Research Article

Arman Oganisian*, Nandita Mitra, and Jason A. Roy

Supplement for Hierarchical Bayesian Bootstrap for Heterogenous Treatment Effect Estimation

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

1 Identification of HTE

Here we identify the HTE in the point-treatment setting discussed in the paper. Recall the HTE is the average treatment effect within stratum v, $\Psi(v) = E[Y^1 \mid V = v] - E[Y^0 \mid V = v]$. Consider the term $E[Y^a \mid V = v]$ and now iterate expectation over W:

$$E[Y^a \mid V = v] = \int\limits_{\mathcal{W}} E[Y^a \mid w, V = v] dP_v(w)$$

Now we assume conditional ignorability. Specifically that within stratum v, once we condition on confounders W, treatment assignment is independent of potential outcome, $Y^a \perp A \mid W, V = v$. This implies that $E[Y^a \mid w, V = v] = E[Y^a \mid A = a, w, V = v]$,

$$E[Y^a \mid V = v] = \int\limits_{\mathcal{W}} E[Y^a \mid A = a, w, V = v] dP_v(w)$$

Now, we assume consistency. That is, the outcome actually observed under treatment assignment $A = a$ actually equals the outcome that would occur under treatment $A = a$, i.e. $Y^a = Y$. This would be violated if, for instance, there is non-adherence to treatment assignment. This yields,

$$E[Y^a \mid V = v] = \int\limits_{\mathcal{W}} E[Y \mid A = a, w, V = v] dP_v(w)$$

So we have identified each term of $\Psi(v)$ as a regression averaged over $P_v(w) = P(w \mid V = v)$. Note that we implicitly make a positivity and non-adherence assumption. By conditioning on $A = a$ within W and V, we are assuming that treatment probability is bounded $0 < P(A = 1 \mid w, v) < 1$ or else we would be conditioning on a zero-probability even. Causally, it would suggest that there is some level and W within stratum V where we only observed patients assigned to one of the two treatments. We cannot estimate a causal contrast between the two groups in this region of the data without (likely incorrect) extrapolation. Moreover, for a particular sample we have assumed that each subjects potential outcome $Y_i^{a_i}$ is unaffected by others' treatment assignment. If subject j's treatment assignment impacts subject i's potential outcome, then we would have had to index the potential outcome with this treatment as well, $Y_i^{a_i, a_j}$.

***Corresponding author: Arman Oganisian,** Brown University, Department of Biostatistics, Providence, RI, USA, e-mail: arman_oganisian@brown.edu
Nandita Mitra, University of Pennsylvania, Department of Biostatistics, Epidemiology, and Informatics, Philadelphia, PA, USA
Jason A. Roy, Rutgers University, Department of Biostatistics and Epidemiology, Piscataway, NJ, USA

2 Posterior Derivations

Here we provide a derivation of the posterior distribution of each P_v using Dirichlet Distributions - the finite-dimensional analogue of the Dirichlet Process. This is to supplement the conjugacy results used in the main text. Suppose our model for the conditional covariate distribution, $P_v(W) = P(W \mid V = v)$, is

$$P_v(W \mid \pi^v) = \sum_{i=1}^{n} \pi_i^v \cdot \delta_{W_i}(W)$$

We have K such distributions for each of the K levels of V. Consider the Dirichlet prior on each $\pi^v = (\pi_1^v, \pi_2^v, \dots, \pi_n^v)$ conditional on the $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ and α

$$\pi^v \mid \pi, \alpha_v \sim Dir(\alpha_v \pi)$$

Now place Dirichlet hyperprior on π:

$$\pi \mid \gamma \sim Dir(\gamma 1_n)$$

Note that the HBB corresponds to setting $\gamma = 0$ and that α_v is user-specified but we will leave γ as it is for now. So the joint posterior is

$$p(\pi^1, \pi^2, \dots \pi^K, \pi \mid \alpha_v, \gamma, W, V) \propto \Big\{ \prod_{v=1}^{K} \frac{\Gamma(\sum_{i=1}^{n} \alpha_v \pi_i)}{\prod_{i=1}^{n} \Gamma(\alpha_v \pi_i)} \prod_{i=1}^{n} (\pi_i^v)^{\alpha_v \pi_i + \delta_v(V_i) - 1} \Big\} p(\pi \mid \gamma) \tag{1}$$

