

**Title: ‘Shaking the Ladder’ reveals how analytic choices can influence associations in nutrition epidemiology: beef intake and coronary heart disease as a case study**

**Author Names:** Colby J. Vorland, Lauren E. O’Connor, Beate Henschel, Cuiqiong Huo, James M. Shikany, Carlos A. Serrano, Robert Henschel, Stephanie L. Dickinson, Keisuke Ejima, Aurelian Bidulescu, David B. Allison, Andrew W. Brown

**Author Affiliations:**

Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA (CJV, BH, CH, CAS, SLD, AB, DBA)

Beltsville Human Nutrition Research Center, Agricultural Research Service, United States Department of Agriculture (LEO)

Division of Preventive Medicine, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA (JS)

UITs Research Technologies, Indiana University, Bloomington, IN, USA (RH)

Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (KE)

Arkansas Children’s Research Institute, Little Rock, AR, USA (AWB)

Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, AR, USA (AWB)

**Corresponding Author:** Andrew W. Brown; 13 Children’s Way, Slot 842, Little Rock, AR 72202; [AWBrown@uams.edu](mailto:AWBrown@uams.edu); (501) 364-2730

**Abbreviations:** Coronary Heart Disease (CHD); REasons for Geographic and Racial Differences in Stroke (REGARDS); hazard ratio (HR)

1 **Abstract**

2 *Background*

3 Many analytic decisions are made when analyzing an observational dataset, such as  
4 how to define an exposure or which covariates to include and how to configure them.  
5 Modelling the distribution of results for many analytic decisions may illuminate how  
6 instrumental decisions are on conclusions in nutrition epidemiology.

7 *Objective*

8 We explored how associations between self-reported dietary intake and a health  
9 outcome depend on different analytical decisions, using self-reported beef intake from a  
10 food frequency questionnaire and incident coronary heart disease as a case study.

11 *Design*

12 We used REasons for Geographic and Racial Differences in Stroke (REGARDS) data,  
13 and various selected covariates and their configurations from published literature to  
14 recapitulate common models used to assess associations between meat intake and  
15 health outcomes. We designed three model sets: in the first and second sets (self-  
16 reported beef intake modeled as continuous and quintile-defined, respectively), we  
17 randomly sampled 1,000,000 model specifications informed by choices used in the  
18 published literature, all sharing a consistent covariate base set. The third model set  
19 directly emulated existing covariate combinations.

20 *Results*

21 Few models (<1%) were statistically significant at  $p < 0.05$ . More hazard ratio (HR) point  
22 estimates were  $> 1$  when beef was polychotomized via quintiles (95% of models) vs.  
23 continuous intake (79% of models). When covariates related to race or multivitamin use  
24 were included in models, HRs tended to be shifted towards the null with similar  
25 confidence interval widths compared to when they were not included. Models emulating  
26 existing published associations were all above HR of 1.

### 27 *Conclusions*

28 We quantitatively illustrated the impact that analytical decisions can have on HR  
29 distribution of nutrition-related exposure/outcome associations. For our case study,  
30 exposure configuration resulted in substantially different HR distributions, with inclusion  
31 or exclusion of some covariates being associated with higher or lower HRs.

32

33 This project was registered at OSF: <https://doi.org/10.17605/OSF.IO/UE457>

34 **Keywords:** analytic flexibility; multiverse; epidemiology; beef; coronary heart disease

35

## 36 Introduction

37 ‘Shaking the Ladder’ is a phrase borrowed  
38 from Sam Savage who wrote: “The last thing  
39 you do before climbing on a ladder to paint  
40 the side of your house is to give it a good  
41 shake. By bombarding it with random  
42 physical forces, you simulate how stable the  
43 ladder will be when you climb on it. You can  
44 then adjust it accordingly so as to minimize  
45 the risk that it falls down with you on it.” (2)  
46 Following Savage’s analogy, just as we  
47 would shake a ladder to test its stability  
48 before trusting it, we must rigorously evaluate  
49 how our analytical choices influence our  
50 conclusions in nutritional epidemiology.  
51 Investigating the associations of foods and  
52 nutrients with chronic disease endpoints is a  
53 challenging line of scientific inquiry. One of these challenges is the many reasonable  
54 decisions that investigators face when defining their exposure and outcome, and  
55 numerous analytical decisions such as how to configure the exposure, covariates, and  
56 model selections. For instance, covariates could be included or excluded, or defined as  
57 continuous, categorical, ordinal, or other ways (what we will refer to as covariate  
58 configuration). With each decision point, the combinations of defensible analytical

### **Glossary of terms**

**Covariate inclusion:** Whether a covariate is included in a model.

**Covariate configuration:** How covariates are defined in a model; for example, as continuous, categorical, ordinal, etc.

**Model specification:** All choices that go into a model, such as covariate inclusion, covariate configuration, and model type.

**Quantile:** Cut points that divide a distribution into intervals with equal likelihood.

**Quintile:** A quantile that divides a distribution into five intervals (i.e., there are four quintiles that make five intervals).

**Effect size:** “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest.” (1)

59 decisions increase exponentially. Several studies demonstrated that if different sets of  
60 investigators were asked to analyze the same data set, their analysis approaches can  
61 vary substantially, sometimes resulting in vastly different conclusions (3-8).

62 Flexibility in analytical choices has been described as a garden of forking paths (9), or  
63 investigator degrees of freedom (10), among other names. Different analysis  
64 approaches may be responsible for some inconsistency in results in nutritional  
65 epidemiology research, although this has not been explored as extensively as other  
66 fields. Given that a set of decisions represents one of many reasonable potential  
67 approaches to analyzing the data, one analysis may lead to a conclusion that is  
68 represented by a minority of those approaches. Many foods and nutrients have both  
69 positive and negative associations with disease outcomes in the published literature  
70 (11), thus it is paramount to explore the degree to which this may be explained by  
71 analysis strategies.

72 One option to explore this phenomenon is to run many models with defensible analytic  
73 choices and report the distribution of results. This ‘multiverse’-style (12, 13) approach  
74 (similar to “specification curve analysis” (14), or “vibration of effects” (15)) can be used  
75 to explore the distribution of association estimates between an exposure and an  
76 outcome for many analytical paths, in turn allowing us to assess what influence the  
77 analytical decisions have on estimating the associations. The concept has been applied  
78 to several nutritional questions (15, 16) that focus on covariate inclusion and exclusion;  
79 however, additional choices such as the configuration of the nutritional exposure and  
80 covariates add additional flexibility. There is poor reporting in nutritional epidemiology  
81 for how covariate selection and configuration are decided (17), which raises questions

82 about whether these methods are being systematically employed or if the selection and  
83 configuration processes are somewhat arbitrary and other choices also defensible.

84 Our objective was to evaluate to what degree associations between self-reported  
85 nutritional intake and health outcomes depend on different analytical decisions (e.g.,  
86 exposure configuration, covariate inclusion and configuration, subject inclusion and  
87 exclusion criteria). Because covariate inclusion and configuration are not well reported  
88 in nutrition epidemiology, we aimed to evaluate the consequences of not carefully  
89 considering these. In contrast to previous approaches (e.g., specification curve,  
90 multiverse analysis), in which models are selected based on theory to explore the  
91 robustness of results for a particular research question, we randomly selected models  
92 based on existing published variable choices, and therefore the research questions  
93 represented by each model may change in subtle ways. We specifically use the case  
94 study of beef consumption and incident coronary heart disease (CHD). The beef-CHD  
95 relation is particularly appropriate for this approach because there is significant  
96 disagreement in the literature on the relationship between red meat intake and CHD  
97 (18-27); thus, analytical flexibility may be one explanation for this disagreement. Our  
98 analysis serves as a case study for how this approach can test the influence of  
99 analytical decisions on diet-outcome associations in nutrition specifically, and in  
100 observational association studies generally.

101

## 102 [Subjects and Methods](#)

### 103 *Study Sample*

104 We used data from the REasons for Geographic and Racial Differences in Stroke  
105 (REGARDS) prospective cohort (28). REGARDS is a national, longitudinal cohort of  
106 30,239 Black and White women and men ages 45 and older, who were recruited from  
107 2003-2007. After excluding 56 participants with data anomalies, we utilized data from  
108 30,183 participants. Participants' CHD status was last updated in 2018. Participants  
109 with a history of CHD or cancer at baseline were excluded from our analyses. **Figure 1**  
110 and **Supplemental Table 2** describe additional participant exclusions, such as those  
111 based on self-reported energy intake cutoffs (varying methods to exclude  
112 extreme/implausible data).

#### 113 *Exposure, Outcome, and Covariate Selection Process*

114 Self-reported beef consumption was originally estimated via the Block 98 food  
115 frequency questionnaire (FFQ). We defined beef intake using gram weight estimates  
116 (using the variables 'hamburger', 'beefroast' 'beeffattrimmed', 'beeffatnotrim' based on  
117 the FFQ items "hamburgers, cheeseburgers, meat loaf, at home or in a restaurant" and  
118 "beef steaks"). Values in the hamburger variable were multiplied by a proportion of 0.59  
119 to refine the estimation of beef content. This proportion was determined from the Food  
120 and Nutrient Database for Dietary Studies 2017-2018 data (29). The outcome of  
121 incident CHD was defined as myocardial infarction event or acute CHD death.

122 Our inclusion of covariates and their configurations was informed by prior  
123 literature, allowing us to indirectly crowdsource expert choices in covariate inclusion and  
124 configuration that had also passed peer review. The prior analyses were identified from  
125 1) a previous systematic review of prospective cohort studies of red meat and CVD  
126 outcomes (30), 2) a selection of observational studies assessing red meat or beef and

127 CVD outcomes known to coauthors or identified through literature searching, and 3)  
128 previous analyses using the REGARDS dataset by a coauthor (JS). All references are  
129 listed in documents attached to our preregistration:  
130 <https://doi.org/10.17605/OSF.IO/UE457>. Covariates and their configurations identified  
131 from studies not using the REGARDS data were matched as closely as possible to  
132 REGARDS variables. History of chronic obstructive pulmonary disease and sleep  
133 outcomes in the literature sampling did not have a close match within the REGARDS  
134 dataset and were not included in models. Configurations included categorical,  
135 continuous, or ordinal via quintiles or sex-specific median. A complete list of included  
136 variables, their configurations, as well as their corresponding REGARDS variable  
137 names, and variables unable to be matched to REGARDS variables, is available in the  
138 following repository: <https://doi.org/10.17605/OSF.IO/SY96K>.

139 Three sets of models were developed. First, we created a random sample of  
140 1,000,000 model combinations, based on variables that appear in previous literature,  
141 where self-reported beef intake was defined as either continuous (model set 1) or in  
142 quintile defined categories (model set 2). Then, we emulated prior literature to try to  
143 reproduce existing variable choice combinations exactly as they have been previously  
144 published as specific, expert, pre-specified analytical examples in the REGARDS  
145 dataset (model set 3); these models were reproduced in the REGARDS set (i.e., they  
146 were not randomly sampled). In the first and second model set, age, sex, energy intake,  
147 size of census tract, and REGARDS region were included in all models, consistent with  
148 the prior literature; thus, we decided it would be unreasonable that expert analysts  
149 would define a model without them. For food- and nutrient-related variables, their



150 inclusion was randomly varied in models, but when included we used the same  
151 configuration for all such included variables because we did not believe analysts would  
152 consider models reasonable if configurations differed among these variables. For other  
153 variables, we randomly varied their inclusion and configuration. Direct comparisons  
154 between hazard ratios (HRs) derived from continuous versus quantile exposure  
155 definitions can be difficult; therefore, we decided to express continuous beef intake per  
156 50g unit increase, which was comparable to the difference in mean reported intakes in  
157 the highest 20% versus lowest 20% of participants (50.01g). Covariate inclusion and  
158 configurations for these models are described in the 'statistical analysis and  
159 visualization' section. **Table 1** summarizes which variables were kept constant in all  
160 models, and which were varied.

