Correlated Confounding Variables Are Not Easily Controlled for in Large Survey Research

William H. Press
Computer Science Department and
Oden Institute for Computational Engineering and Sciences
The University of Texas at Austin

December 2, 2025

Abstract

Results in epidemiology and social science often require the removal of confounding effects from measurements of the pairwise correlation of variables in survey data. This is typically accomplished by some variant of linear regression (e.g., "logistic" or "Cox proportional"). But, knowing whether all possible confounders have been identified, or are even visible (not latent), is in general impossible. Here, we exhibit two examples that frame the issue. The first example proposes a highly unlikely hypothesis on drug use, draws data from a large, respected survey, and succeeds in "proving" the implausible hypothesis, despite regressing out more than 20 confounding variables. The second constructs a "metamodel" in which a single (by hypothesis unmeasurable) latent variable affects many mutually correlated confounders. From simulations, we derive formulas for the magnitude of spurious association that persists even as increasing numbers of confounders are regressed out. The intent of these examples is for them to serve as cautionary tales.

Significance Statement

"Correlation does not imply causation," is a verity much repeated, even by investigators who disregard it. Still, a correlation between two measurable features can indicate that one *may* be in the other's causal chain. However, not all correlations are equally compelling. The association of two variables may be solely due to so-called confounding variables that influence both without the two having any direct relationship. There exist established statistical techniques to control for such confounders. Here, first in a real-data case study and then in an idealized "metamodel", we show and quantify the pitfalls of such standard techniques. These are cautionary examples of effects understood in theory, but too often ignored in practice.

1 Introduction

Since the reader may find the title either tautological or else meaningless across the huge range of possible survey research, some framing and explanation is necessary. Throughout, we use the term "confounding" in its statistical meaning, namely a variable that influences both the dependent variable and independent variable in a study, causing a spurious association between them that is not evidence of causality.[1]

As an example of the more generally dangerous "undisclosed flexibility in data collection and analysis" [2], the pitfalls of compensating (or not compensating) for covariate variables have been widely discussed [3]. False positive [4] or false negative [5] results may be produced. General methodologies (e.g., "deconfounders") may [6] or may not [7] be mitigations. Other methods for inferring causality may be brought to bear [8, 9]. This paper joins the cautionary chorus in contributing two very specific examples, one based on actual data, the other on a simple model.

1.1 Motivating This Paper

In January, 2025, an Advisory by the U.S. Surgeon General [10] flagged alcohol use as "a leading preventable cause of cancer in the United States." Among statistics put forth in the Advisory was that that the relative risk of breast cancer in women increases by $\sim 10\%$ with an additional ~ 1 drink per day, and increases further with greater alcohol consumption. These findings were widely reported in the media [11, 12, 13]. A national network news broadcast summarized one finding as, "over 16 percent of all breast cancer cases in the U.S. in 2019 were alcohol-related." [14] The near-simultaneous release of a National Academies report [15] on the same subject added some support and some confusion: That study concluded "with moderate certainty" that alcohol was associated with a higher risk of breast cancer, but that the dose-dependence of the association could be concluded only with "low certainty".

The Advisory cited studies implicating alcohol in multiple cancer types, including oral cavity, pharynx, larynx, esophagus, breast (women), liver, colon, and rectum. Among the cited studies, one of the largest and most thorough was an Australian study of 226,162 participants that reported hazard ratios and their 95% confidence limits (roughly, fractional increases in risk) for 26 types of cancer associated with a 7 drink weekly increase in alcohol consumption [16]. Of the 26 types, 7 were statistically significant with 95% confidence and p-values ≤ 0.05 (liver, esophagus, upper digestive tract, pharynx and larynx, colon, breast, colorectum).

Recognizing that various demographic, socioeconomic, and lifestyle variables might confound measurement of the direct relationship between alcohol and cancer, the Australia study removed by Cox proportional regression (e.g., [17]) an extensive list of potentially confounding variables including sex, education, household income, health insurance status, partner status, country of birth, smoking status and intensity, body mass index, physical activity level, remoteness, and (for individual types of cancers) appropriate subsets of time spent outdoors, skin-tone, diet, parity, menopausal status,

hormonal contraception use, and aspirin use. Figure 1 shows findings from the Australia study, re-plotted in a format in common with later figures in this paper.

