause-specific excess deaths associated with underweight, overweight, and obesity. J Am Med Assoc. 2007; 298:2028–37.
24. Ford ES, Zhao G, Tsai J, Li C. Vitamin D and all-cause mortality among adults in USA: findings from the National Health and Nutrition Examination Survey Linked Mortality Study. Int J Epidemiol. 2011; 40:998–1005. [PubMed: 21266455]
25. Patel CJ, Rehkopf DH, Leppert JT, Bortz WM, Cullen MR, Chertow G, et al. Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey. Int J Epidemiol. 2013; 42:1795–810. [PubMed: 24345851]
26. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS ONE. 2010; 5:e10746. [PubMed: 20505766]
27. Therneau, T.; Grambsch, P. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2010.
28. Lumley, T. Complex Surveys: A guide to analysis using R. Hoboken: Wiley; 2010.
29. Patsopoulos NA, Tatsioni A, Ioannidis JP. Claims of sex differences: an empirical assessment in genetic associations. J Am Med Assoc. 2007; 298:880–93.
30. Gaziano JM. Vitamin E and cardiovascular disease: observational studies. Ann N Y Acad Sci. 2004; 1031:280–91. [PubMed: 15753154]
31. Sung L, Greenberg ML, Koren G, Tomlinson GA, Tong A, Malkin D, et al. Vitamin E: the evidence for multiple roles in cancer. Nutr Cancer. 2003; 46:1–14. [PubMed: 12925298]
32. Song Y, Cook NR, Albert CM, Van Denburgh M, Manson JE. Effects of vitamins C and E and beta-carotene on the risk of type 2 diabetes in women at high risk of cardiovascular disease: a randomized controlled trial. Am J Clin Nutr. 2009; 90:429–37. [PubMed: 19491386]
33. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet. 2002; 360:23–33. [PubMed: 12114037]
34. Yusuf S, Dagenais G, Pogue J, Bosch J, Sleight P. Vitamin E supplementation and cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. N Engl J Med. 2000; 342:154–60. [PubMed: 10639540]
35. Virtamo J, Pietinen P, Huttunen JK, Korhonen P, Malila N, Virtanen MJ, et al. Incidence of cancer and mortality following alpha-tocopherol and beta-carotene supplementation: a postintervention follow-up. JAMA. 2003; 290:476–85. [PubMed: 12876090]
36. Vivekananthan DP, Penn MS, Sapp SK, Hsu A, Topol EJ. Use of antioxidant vitamins for the prevention of cardiovascular disease: meta-analysis of randomised trials. Lancet. 2003; 361:2017–23. [PubMed: 12814711]

37. Miller ER 3rd, Pastor-Barriuso R, Dalal D, Riemersma RA, Appel LJ, Guallar E. Meta-analysis: high-dosage vitamin E supplementation may increase all-cause mortality. Ann Intern Med. 2005; 142:37–46. [PubMed: 15537682]
38. Wright ME, Lawson KA, Weinstein SJ, Pietinen P, Taylor PR, Virtamo J, et al. Higher baseline serum concentrations of vitamin E are associated with lower total and cause-specific mortality in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study. Am J Clin Nutr. 2006; 84:1200–7. [PubMed: 17093175]
39. Gotzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. BMJ. 2006; 333:231–4. [PubMed: 16854948]
40. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007; 370:1453–7. [PubMed: 18064739]
41. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005; 2:e124. [PubMed: 16060722]
42. Altham P. Improving the precision of estimation by fitting a model. J R Stat Soc B. 1984; 46:118–9.
43. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004; 66:411–21. [PubMed: 15184705]
44. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Stat Med. 1989; 8:771–83. [PubMed: 2672226]
45. Copas JB. Regression, Prediction and Shrinkage. J R Stat Soc B. 1983; 45:311–54.
46. Hoeting J, Madigan D, Raftery AE, CTV. Bayesian Model Averaging: A Tutorial. Statistical Science. 1999; 14:382–417.
47. Claeskins, G.; Hjort, N. Model Selection and Model Averaging. Cambridge: Cambridge University Press; 2008.
48. Grömping U. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. The American Statistician. 2007; 61:139–47.
49. Clyde M. Model uncertainty and health effect studies for particulate matter. Environmetrics. 2000; 11:745–63.
50. Berry D. Multiplicities in cancer research: ubiquitous and necessary evils. J Natl Cancer Inst. 2012; 104:1124–32. [PubMed: 22859849]
51. Berry D. The difficult and ubiquitous problems of multiplicities. Pharm Stat. 2007; 6:155–60. [PubMed: 17879328]
52. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008; 3:e3081. [PubMed: 18769481]
53. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ. 2005; 330:753. [PubMed: 15681569]
54. Furnival GM, Wilson RW Jr. Regressions by Leaps and Bounds. Technometrics. 1974; 16:499–511.
55. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med. 2012; 31:4401–15. [PubMed: 23015364]
56. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the Impact of Database Heterogeneity on Observational Study Results. Am J Epidemiol. 2013; 178:645–51. [PubMed: 23648805]
57. Pittas AG, Chung M, Trikalinos T, Mitri J, Brendel M, Patel K, et al. Systematic review: Vitamin D and cardiometabolic outcomes. Ann Intern Med. 2010; 152:307–14. [PubMed: 20194237]
58. Glaeser EL. Researcher incentives and empirical methods. 2006