### 161 *Statistical Analysis and Visualization*

162 Cox proportional hazards regression models were used, with time from enrollment as  
163 the underlying time metric within each of the analyses, censoring date of CHD  
164 diagnosis, date of death, date of withdrawal, or date of last follow-up. Sample size was  
165 allowed to vary on a complete-case basis depending on which covariates were included  
166 in each model. Missing data were not imputed. This approach was consistent with the  
167 sampled prior literature. The proportion of missingness for any given covariate is shown  
168 in **Supplemental Table 3**. The total number of possible combinations of covariates and  
169 configurations was far beyond computational capabilities (see results); thus, model sets  
170 1 and 2 randomly sampled 1,000,000 variable combinations total (500,436 for beef as  
171 continuous and 499,564 for beef as quintiles). Covariates were first sampled for  
172 inclusion or exclusion; if the covariate was included, the configurations were equally

173 sampled. For example, in the case of a food variable such as dairy intake, it had a 50%  
174 chance of being included; if it was included, then each of the three configurations  
175 (continuous, sex-specific median or quintile) were sampled with equal probability (1 out  
176 of 3 conditional on being included). For model set 3, all models were computed as  
177 closely as possible to emulate the prior literature (see <https://osf.io/sy96k> for the  
178 models).

179 Code to run the analysis was developed and tested on a small scale and later  
180 parallelized for the full 1,000,000 model runs. Briefly, code consisted of a ‘for loop’  
181 iterating through model runs and saving the model output. After a loop dependency  
182 analysis, we found that the loop did not depend on any other “outside” data including  
183 dependencies between models, so we parallelized by modifying the code to run small  
184 subsets of the total models (500 models) and run this code multiple times so that these  
185 subsets were run in parallel. Lastly, all subsets of results were combined using a Linux  
186 shell script. Parallel code was run on Carbonate, which is Indiana University's large-  
187 memory computer cluster, designed to support data-intensive computing (31).

188 Model results were visualized and further analyzed in different ways. Distribution of  
189 HRs, z-scores and p-values were plotted in histograms. To visualize the impact of  
190 variable configuration on beef hazard ratios we created specification curve plots. In the  
191 first step, we did this for the beef variable itself, but then also for all other covariates  
192 included in the model. Specification curve plots show the distribution of the estimates of  
193 the association of interest and the impact of analytic decisions on those estimates by  
194 showing the distribution of estimates for each analytic decision. Bivariate scatterplots of  
195 the beef HRs and its 95% confidence interval (CI) widths were created to show possible

196 relationships of variable inclusion/exclusion on estimates and their precision.  
197 Additionally, to show data density, we plotted bivariate KDE curves. Model meta-  
198 information such as sample size, number of covariates, and number of CHD events  
199 were plotted showing their density by significance of the beef HR. Descriptive statistics  
200 of the beef HR, its z-scores, its 95% CI, and its p-values were calculated overall and by  
201 beef configuration for model set 1 and 2 and the models from the literature. Impact of  
202 covariate configuration (including exclusion) on the beef HR was assessed in  
203 multivariate logistic regression models that adjusted for all included covariates and their  
204 configurations at the same time. Odds ratios and 95% CI are presented. Lastly, a series  
205 of two-sample Kolmogorov-Smirnov statistics (e.g., D statistics) were calculated to  
206 quantify the distance between the cumulative distributions of HR by inclusion/exclusion  
207 of covariates. Significance tests with p-values were not used for the K-S test because  
208 the samples are not independent and identically distributed. Cross-correlation  
209 coefficient and likeness measures between KDEs were calculated as defined in (32).  
210 Higher values for both indicate higher overlap between both KDEs. For all covariate  
211 specific plots, we show the results for four selected covariates in the main text; the  
212 remaining plots for all covariates can be found in supplemental materials.

213 SAS [version 9.4] was used to prepare the dataset for analysis, R [version 4.1.1] was  
214 used on a x86 64-bit Linux cluster for the models, and R [version 4.2.3] and RStudio  
215 [version 2023.03.0] were used for analyses and to produce visualizations.

216 *Ethics*

217 This study was approved by the Indiana University Institutional Review Board (#11227).

218 The REGARDS study was previously approved by all associated institutional review  
219 boards (33).

#### 220 *Power calculation*

221 We estimated that we would have sufficient power to detect most associations with HR  
222 > 1.1 with a sample size of 20,000 or lower (**Supplemental Figure 1**) using incident  
223 CVD from the lowest quartile of estimated red meat consumption from Zhong et al. (24)  
224 as the reference hazard. Thus, REGARDS provided a sufficiently large sample for small  
225 HRs.

#### 226 *Inference Criteria*

227 We used  $p < 0.05$  as a threshold of statistical significance within any given analysis,  
228 consistent with standard practice in the prior nutritional epidemiology literature. We did  
229 not correct for multiple comparisons, because each analysis represents one theoretical  
230 independent choice of many that an analyst could make.

#### 231 *Changes after Preregistration*

232 Our project was preregistered at OSF: <https://doi.org/10.17605/OSF.IO/UE457>. We  
233 describe changes post-registration and our reasoning in **Supplemental Table 1**.

## 234 **Results**

#### 235 *Calculation of Model Possibilities*

236 For model sets 1 and 2, with 2 beef configurations, 7 exclusion criteria configurations  
237 based on self-reported energy intake cutoffs (**Supplemental Table 2**), and 34

238 covariates with 117 configurations, we calculated over 4.16 quadrillion total possible  
239 model combinations. Running all possible combinations would not have been feasible  
240 even with the use of parallel high-performance computing resources available to the  
241 study team. Additionally, storing, analyzing, and presenting model results would have  
242 been challenging if not impossible.

243 Comparing results (e.g., distribution of HRs and covariate associations) from an initial  
244 test run using 10,000 models and the results shown herein of 1,000,000 models, we are  
245 not convinced that more insights will be gleaned from an even greater number of  
246 samples.

#### 247 *Model Summaries*

248 **Supplemental Table 3** shows covariates and their configurations as defined using  
249 REGARDS data, along with the number of missing values for each. **Table 2** shows the  
250 mean, median, 5<sup>th</sup> to 95<sup>th</sup> percentile range, and min and max values for HRs, p-values  
251 and z-scores in model sets 1 and 2. **Figure 2** shows distributions of HRs, significant  
252 HRs (with  $p < 0.05$ ), z-scores, and p-values by beef configuration.

253 As shown in **Figure 2** (top panel) and **Table 2**, the proportion of models with HR greater  
254 than 1.0 was much higher when beef intake was expressed in quintile defined  
255 categories (right, 95.2%) compared to expressed as a continuous variable (per 50g  
256 intake; left, 78.6%). Of the 9556 significant beef HRs, only 38.7% (3695) came from  
257 models using beef as a continuous variable while the other 61.3% (5861) came from the  
258 quintile models. This is further illustrated in the specification curve (**Figure 3**) in which  
259 HRs are ranked from lowest to highest; the vertical dashed line (in top plot) shows that

260 131,205 of 1,000,000 models were less than an HR of 1. The bottom plot of **Figure 3**  
261 shows 1) the distributions of the ranked HR with the associated exposure  
262 configurations: continuous or quintile beef intake, and 2) that the statistically significant  
263 ( $p < 0.05$ ) HRs appear lower along the ranking for continuous beef intake compared to  
264 beef intake in quintile defined categories. Despite these differences, overall, very few  
265 models produced statistically significant associations in either approach (9556/1000000  
266 models=0.96% of all models), and these significant associations were associated with  
267 higher HRs (all were above 1.0).

#### 268 *Influence of Covariates on HRs and Precision*

269 We used the same HRs that were plotted in Figures 2 and 3 to generate additional plots  
270 to highlight the influence of the covariate selection and configuration. **Figure 4** shows  
271 the ranked HRs for continuous beef (left) and beef in quintile defined categories (right)  
272 by four selected covariates (race, income, education, and multivitamin use). We  
273 selected these variables as examples to highlight because their inclusion/exclusion and  
274 configuration showed a range of strong to weak influences on HRs. HRs that came from  
275 model specifications that adjusted for race or years of multivitamin use tended towards  
276 smaller HRs while HRs from models not adjusting for race tended towards higher HRs.  
277 Note that these shifts reflect the impact on the beef coefficient with inclusion or  
278 configuration of the covariate, not the coefficient for the covariate itself (e.g., the  
279 influence of a specific race or multivitamin use on the HR). For income, we observed  
280 that either not adjusting for income or adjusting for income using four categories  
281 appears to have not much impact on the size of the HR but adjusting for income as a  
282 continuous measure tended towards higher HRs at the right tail. Lastly, the covariate

283 education is an example where the distributions of beef HRs do not appear to differ  
284 much when adjusting or not adjusting in the analysis.

285 We explored the results for these four covariates further in **Figure 5** by plotting HRs  
286 across the x-axis and the 95% confidence interval width of the HRs across the y-axis,  
287 the latter representing precision of the estimate, for the same set of covariates. When  
288 multivitamin use (bottom right) was excluded from the beef as continuous models, HRs  
289 and confidence interval width tended to be shifted above 1 compared to inclusion, but  
290 with similar precision, whereas inclusion (i.e., adjusting for years of multivitamin use)  
291 tended to be centered around 1. The trends were similar in the quintile approach,  
292 though inclusion or exclusion models remained with densities higher than an HR of 1  
293 and less precision (i.e., wider confidence intervals) as compared to the continuous beef  
294 models. For race (top left), the results are much more diffuse for inclusion and exclusion  
295 for both continuous and quintile beef models. For income (top right), the density curves  
296 for exclusion and categorical income adjustment (labeled 'Income\_4cat') are almost  
297 identical and on top of each other, while the curve for continuous income adjustment  
298 (labeled 'Income') is shifted away from the null with less precision. Lastly, for education,  
299 the density curves for inclusion vs exclusion do not show visual differences. The plots  
300 for all 34 covariates are shown in **Supplemental File 1**. In order to analytically assess  
301 differences in the distributions of HRs by inclusion/exclusion of covariates, we ran a  
302 series of Kolmogorov-Smirnov tests (**Table 4**). With the high number of different models  
303 we ran, we had large sample sizes, so these tests had high power to detect even minor  
304 differences. All beef hazard ratio distributions were significantly different when including  
305 a covariate compared to excluding a covariate. When using a Bonferroni-corrected p-

306 value of 0.0017 (0.05/29 tests), inclusion vs exclusion of two covariates (“grains” and  
307 “monofat”) was no longer significantly associated with the HR distributions in the  
308 continuous beef configuration, while all other tests remained significant. Visually  
309 inspecting the distributions of beef HRs by configurations of covariates, we observed  
310 the largest shifts for the following covariates in the beef as quintile models: race,  
311 income, multivitamin use, history of diabetes, history of stroke, physical activity, fiber,  
312 and fruit intake (**Supplemental File 3**).

