Partial Identification of Dose Responses with Hidden Confounders (Supplementary Material)

Myrl G. Marmarelis[1] **Elizabeth Haddad**[2] **Andrew Jesson**[3]
Neda Jahanshad[2] **Aram Galstyan**[1] **Greg Ver Steeg**[1]

[1]USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292
[2]USC Stevens Neuroimaging and Informatics Institute, 4676 Admiralty Way, Marina del Rey, CA 90292
[3]University of Oxford, OATML, 14 Parks Road, Oxford, UK OX1 3AQ

A COMPLETING THE DERIVATIONS

Consider Equation 3.A:

$$\int_0^1 w_t(\tau)\tilde{p}(y_t|\tau,x)p(\tau|x)\,\mathrm{d}\tau = \underbrace{p(y_t|t,x)\int_0^1 w_t(\tau)p(\tau|x)\,\mathrm{d}\tau}_{(A.0)}$$
$$+ \underbrace{g_1(y_t|t,x)\int_0^1 w_t(\tau)(\tau-t)p(\tau|x)\,\mathrm{d}\tau}_{(A.1)} + \underbrace{g_2(y_t|t,x)\int_0^1 w_t(\tau)\frac{(\tau-t)^2}{2}p(\tau|x)\,\mathrm{d}\tau}_{(A.2)},$$
$$\text{where} \quad g_k(y_t|t,x) \coloneqq \partial_\tau^k p(y_t|\tau,x)|_{\tau=t}. \quad (10)$$

Lightening the notation with a shorthand for the weighted expectations, $\langle\cdot\rangle_\tau \coloneqq \int_0^1 w_t(\tau)(\cdot)p(\tau|x)\,\mathrm{d}\tau$, it becomes apparent that we must grapple with the pseudo-moments $\langle 1\rangle_\tau$, $\langle\tau-t\rangle_\tau$, and $\langle(\tau-t)^2\rangle_\tau$. Note that t should not be mistaken for a "mean" value.

Furthermore, we have yet to fully characterize $g_k(y_t|t,x)$. Observe that

$$p(y_t|\tau,x) = \frac{p(\tau|y_t,x)p(y_t|x)}{p(\tau|x)} \quad \iff \quad \partial_\tau p(y_t|\tau,x) = p(y_t|x)\cdot\frac{\partial}{\partial\tau}\frac{p(\tau|y_t,x)}{p(\tau|x)}.$$

The $p(y_t|x)$ will be moved to the other side of the equation as needed; by Equation 6,

$$\frac{\partial}{\partial\tau}\frac{p(\tau|y_t,x)}{p(\tau|x)} = \frac{\partial}{\partial\tau}\Lambda(\tau|y_t,x).$$

Expanding,

$$= \frac{\partial}{\partial\tau}\exp\left\{\int_0^\tau \gamma(\tau|y_t,x)\,\mathrm{d}\tau\right\} \quad = \quad \gamma(\tau|y_t,x)\exp\left\{\int_0^\tau \gamma(\tau|y_t,x)\,\mathrm{d}\tau\right\}$$
$$= (\gamma\Lambda)(\tau|y_t,x).$$

Appropriate bounds will be calculated for $g_2(y_t|t,x)$ next, utilizing the finding above as their main ingredient. Let

$$\tilde{g}_k(y_t|t,x) \coloneqq p(y_t|x)^{-1}g_k(y_t|t,x) = \left(\frac{\partial}{\partial\tau}\right)^k \frac{p(\tau|y_t,x)}{p(\tau|x)}\bigg|_{\tau=t.}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

The second derivative may be calculated in terms of the ignorance quantities γ, Λ:

$$\begin{aligned}\tilde{g}_2(y_t|t,x) =& \partial_\tau \gamma(\tau|y_t,x)\Lambda(\tau|y_t,x) \\ =& \gamma(\tau|y_t,x)^2\Lambda(\tau|y_t,x) + \dot{\gamma}(\tau|y_t,x)\Lambda(\tau|y_t,x) \\ =& (\gamma^2+\dot{\gamma})\Lambda(\tau|y_t,x).\end{aligned}$$