## What is new?

### Key Findings

We have developed a standardized, automated procedure to estimate the "Vibration of Effects" (VoE) to measure the instability of observational associations between variables under different modeling scenarios; for example, the vitamin E variant α-tocopherol had a VoE that indicated both higher *and* lower risk for mortality.

### What this adds to what is known

Model selection influences how we infer from associations, such as hazard ratios and p-values. In this report, we provide a way to measure and visualize how much model selection influences variability in hazard ratios and p-values.

### Implications

The VoE can offer a proper picture of the instability of postulated associations and allow more judicious claims for associations in observational studies; VoE can be widely adopted in observational studies.

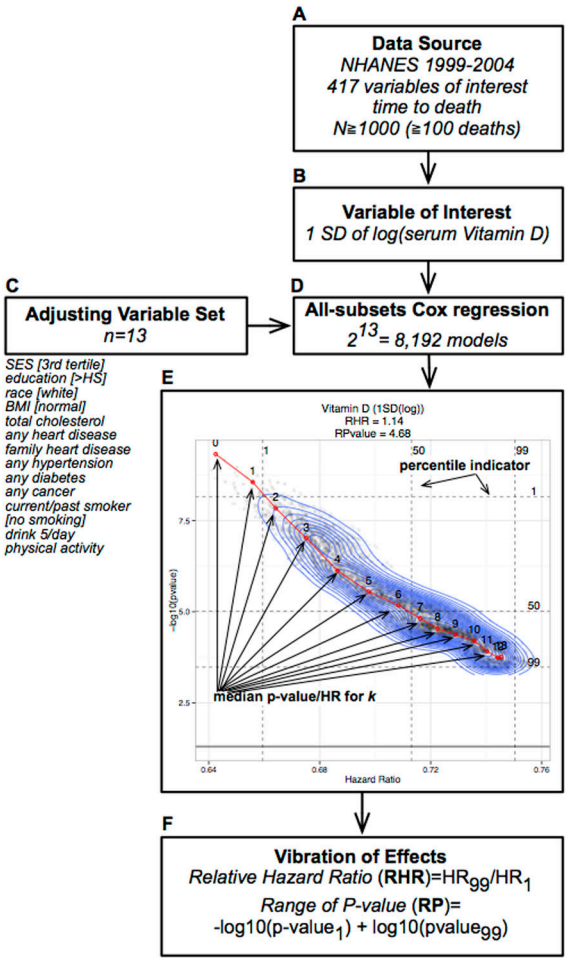


**Fig. 1.**

VoE computation schematic. (*A*) Data source. (*B*) Choose a variable of interest. (*C*) Construct a set of adjustment variables from a set of 13 socioeconomic, demographic, or health-related variables. Reference level is in the square brackets. (*D*) All-subsets Cox regression for each 8,193 models. (*E*) Visualization ("volcano plot") of −log10(p-value) versus effect size (HR). The median HR and p-value the number of adjustment variables (*k*) in the model is in red. The 1st, median, and 99th percentile of the −log10(pvalue) and HR are depicted in the dotted line. (*F*) Compute VoE summary statistics, the Relative Hazard Ratio (RHR) and Relative P-value (RP).