The objective is to sample the π^v. To do this, we sample from the joint and simply ignore draws of π. Note that the joint can be expressed as a marginal posterior for π and independent conditional posteriors for π^v

$$p(\pi^1, \pi^2, \dots \pi^K, \pi \mid \alpha_v, \gamma, W, V) = \{ \prod_{v=1}^{K} p(\pi^v \mid \pi, \alpha_v, \gamma, W, V) \} p(\pi \mid \alpha_v, \gamma, W, V)$$

Thus to sample from the joint, we first sample π from the marginal posterior. Then conditional on π, we can sample the π^v independently. These are exactly Step 1 and 2, respectively, in the algorithm of Section 3.1. We now derive this marginal posterior and then turn to the conditional posteriors of π^v. To get the marginal, integrate out each of the π^v in (1)

$$\begin{aligned} p(\pi \mid \alpha_v, \gamma, W, V) &\propto \Big\{ \prod_{v=1}^{K} \int_{\Pi_v} \frac{\Gamma(\sum_{i=1}^{n} \alpha_v \pi_i)}{\prod_{i=1}^{n} \Gamma(\alpha_v \pi_i)} \prod_{i=1}^{n} (\pi_i^v)^{\alpha_v \pi_i + \delta_v(V_i) - 1} d\pi^v \Big\} p(\pi \mid \gamma) \\ &\propto \Big\{ \prod_{v=1}^{K} \frac{\Gamma(\alpha_v)}{\prod_{i=1}^{n} \Gamma(\alpha_v \pi_i)} \frac{\prod_{i \in S_v}^{n} \Gamma(\alpha_v \pi_i + 1) \prod_{i \notin S_v}^{n} \Gamma(\alpha_v \pi_i)}{\Gamma(\alpha_v + n_v)} \Big\} p(\pi \mid \gamma) \end{aligned}$$

Above, Π_v is the n-dimensional simplex we integrate over. This result follows because the integral is over the kernel of a Dirichlet distribution, with concentration parameter vector $\alpha_v \pi_i + \delta_v(V_i)$ and recognizing that $\sum_{i=1}^{n} \alpha_v \pi_i = \alpha_v$ since π_i sum to 1. Continuing the derivation, we cancel like terms from the numerator and denominators and note that $\Gamma(\alpha_v \pi_i + 1) = \alpha_v \pi_i \Gamma(\alpha_v \pi_i)$. Therefore, $\frac{\Gamma(\alpha_v \pi_i + 1)}{\Gamma(\alpha_v \pi_i)} = \alpha_v \pi_i$ and we have

$$p(\pi \mid \alpha_v, \gamma, W, V) \propto \Big\{ \prod_{v=1}^{K} \frac{\Gamma(\alpha_v) \alpha_v^{n_v}}{\Gamma(\alpha_v + n_v)} \Big\} (\prod_{i=1}^{n} \pi_i) p(\pi \mid \gamma)$$

Now, note that in the last line the term in brackets is constant with respect to π, so we can eliminate it and maintain proportionality. Then, substituting the prior $p(\pi \mid \gamma = 0) = Dir(0_n) \propto \prod_{i=1}^{n} \pi_i^{-1}$,

$$p(\pi \mid \alpha_v, \gamma, W, V) \propto (\prod_{i=1}^{n} \pi_i) \prod_{i=1}^{n} \pi_i^{-1} \propto \prod_{i=1}^{n} \pi_i^{1-1}$$

This is the kernel of $Dir(1_n)$ - the posterior of Rubin's bootstrap. Thus, to draw from this marginal posterior, we can draw $\pi \sim Dir(1_n)$. This is the distribution we sample from in Step 1 of the algorithm in Section 3.1.