And finally we address $\tilde{p}(y_t|x)$. Carrying over the components of Equation 10 into Equation 3,

$$\begin{aligned}\tilde{p}(y_t|x) &= \frac{p(y_t|t,x)\langle 1\rangle_\tau}{\langle \Lambda(\tau|y_t,x)\rangle_\tau - \tilde{g}_1(y_t|t,x)\langle \tau - t\rangle_\tau - \tilde{g}_2(y_t|t,x)\langle (\tau-t)^2\rangle_\tau} \\ &= \frac{p(y_t|t,x)}{\mathbb{E}_\tau[\Lambda(\tau|y_t,x)] - (\gamma\Lambda)(t|y_t,x)\,\mathbb{E}_\tau[\tau - t] - \frac{1}{2}((\dot{\gamma}+\gamma^2)\Lambda)(t|y_t,x)\,\mathbb{E}_\tau[(\tau-t)^2]},\end{aligned} \tag{11}$$

where these expectations $\mathbb{E}_\tau[\cdot]$ are with respect to the implicit distribution $q(\tau|t,x) \propto w_t(\tau)p(\tau|x)$. The first term in the denominator, $\mathbb{E}_\tau[\Lambda(\tau|y_t,x)]$, may be approximately bounded by the same Algorithm 1.

B HOW TO CALIBRATE THE WEIGHING SCHEME

We present an argument based on the absolute error of the approximation in Equation 2, specifically for Beta propensities. The following applies to both Beta and Balanced Beta, $0 < t < 1$.

Suppose that the the second derivative employed in the Taylor expansion is Q-Lipschitz, so that $\left|\partial_\tau^3 p(y_t|\tau,x)\right| \leq Q$. Denote the remainder as $\rho(y_t|\tau,x)$. By Taylor's theorem,

$$|\rho(y_t|\tau,x)| \leq \frac{|\tau-t|^3}{6}Q.$$

The approximated quantity (part A) in Equation 3 is the following integral, which ends up becoming the numerator in Equation 4:

$$\int_0^1 w_t(\tau)\tilde{p}(y_t|\tau,x)p(\tau|x)\,\mathrm{d}\tau \;=\; \int_0^1 w_t(\tau)\big[p(y_t|\tau,x) + \rho(y_t|\tau,x)\big]p(\tau|x)\,\mathrm{d}\tau.$$

The absolute error of this integral is therefore

$$\left|\int_0^1 w_t(\tau)\rho(y_t|\tau,x)p(\tau|x)\,\mathrm{d}\tau\right| \;\leq\; \frac{1}{6}Q\underbrace{\int_0^1 w_t(\tau)p(\tau|x)|\tau-t|^3\,\mathrm{d}\tau}_{:=\,J,\text{ which upper-bounds the error.}} \quad \text{by the remainder theorem.}$$

Let $A = \alpha - 1 + rt$ and $B = \beta - 1 + r(1-t)$, where (α,β) parametrize the nominal propensity and r is the precision of the Beta trust-weighing scheme. The trust-propensity combination is

$$w_t(\tau)p(\tau|x) = \frac{\tau^A(1-\tau)^B}{c_t\,\mathbb{B}(\alpha,\beta)}, \quad \text{where } c_t = t^{rt}(1-t)^{r(1-t)}.$$

Hence, the error bound reduces to

$$\begin{aligned}J \;&= [c_t\,\mathbb{B}(\alpha,\beta)]^{-1}\int_0^1 \tau^A(1-\tau)^B|\tau-t|^3\,\mathrm{d}\tau \\ &= [c_t\,\mathbb{B}(\alpha,\beta)]^{-1}\left[\underbrace{\frac{\Gamma(A+1)\Gamma(B+1)}{\Gamma(A+B+5)}U_3(A,B,t)}_{\text{first term}} \;+\; \underbrace{\frac{\Gamma(A+1)}{\Gamma(A+5)}12t^{A+4}(1-t)^{B+4}\,{}_2F_1(4,A+B+5,A+5;\,t)}_{\text{second term}}\right],\end{aligned}$$

where $U_3(A,B,t)$ is a cubic polynomial in A, B, and t. Notice that even though the quantity is symmetric about $(A,B,t) \mapsto (B,A,1-t)$, the form does not appear so. We shall focus on the relation of the error bound entirely with A and α, then justify the analogous conclusion for B and β by the underlying symmetry of the expression.