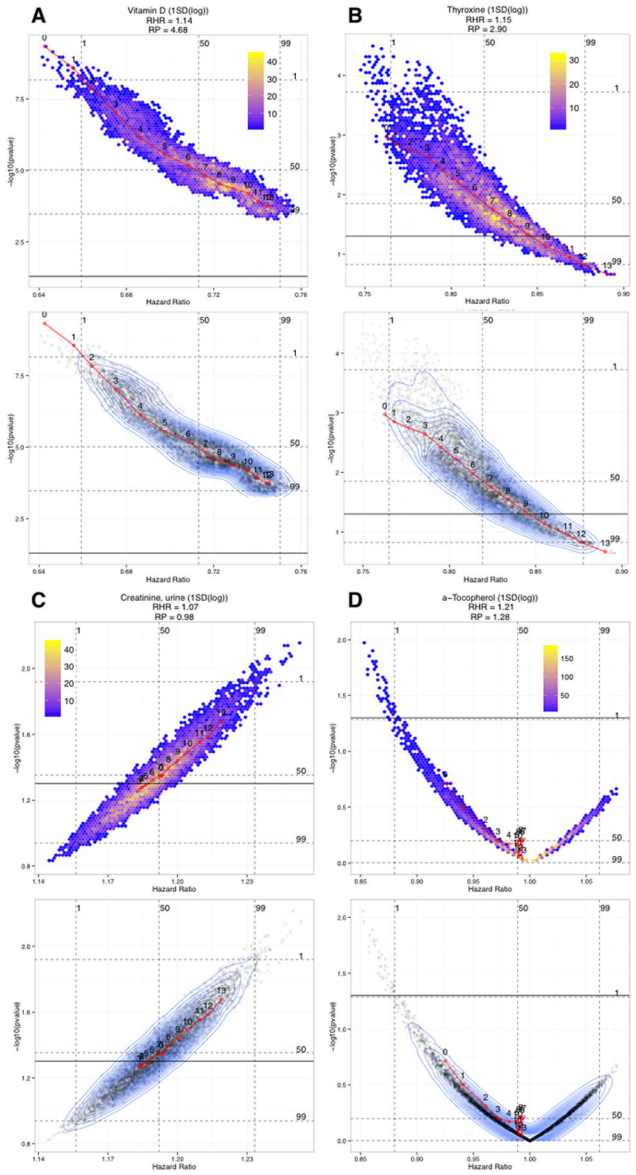


**Fig. 2.**

Volcano plots visualizing the VoE for four examples, (*A*) Serum Vitamin D, (*B*) Serum Thyroxine, (*C*) Urinary Creatinine, (*D*) Serum α-Tocopherol. 2D histogram representation in upper panel and contour scatter plot is in lower panel. All effects are for a 1SD change in logged level of variable interest.

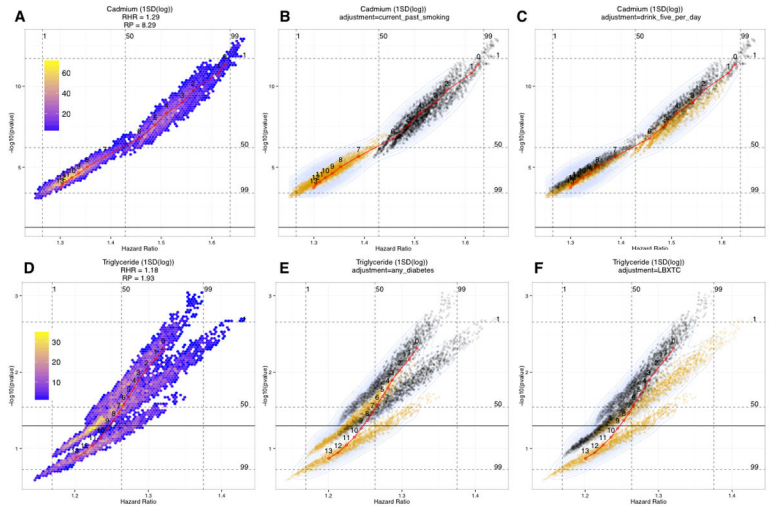


**Fig 3.**

Volcano plots visualizing VoE for three examples with multiple "modes". (*A*) The 2D histogram for 1SD increase of the logarithm of serum cadmium, (*B*) Volcano scatter plot with of serum cadmium if smoking was included in the model (yellow) or smoking not included in model (black). (*C*) Volcano scatter plot for serum cadmium models with drink five per day (yellow) or models without drink five per day (black). *(D)* The 2D histogram for 1SD increase of the logarithm of serum triglycerides, (*E*) With total cholesterol included in the model (yellow) or total cholesterol not included in model (black). (*F*) With any diabetes (yellow) or models without any diabetes (black).