Now, the conditional posterior of each π^v conditional on π is much simpler. Just absorb all terms not involving π_i^vin (1) into the proportionality constant and we have

$$p(\pi^v \mid \pi, \alpha_v, \gamma, W, V) \propto \prod_{i=1}^{n} (\pi_i^v)^{\alpha_v \pi_i + \delta_v(V_i) - 1}$$

Which is proportional to a $\pi^v \sim Dir\Big(\alpha_v\pi_1 + \delta_v(V_1), \alpha_v\pi_2 + \delta_v(V_2), \dots, \alpha_v\pi_n + \delta_v(V_n)\Big)$. This is the distribution we sample from in Step 2 of the algorithm in Section 3.1.

3 Simulation Details

3.1 Main manuscript simulation

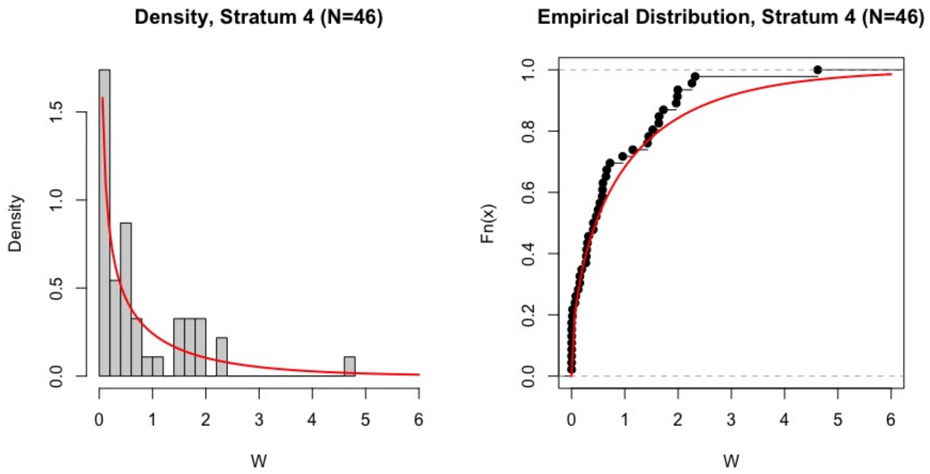

Fig. 1: histogram of the observed covariate values in stratum 4 (the sparse stratum) from a single simulation run in the gamma setting. The red line shows the true gamma density we simulated from. Values of $W \in [2.32, 4.62]$ are plausible - under the true gamma density, there is $\approx 10\%$ probability on this interval. However, we observe no data in this interval within stratum 4 due to the small sample. Unlike the empirical distribution, the HBB borrows points in this interval that appear in the other strata.

Here we provide details for the simulation study in Section 4 of the main manuscript. In each setting, we simulate 1000 data sets with $n = 300$ subjects as follows. For $i = 1, \dots, 300$

1. Simulate stratum allocation:
$$V_i \sim Multinom(1; \frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10})$$
The parameter vectors gives the probability of assignment to stratum 1, 2, 3, and 4, respectively.
2. Simulate 10-dimensional confounder vector $W_i = (W_i^p)_{p=1:10}$,
$$W_i \mid V_i = v \sim p(W \mid V = v)$$
The form of $p(W \mid V = v)$ varies with simulation setting and is specified below.
3. Simulate treatment assignment, A_i, from Bernoulli with probability
$$P(A = 1 \mid W_i, V_i = v) = expit(\eta_v + W_i'\beta)$$

4. Simulate binary outcome, Y_i, from a Bernoulli with probability

$$P(Y=1 \mid W_i, V_i = v) = expit(-1 + \gamma_v + W_i'\theta + a_v A_i)$$

Note in the above that W_i impacts both treatment assignment (via β) and outcome (via θ) - so it is a confounder. Similarly, V_i impacts both treatment assignment (via η_v) and outcome (via γ_v). Note that the conditional treatment effect, a_v, varies across stratum - so this is a complex scenario with treatment effect heterogeneity across strata. This yields a simulated data set $\{Y_i, A_i, W_i, V_i\}_{i=1:n}$. We simulate 1000 such data sets across four settings. Figure 1 provides an illustration of the distribution of W within stratum 4.