The Gaussian hypergeometric function in the second term can be expressed as

$$\sum_{i=0}^{\infty} \frac{(4)_i (A+B+5)_i}{(A+5)_i} \frac{t^i}{i!} = \sum_{i=0}^{\infty} (4)_i \underbrace{\left(\frac{A+B+5}{A+5}\right)\left(\frac{A+B+6}{A+6}\right)\cdots}_{\text{length } i} \frac{t^i}{i!}$$

$$= \sum_{i=0}^{\infty} \frac{(4)_i}{i!} \left(1+\frac{B}{A+5}\right)\left(1+\frac{B}{A+6}\right)\cdots t^i, \quad \text{where } \frac{(4)_i}{i!} = \frac{(i+2)(i+3)(i+4)}{3!}.$$

by using the definition of the Pochhammer symbol $(x)_i = x(x+1)\dots(x+i-1)$. In terms of $A \to \infty$, the whole second term in J is $\mathcal{O}(A^{-4})$ due to the fraction of Γ functions. The first term in J is

$$\mathcal{O}(A^{-(B+4)} B^{-(A+4)}) \cdot U_3(A,B,t) = \mathcal{O}(A^{-B-1} B^{-A-1})$$

by Stirling's approximation of $\Gamma(x) = \mathcal{O}(x^{x-\frac{1}{2}})$. Clearly, a small $B > 0$ might cause the first term in J to explode with large A due to the $\mathcal{O}(B^{-A-1})$ part. This could occur with high α, low β, and low r—it is an instance of a high-precision propensity and low-precision weighing scheme destroying the upper error bound. Hence follows an argument for having r match the propensity's precision, to avoid these cases.

As mentioned earlier, the same argument flows for large B and small A, while swapping $t \mapsto (1-t)$.

C CORRECTNESS OF ALGORITHM 1

The algorithm functions by incrementally reallocating mass (relative, in the weights) to the righthand side, from a cursor beginning on the lefthand side of the "tape".

Proof. Firstly we characterize the indicator quantity Δ_j. Differentiate the quantity to be maximized with respect to w_j;

$$\begin{aligned}
\frac{\partial}{\partial w_j} \frac{\sum_i w_i f_i}{\sum_i w_i} &= \frac{f_j}{\sum_i w_i} - \frac{\sum_i w_i f_i}{\left(\sum_i w_i\right)^2} \\
&= \frac{f_j \sum_i w_i - \sum_i w_i f_i}{\left(\sum_i w_i\right)^2} \\
&\propto \underbrace{\sum_i w_i (f_j - f_i)}_{\coloneqq \Delta_j} \quad \text{up to some positive factor.}
\end{aligned}$$

Hence, Δ_j captures the sign of the derivative.

We shall proceed with induction. Begin with the first iteration, $j = 1$. No weights have been altered since initialization yet. Therefore we have

$$\Delta_1 = \sum_i \overline{w}_i (f_1 - f_i).$$

Since $\forall i,\ f_1 \le f_i$ due to the prior sorting, Δ_1 is either negative or zero. If zero, trivially terminate the procedure as all function values are identical.

Now assume that by the time the algorithm reaches some $j > 1$, all $w_k = \underline{w}_k$ for $1 \le k < j$. In other words,

$$\Delta_j = \sum_{i<j} \underline{w}_i \underbrace{(f_j - f_i)}_{(+)} + \sum_{i>j} \overline{w}_i \underbrace{(f_j - f_i)}_{(-)}.$$

Per the algorithm, we would flip the weight $w_j \leftarrow \underline{w}_j$ only if $\Delta_j < 0$. In that case,

$$\sum_{i<j} \underline{w}_i (f_j - f_i) < \sum_{i>j} \overline{w}_i (f_i - f_j), \quad \text{where both sides are non-negative.}$$

Notice that the above is not affected by the current value of w_j. This update can only increase the current estimate because the derivative remains negative and the weight at j is non-increasing. We *must* verify that the derivatives for the previous

weights, indexed at $k < j$, remain negative. Otherwise, the procedure would need to backtrack to possibly flip some weights back up.

More generally, with every decision for weight assignment, we seek to ensure that the condition detailed above is not violated for any weights that have been finalized. That includes the weights before j, and those after j at the point of termination. Returning from this digression, at $k < j$ after updating w_j,

$$\Delta_k = \sum_{i \leq j} \underline{w}_i (f_k - f_i) + \sum_{i > j} \overline{w}_i (f_k - f_i).$$

To glean the sign of this, we refer to a quantity that we know.