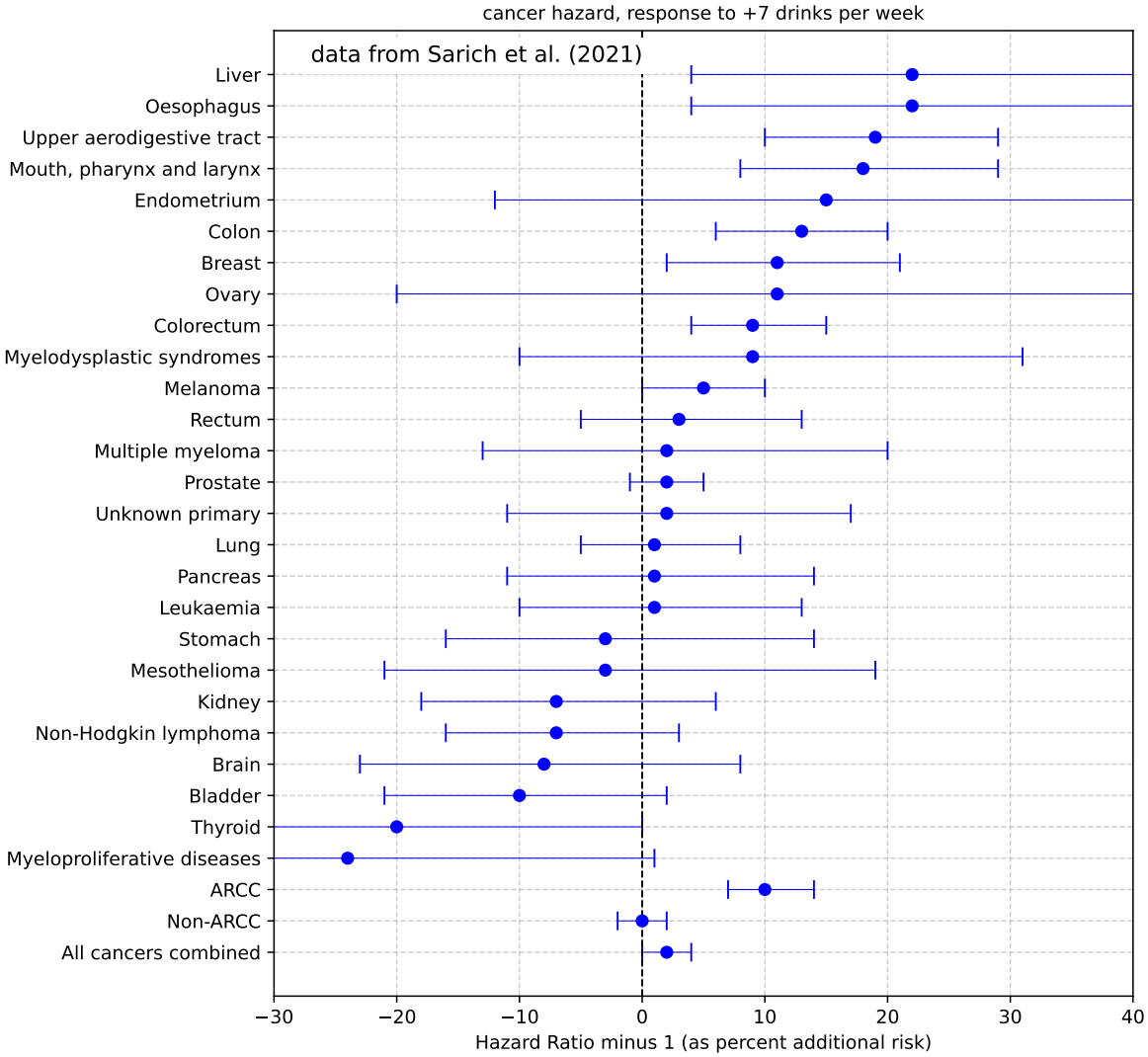

Figure 1: From Sarich et al. (ref. [16]), cancer hazard ratios (roughly, fractional increased risk) associated with increased alcohol use in an Australian study of 226,162 participants. The effects of a large number of potentially confounding variables (see text for partial list) were removed in these authors' analysis by a standard regression technique. The error bars are 95% confidence intervals.

Nothing in this paper is intended to cast doubt on the fact that alcohol consumption causes otherwise preventable cancers. As both the Surgeon General's and National Academies' reports note, causal biochemical pathways are well established: alcohol is metabolized to acetaldehyde, a carcinogen. Alcohol's production of reactive oxygen species can produce inflammation and activate cancer pathways. Alcohol is known to alter hormone levels that are implicated in breast cancer. Other effects are likewise well documented. The Advisory explicitly met the accepted Bradford Hill criteria [18] for

whether an association merits consideration as causal.

Of concern here is the more general question of whether, in purely survey-based epidemiological and demographic studies, the reporting of derived relative risks for multiple endpoints (here, different cancer types as in Figure 1) should be viewed as quantitatively reliable in the face of removal by regression of many related confounding variables whose causal relationships are likely complex and generally unknown.

Such regressions are not uncommon. A much publicized study of the cancer risk in 92,000 French adults posed by food additives [19] controlled for age, BMI, height, physical activity, smoking status, number of cigarettes smoked, educational level, family history of cancer, energy intake from lipids and (separately) sugars, sodium, fibre, dairy products; and, for breast cancer, a half dozen further variables.

There is, of course, no universal answer to the question of reliability. Here, to illuminate some of the issues involved, we construct two examples: The first example proposes a highly unlikely hypothesis on drug use, draws data from a large, respected survey [20], and succeeds in "proving" the implausible hypothesis, despite regressing out more than 20 confounding variables. The second constructs a "metamodel" in which a single (by hypothesis unmeasurable) latent variable affects many mutually correlated confounders.

1.2 Fictitious Case Study on Real Data

We posit a research project on the misuse of the inhalants listed in Table 1. An investigator entertains the hypothesis that inhalant use is directly attributable to alcohol consumption. Here attributable means not the result of some confounding variable that causes both inhalant and alcohol use, but rather a statistically significant association that remains after all such confounders are accounted for—therefore one that might prove to be causal. While seemingly unlikely, the hypothesis is not immediately dismissible: Perhaps individuals tend to use inhalants preferentially when they are drunk; or perhaps inhalation of alcohol fumes in the course of drinking in some way rewires the brain to favor other inhalant use. At any rate, the investigator sets out to test the hypothesis on actual survey data.

It is an important point that, for this fictitious case study, the hypothesis of direct association is on its face unlikely, while the existence of correlated confounding variables is almost certain. We want to explore circumstances under which accepted methods can produce unreliable or misleading results, in this case statistically significant false positives.

While the investigator's hypothesis may be unlikely, his methodology is assumed to be by-the-book. An ideal data source is the (real, not fictitious) National Survey on Drug Use and Health (NSDUH) [20] conducted annually by the Substance Abuse and Mental Health Services Administration (SAMHSA), an agency within the U.S. Department of Health and Human Services (HHS). NSDUH 2023 makes available a Public Use File (PUF) with the responses of 56,705 respondents (anonymized as to identity) in 2,636 columns. (Depending on their answers to previous questions, not all

respondents are asked all questions.)

Inhalant	**Abbreviation**
Amyl nitrite, 'poppers,' locker room odorizers, or 'rush'	AMYLNIT
Correction fluid, degreaser, or cleaning fluid	CLEFLU
Gasoline or lighter fluid	GAS
Glue, shoe polish, or toluene	GLUE
Halothane, ether, or other anesthetics	ETHER
Lacquer thinner, or other paint solvents	SOLVENT
Lighter gases, such as butane or propane	LGAS
Nitrous oxide or 'whippits'	NITOXID
Felt-tip pens, felt-tip markers, or magic markers	FELTMARKR
Spray paints	SPPAINT
Computer keyboard cleaner, also known as air duster	AIRDUSTER

Table 1: List of Inhalants Studied and Their Abbreviations

Among various columns relating to frequency and intensity of alcohol use, the investigator selects as most reliable the response to "total number of days used alcohol in past 12 months" (column ALCYRTOT), reasoning that self-reported numbers of days is likely to be more accurate than self-reported numbers of drinks. ALCYRTOT, in the range $0 \ldots 365$, is thus the independent variable.

The dependent variables of interest are the binary (yes/no) answers to the questions, "Have you ever, even once, inhaled _____ for kicks or to get high?", separately asked for each inhalant listed in Table 1.