313 Although overall only about 1% of the models resulted in statistically significant HRs, we  
314 tested the influence that inclusion and configuration of covariates had on the statistical  
315 significance of the HRs using separate multivariable logistic models for the two beef  
316 configurations. **Supplemental Table 4** displays the results showing the proportion of  
317 significant HRs for each configuration, an odds ratio (OR) for significant HRs using one  
318 of the configurations as the reference group (in most cases: covariate exclusion), and  
319 the p-value for the OR. We observed significant associations for all covariates except  
320 education (type 3 p-value = 0.061) and history of PAD (p = 0.129), indicating that all  
321 other covariates had some influence whether the association between self-reported  
322 beef and CHD was statistically significant. Notably high ORs were found for the  
323 continuous configuration of income (OR=85.42 (95% CI: 73.05, 99.88) for continuous  
324 beef, OR=28.91 (95% CI: 26.04, 32.09) for quintile beef intake), meaning that when  
325 continuous income was included in models, there were higher odds of a significant HR  
326 for the beef-CHD association. In contrast, including the years of multivitamin use in the  
327 model resulted in far fewer significant HRs for beef (OR<0.01 (95% CI: <0.01, <0.01) for



328 continuous beef, OR=0.02 (95% CI: 0.02, 0.03) for quintile beef) compared to excluding  
329 multivitamin use.

330 **Figure 6** shows pairwise density plots for the number of CHD events and sample size  
331 depending on the number of covariates in the model. Models with significant beef HR  
332 ( $p < 0.05$ ) show density curves in red and those not significant ( $p \geq 0.05$ ) in black. There  
333 was a tendency that as more covariates were included in the model, the sample size  
334 was smaller. Given that we let the sample size vary depending on the complete case of  
335 the model specification, this result is expected. With higher sample sizes, the number of  
336 CHD events tended to be higher (**Figure 6b**), and models with significant HRs appear to  
337 have a higher number of CHD events (**Figure 6b**). Finally, models with significant HRs  
338 tended to have a lower number of covariates (**Figure 6c**). Specification curves showing  
339 the distribution of HRs are shown for all 32 covariates in **Supplemental File 3** (beef as  
340 continuous) and **Supplemental File 4** (beef as quintiles).

#### 341 *Emulating existing literature*

342 To benchmark our agnostic, random sampling approach against expert-chosen models,  
343 we reproduced 20 models from the literature (see references in the preregistration).  
344 **Figure 7** shows that the frequency of HRs for these models were all greater than 1.0,  
345 and, from visual inspection, tended to have higher HRs and lower precision when beef  
346 was expressed as quintiles of intake. Two of 20 models were statistically significant;  
347 both of them when beef was expressed as quintiles of intake (**Table 3**). **Figure 8** shows  
348 the cumulative distributions of the HR from the 1,000,000 models from model sets 1 and  
349 2 combined, and the 20 models that we emulated as they appear in the existing  
350 literature. We see that the empirical cumulative distribution function (ECDF) for the

351 existing literature is below the one for model sets 1 and 2, which suggests that the  
352 results from the existing literature are shifted towards higher HRs overall. The strongest  
353 divergence between the two distributions ( $D=0.438$ ) can be observed for HRs below the  
354 median (ECDF=0.50), with a higher concentration of HR between 1.05 and 1.10 for the  
355 results from existing literature.

## 356 Discussion

357 Many analytic choices are needed when analyzing data from observational cohort  
358 studies. Historically, it was only feasible to analyze and report a handful of models,  
359 which represent only a small fraction of possible combinations. Indeed, by identifying  
360 covariates that have been used in the literature, we calculated over four quadrillion  
361 models that could be run to test the association between self-reported beef intake and  
362 incident CHD using the REGARDS dataset. Random sampling from these showed that  
363 HRs varied around the null of 1, and few models were statistically significant.

364 The results from our approach pose challenges to interpretation. Overall, the point  
365 estimates of the HRs were disproportionately above the null (87.9% overall; Table 2);  
366 however, a sizeable proportion of HRs remained less than the null (12.1%).

367 Furthermore, less than 1% of individual models reached classical statistical significance  
368 thresholds of  $p<0.05$ , which is less than expected if results were derived by random  
369 chance. Yet, those that did were all in the deleterious to health direction. Also, all of the  
370 models are dependent (that is, they are based on the same underlying data), and thus  
371 benchmarking the number of statistically significant findings against traditional type I  
372 error metrics may be inappropriate. Therefore, the approach overall leaves some

373 ambiguity regarding the association (let alone the causal relation) between beef and  
374 CHD from these data.

375 A qualitative inspection of our figures suggested that two variables had the greatest  
376 influence on results: years of multivitamin use and race. When each was excluded from  
377 models, HRs tended to be higher and model standard errors smaller. Multivitamin use is  
378 considered among health-related behaviors (34), and race is often considered with  
379 socioeconomic status (SES) (35). Not adjusting for these particular covariates, which  
380 indirectly capture concepts related to health consciousness and socioeconomic status,  
381 may produce more extreme results because of confounding. This raises the possibility  
382 that, even if one has an appropriate data generating process for selecting covariates to  
383 include based on a causal structure, covariate concepts may or may not be measured  
384 among different cohorts, or may be operationalized differently. For instance, SES is  
385 difficult to measure, so correlated indirect measures like race, income, and education  
386 may be used, but are still subject to unmeasured confounding. None of those measures  
387 alone fully capture SES, while adjusting for all of them results in multicollinearity; yet,  
388 choosing only one leads to measurement error and potentially high residual  
389 confounding. Thus, results may differ among models not because of any nefarious  
390 action by an epidemiologist, but because of which variables are available in a dataset  
391 and how they are operationalized or measured.

392 Given the inherent limitations of observational study designs (36, 37), the choice of  
393 covariates significantly influences the derived results. Notably, when we emulated  
394 models from published literature on meat-heart disease associations, all HRs were  
395 above 1, suggesting that those particular covariate choices tend to produce larger HRs

396 than if one takes our agnostic approach. This discordance between our approach and  
397 the replications of other investigators' models may or may not indicate the presence of  
398 publication or selection bias (38, 39) that drives the observation of published models  
399 exhibiting high effect sizes. Another possibility is that the effect size distribution from the  
400 replicated research better represents the underlying relation between beef and CHD.  
401 Without knowledge of the data-generating process, it is impossible to discern between  
402 these two scenarios or others.

403 Hypothesized data-generating processes (and thus any hypothesized causal structure)  
404 are rarely explicitly articulated in the choice of exposures, outcomes, and covariates in  
405 published literature. This leads to a crucial oversight in causal inference. The generation  
406 of a model should ideally be based on a robust mechanistic theory that justifies why a  
407 particular variable is a confounder, mediator, or collider (13). In this context, biases like  
408 collider bias, among others, are of significant concern, particularly when adjusting for  
409 measures such as energy intake (40). The practice of adhering to norms, such as  
410 including covariates for adjustment without a well-founded theoretical basis, might not  
411 be sufficient to account for these biases. We chose our approach because we generally  
412 do not observe that published articles on food- or nutrient-disease associations explicitly  
413 include a causal model with their analysis, and thus we wanted to evaluate the potential  
414 consequences of model selection in a way that emulates the current state of the  
415 literature. Indeed, a sample of 150 nutritional epidemiology studies found that 94% did  
416 not report *a priori* covariate selection, and only 20% reported the selection criteria for all  
417 covariates (17). Simulations have shown that flexibility in covariate selection can  
418 increase the chance of achieving statistical significance (10, 41-44). Together, the lack

419 of a theoretical framework for any of our varied models raises the question of where on  
420 the HR distribution a true causal association may reside.

421 Including more covariates tended to decrease HRs in our models; this is consistent with  
422 accounting for more confounding. Yet, adding more covariates risks misspecification  
423 that could potentially bias results toward the null; however, such misspecification could  
424 also induce spuriously inflated associations, and we intuitively (though without empirical  
425 claim) find it unlikely that additional covariates would systematically bias toward versus  
426 away from the null in our permuted models. Thus, accounting for more covariates  
427 seems to weaken the argument for a causal association between beef and CHD in  
428 these models. Regardless, our approach does not necessarily resolve unmeasured  
429 confounders that systematically bias associations in either direction. For example, in  
430 cohorts from the U.S., higher self-reported consumers of red meat are more likely to  
431 self-report being less physically active, smoking, drinking alcohol, having higher body  
432 weight, and poorer diet quality compared to those who self-report lower red meat  
433 consumption (45-47).

434 Other studies have observed substantial variability in conclusions when different  
435 analysis strategies are used, such as asking different research teams to analyze the  
436 same data set (3, 4, 7, 8). Other methods have approached analysis strategy variability  
437 more systematically to evaluate the robustness of statistical findings to changes in  
438 model specification (specification curve analysis (14), multiverse analysis (12, 13), or  
439 vibration of effects (15)). The latter concept has been applied to nutritional questions to  
440 explore how including and excluding covariates influence the association between

441 alpha-tocopherol and mortality, calcium and femur density, carrots and eyesight, and  
442 vitamin D level and COVID-19 (15, 16).

443 Our analysis is distinct from these approaches in that we varied both covariate inclusion  
444 and exclusion and covariate configuration, as well as exposure configuration.

445 Importantly, our data generating process to select covariates was done agnostically, at  
446 random, which is not the intention of multiverse-style approaches that should carefully  
447 consider the causal structure of the research question to examine the robustness of the  
448 question to analytic decisions (13). We adapted some visualization methods developed  
449 for specification curve and vibration of effects analyses. Because we allowed our  
450 sample size to vary among models (consistent with a common complete-case approach  
451 in nutritional epidemiology), and our research question was not strictly held constant by  
452 nature of allowing model choice to vary, we chose not to compute an average p-value of  
453 all models (12), nor use a bootstrap technique (14). Future work is needed to improve  
454 quantitative interpretations when exploring many models and tease out analytic  
455 decisions that have a higher relative influence on associations.