$$\begin{aligned}
& \sum_{i<j} \underline{w}_i (f_j - f_i) < \sum_{i>j} \overline{w}_i (f_i - f_j) \\
\iff & \sum_{i \leq j} \underline{w}_i (f_k - f_i) < \sum_{i>j} \overline{w}_i (f_i - f_j) + \sum_{i \leq j} \underline{w}_i (f_k - f_j) \\
\iff & \underbrace{\sum_{i \leq j} \underline{w}_i (f_k - f_i) + \sum_{i>j} \overline{w}_i (f_k - f_i)}_{\Delta_k} < \underbrace{\sum_{i>j} \overline{w}_i (f_k - f_j) + \sum_{i \leq j} \underline{w}_i (f_k - f_j)}_{\text{negative.}}
\end{aligned}$$

The remaining fact to be demonstrated is that upon termination, when $\Delta_j \geq 0$, no other pseudo-derivatives $\Delta_{j'},\ j' > j$ are negative. This must be the case simply because $f_{j'} \geq f_j$. $\square$

D ON THE INTRODUCTORY ILLUSTRATION

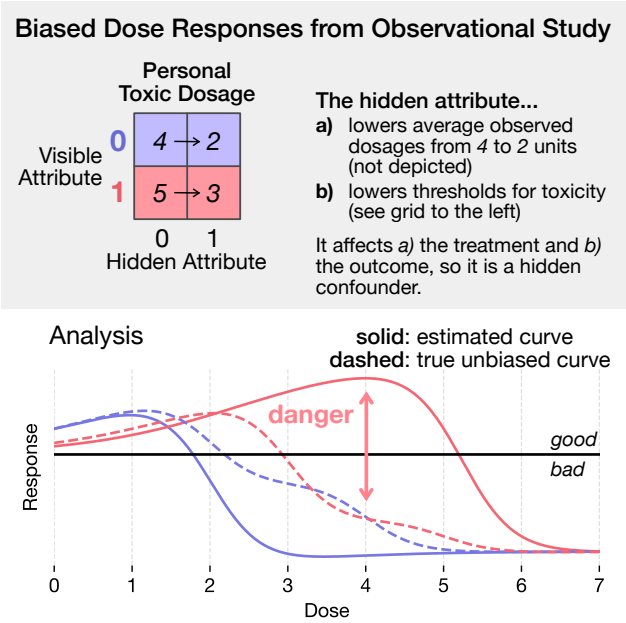

Figure 8. Elaboration on the example in Figure 1. Treatments were exponentially distributed, and the thresholds displayed in the grid controlled the center of the second sigmoid in S^2 due to Taleb [2018]. Two different visible attributes demonstrate how the hidden bias depends on the interplay between propensity and outcome, via the hidden attribute. The blue curve is a little shorter, which allows the vulnerable subgroup's threshold change to be revealed in the data. Estimation minimized the empirical squared error.

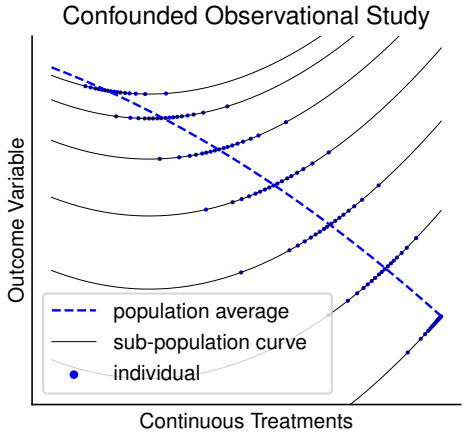

Figure 9. A different example that shows the connection to Simpson's paradox more clearly [Yule, 1903, Simpson, 1951]. When a confounder is distorting the assigned treatments in sub-populations, the overall population-level trend may appear flipped in comparison to each sub-population's dose response.

E DETAILS ON THE BENCHMARK

During each trial, 750 train and 250 test instances of (observed/hidden) confounders, treatment, and outcome were generated. The APO was computed on the test instances. Coverage of the dose-response curve was assessed on a treatment grid of 100 evenly spaced points in $[0, 1]$. The different violation factors Γ that were tested were also from a 100-sized grid in $[0, 2.5]$.

The data-generating process constructed vectors

$$V := \langle \text{visible conf.} \ldots, \text{treatment, hidden conf.} \ldots \rangle \in \mathbb{R}^k$$

where k is the number of confounders plus one, for the treatment. Each of these variables is a projection of the original data with i.i.d normal coefficients. We upscale the middle (i.e. treatment) entry by $(k-1)$ to keep the treatment effect strong enough. Then, we experiment with two functional forms of confounded dose-response curves:

- (linear) mixing vector $\{M_i\}_{i=1}^k \sim$ i.i.d Normal$(0, 1)$. Pre-activation outcome is $u := M \cdot v$.
- (quadratic) matrix $\{M_{ij}\} \sim$ i.i.d Normal$(0, 1)$. Pre-activation outcome is $u := v^{\mathrm{T}} M v$. Unlike a covariance, M is not positive (semi-)definite. The fact that all entries are i.i.d Gaussian implies that there are cases where the off-diagonal entries are much larger in magnitude than the on-diagonal entries, in such a way that cannot occur in a covariance matrix. This induces more confounding and strengthens our benchmark.