Recognizing that confounding variables are likely, the investigator includes in the analysis, and controls for by logistical regression [21, 22], these potentially confounding responses (see Methods for the NSDUH column names and further details): sex, age, race, marital status, family income, education level, government financial aid status, health insurance status, employment status, household size, urban vs. rural residence, self-reported overall health, and body mass index. (We refer to this list as "A".)

2 Results

Figure 2 shows the results of the analysis just described, obtained using a standard Python package for logistic regression models (see Methods). Plotted for each inhalant is its so-called relative risk, the fractional change in probability of the dependent variable under a change of (here) a change from zero to one day per week (on average) with alcohol consumption. Percentages listed after the inhalant abbreviations are the baseline reported prevalence of use, so that the relative risk is the fractional increase relative to these values. Error bars shown are 95% confidence intervals.

Consider first the dark blue values and errorbars, with confounders "A". One sees a highly statistically significant association between alcohol use and inhalant use for all

inhalants tested. Relative risks vary between about 9% and about 23%, with strongly significant differences between the low end (felt-tip markers) and the high end (lighter gasses butane and propane). Notable is that these values are found after removal by regression of the long (the investigator argues, exhaustive) list of possible confounders,"A".

We may be more skeptical than the investigator that the included confounding variables are sufficiently exhaustive. Experience may tell us that there are recognizable personality types like "thrill seeking" or "addiction prone," not perfectly captured by the socioeconomic and health-related confounders included by the investigator.

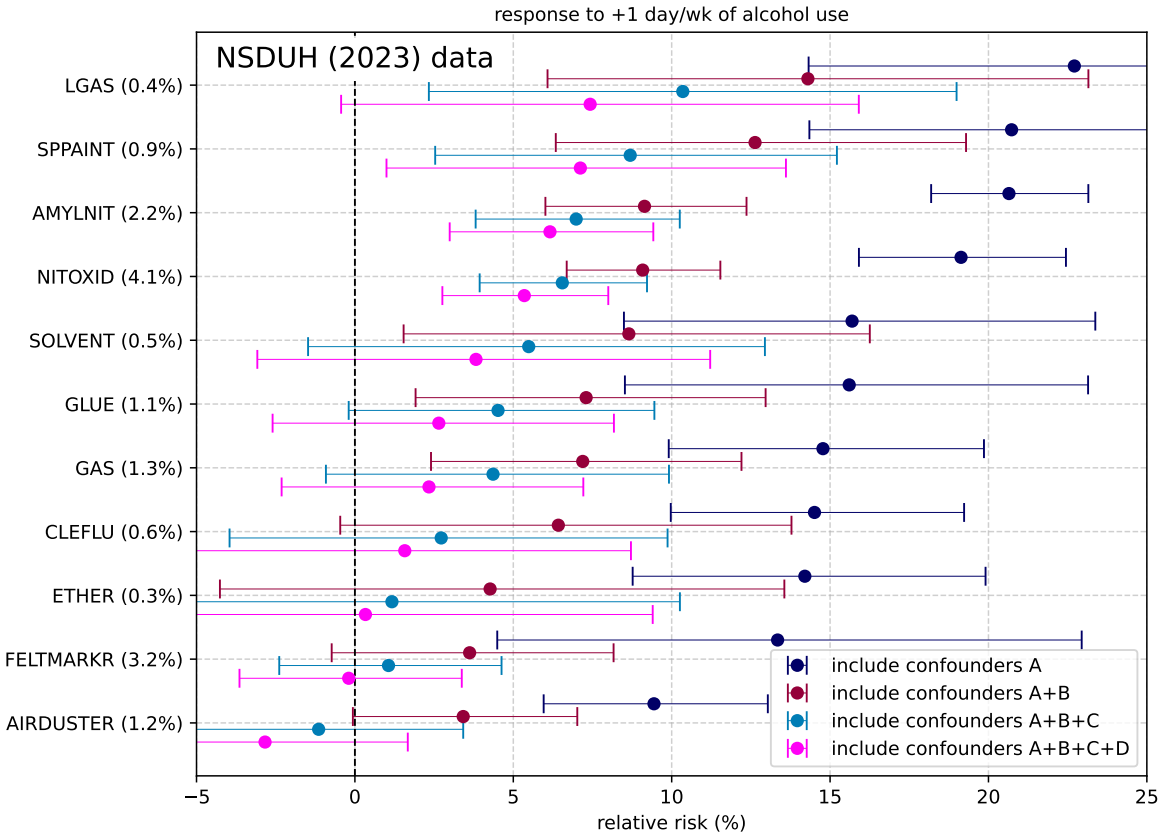

Figure 2: Change in the use of inhalants (Table 1) with increase in alcohol consumption, as predicted by logistic regression after correction for multiple potentially confounding demographic and socioeconomic factors (termed "A", see text). Shown are relative risks associated with one additional day per week of alcohol use. Error bars are 95% confidence intervals. "B" adds marijuana use as a confounding variable. "C" adds 7 hallucinogens as confounders. "D" adds cigarette use as a confounder. As "A" through "D" are successively included in the analysis, the apparent associations with alcohol are successively reduced, most to statistical non-significance. The remaining signals may be a genuine direct association, or may merely indicate that not all confounders are known.

Several columns in the NSDUH questionaire pertain to marijuana use, including (MJEVER) whether ever used, with a 45% positive response, and "total number of days used marijuana/cannabis in past 12 months" (MJYRTOT), the exact analog of our

independent variable ALCYRTOT for alcohol. We reason that adding these columns to the regression as confounders should correct for addictive (or other related) personality types that could also be associated with alcohol use.

Results are shown as the brown values and errorbars in Figure 2, labeled as "A+B". We see that indeed all of the relative risks associated with alcohol use are reduced, roughly proportionally, with four now not statistically significant at the 95% confidence level. Lighter gas, spray paint, amyl nitrite, and nitrous oxide use remain highly significantly associated with alcohol use.

But, we can continue in this manner. Blue values and errorbars labeled "A+B+C" show the result of adding reported use of each of 7 hallucinogens as confounders (LSD, PCP, peyote, ecstacy/molly, ketamine, dmt/amt/foxy, and Salvia divinorum). The relative risks are all slightly further reduced, again roughly in proportion. Three additional inhalants lose statistical significance. Finally for this analysis, the pink values and errorbars, labeled "A+B+C+D", show the result of adding cigarette use and intensity in the previous 30 days (CIGEVER and CIG30USE) as confounders, producing additional reductions in relative risk. Lighter gas, originally the largest apparent signal, now loses statistical significance, in part because of large error bars, a consequence of its small prevalence in the population. Howeve spray paints, amyl nitrite, and nitrous oxide remain statistically significant (95% confidence interval).