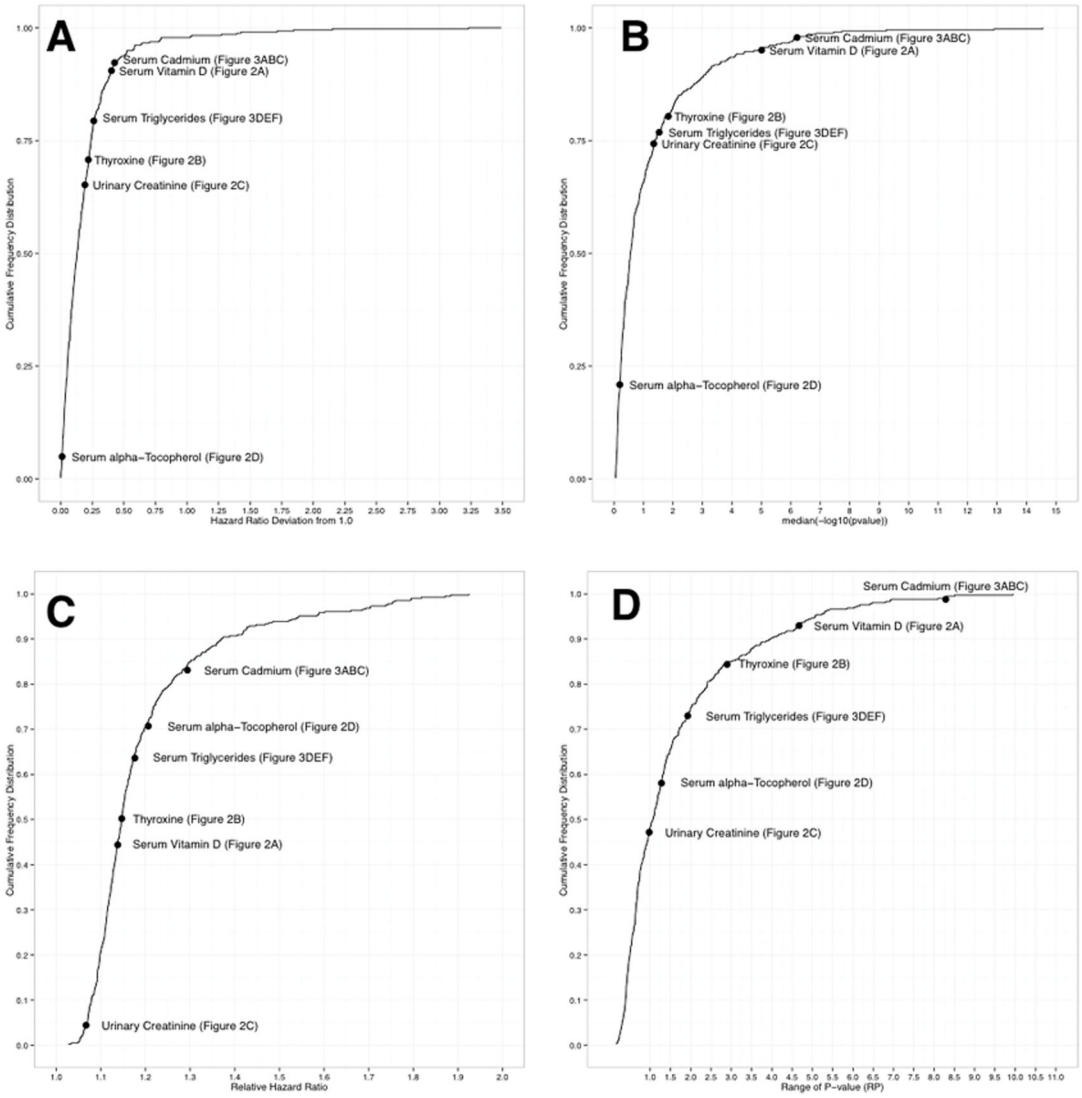


**Fig. 4.**

Cumulative distributions of VoE for 417 variables. (*A*) Absolute deviation of HR from 1, (*B*) log10(pvalue), (*C)* Relative Hazard Ratio (RHR), (*D*) Relative P-value (RP). Examples shown in figures 1–3 are shown in the distribution.