The actual outcome is Bernoulli with probability $u^\star := \phi\big((u-m)/s\big)$, wherein ϕ is the standard normal CDF, location parameter m is the sample median, and scale s is the sample mean absolute deviation from the median. If u were normal, s would be expected to be a bit smaller than σ, by a factor of $\sqrt{2/\pi}$. Generally $u^\star$ is no longer uniformly distributed (on margin) because we use s, and instead it gravitates towards zero or one. Since the estimated outcome models use logistic sigmoid activations, there is already an intentional measure of model mismatch present in this setup.

See Table 4 for results under all the settings considered.

The linear outcome and propensity predictors were estimated by maximum likelihood using the ADAM gradient-descent optimizer, with learning rate 10^1, 4 batches, and 50 epochs throughout. For the outcome, we used a sigmoid activation stretched horizontally by 10^2 for smooth training. For the propensity, similarly, we stretched a sigmoid horizontally and vertically, gating the output in order to yield Beta parameters within $(0, 10^2)$.

Data sources. The datasets `brain` and `blood` both came from the UK Biobank, which is described in the case study of §5. The two datasets are taken from disjoint subsets

of all the available fields, one pertaining to parcelized brain volumes (via MRI) and the other to blood tests. The `pbmc` dataset came from single-cell RNA sequencing, a modality that is exploding in popularity for bioinformatics. PBMC data are a commonly used benchmark in the field [Kang et al., 2018]. Finally, the `mftc` dataset consisted of BERT embeddings for morally loaded tweets [Hoover et al., 2020, Mokhberian et al., 2020].

Dataset	Sample Size	Dimension
`brain`	43,069	148
`blood`	31,811	42
`pbmc`	14,039	16
`mftc`	17,930	768

Table 3. Characteristics of the various datasets employed in our experiments.

Model mismatch varied with how approximately linear the true dose responses were. As expected, there was a significant negative correlation between model likelihood and divergence cost, so poorer fits had higher costs for coverage.

F DETAILS ON THE BIOBANK STUDY

The application number used to access data from the UK Biobank will be mentioned in the de-anonymized manuscript. The measured outcomes were cortical thicknesses and subcortical volumes, the latter normalized by intracranial volume, obtained via structural Magnetic Resonance Imaging (MRI). The results in the main text (§5) focused on the cortical thicknesses, for brevity. Input variables comprising the covariates and DQS treatments are listed in Table 5. Inputs were normalized in the unit interval, and outputs were z-scored.

Training the models. The outcome predictors with 40 inputs and 48 outputs were implemented as multilayer perceptions with three hidden layers of width 32, and single-skip connections. They used Swish activation functions and a unit dropout rate of 0.1. The ADAM optimizer with learning rate 5×10^{-3} was was run for 10^4 epochs. The data were split into four non-overlapping test sets, with separate ensembles of 16 predictors trained for each split. Training sets were bootstrap-resampled for each estimator in the ensemble. The propensity was formulated as a linear model outputting Beta parameters within $(0, 64)$, trained in a similar fashion. Finally, CAPOs were partially identified using the set of models from the train-test split for which the data instance belonged to the test set.

Additional figures. This exploratory study includes plots of relative effects on the various brain regions, shown in Figures 10 & 11. We plan on studying the differential effects of diet on the brain further.