456 There are limitations to our work that may be resolved in future research using these  
457 methods. For one, not all variables that we identified or their configurations in the  
458 literature could be exactly matched to REGARDS variables. Additional publications  
459 identified using different search strategies may identify additional variables or  
460 configurations to include in the analyses. Likewise, not all our modeling choices can be  
461 translated to different datasets to look at the same question. Many variables, including  
462 beef as our exposure of interest, were self-reported, and it is not clear how accurately  
463 intake is captured (36). In addition, we could not identify sufficient existing literature on

464 beef *per se* and cardiovascular outcomes, so we used those on red meat more broadly,  
465 with the assumption that modeling choices would not differ. Further, two covariates  
466 used in previous literature, history of chronic obstructive pulmonary disease and sleep  
467 outcomes, did not closely match a variable in the REGARDS dataset, and therefore  
468 were missing in the distributions of results. Other modeling choices may be made by  
469 other analysts that may expand the decision tree even further and are not reflected in  
470 our analyses, such as excluding participants with a history of cancer at baseline (our  
471 rationale being that stronger associations may have been observed due to cardiotoxicity  
472 and cardiovascular deterioration in individuals with cancer); or using the 'energy  
473 adjustment' method (48) instead of including energy as a covariate. A particular  
474 challenge was to identify a dataset that permitted reasonable assessment of beef  
475 consumption specifically, rather than confounding the exposure of interest with other red  
476 or processed meats. Our estimation of self-reported beef from the FFQ used by  
477 REGARDS, using a proportion derived from 2017-2018 data from the Food and Nutrient  
478 Database for Dietary Studies was yet another point where various calculations may be  
479 considered reasonable and add additional model combinations, as well as the various  
480 ways to define beef such as total, unprocessed, processed, etc. (49). Although we used  
481 published literature to inform our covariable selection process, this does not necessarily  
482 mean that these covariates are those that all epidemiologists would deem as  
483 reasonable to include in models. In addition, some model combinations as randomly  
484 sampled may be less likely to be selected by epidemiologists than others, and thus our  
485 model distributions do not reflect models that would be weighted as more reasonable  
486 than others. Yet, because we included a subset of covariates in all of our models that

487 are commonly included in observational studies, we believe that our models are all  
488 within the possibility of what qualified analysts might use. Our permutation approach  
489 currently has limitations in how many models can be evaluated because of  
490 computational limitations. Indeed, we discovered that only a fraction of the total possible  
491 models (quadrillions) can be feasibly run with current resources. We therefore leave  
492 open future investigations to run more or targeted sampling to refine the distributions or  
493 further investigate features of the HR distribution space. Even then, we could have  
494 varied more choices in our model and increased the model space exponentially, such  
495 as using additional covariates based on different sets of literature, how beef is defined,  
496 whether certain variables should be recoded or not, whether each covariate's chance of  
497 being excluded is the same percentage as each included configuration, and so on.  
498 Finally, each way to express a model changes the research question in subtle ways,  
499 and thus we wish to emphasize that our approach does not necessarily assess the  
500 robustness of a particular question, but rather how it may vary when expressed in  
501 different ways (13, 40).

502 When there are not strong theory-based reasons to utilize specific statistical models for  
503 nutrition epidemiology questions, the approach we present herein may be useful to  
504 increase transparency and assess the distribution of results across many possible  
505 models. This approach may be facilitated by incorporating into standard workflows, and  
506 improving the availability of datasets used for nutrition epidemiology research questions  
507 (36).



## 508 Acknowledgements

509 The authors acknowledge the Indiana University Pervasive Technology Institute for  
510 providing supercomputing, storage, and consulting resources that have contributed to  
511 the research results reported within this paper. This research was supported in part by  
512 Lilly Endowment, Inc., through its support for the Indiana University Pervasive  
513 Technology Institute.

514 Authors' contributions: CJV, LEO, BH, AWB designed research; CJV, LEO, BH, CH, JS,  
515 CAS, RH, SLD, KE, AB, DBA, AWB conducted research; BH, CAS, SLD analyzed data  
516 or performed statistical analysis; CJV, LEO, BH, AWB wrote paper; AWB had primary  
517 responsibility for final content. All authors critically reviewed, edited, and approved the  
518 final manuscript.

## 519 *Conflicts of interest*

520 In the 36 months prior to the initial submission, Dr. Vorland has received honoraria from  
521 The Obesity Society and The Alliance for Potato Research and Education. In the 36  
522 months prior to the initial submission, Dr. Allison has received personal payments or  
523 promises for same from: Amin Talati Wasserman for KSF Acquisition Corp (Glanbia);  
524 Clark Hill PLC; General Mills; Kaleido Biosciences; Law Offices of Ronald Marron;  
525 Medpace/Gelesis; Novo Nordisk Fonden; Sports Research Corp.; USDA; and Zero  
526 Longevity Science (as stock options). Donations to a foundation have been made on his  
527 behalf by the Northarvest Bean Growers Association. The institution of Dr. Vorland, Ms.  
528 Henschel, Mr. Serrano, Ms. Dickinson, and Dr. Allison, Indiana University, and the  
529 Indiana University Foundation have received funds or donations to support their  
530 research or educational activities from: Alfred P. Sloan Foundation; Alliance for Potato

531 Research and Education; American Egg Board; Arnold Ventures; Eli Lilly and Company;  
532 Mars, Inc.; National Cattlemen’s Beef Association; Pfizer, Inc.; National Pork Board;  
533 USDA; Soleno Therapeutics; WW (formerly Weight Watchers); and numerous other for-  
534 profit and non-profit organizations to support the work of the School of Public Health  
535 and the university more broadly. Dr. O’Connor’s research is funded by internal funds at  
536 the Agricultural Research Service, USDA and the National Cancer Institute, NIH as well  
537 as external funds from the National Institute of Agricultural, USDA and the Beef  
538 Checkoff. Dr. O’Connor also served unpaid on the National Pork Board - Real Pork  
539 Research Advisory 2nd Advisory Council. In the past 36 months, Dr. Brown has  
540 received travel expenses from Alliance for Potato Research and Education,  
541 International Food Information Council, and Soy Nutrition Institute Global; speaking  
542 honoraria from Alliance for Potato Research and Education, Calorie Control Council,  
543 Eastern North American Region of the International Biometric Society, International  
544 Food Information Council Foundation, Potatoes USA, Purchaser Business Group on  
545 Health, The Obesity Society, and University of Arkansas for Medical Sciences;  
546 consulting payments from National Cattlemen’s Beef Association, and Soy Nutrition  
547 Institute Global; and grants through his institution from Alliance for Potato Research &  
548 Education, American Egg Board, National Cattlemen’s Beef Association, NIH/NHLBI,  
549 NIH/NIDDK, NIH/NIGMS, and NSF/NIH. He has been involved in research for which his  
550 institution or colleagues have received grants or contracts from ACRI, Alliance for  
551 Potato Research & Education, Gordon and Betty Moore Foundation, Hass Avocado  
552 Board, Indiana CTSI, NIH/NCATS, NIH/NCI, NIH/NIA, NIH/NIGMS, NIH/NLM, and

553 UAMS. His wife is employed by Reckitt. Other authors report no disclosures in the last  
554 36 months prior to the initial submission.

### 555 [Data and Code Availability](#)

556 Researchers who wish to reproduce our analyses can submit a project proposal to the  
557 REGARDS team (28). Code to reproduce our analyses is publicly available:

558 <https://osf.io/sy96k/>

### 559 [Funding](#)

560 Funded by the Beef Checkoff. Supported in part by NIH grants R25DK099080,  
561 R25HL124208, and R25GM141507. The assertions expressed are those of the authors  
562 and not necessarily those of the NIH, USDA, or any other organization.

### 563 [Acknowledgements](#)

564 The REGARDS study is supported by cooperative agreement U01 NS041588 co-  
565 funded by the National Institute of Neurological Disorders and Stroke (NINDS) and the  
566 National Institute on Aging (NIA), National Institutes of Health, Department of Health  
567 and Human Service. Additional funding for REGARDS CHD outcomes was provided by  
568 R01HL080477. The content is solely the responsibility of the authors and does not  
569 necessarily represent the official views of the NINDS, NHLBI, or the NIA.

570 Representatives of the NINDS were involved in the review of the manuscript but were  
571 not directly involved in the collection, management, analysis or interpretation of the  
572 data. The authors thank the other investigators, the staff, and the participants of the  
573 REGARDS study for their valuable contributions. A full list of participating REGARDS  
574 investigators and institutions can be found at: <https://www.uab.edu/soph/regardsstudy/>

## Tables

**Table 1.** Model sets

	Model sets		
Variable choices	Model set 1	Model set 2	Model set 3
Outcome	CHD		
Exclusion criteria	History of CHD or cancer		
Base model <sup>1</sup>	'Basic'		
Measure of association	HR		
Exposure definition	Beef intake <sup>2</sup>		
Exposure configuration	Continuous	Quintiles	Quintiles and Continuous
Energy cutoff	Vary cutoffs per <b>Supplemental Table 2</b>		According to published papers <sup>3</sup>
Covariates	Inclusion/Exclusion and configuration when included		According to published papers <sup>3</sup>

Grey = Consistent across model sets; White = Varies across model sets for the exposure configuration; Blue = Varies within each model set. <sup>1</sup> 'Base model' covariates that were always included were: age (3 different configurations, one at a time), gender, calorie intake (3 different configurations, one at a time), size of census tract, and REGARDS region. <sup>2</sup> Beef was defined using gram weight estimates (using the following variables: 'hamburger', 'beefroast' 'beeffattrimmed', 'beeffatnotrim'; based on the FFQ items "hamburgers, cheeseburgers, meat loaf, at home or in a restaurant" and "beef steaks"); values in the hamburger variable were multiplied by a proportion of 0.59 to refine the estimation of beef content. This proportion was determined from the Food and Nutrient Database for Dietary Studies 2017-2018 data (29). <sup>3</sup> We used the energy cutoff that aligned closest to one of our 7 different configurations for energy cutoffs. <sup>3</sup> We used covariate configurations that aligned closest with one of our covariate configurations.

**Table 2.** Descriptive statistics for model sets 1 and 2.

	<b>Overall (n=1,000,000)</b>	<b>Beef as Continuous<sup>1</sup> (n=499,564, Model Set 1)</b>	<b>Beef as Quintiles<sup>2</sup> (n=500,436, Model Set 2)</b>
<b>Hazard ratios (HR) for beef</b>			
• Mean HR	1.09	1.04	1.13
• Median HR	1.08	1.04	1.13
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile HR <sup>3</sup>	0.97, 1.23	0.96, 1.13	1.00, 1.26
• Min, max HR	0.85, 1.49	0.85, 1.25	0.85, 1.49
• N (%) of HR > 1.00 <sup>4</sup>	868,796 (87.9%)	392,426 (78.6%)	476,370 (95.2%)
• N (%) of significant HR (p<0.05) <sup>5</sup>	9,556 (0.96%)	3,695 (0.74%)	5,861 (1.17%)
• N (%) of significant HR > 1.00	9,556 (0.96%)	3,695 (0.74%)	5,861 (1.17%)
<b>95% Confidence interval width for beef HR</b>			
• Mean width	0.49	0.35	0.64
• Median width	0.51	0.34	0.63
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile of width	0.32, 0.71	0.32, 0.38	0.56, 0.73
• Min, max width	0.29, 0.95	0.29, 0.44	0.46, 0.95
<b>Significance</b>			
• Mean p-value	0.50	0.56	0.44
• Median p-value	0.48	0.57	0.40
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile p-value	0.11, 0.94	0.13, 0.96	0.10, 0.91
• Min, max p-value	0.00, 1.00	0.00, 1.00	0.00, 1.00
<b>z-score</b>			
• Mean z-score	0.68	0.52	0.84
• Median z-score	0.70	0.52	0.84
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile z-score	-0.31, 1.61	-0.47, 1.53	0.01, 1.66
• Min, max z-score	-1.67, 2.97	-1.67, 2.97	-1.08, 2.92

<sup>1</sup> Per 50g self-reported, estimated intake. <sup>2</sup> Hazard ratio of highest versus lowest. <sup>3</sup> 5% and 95% represent the actual 5<sup>th</sup> percent and 95<sup>th</sup> percent of the ranked distribution of HRs from the models fit, not a confidence interval around the HR. <sup>4</sup> comparing the number of positive associations by beef configuration: Chi-Square test:  $\chi^2(1)=60707$ ,  $p<0.001$ . <sup>5</sup> significant at  $p<0.05$ , comparing number of significant results by beef configuration: Chi-Square test:  $\chi^2(1)=492$ ,  $p<0.001$ .