We did not identify other columns in NSDUH as plausible confounding variables. A conclusion from this very extensive data set is that spray paints, amyl nitrite, and nitrous oxide inhalent use is directly (i.e., not just via confounding socioeconomic and lifestyle-related variables) associated with alcohol use with high statistical significance, and that several other inhalants may also be, at lower levels of association and with lower statistical confidence.

That may be the conclusion, but should we rely on it? A danger sign is the sequence of reductions in the magnitude of the associations seen as additional confounding variables are introduced into the analysis. Have we overlooked confounding variables that could whittle away the apparent direct association to statistical insignificance? Or, what if such variables exist, but were not included in the survey, hence not accessible to us?

2.1 Metamodel for Mutually Correlated Confounding Variables

A metamodel, so called, is a model of models. Each line in Figure 2 is a model, specifically the best-fit logistic regression model to a multivariate set of survey variables (including one of primary interest) to a dependent variable. Our metamodel, by contrast, abstracts both the data and the model to a simple set of idealizations. Characteristics seen in the metamodel may then shed light on possible characteristics of real models and their data.

Complex surveys ask questions many of which will show mutual pairwise correlations, not because the one answer directly causes the other, but because both are

embedded in some complicated, unknown causal structure. The simplest abstraction of this with (importantly) no actual direct causal relation among the measured variables is, for a population indexed by i with responses R_{ij} indexed by j,

$$\begin{aligned} \forall i: &\quad Q_i \in \{-1,1\} \sim \text{Bernoulli } \{0.5, 0.5\} \\ \forall i,j: &\quad R_{ij} \in \{0,1\} = (Q_i Y_{ij} + 1)/2 \text{ where } Y_{ij} \in \{-1,1\} \sim \text{Bernoulli } \{1-p, p\} \end{aligned} \tag{1}$$

with p a fixed probability in the range $0.5 < p < 1$. Putting this in words, each respondent randomly carries a Bernoulli latent variable; and that respondent's responses to all questions randomly agree with their latent variable with probability $p > 0.5$, that is, more often than not but still randomly.

The pairwise correlation r between any two responses is readily calculated to be

$$r = (2p-1)^2 = b^2, \quad b \equiv 2p - 1 \tag{2}$$

where where $-1 < b < 1$, parametrizes p as a quantity termed "bias". Notice that even fairly large values of p, say 0.75, yield only small values of r, 0.25 and even smaller values of r^2, the fraction of one variable's variance explained by another, 0.0625, which would be barely noticeable in a scatter plot of the two variables. While equation (1) posits all positive mutual correlations, one can easily extend the model to have two groups of columns, with mutually positive correlations within each group and negative correlations cross-groups.

Selecting any one of its equivalent columns, label it 0, as a dependent variable of interest ("inhalant use"), and any other column as the independent variable or predictor of interest ("alcohol use"), label it 1, the metamodel finds the best-fitting logistic regression for explaining the former by the latter. Recognizing the existence some number $k-1$ of confounding variables, their columns (2 through k) are also included in the fitted model,

$$\text{logit } (p_i) \equiv \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 R_{i1} + \sum_{j=2}^{k} \beta_j R_{ji} \tag{3}$$

where $p_i = \text{Prob } (R_{0i} = 1)$ predicts the dependent variable 0 for respondent i. The coefficient β_1 of the predictor in the fitted model, commonly termed the logit coefficient or log-odds coefficient, also with its standard error σ_1, is the output of one realization of the metamodel.

We mention in passing that logistic regression (equation 3) is different from the Cox proportional regression used in (e.g.) [16] and [19]. However, the two methods are closely related. The Cox model posits a factorable time function, but is otherwise (up to the distinction between log and logit, negligible for small p_i's) a very similar regression for coefficients β_j,

$$h(t \mid \{R_{ij}\}) = h_0(t) \exp(\beta_1 R_{i1} + \sum_{j=2}^{k} \beta_j R_{ji}). \tag{4}$$

In the metamodel, the selected independent variable is, by construction, just another confounding variable. Even in studies of real data, however, regressions akin to equations (3) or (4) allocate their explanatory coefficients β_i with independent and confounding variables treated mathematically equivalently; that is the defining idea of "regressing out" the confounders. Identifying a particular variable as "independent" is a matter of investigator choice, not mathematics.

We are interested in the average values of β_1 and its σ_1 over many realizations of the model, which will depend only the model parameters p, k, and $N \gg 1$ (the number of respondents). We can define these averages,

$$\mathbb{E}(\beta_1) \equiv \widehat{\beta_1} = \widehat{\beta_1}(p,k), \quad \mathbb{E}(\sigma_1) \equiv \widehat{\sigma_1} = \widehat{\sigma_1}(p,k,N) \tag{5}$$

While it may be possible to characterize these "universal" functions $\widehat{\beta_1}(p,k)$ and $\widehat{\sigma_1}(p,k,N)$ analytically in some limits, it is also easy to approximate them by simulation for any desired values. Doing so (see Methods) we have found purely empirical functional forms that are adequate approximations for present purposes,

$$\begin{aligned} \widehat{\beta_1}(p,k) &\approx \frac{3b^2}{k} \\ \widehat{\sigma_1}(p,k,N) &\approx N^{-1/2}(4+12\,b^5)\left(\frac{k-(1+b)/4}{k}\right) \end{aligned} \tag{6}$$

Figure 3 displays the results of equations (6) for a range of values of mutual correlation r (equivalent to ranges of p or b, see equation 2) and number of confounders $(k-1)$ in essentially the same format as Figure 1, and with the same x-axis scale. Errorbars in Figure 2 are normalized to the same population size as Figure 1. In the ranges of r and k shown one sees a pattern strikingly similar to that in Figure 1, despite the fact that the metamodel has no actual direct causal associations, only confounding variables with mutual correlations r ranging from 0.01 to 0.15.