The covariate distribution $p(W \mid V)$ has a different family governed by different parameters in each of the two settings:

1. $W_i^p \mid V = v \sim N(\mu_v, 1)$ where $\mu_v \in \{-2, 0, 2, 4\}$ for $v = 1, \dots 4$, respecting order. Marginal of V, the distribution of W is a location mixture of normals.
2. $W_i^p \mid V = v \sim Gam(shape = \frac{1}{2}\tau_v, rate = \frac{1}{2})$. Here $\tau_v \in \{8, 6, 4, 1\}$ for $v = 1, \dots 4$, respecting order.

Both settings share these simulation parameters:

- Set $\beta = \theta = (1, -1, 1, -1, 1, -1, 1, -1, 1, -1)$.
- Set $\eta_v \in (0, -.5, .5, .5)$ for $v = 1, \dots, 4$ in order.
- $\gamma_v \in (-.1, -.5, .1, .5)$ for $v = 1, \dots, 4$ in order.
- $a_v \in (1, -1.5, 1, 1.5)$ for $v = 1, \dots, 4$ in order.

Using each simulated dataset, we specify the following logistic regression

$$P(Y \mid A, W, V = v) = expit\Big(\omega_0 + \omega_v + W'\omega_W + \omega_v^* A\Big)$$

Normal priors with mean zero and standard deviation 3 were placed on each parameter. We obtain $M = 5000$ posterior samples $\{\omega_0, \omega_1^{(m)}, \dots, \omega_4^{(m)}, \omega_W^{(m)}, \omega_1^{*(m)}, \dots, \omega_4^{*(m)}\}_{m=1:M}$ after discarding the first 5000 draws as burn-in. Sampling was done via hamiltonian monte carlo as implemented in Stan. These samples were combined with HBB as described in Section 3.1.

3.2 Additional simulation results exploring interactions

In addition to the simulation results above, we ran another set exploring the impact of strong interactions between W and A within each stratum of V. We specify a logistic outcome model as in the previous simulation, but now with interactions terms between A and W included. We run 1000 simulations with $N = 500$ each and report the results in Table 1. Note that on average across simulations, stratum 4 has $(1/10) * 500 = 50$ subjects to estimate an outcome model with 22 (intercept, 10 W main effect coefficients, a main treatment effect coefficient, and 10 $W - A$ interaction coefficients) parameters. So this is rather severe sparsity setting. Focusing on the performance in this sparse stratum, we see that the causal effect estimate using HBB outperforms the other confounder distribution estimates in terms of MSE, however the 95|% posterior interval has higher 95% frequentist coverage.

The HBB borrows more information about W from other strata, thus the outcome model extrapolates the - leading to significant posterior uncertainty. This is reflected in the much wider posterior interval width of the HBB in each of the settings. On the flip side, the empirical and BB estimates yield 95% intervals with undercoverage. Because insufficient information about W exists, our estimate doesn't capture the full range across which W modifies the treatment effect A in stratum 4. Thus, producing intervals that are too narrow.

Tab. 1: Additional Simulation results: Relative (Rel.) MSE, absolute bias, empirical variance of the posterior mean along with the width and coverage of the 95% credible interval across 1,000 simulation runs. MSE is computed as average of the squared difference between posterior mean and truth across simulations. Empirical variance is computed as the variance of the 1,000 posterior mean causal effect estimates. Data were generated as described in the main text, but with interactions between treatment and confounders within strata.