**Table 3:** Descriptive statistics for models emulating existing literature.

	<b>Overall (n=20)</b>	<b>Beef as Continuous<sup>1</sup> (n=6)</b>	<b>Beef as Quintiles<sup>2</sup> (n=14)</b>
<b>Hazard ratios (HR) for beef</b>			
• Mean HR	1.12	1.07	1.15
• Median HR	1.09	1.08	1.13
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile HR <sup>3</sup>	1.05, 1.27	1.02, 1.09	1.08, 1.27
• Min, max HR	1.02, 1.27	1.02, 1.09	1.08, 1.27
• N (%) of HR > 1.00 <sup>4</sup>	20 (100%)	6 (100%)	14 (100%)
• N (%) of significant HR (p<0.05) <sup>5</sup>	2 (10%)	0 (0%)	2 (14%)
• N (%) of significant HR > 1.00	2 (10%)	0 (0%)	2 (14%)
<b>95% Confidence Interval width for beef HR</b>			
• Mean width	0.48	0.31	0.55
• Median width	0.55	0.30	0.58
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile of width	0.29, 0.62	0.29, 0.32	0.37, 0.63
• Min, max width	0.29, 0.63	0.29, 0.32	0.37, 0.63
<b>Significance</b>			
• Mean p-value	0.33	0.37	0.31
• Median p-value	0.30	0.30	0.31
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile p-value	0.04, 0.71	0.23, 0.81	0.04, 0.60
• Min, max p-value	0.04, 0.81	0.23, 0.81	0.04, 0.60
<b>z-score</b>			
• Mean z-score	1.07	0.93	1.13
• Median z-score	1.04	1.04	1.02
• 5 <sup>th</sup> , 95 <sup>th</sup> percentile z-score	0.38, 2.08	0.24, 1.21	0.52, 2.10
• Min, max z-score	0.24, 2.10	0.24, 1.21	0.52, 2.10

<sup>1</sup> Per 50g self-reported, estimated intake. <sup>2</sup> Hazard ratio of highest versus lowest. <sup>3</sup> 5% and 95% represent the actual 5<sup>th</sup> percent and 95<sup>th</sup> percent of the ranked distribution of HRs from the models fit, not a confidence interval around the HR. <sup>4</sup> comparing the number of positive associations by beef configuration: Chi-Square test:  $\chi^2(1)=60707$ ,  $p<0.001$ . <sup>5</sup> significant at  $p<0.05$ , comparing number of significant results by beef configuration: Chi-Square test:  $\chi^2(1)=492$ ,  $p<0.001$ .

**Table 4:** Results from Kolmogorov-Smirnov test comparing the distributions of HR for beef when including and excluding specific covariate.

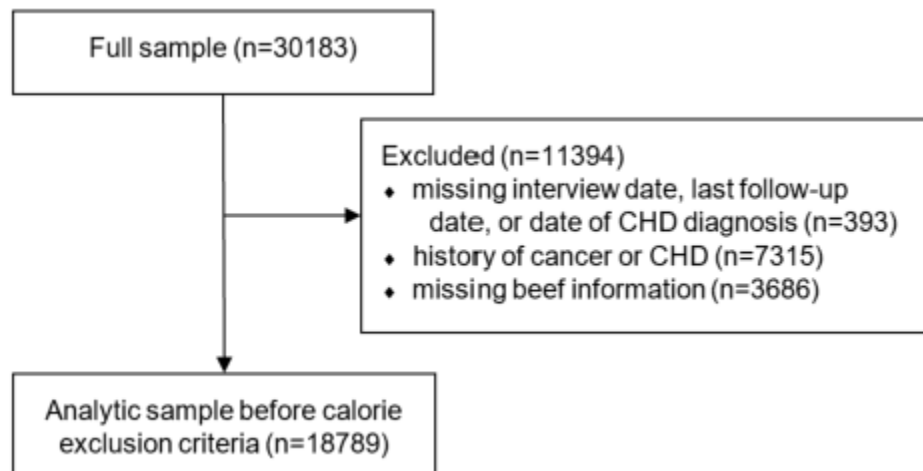
Covariate	Beef as Continuous				Beef as Quintiles			
	D statistic <sup>a</sup>	p value	Likeness	Cross-correlation	D statistic <sup>a</sup>	p value	Likeness	Cross-correlation
race	0.1815	<0.001	0.88	0.98	0.3190	<0.001	0.89	0.98
education	0.0297	<0.001	0.90	0.98	0.0058	<0.001	0.92	0.99
income	0.1329	<0.001	0.82	0.97	0.1419	<0.001	0.86	0.99
relationship	0.0290	<0.001	0.93	0.99	0.0502	<0.001	0.95	1.00
smoking	0.0166	<0.001	0.87	0.97	0.0696	<0.001	0.89	0.98
alcohol	0.0204	<0.001	0.87	0.97	0.1163	<0.001	0.73	0.88
physact	0.1027	<0.001	0.92	0.99	0.1611	<0.001	0.77	0.91
sedent	0.0177	<0.001	0.90	0.98	0.0596	<0.001	0.91	0.99
multivit	0.6308	<0.001	0.71	0.87	0.1910	<0.001	0.83	0.96
subjhealth	0.0394	<0.001	0.96	1.00	0.0236	<0.001	0.80	0.93
pain	0.0368	<0.001	0.94	0.99	0.0459	<0.001	0.89	0.98
hypertension	0.0079	<0.001	0.91	0.99	0.0417	<0.001	0.86	0.97
histhyperlip	0.0175	<0.001	0.91	0.98	0.0649	<0.001	0.83	0.95
histdiab	0.1374	<0.001	0.93	0.99	0.1828	<0.001	0.73	0.88
histafib	0.0244	<0.001	0.95	1.00	0.0234	<0.001	0.91	0.98
histpad	0.0348	<0.001	0.96	1.00	0.0090	<0.001	0.95	1.00
histcvd	0.0102	<0.001	0.88	0.98	0.0073	<0.001	0.97	1.00
stroke	0.1459	<0.001	0.88	0.97	0.1072	<0.001	0.58	0.71
weight	0.0863	<0.001	0.94	1.00	0.1711	<0.001	0.85	0.96
HEI	0.0656	<0.001	0.92	0.99	0.0190	<0.001	0.90	0.98
fiber	0.0424	<0.001	0.93	0.99	0.0906	<0.001	0.84	0.96
satfat	0.0349	<0.001	0.96	1.00	0.0224	<0.001	0.85	0.96
monofat	0.0053	0.002	0.88	0.97	0.0260	<0.001	0.93	0.99
polyfat	0.0106	<0.001	0.92	0.99	0.0092	<0.001	0.87	0.97
wholegrains	0.0078	<0.001	0.89	0.98	0.0347	<0.001	0.87	0.97
grains	0.0045	0.014	0.88	0.98	0.0096	<0.001	0.97	1.00

fruit	0.1391	<0.001	0.91	0.98	0.2397	<0.001	0.92	0.99
veggies	0.1290	<0.001	0.92	0.99	0.1180	<0.001	0.83	0.95
dairy	0.0530	<0.001	0.93	0.99	0.0630	<0.001	0.94	0.99