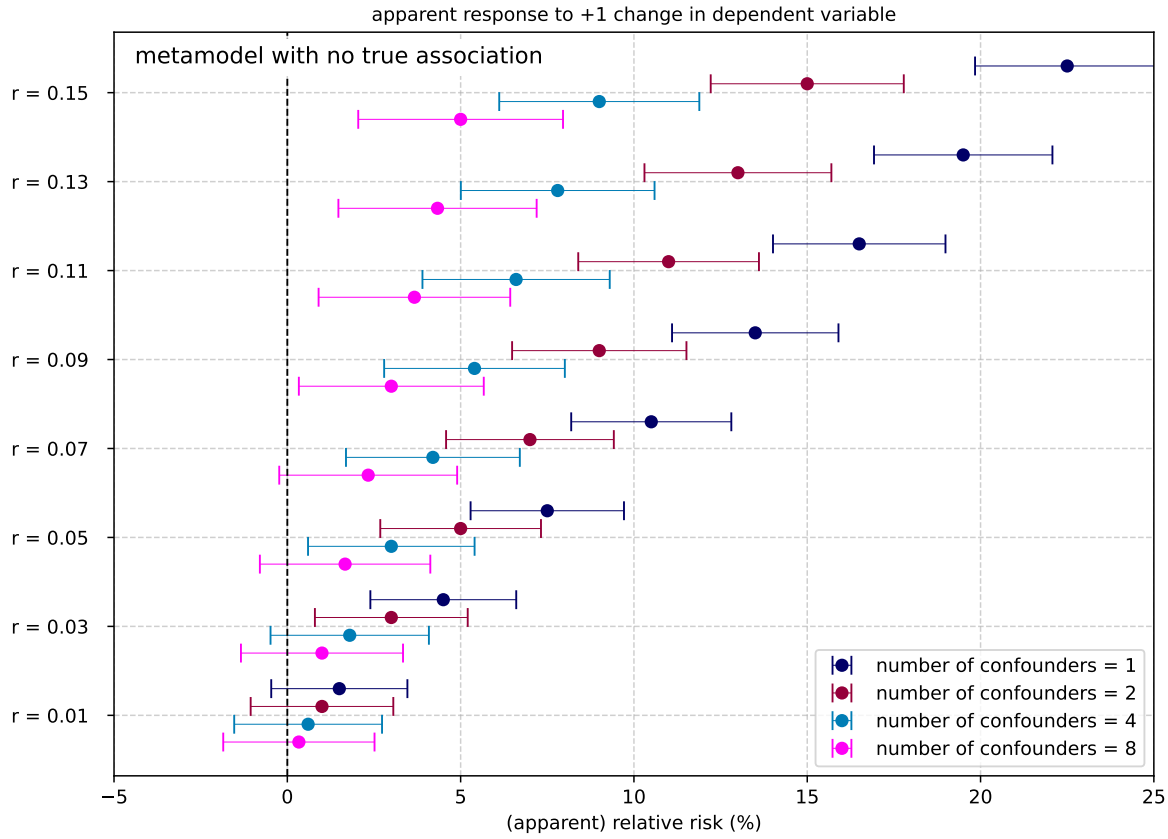

Figure 3: Results from metamodel described in the text. Every member of a large fictitious population of size N has multiple binary features, none causally related, but all mutually correlated with constant pairwise correlation coefficient r. An investigator takes one feature as the dependent variable and seeks to measure the direct influence of another feature on it, not knowing that the true answer is zero. Recognizing correlated variables as confounders, the investigator removes $n = 1, 2, 4,$ or 8 of them by regression. The figure shows the measured direct association (plotted as relative risk as in Figure 2) for various values of r and n. Error bars are 95% confidence intervals for $N \approx 50,000$, as in Figure 2, but would scale as $N^{-1/2}$.

3 Discussion

Figures 2 and 3 are not directly comparable line-by-line: Where the metamodel assumes a single causal latent variable yielding a single value of mutual correlation among many observables, the variables in actual survey data are related by a complicated causal web of many observable and latent variables. Still, the lines in the metamodel may correspond to magnitudes of confounding in some rough average sense.

Figures 2 and 3 are both cases where the measured direct associations are spurious—by construction in the metamodel and with near certainty for the inhalant hypothesis. What should we expect to see in the case of a true direct association? We can extend the metamodel to give some indication of this. Instead of drawing values for the dependent