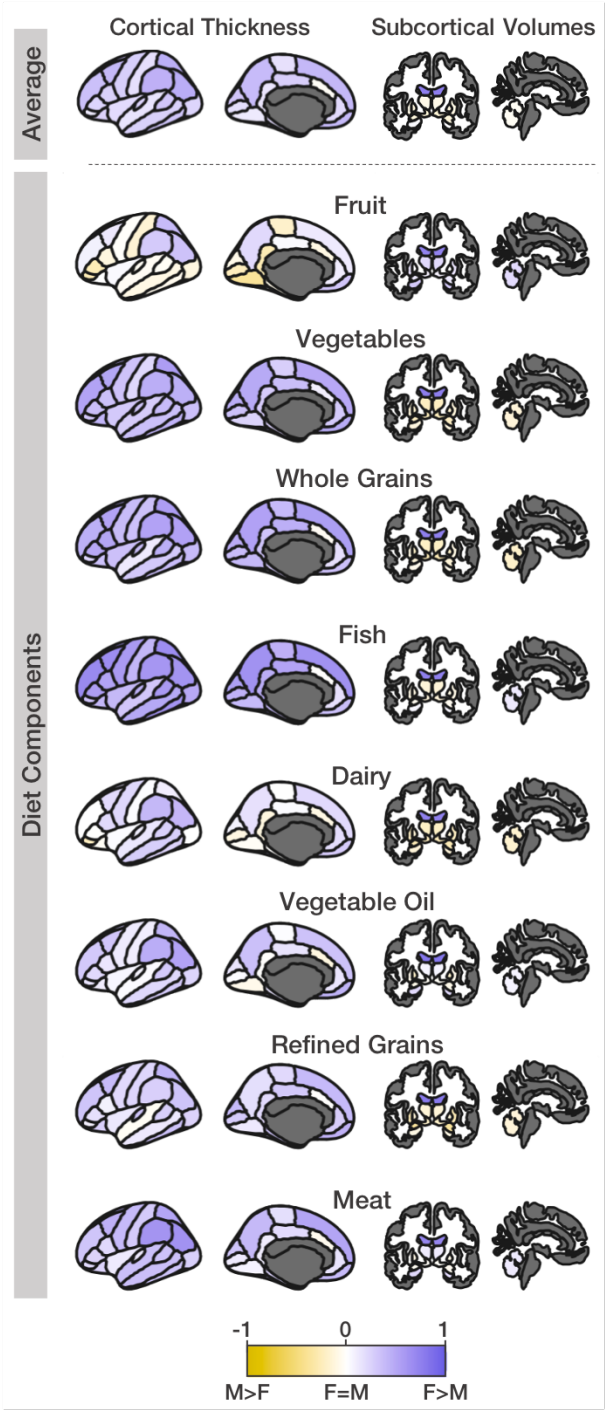

Figure 10. Normalized effect differences between males and females for the overall average diet score and stratified by individual diet components. The lefthand columns depict individual effects across all cortical thickness parcellations and the righthand side shows subcortical regional volumes. Females show generally larger effects across most diet components.

Benchmarks \ Scores			brain		blood		pbmc		mftc	
			mean	median	mean	median	mean	median	mean	median
linear	2 confounders	δMSM	**94**	**71**	**86**	**63**	**105**	**75**	**69**	**59**
		CMSM	291	253	261	228	288	259	243	204
		uniform	116	82	104	71	128	83	78	66
		binary MSM	116	90	104	73	127	94	91	73
	6 confounders	δMSM	**63**	**39**	**63**	**33**	**77**	**44**	**47**	**31**
		CMSM	177	111	186	117	198	136	167	105
		uniform	68	41	68	36	83	47	51	33
		binary MSM	177	176	173	163	188	195	168	160
	10 confounders	δMSM	**57**	**31**	**61**	**35**	**72**	**31**	**43**	**27**
		CMSM	151	81	146	84	158	84	126	74
		uniform	58	32	63	37	73	33	45	28
		binary MSM	177	181	182	190	172	170	184	191
quadratic	2 confounders	δMSM	**170**	**151**	**160**	**139**	**180**	**160**	**159**	**144**
		CMSM	301	275	283	263	299	274	270	248
		uniform	198	180	190	166	212	188	190	167
		binary MSM	205	186	192	169	217	198	190	173
	6 confounders	δMSM	**138**	**103**	**145**	**120**	**155**	**134**	**140**	**112**
		CMSM	216	171	220	193	239	223	222	198
		uniform	171	118	181	149	189	158	177	132
		binary MSM	217	231	227	257	230	266	224	249
	10 confounders	δMSM	**138**	**101**	**141**	**100**	**138**	**104**	**144**	**117**
		CMSM	186	173	188	165	205	178	182	165
		uniform	158	116	162	108	157	117	167	140
		binary MSM	211	241	213	240	222	258	214	242

Table 4. The full array of experiments. Underlined settings are those shown in Table 2.