a) Critical value for D: 0.0038

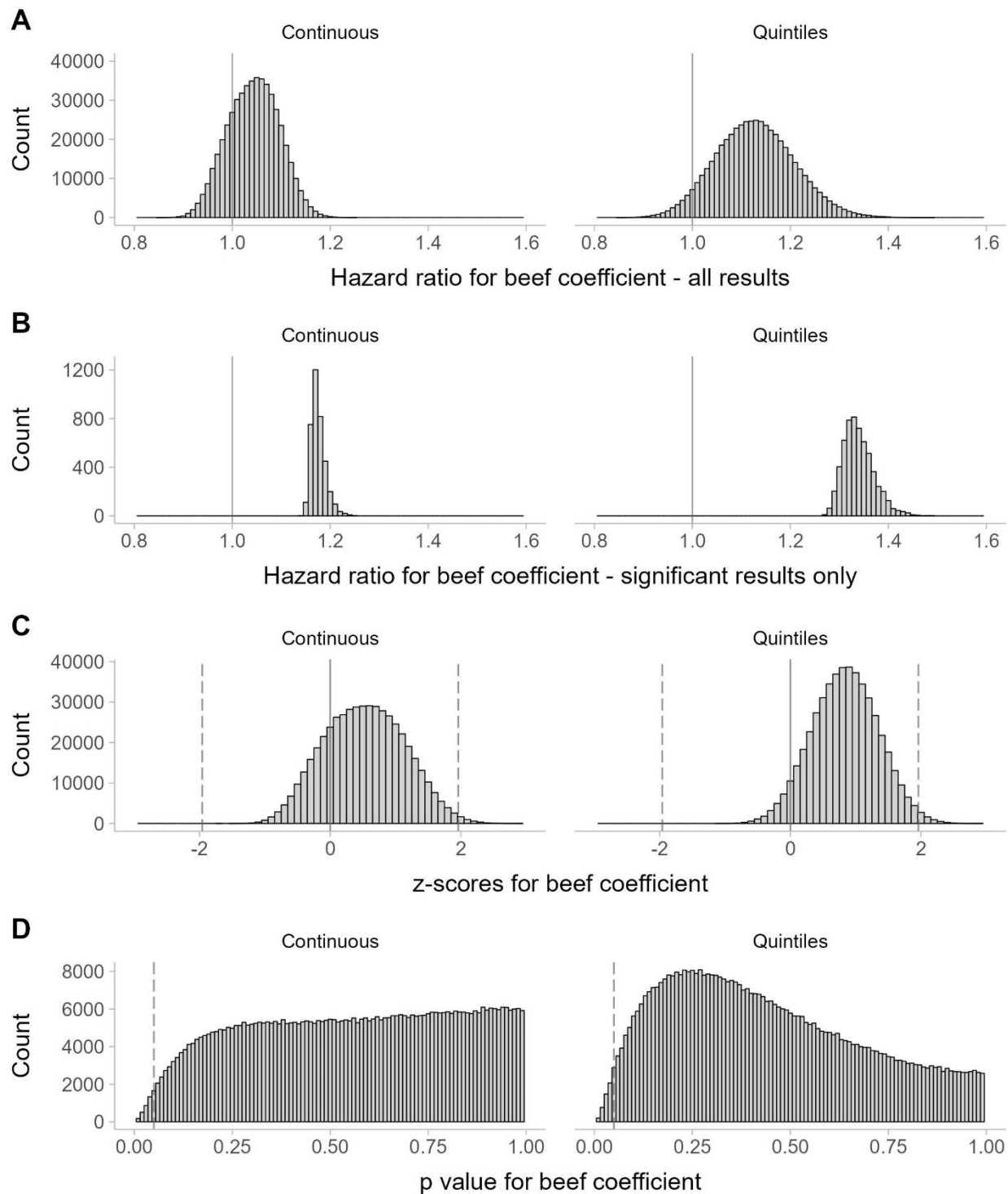


## Figures



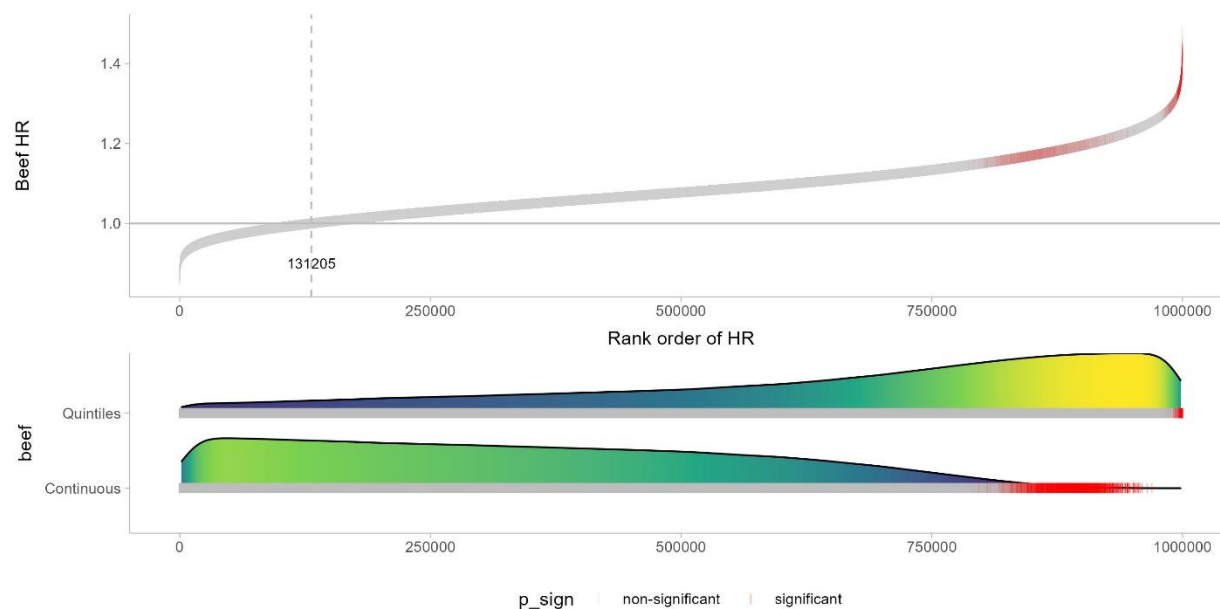
**Figure 1.** Sample size flow chart.

The number of participants in the full sample, number after participant exclusions and before energy intake exclusions, and reasons for exclusion.

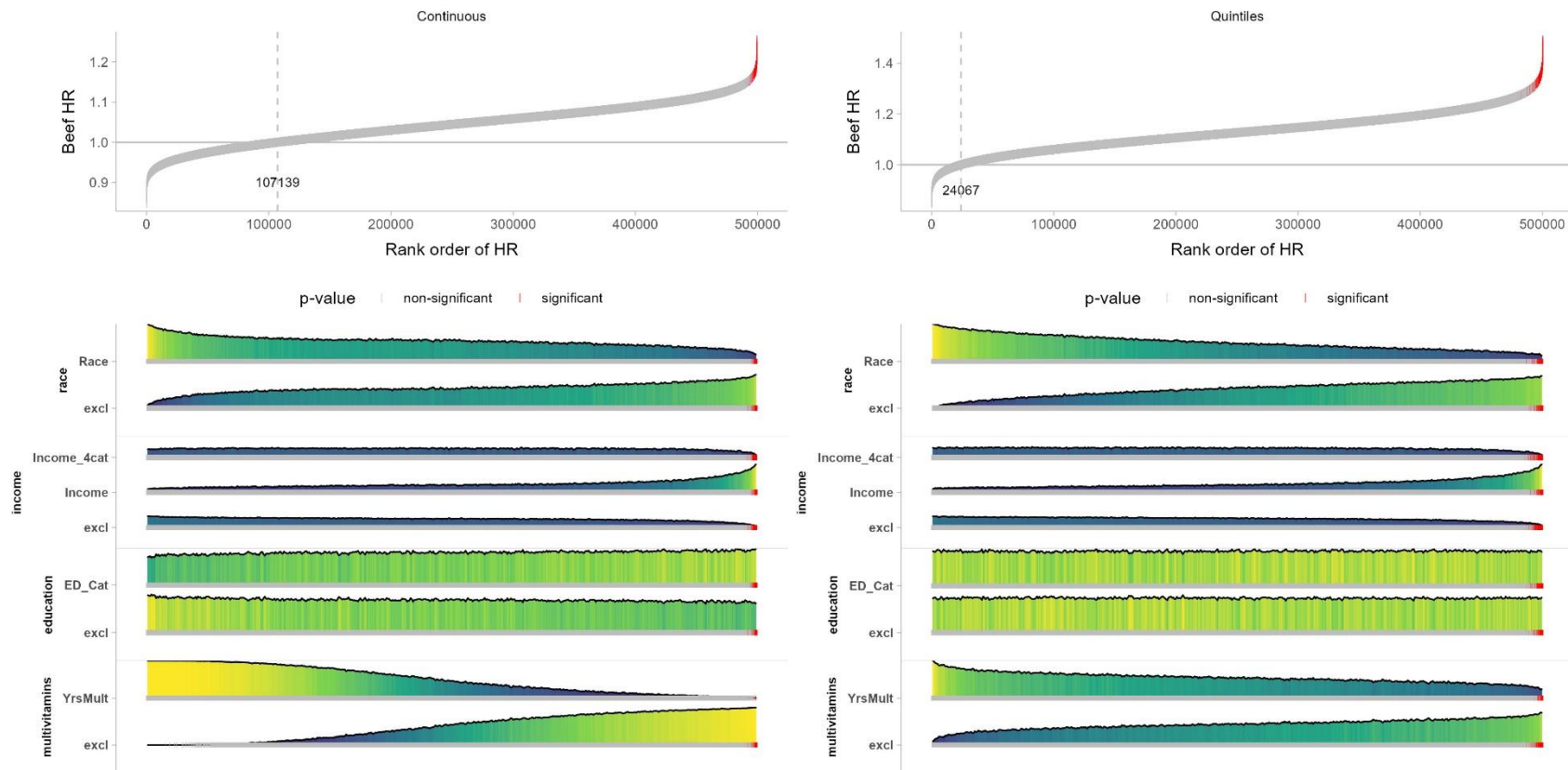


**Figure 2.** Frequency of Hazard ratios all model results (row A), Hazard ratios from significant model results ( $p < 0.05$ , row B), z scores from all model results (row C), and p-values from all model results (row D) for coefficient for beef intake for model set 1 when beef is expressed as continuous (left; per 50g), and model set 2 when beef is expressed as quintiles (right; highest vs. lowest). y-axis shows the number of models, the scale in the rows B and D is smaller to better show the distribution for significant hazard ratios

and p-values. Vertical dashed lines represent  $z=|1.96|$ , values outside the  $\pm 1.96$  range are considered significant at  $p < 0.05$  (row C),  $p = 0.05$  (row D).

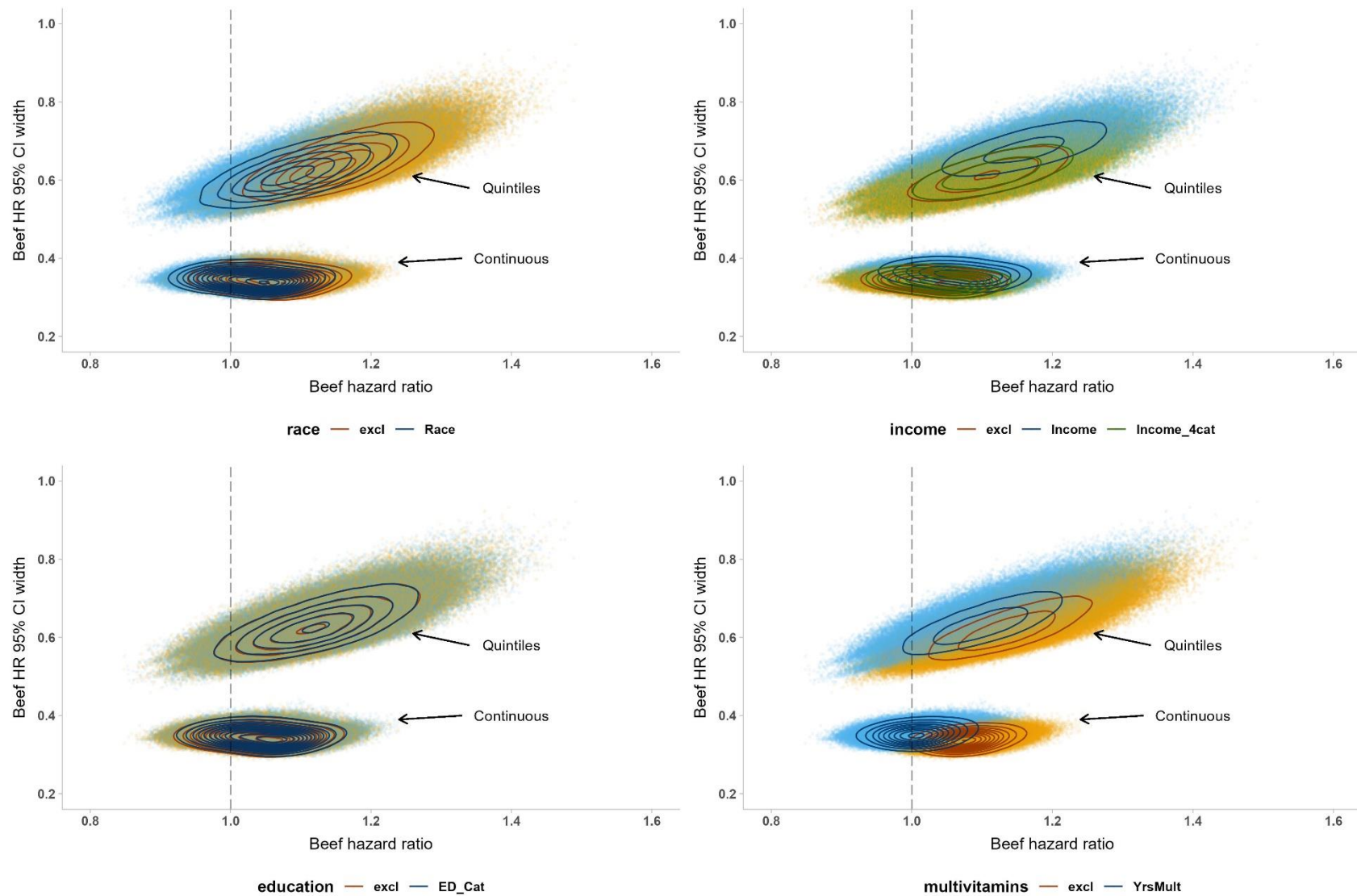


**Figure 3.** Specification curve that shows the distribution of HRs for the association between self-reported beef consumption and CHD for model sets 1 and 2. Each of the 1,000,000 model combinations is represented by a thin vertical bar in either gray color (if  $p \geq 0.05$ ) or red (if  $p < 0.05$ ). Curves in the bottom plot show distribution of HR by beef intake configuration, with the same gray or red thin vertical bars as above. A bar in the top plot can be traced down to the bottom plot. Color represents the density in models along the HR distribution (yellow=more models, dark blue=fewer models).