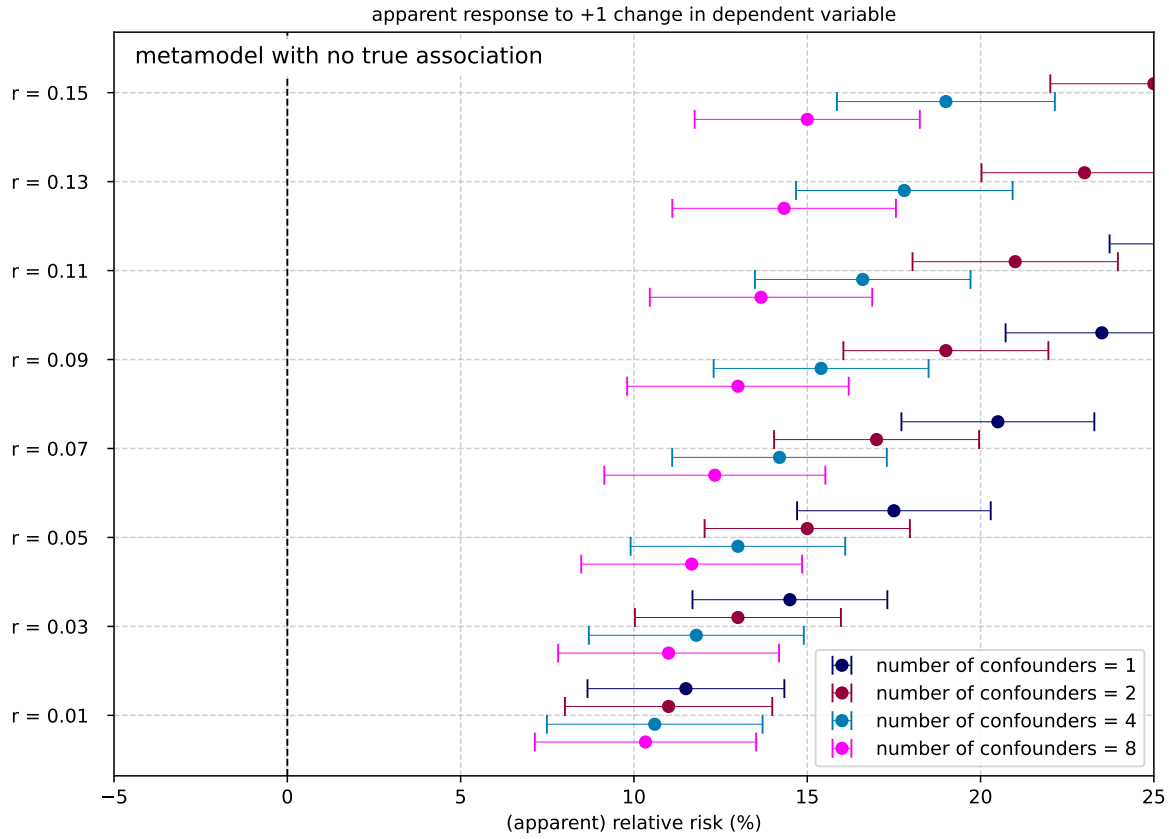

Figure 4: Like Figure 3, except that a value 0.10 is added to the logit probability in drawing values of the dependent variable.

variables R_{i0} with correlated probabilities p_i, we can draw them with probabilities

$$p_i' = \text{logit}^{-1}[\text{logit}(p_i) + \beta' R_{i1}] \tag{7}$$

which makes the dependent variable (index 0) causally related to the independent variable (index 1) by a logit increment β'. Figure 4 shows the results for $\beta' = 0.10$. As expected, Figure 4 and Figure 3 differ in first order by an increase of 10% (0.10) in all the relative risks, with also some changes in the size of error bars. Worth noting is that confounding variables with larger mutual correlations (top line in Figure 4) significantly bias the observed relative risk to larger than its true (i.e., causal) value.

While the coefficients β_i in equations (3) or 4) enter the models in linear combination, there is no reason to think that nonlinear regressions (e.g., [23]) would produce qualitatively different findings. Any type of regression will allocate some weight to both the desired independent variable and any included confounding variables that have a statistical association (linear or not) with the dependent variable. And, as in the cases considered above, the independent variable may be just another confounder whose association is not causal.

For actual studies such as [16] and [19], one could contemplate constructing a series of models that add confounding variables sequentially, one by one or in groups. It is not

clear, however, that much would be learned. Convergence to a putative true association could be real, or it could simply indicate that an important latent confounding variable is not present in the survey data. Here, we purposely constructed clean examples of pure confounding.

The general conclusion that can be drawn, however unsurprising, is that the quantitative measurement of small associations in the presence of a complex web of confounding associations is a fraught exercise, to be approached with great caution.

4 Methods and Materials

4.1 Data Source and Curation

The NSDUH 2023 Public Use File (delimited) [20] and PUF Codebook [24] were downloaded from `https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/datafiles`.

Seventy-nine columns were identified as potentially interesting (see SI for list). For each column, responses that were not natively ordinal (i.e, logically ordered) were mapped to the most logical or most neutral ordinal value. For example, 99 for No Response might, for most questions, be mapped to 0, for "none" or "no". Categorical variables (e.g., race) were mapped to consecutive integers from 0 and tagged for one-hot encoding. See SI for the complete list of mappings.

4.2 Model Fitting

A single column, one of those listed in Table 1, was chosen to be the dependent variable. The independent variables, all treated equivalently, were ALCYRTOT plus, for the four cases shown in Figure 2, the confounding variable columns listed in Table 2.

The logistic regression model best fitting the data was in all cases found by the standard package "statsmodels" [25], coded as,

```
import statsmodels.api as sm
model = sm.Logit(arr[:,0], arr[:,1:k+1])
result = model.fit(disp=0)
coef = result.params
std_err = result.bse
```

That the returned coefficients are themselves relative risks for a dependent variable upon changing a single independent variable by one unit can be seen by

$$\begin{aligned} \text{relative risk} \equiv \frac{p'}{p} - 1 &= \frac{\text{logit}^{-1}[\text{logit}(p) + \beta_1]}{\text{logit}^{-1}[\text{logit}(p)]} - 1 \\ &= \frac{\exp(\beta_1)}{1 + [\exp(\beta_1) - 1]\, p} - 1 \approx \beta_1(1-p) \approx \beta_1 \ , \end{aligned} \tag{8}$$

where the approximate equalities are for $\beta_1 \ll 1$ and $p \ll 1$.

Case	Confounding Variables Included
A	IRSEX, ANYHLTI2, GOVTPROG, IREDUHIGHST2, IRFAMIN3, BMI2, HEALTH2, AGE3, COUTYP4, IRHHSIZ2 (demographic and health variables)
B	all A, plus MJYRTOT, MJEVER (marijuana-use variables)
C	all A and B, plus LSD, PCP, PEYOTE, ECSTMOLLY, KETMINESK, DMTAMTFXY, SALVIADIV (hallucinogen-use variables)
D	all A, B, and C, plus CIGEVER, CIG30USE (cigarette-use variables)

Table 2: Confounding variables included in the analyses A, B, C, and D shown in Figure 2. (For variable-name decoding, see NSDUH Codebook at [24].)

4.3 Empirical Fitting Formulas for the Metamodel

Five hundred synthetic data sets of size $N = 10000$ were generated according to equation (1) and then fit by the code above (same code as used for the real data). The resulting logit coefficients were averaged over equivalent columns and all data sets and plotted as functions of p and k. (See SI for these plots.) The simple fitting formulas, equation (6), were devised by trial and error. SI Figures SI.2 and SI.2 show the simulation and fitting formula results.

Acknowledgments

I am grateful to David Donoho for a close reading of this paper and for suggesting multiple references to the literature. Donoho also outlined a more rigorous, purely analytical approach to several issues treated above only by simulation, an elegant approach that would be his work, not mine, hence not in scope for this paper.

References

[1] Wikipedia contributors (2025) Confounding (*Wikipedia, The Free Encyclopedia*). Available at: `https://en.wikipedia.org/wiki/Confounding`,.

[2] Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11):1359–1366.

[3] Yu X, et al. (2024) Misstatements, misperceptions, and mistakes in controlling for covariates in observational research. *eLife* 13:e82268.

[4] Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8):904–909.

[5] van Zwieten A, Dai J, Blyth FM, Wong G, Khalatbari-Soltani S (2024) Overadjustment bias in systematic reviews and meta-analyses of socio-economic inequalities in health: A meta-research scoping review. *International Journal of Epidemiology* 53(1).