G SOURCE-CODE AVAILABILITY

Please visit `https://github.com/marmarelis/TreatmentCurves.jl`. Also in the `scripts/` subdirectory of the supplementary source, the `synthetic.jl` file recreates the semi-synthetic benchmarks, and the `biobank.jl` file sets up the case study.

REFERENCES

J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11 (8):1057–1071, 2020.

H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94, 2018.

N. Mokhberian, A. Abeliuk, P. Cummings, and K. Lerman. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 206–219. Springer, 2020.

M. A. Said, N. Verweij, and P. van der Harst. Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the uk biobank study. *JAMA cardiology*, 3(8):693–702, 2018.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

N. N. Taleb. (anti) fragility and convex responses in medicine. In *Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems 9*, pages 299–325. Springer, 2018.

G. U. Yule. NOTES ON THE THEORY OF ASSOCIATION OF ATTRIBUTES IN STATISTICS. *Biometrika*, 2(2):121–134, 02 1903. ISSN 0006-3444. doi: 10.1093/biomet/2.2.121. URL `https://doi.org/10.1093/biomet/2.2.121`.

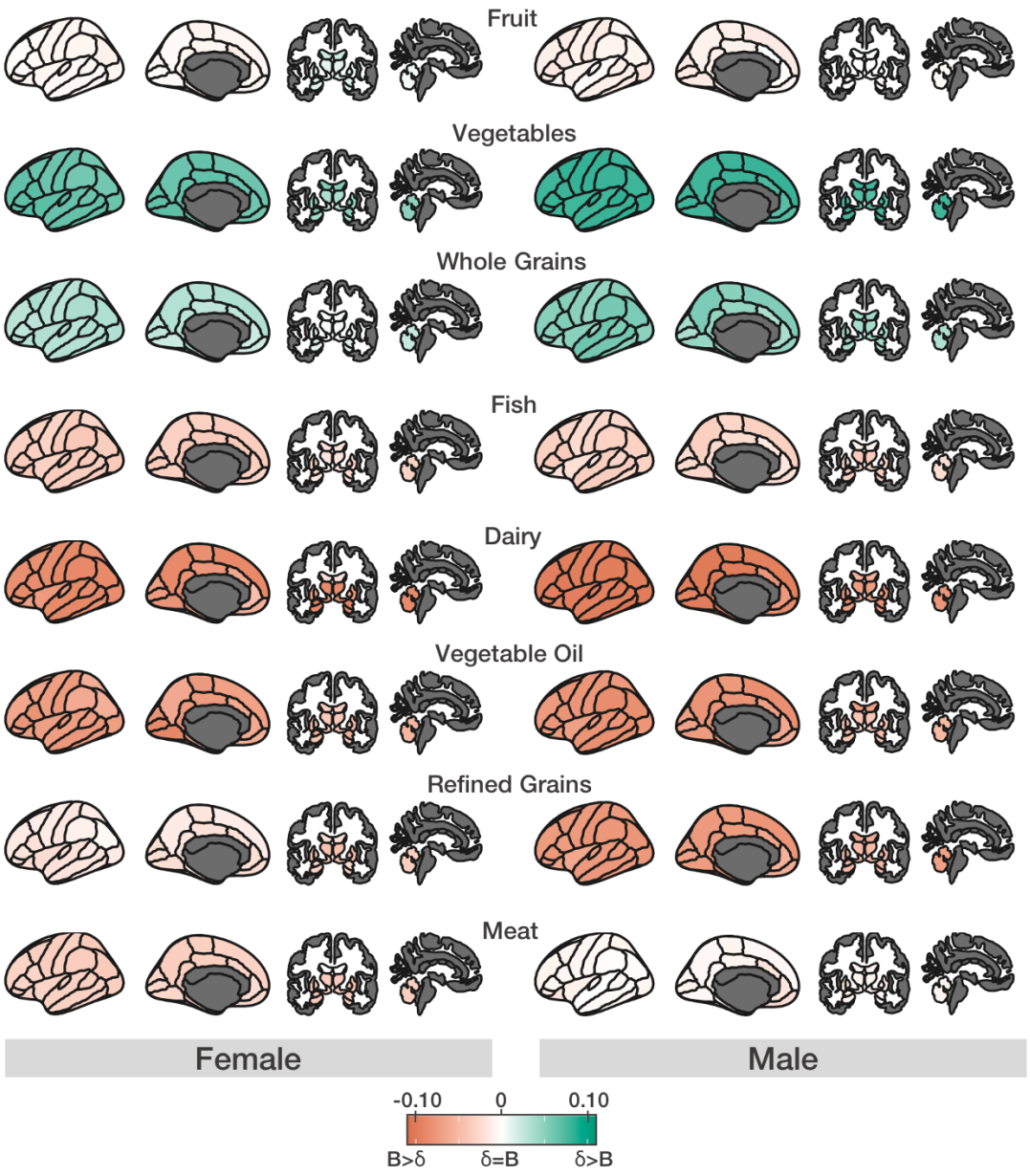

Figure 11. Normalized effect differences comparing the δMSM against a shoehorned binary MSM ("δ" vs. "B") stratified by sex. Note differences in relative feature importance, where continuous modeling ranks vegetables and whole grains to be the most important compared to the binary model which emphasizes dairy, vegetable oils, refined grains (primarily for males) and fish.

P. Zhuang, X. Liu, Y. Li, X. Wan, Y. Wu, F. Wu, Y. Zhang, and J. Jiao. Effect of diet quality and genetic predisposition on hemoglobin a1c and type 2 diabetes risk: gene-diet interaction analysis of 357,419 individuals. *Diabetes Care*, 44(11):2470–2479, 2021.

Variable	Features	Classifications	Data Field ID
Demographics	Age at scan	-	21003
	Sex	Male/Female	31
	Townsend Deprivation Index	-	189
	ApoE4 copies	0, 1, 2	-
Education	College/University	Yes/No	6138
Physical Activity/ Body Composition	American Heart Association (AHA) guidelines for weekly physical activity	Ideal ($\geq$150 min/week moderate or $\geq$75 min/wk vigorous or 150 min/week mixed); Intermediate (1–149 min/week moderate or 1–74 min/week vigorous or 1–149 min/week mixed); Poor (not performing any moderate or vigorous activity)	884, 904, 894, 914