**Figure 4.** Specification curves that show the distribution of HRs for the association between self-reported beef consumption and CHD, when beef is expressed as a continuous variable (left), and as quintiles of intake (right). Each combination of covariates is represented if a vertical line were traced from any point on the curve on top down through each variable underneath. Red lines represent models that were  $p < 0.05$ . Variables: race (White; Black); income ('Income\_4cat': <\$20K, \$20K-\$35K, \$35K-75K, \$75K+, Refused; 'Income': 1 (<5K), 2 (5-10K), 3 (10-15K), 4 (15-20K), 5 (20-25K), 6 (25-35K), 7 (35-50K), 8 (50-75K), 9 (75-150K), 10 (>150K)); education ('ED\_Cat': <HS, HS, Some College, College+); multivitamins ('YrsMult': Years took multivitamins (0=No vitamins taken in past year, 1=Less than 1 year, 2=1

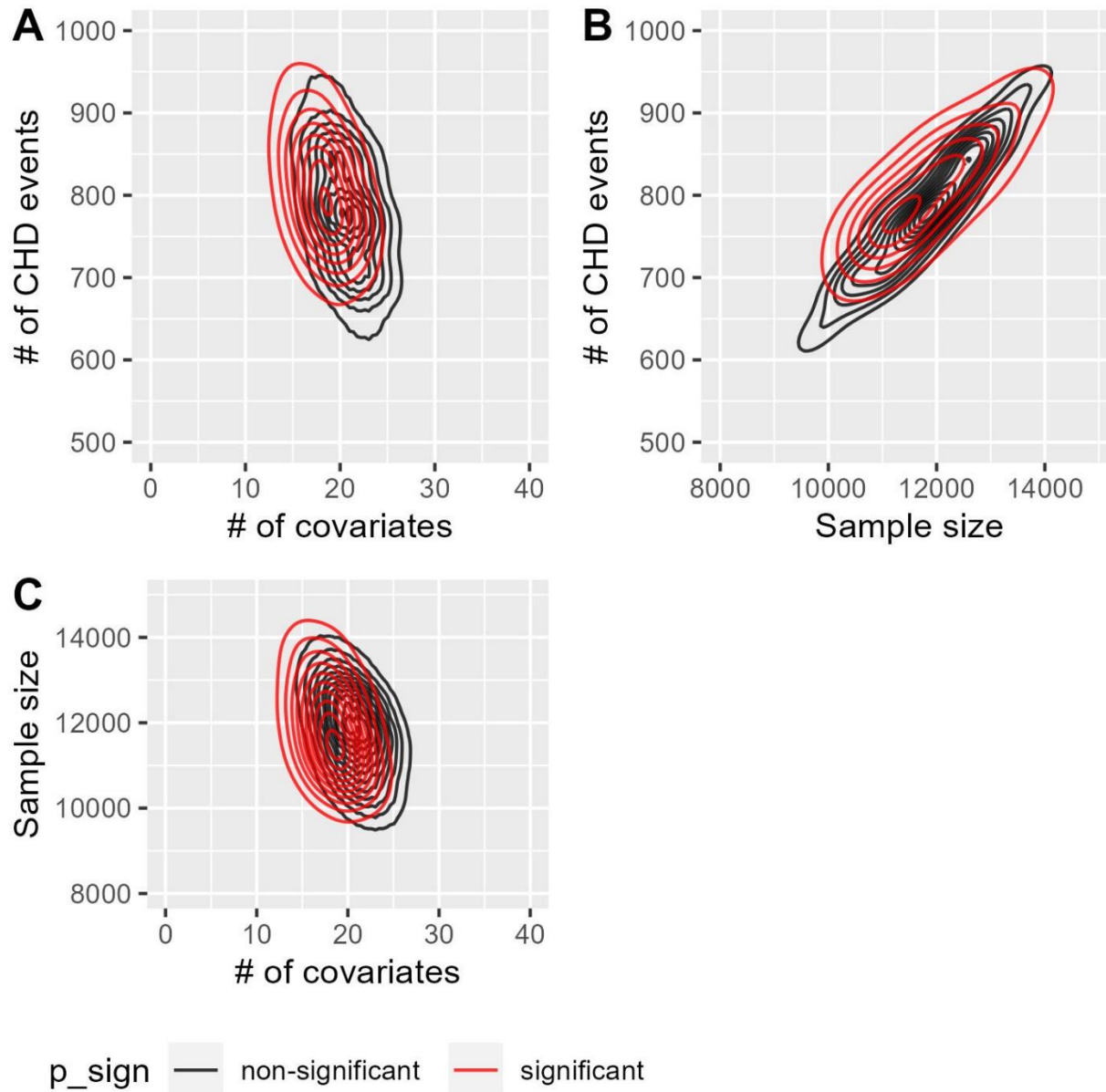
Year, 3=2 Years, 4=3-4 Years, 5=5-9 Years, 6=10+ Years)). Color represents the density in models along the HR distribution (yellow=more models, dark blue=fewer models).



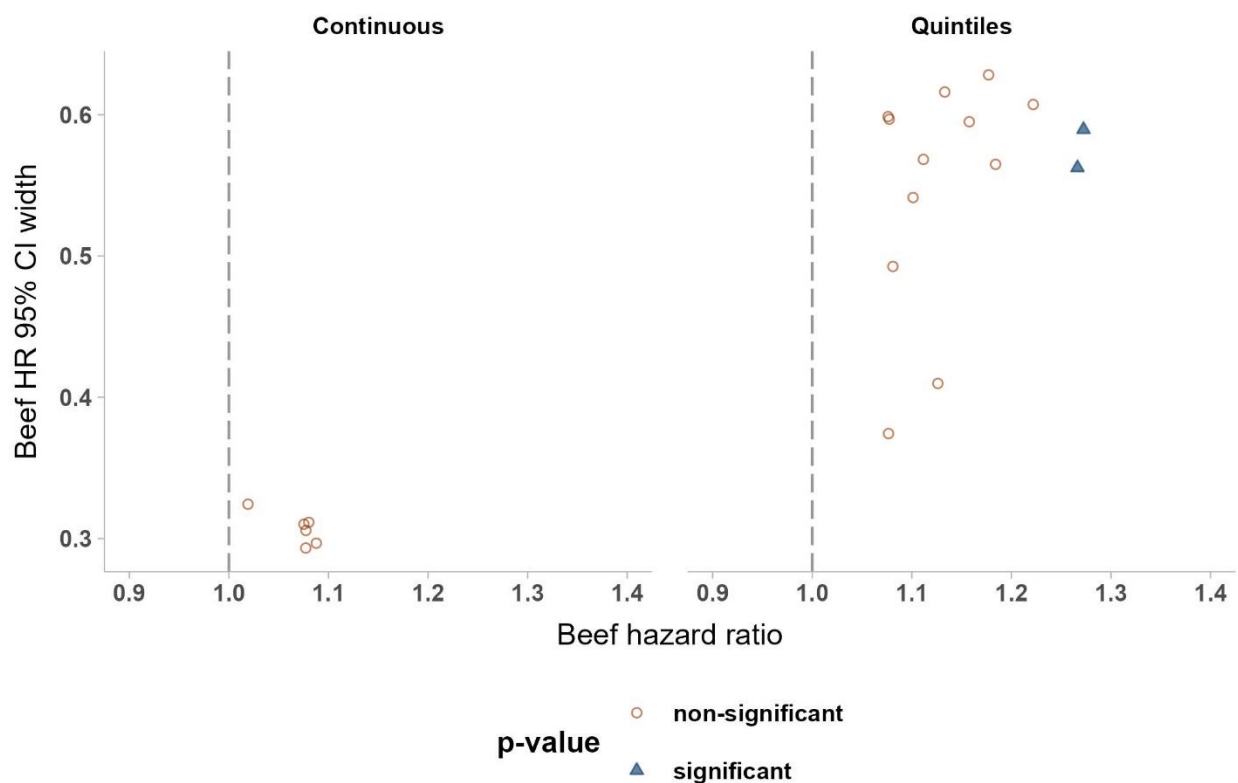
**Figure 5.** Plots that show the distribution of HRs vs. 95% CI width for four selected covariates (top left: race; top right: income, bottom left: education, bottom right: years of multivitamin use) when inclusion/exclusion and configuration is

varied, when self-reported beef consumption is expressed as continuous and as quintiles of intake. Lines are contour lines from a kernel density estimation using a normal distribution kernel; kernel smoothing was done over 200 grid points. Kernel density estimates were made using the MASS package with the `kde2d` function, as described (50).

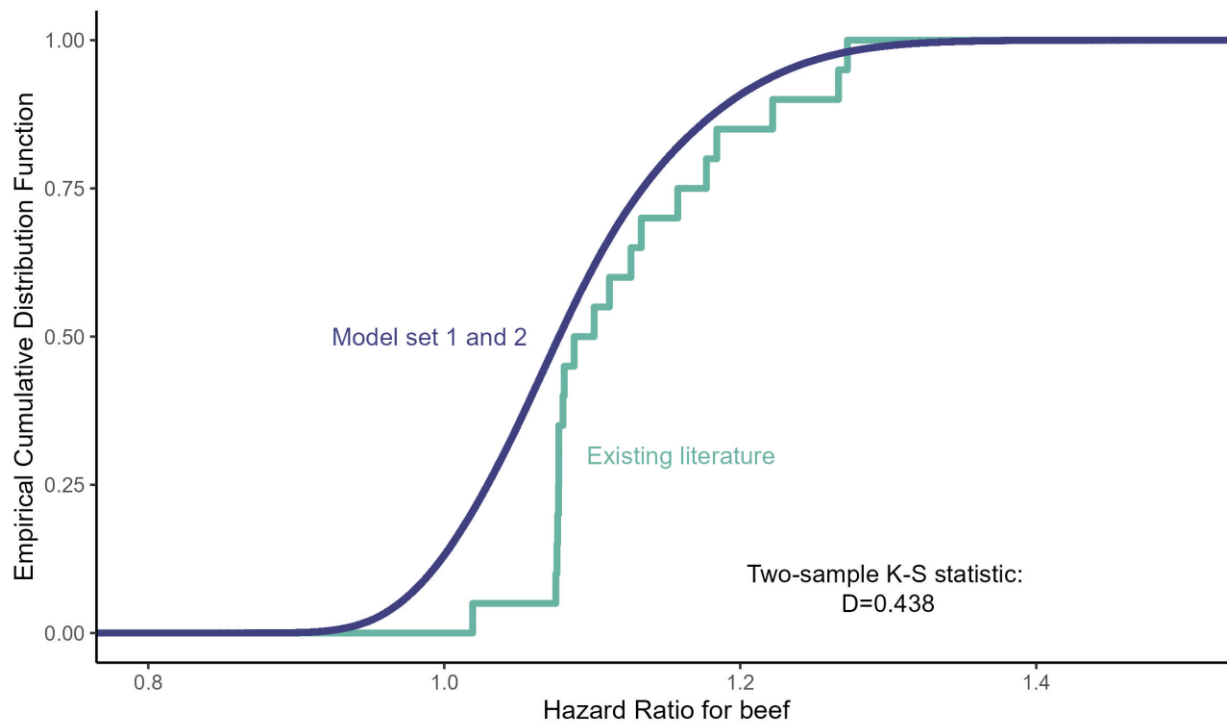




**Figure 6:** Pairwise density plots of number of CHD events (A, B), sample size (B, C), and number of covariates in the model (A, C). Density curves in red color represent models with significant coefficients for beef ( $p < 0.05$ ).