[6] Wang Y, Blei DM (2018) The blessings of multiple causes. *arXiv*.

[7] Grimmer J, Knox D, Stewart BM (2020) Naïve regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *Journal of Machine Learning Research*.

[8] Pearl J (2013) Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* 1(1):155–170.

[9] Etminan M, Brophy JM, Collins G, Nazemipour M, Mansournia MA (2021) To adjust or not to adjust: The role of different covariates in cardiovascular observational studies. *American Heart Journal* 237:62–67.

[10] Office of the U.S. Surgeon General (2025) Alcohol consumption and cancer risk: A surgeon general's advisory, (U.S. Department of Health and Human Services), Technical report. Available at: `https://www.hhs.gov/sites/default/files/oash-alcohol-cancer-risk.pdf`.

[11] Rabin RC (2025 January 3) Surgeon general calls for cancer warnings on alcohol. *The New York Times*. Available at: `https://www.nytimes.com/2025/01/03/health/alcohol-surgeon-general-warning.html`.

[12] Rumney E (2025 January 3) US surgeon general urges cancer warnings for alcoholic drinks. *The Washington Post*. Available at: `https://www.reuters.com/world/us/us-surgeon-general-urges-cancer-warnings-alcoholic-drinks-2025-01-03`.

[13] Reddy S (2025 February 24) Few are aware alcohol raises the risk of certain cancers. *The Wall Street Journal*. available at `https://www.wsj.com/health/wellness/your-happy-hour-habits-could-raise-your-cancer-risk-acc93381`.

[14] PBS NewsHour (2025 January 3) U.S. surgeon general explains why he's calling for cancer warnings on alcohol. Available at: `https://www.pbs.org/newshour/show/u-s-surgeon-general-explains-why-hes-calling-for-cancer-warnings-on-alcohol`.

[15] National Academies of Sciences, Engineering, and Medicine (2024) *Review of Evidence on Alcohol and Health*. (The National Academies Press, Washington, DC). Available at: `https://nap.nationalacademies.org/catalog/28582/review-of-evidence-on-alcohol-and-health`.

[16] Sarich P, et al. (2021) Alcohol consumption, drinking patterns and cancer incidence in an australian cohort of 226,162 participants aged 45 years and over. *British Journal of Cancer* 124(2):513–523. Available at: https://doi.org/10.1038/s41416-020-01101-2.

[17] Klein JP, Moeschberger ML (2003) *Survival Analysis: Techniques for Censored and Truncated Data*, Statistics for Biology and Health. (Springer), 2nd edition.

[18] Wikipedia contributors (2025) Bradford Hill criteria (*Wikipedia, The Free Encyclopedia*). Available at: https://en.wikipedia.org/wiki/Bradford_Hill_criteria.

[19] Sellem L, et al. (2024) Food additive emulsifiers and cancer risk: Results from the French prospective NutriNet-Santé cohort. *PLOS Medicine* 21(2):e1004338.

[20] Substance Abuse and Mental Health Services Administration (SAMHSA) (2023) National Survey on Drug Use and Health (NSDUH). Available at: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health.

[21] Wikipedia contributors (2025) Logistic regression (*Wikipedia, The Free Encyclopedia*). Available at: https://en.wikipedia.org/wiki/Logistic_regression.

[22] Hosmer DW, Lemeshow S, Sturdivant RX (2013) *Applied Logistic Regression*. (Wiley, Hoboken, NJ), 3rd edition.

[23] Seber GAF, Wild CJ (1989) *Nonlinear Regression*, Wiley Series in Probability and Statistics. (Wiley), 1st edition.

[24] Substance Abuse and Mental Health Services Administration (SAMHSA) (2023) NSDUH 2023 Codebook. Available at: https://www.samhsa.gov/data/system/files/media-puf-file/NSDUH-2023-DS0001-info-codebook_v1.pdf.

[25] Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with Python in *9th Python in Science Conference*. https://www.statsmodels.org.

Supporting Information

Data Source and Curation

The columns selected as potentially interesting and their response mappings (see main text) are given in the following table. As an example, the entry "ALCEVER ORD 2:0, 85-97:0" means that ALCEVER, with tabulated responses 1 (yes), 2 (no), 85 (bad data), 94 (don't know), and 97 (refused) are mapped by us to the ordinal values 0 (no

or other) and 1 (yes). For variable-name decoding, see NSDUH Codebook at the main text's ref. [24].

Column	Type	Mappings
IRSEX	CAT	1:0, 2:1
IREDUHIGHST2	ORD	
EDUHIGHCAT	ORD	5:0
IRFAMIN3	ORD	
ANYHLTI2	ORD	2:0, 94-98:0, 1:1
IRMARIT	CAT	99:5
CIGEVER	ORD	2:0, 1:1
CIG30USE	ORD	91-98:0
CIG30AV	ORD	91-98:0
BMI2	ORD	
COUTYP4	ORD	3:0, 2:1, 3:2
AGE3	ORD	
NEWRACE2	CAT	3-4:3, 5:4, 6:5, 7:6
ALCEVER	ORD	2:0, 85-97:0
ALCTRY	ORD	985-998:80
ALCYRTOT	ORD	985-998:0
ALCDAYS	ORD	85-98:0
ALCUS30D	ORD	975:4, 985-998:0
ALCBNG30D	ORD	80-98:0
IRALCFY	ORD	991-993:0
HEALTH	ORD	94-97:3
HEALTH2	ORD	
IRWRKSTAT18	CAT	99:0
IRHHSIZ2	ORD	
IRKI17_2	ORD	
GOVTPROG	ORD	2:0
NICVAPEVER	ORD	2:0, 94-97:0
NICVAPAGE	ORD	985-998:80
NICVAP30N	ORD	91-98:0
ILLYR	ORD	
SNYATTAK	ORD	85-99:1
SNRLGSVC	ORD	85-99:0
SNRLGIMP	ORD	1-2:0, 85-99:1, 3-4:2
IRDSTNRV30	ORD	99:5
IRDSTHOP30	ORD	99:5
IRIMPSOC	ORD	99:0
HTINCHE2	ORD	985-998:66
WTPOUND2	ORD	9985-9998:150
MJEVER	ORD	2:0, 94-97:0