	Waist to Hip Ratio (WHR)	-	48,49
	Normal WHR	Females: ≤ 0.85; Males ≤ 0.90	48,49
	Body Mass Index (BMI)	-	23104
	Body fat percentage	-	23099
Sleep	Sleep 7-9 Hours a Night	-	1160
	Job Involves Night Shift Work	Never/Rarely	3426
	Daytime Dozing/Sleeping	Never/Rarely	1220
Diet	DQS 1 - Fruit	-	1309, 1319
	DQS 2 - Vegetables	-	1289, 1299
	DQS 3 - Whole Grains	-	1438, 1448, 1458, 1468
	DQS 4 - Fish	-	1329, 1339
	DQS 5 - Dairy	-	1408, 1418
	DQS 6 - Vegetable Oil	-	1428, 2654, 1438
	DQS 7 - Refined Grains	-	1438, 1448, 1458, 1468
	DQS 8 - Processed Meats	-	1349, 3680
	DQS 9 - Unprocessed Meats	-	1369, 1379, 1389, 3680
	DQS 10 - Sugary Foods/Drinks	-	6144
	Water intake	Glasses/day	1528
	Tea intake	Cups/day	1488
	Coffee intake	Cups/day	1498
	Fish Oil Supplementation	Yes/No	20084
	Vitamin/Mineral Supplementation	Multivitamin (with iron/ calcium/ multimineral)/ Vitamins A, B6, B12, C, D, or E/ Folic acid/ Chromium/ Magnesium/ Selenium/ Calcium/ Iron/ Zinc/ Other vitamin	20084
	Variation in diet	Never/Rarely; Sometimes; Often	1548
	Salt added to food	Never/Rarely; Sometimes; Usually; Always	1478
Smoking	Smoking status	Never; Previous; Current	20116
Alcohol	Alcohol Frequency	Infrequent (1–3 times a month, special occasions only, or never); Occasional (1–2 a week or 3–4 times a week), Frequent (daily/almost daily and ICD conditions F10, G312, G621, I426, K292, K70, K860, T510)	1558/ICD
Social Support	Leisure/social activities	Sports club/gym; pub/social; social/religious; social/adult education; other social group	6160
	Frequency of Friends/Family Visits	Twice/week or more	1031
	Able to Confide in Someone	Almost Daily	2110

Table 5. Variables, features, classifications, and respective data fields use in the models. Diet quality scores (DQS) ranging from 0–10 for 10 components were computed using the same coding scheme as in Said et al. [2018], Zhuang et al. [2021]. Leisure/social activity classifications served as their own binary variables. Our results omitted DQS #8 & #10 because they were not even approximately continuous, taking on only a few discrete values.