**Figure 7.** Scatterplot of HRs vs. 95% CI width for when self-reported beef consumption is expressed as continuous (left), and as quintiles of intake (right) for models emulating existing literature. Different symbols represent statistical significance at  $p < 0.05$  (filled triangle) versus non-significance ( $p \geq 0.05$ , circles)



**Figure 8.** Comparison of the cumulative distribution between HRs from model sets 1 and 2 (combined) with HRs from models emulating existing literature. K-S statistic: Kolmogorov-Smirnov statistic.

## **Supplemental File Descriptions**

**Supplemental Tables and Figures.** Supplemental tables and figures referenced within the text.

**Supplemental File 1.** Plots that show the distribution of HRs vs. SE for each covariate when inclusion/exclusion is varied, when self-reported beef consumption is expressed as continuous (left) and as quintiles of intake (right). Lines are percentile contours from a kernel density estimation using a normal distribution kernel.

**Supplemental File 2.** Plots that show the distribution of HRs vs. SE for each covariate when configuration is varied, when self-reported beef consumption is expressed as continuous (left) and as quintiles of intake (right). Lines are percentile contours from a kernel density estimation using a normal distribution kernel.

**Supplemental File 3.** Specification curves that show the distribution of HRs for the association between self-reported beef consumption and CHD, when beef is expressed as a continuous variable. Each combination of covariates is represented if a vertical line were traced from any point on the curve on top down through each variable underneath.

**Supplemental File 4.** Specification curves that show the distribution of HRs for the association between self-reported beef consumption and CHD, when beef is expressed as quintiles of intake. Each combination of covariates is represented if a vertical line were traced from any point on the curve on top down through each variable underneath.

## References

1. Kelley K, Preacher KJ. On effect size. *Psychological methods*. 2012;17(2):137.
2. Savage SL, Danziger J. *The flaw of averages : why we underestimate risk in the face of uncertainty*. 1st edition ed. Hoboken, New Jersey: John Wiley & Sons, 2009.
3. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*. 2018;1(3):337-56. doi: 10.1177/2515245917747646.
4. Breznau N, Rinke EM, Wuttke A, Nguyen HH, Adem M, Adriaans J, Alvarez-Benjumea A, Andersen HK, Auer D, Azevedo F. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*. 2022;119(44):e2203150119.
5. Dutilh G, Annis J, Brown SD, Cassey P, Evans NJ, Grasman RP, Hawkins GE, Heathcote A, Holmes WR, Kryptos A-M. The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic bulletin & review*. 2019;26:1051-69.
6. Bastiaansen JA, Kunkels YK, Blaauw FJ, Boker SM, Ceulemans E, Chen M, Chow S-M, de Jonge P, Emerencia AC, Epskamp S. Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of psychosomatic research*. 2020;137:110211.
7. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020;582(7810):84-8.

8. Menkveld AJ, Dreber A, Holzmeister F, Huber J, Johannesson M, Kirchler M, Neusüss S, Razen M, Weitzel U. Non-standard errors. 2021.
9. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. . Department of Statistics, Columbia University. 2013.
10. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci.* 2011;22(11):1359-66. doi: 10.1177/0956797611417632.
11. Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. *The American journal of clinical nutrition.* 2013;97(1):127-34.
12. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science.* 2016;11(5):702-12.
13. Del Giudice M, Gangestad SW. A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science.* 2021;4(1):2515245920954925.
14. Simonsohn U, Simmons JP, Nelson LD. Specification curve: Descriptive and inferential statistics on all reasonable specifications. Available at SSRN 2694998. 2015.
15. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of clinical epidemiology.* 2015;68(9):1046-58.
16. Tierney BT, Anderson E, Tan Y, Claypool K, Tangirala S, Kostic AD, Manrai AK, Patel CJ. Leveraging vibration of effects analysis for robust discovery in observational biomedical data science. *PLoS biology.* 2021;19(9):e3001398.

17. Zeraatkar D, Cheung K, Milio K, Zworth M, Gupta A, Bhasin A, Bartoszko JJ, Kiflen M, Morassut RE, Noor ST. Methods for the selection of covariates in nutritional epidemiology studies: a meta-epidemiological review. *Current developments in nutrition*. 2019;3(10):nzz104.
18. Klurfeld DM. Research gaps in evaluating the relationship of meat and health. *Meat Sci*. 2015;109:86-95. doi: 10.1016/j.meatsci.2015.05.022.
19. O'Connor LE, Kim JE, Campbell WW. Total red meat intake of  $\geq 0.5$  servings/d does not negatively influence cardiovascular disease risk factors: a systemically searched meta-analysis of randomized controlled trials. *Am J Clin Nutr*. 2017;105(1):57-69. doi: 10.3945/ajcn.116.142521.
20. Satija A, Malik VS, Willett WC, Hu FB. Meta-analysis of red meat intake and cardiovascular risk factors: methodologic limitations. *Am J Clin Nutr*. 2017;105(6):1567-8. doi: 10.3945/ajcn.117.153692.
21. O'Connor LE, Kim JE, Campbell WW. Reply to A Satija et al. *Am J Clin Nutr*. 2017;105(6):1568-9. doi: 10.3945/ajcn.117.154625.
22. Gifford CL, O'Connor LE, Campbell WW, Woerner DR, Belk KE. Broad and Inconsistent Muscle Food Classification Is Problematic for Dietary Guidance in the U.S. *Nutrients*. 2017;9(9). doi: 10.3390/nu9091027.
23. Guasch-Ferre M, Satija A, Blondin SA, Janiszewski M, Emlen E, O'Connor LE, Campbell WW, Hu FB, Willett WC, Stampfer MJ. Meta-Analysis of Randomized Controlled Trials of Red Meat Consumption in Comparison With Various Comparison Diets on Cardiovascular Risk Factors. *Circulation*. 2019;139(15):1828-45. doi: 10.1161/CIRCULATIONAHA.118.035225.
24. Zhong VW, Van Horn L, Greenland P, Carnethon MR, Ning H, Wilkins JT, Lloyd-Jones DM, Allen NB. Associations of processed meat, unprocessed red meat, poultry, or fish intake with incident cardiovascular disease and all-cause mortality. *JAMA internal medicine*. 2020;180(4):503-12.

25. Neuhouser ML. Red and processed meat: more with less? *Am J Clin Nutr.* 2020;111(2):252-5. doi: 10.1093/ajcn/nqz294.
26. Johnston BC, Guyatt GH. Causal inference, interpreting and communicating results on red and processed meat. *Am J Clin Nutr.* 2020;111(5):1107-8. doi: 10.1093/ajcn/nqaa043.
27. Neuhouser ML. Reply to BC Johnston and GH Guyatt. *Am J Clin Nutr.* 2020;111(5):1108-9. doi: 10.1093/ajcn/nqaa038.
28. *REGARDS - REasons for Geographic and Racial Differences in Stroke* [Internet]. Available from: <https://www.uab.edu/soph/regardsstudy/>.
29. U.S. Department of Agriculture ARS. *Food and Nutrient Database for Dietary Studies (FNDDS)* [Internet]. Available from: <https://data.nal.usda.gov/dataset/food-and-nutrient-database-dietary-studies-fndds>.
30. Wang X, Lin X, Ouyang YY, Liu J, Zhao G, Pan A, Hu FB. Red and processed meat consumption and mortality: dose–response meta-analysis of prospective cohort studies. *Public health nutrition.* 2016;19(5):893-905.
31. *About Carbonate at Indiana University* [Internet]. Available from: <https://kb.iu.edu/d/aolp>.
32. Sundell K, Saylor J. Two-dimensional quantitative comparison of density distributions in detrital geochronology and geochemistry. *Geochemistry, Geophysics, Geosystems.* 2021;22(4):e2020GC009559.
33. Howard VJ, Cushman M, Pulley L, Gomez CR, Go RC, Prineas RJ, Graham A, Moy CS, Howard G. The reasons for geographic and racial differences in stroke study: objectives and design. *Neuroepidemiology.* 2005;25(3):135-43. doi: 10.1159/000086678.
34. Touvier M, Kesse E, Volatier J-L, Clavel-Chapelon F, Boutron-Ruault M-C. Dietary and cancer–related behaviors of vitamin/mineral dietary supplement users in a large cohort of French women. *European journal of nutrition.* 2006;45:205-14.



35. Williams DR, Mohammed SA, Leavell J, Collins C. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. *Annals of the new York Academy of Sciences*. 2010;1186(1):69-101.
36. Brown AW, Aslibekyan S, Bier D, Ferreira da Silva R, Hoover A, Klurfeld DM, Loken E, Mayo-Wilson E, Menachemi N, Pavela G. Toward more rigorous and informative nutritional epidemiology: The rational space between dismissal and defense of the status quo. *Critical Reviews in Food Science and Nutrition*. 2021:1-18.
37. Maki KC, Slavin JL, Rains TM, Kris-Etherton PM. Limitations of observational evidence: implications for evidence-based dietary recommendations. *Advances in nutrition*. 2014;5(1):7-15.
38. Song F, Hooper L, Loke YK. Publication bias: what is it? How do we measure it? How do we avoid it? *Open Access Journal of Clinical Trials*. 2013:71-81.
39. Steiner PM, Cook TD, Shadish WR, Clark MH. The importance of covariate selection in controlling for selection bias in observational studies. *Psychological methods*. 2010;15(3):250.
40. Tomova GD, Arnold KF, Gilthorpe MS, Tennant PW. Adjustment for energy intake in nutritional research: a causal inference perspective. *The American journal of clinical nutrition*. 2022;115(1):189-98.
41. Stefan A, Schönbrodt F. Big little lies: A compendium and simulation of p-hacking strategies. 2022.
42. Sturman MC, Sturman A, Sturman CJ. Uncontrolled control variables: The extent that a researcher's degrees of freedom with control variables increases various types of statistical errors. *Journal of Applied Psychology*. 2021.
43. Christensen JD, Orquin JL, Perkovic S, Lagerkvist CJ. Preregistration is important, but not enough: Many statistical analyses can inflate the risk of false-positives. 2021.

44. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in medicine*. 2004;23(7):1159-78.
45. Ley SH, Sun Q, Willett WC, Eliassen AH, Wu K, Pan A, Grodstein F, Hu FB. Associations between red meat intake and biomarkers of inflammation and glucose metabolism in women. *The American journal of clinical nutrition*. 2014;99(2):352-60.
46. Al-Shaar L, Satija A, Wang DD, Rimm EB, Smith-Warner SA, Stampfer MJ, Hu FB, Willett WC. Red meat intake and risk of coronary heart disease among US men: prospective cohort study. *bmj*. 2020;371.
47. Etemadi A, Sinha R, Ward MH, Graubard BI, Inoue-Choi M, Dawsey SM, Abnet CC. Mortality from different causes associated with meat, heme iron, nitrates, and nitrites in the NIH-AARP Diet and Health Study: population based cohort study. *bmj*. 2017;357.
48. National Cancer Institute. *Learn More about Energy Adjustment* [Internet]. Available from: <https://www.dietassessmentprimer.cancer.gov/learn/adjustment.html>.
49. O'Connor LE, Gifford CL, Woerner DR, Sharp JL, Belk KE, Campbell WW. Dietary meat categories and descriptions in chronic disease research are substantively different within and between experimental and observational studies: a systematic review and landscape analysis. *Advances in Nutrition*. 2020;11(1):41-51.
50. Venables WR, Ripley B. BD (2002). *Modern Applied Statistics with S*. Edtion ed. New York: Springer Science & Business Media, 2002:130.