Continued on next page

Column	Type	Mappings
MJDAY30A	ORD	91-98:0
MRJYR	ORD	
IRMJFY	ORD	991-993:0
BLNTEVER	ORD	2:0, 4:0, 11:1, 85-98:0
COCEVER	ORD	2:0, 94-97:0
COCYR	ORD	
CRKEVER	ORD	2-98:0
CRKYR	ORD	
HEREVER	ORD	2-97:0
HERYR	ORD	
LSD	ORD	2-97:0
LSDYR	ORD	
PCP	ORD	2-97:0
PEYOTE	ORD	2-97:0
MESC	ORD	2-97:0
PSILCY	ORD	2-97:0
ECSTMOLLY	ORD	2-97:0
KETMINESK	ORD	2-97:0
DMTAMTFXY	ORD	2-97:0
SALVIADIV	ORD	2-97:0
AMYLNIT	ORD	2-97:0
CLEFLU	ORD	2-97:0
GAS	ORD	2-97:0
GLUE	ORD	2-97:0
ETHER	ORD	2-97:0
SOLVENT	ORD	2-97:0
LGAS	ORD	2-97:0
NITOXID	ORD	2-97:0
FELTMARKR	ORD	2-97:0
SPPAINT	ORD	2-97:0
AIRDUSTER	ORD	2-97:0
METHAMEVR	ORD	2-97:0
METHAMYR	ORD	
FENTANYYR	ORD	
OXCNNMYR	ORD	2-98:0
TRQNMYR	ORD	
STMNMYR	ORD	
SEDNMYR	ORD	
OPINMYR	ORD	
MJYRTOT	ORD	985-998:0

Metamodel Simulation Results

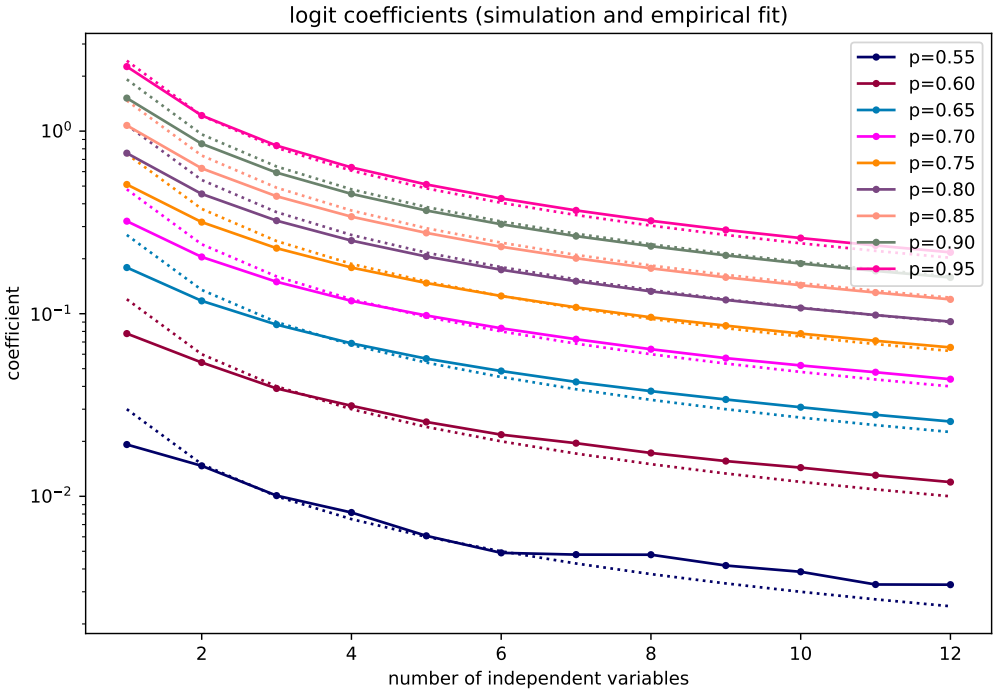

Figure SI.1: Logit coefficient of apparent relative risks in the metamodel and (dotted) empirical fitting formula, main text equation (6).

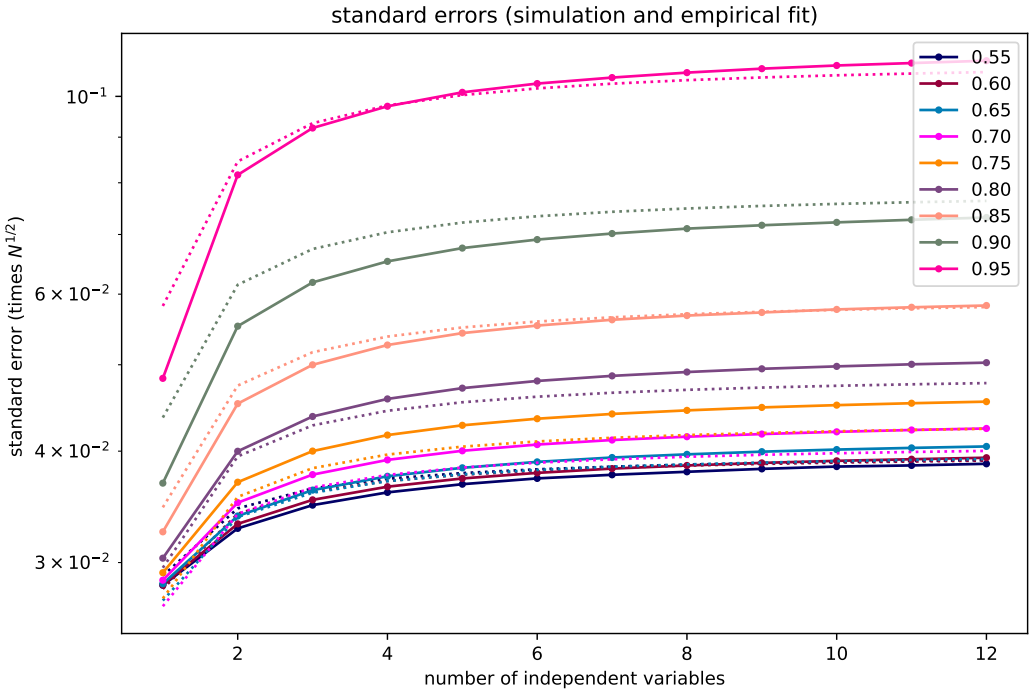

Figure SI.2: Standard error of logit coefficient of apparent relative risks in the metamodel and (dotted) empirical fitting formula, main text equation (6).