		Gaussian Mixture					Gamma Mixture				
	Model	Rel. MSE	Bias	Var.	Width	Cov.	Rel. MSE	Bias.	Var.	Width	Cov.
Stratum 1	Emp.	0.95	0.000	0.003	0.188	0.934	0.80	0.011	0.003	0.195	0.904
	BB	0.95	0.000	0.003	0.197	0.948	0.80	0.011	0.003	0.205	0.918
	HBB	1	0.005	0.003	0.241	0.981	1	0.015	0.004	0.240	0.921
	Oracle	0.96	0.002	0.003	0.190	0.929	0.96	0.002	0.004	0.247	0.94
Stratum 4	Emp.	1.84	0.013	0.010	0.346	0.925	1.80	0.080	0.014	0.426	0.878
	BB	1.84	0.013	0.010	0.368	0.931	1.80	0.080	0.014	0.434	0.881
	HBB	1	0.025	0.005	0.534	1	1	0.021	0.011	0.460	0.972
	Oracle	1.91	0.012	0.010	0.351	0.923	0.94	0.008	0.010	0.452	0.974

3.3 Additional simulation results exploring homogenous confounder distributions

In additional to the simulations in the main text described in Section 3.1, we ran an experiment that considered a scenario in which the true confounder distribution is the same across strata. Specifically $P_v(W) = N_{10}(\bar{0}_{10}, I_{10})$ for each v. All other settings are the same as described in Section 3.1 and the results are described in Table 2.

Tab. 2: Additional Simulation results: exploring homogenous $N_{10}(\bar{0}_{1}0, I_{10})$ confounder distributions across strata. Results reported across 1,000 simulation runs.

		Standard normal covariates in each stratum				
	Model	Rel. MSE	Bias	Var.	Width	Cov.
Stratum 4	Emprical	1.05	0.005	0.014	0.46	0.955
	BB	1.05	0.005	0.014	0.48	0.956
	HBB	1	0.002	0.013	0.46	0.949
	Oracle	0.99	0.002	0.013	0.46	0.949

This setting is a favorable for HBB relative to separate estimation because the HBB's partial-pooling leverages more information from the other strata and - because all the stratum-specific distributions are the same - this does not come at the expense of additional finite-sample bias. All methods perform similarly since the confounder distribution is the same across strata. HBB has slightly lower finite-sample bias and more precise (it produces an interval with 95% coverage, but with a narrower interval on average than BB).

4 Data Analysis Details

Here we provide additional details about the data analysis in the main text. In the parametric Poisson model, we include the following covariates for each stratum except gynecological cancer.

- treatment: binary with one indicating proton.
- race: categorical with levels white, black, and other.
- sex: binary with one indicating male.
- insurance: categorical with levels medicare, private, and other.
- body-mass index: normalized.
- age: normalized
- charlson index: logged.

Tab. 3: Summary statistics of covariates across cancer strata. Sample average and standard deviation - $N, (\%)$ - are reported for continuous covariates - $avg.(sd)$. Count and proportions are reported for categorical covariates. The abbreviations are gynecological (gyn), pancreas/duodenum/hepatobiliary (p/d/h), esophagus/gastric (e/g), and head/neck (h&n).

	Cancer Type							
	Anal	**Brain**	**E/G**	**Gyn**	**H&N**	**Lung**	**P/D/H**	**Rectum**
Variable	**N=80**	**N=231**	**N=148**	**N=34**	**N=435**	**N=325**	**N=91**	**N=124**
Male	**29 (36.2)**	**140 (60.6)**	**120 (81.1)**	**0 (0.0)**	**336 (77.2)**	**167 (51.4)**	**54 (59.3)**	**78 (62.9)**
Race								
Black	**15 (18.8)**	**14 (6.1)**	**13 (8.8)**	**18 (52.9)**	**54 (12.4)**	**82 (25.2)**	**24 (26.4)**	**29 (23.4)**
White	**64 (80.0)**	**209 (90.5)**	**131 (88.5)**	**16 (47.1)**	**362 (83.2)**	**224 (68.9)**	**64 (70.3)**	**89 (71.8)**
Other	**1 (1.2)**	**8 (3.5)**	**4 (2.7)**	**0 (0.0)**	**19 (4.4)**	**19 (5.8)**	**3 (3.3)**	**6 (4.8)**
Insurance								
Medicare	**21 (26.2)**	**59 (25.5)**	**80 (54.1)**	**9 (26.5)**	**103 (23.7)**	**155 (47.7)**	**39 (42.9)**	**40 (32.3)**
Private	**55 (68.8)**	**166 (71.9)**	**63 (42.6)**	**21 (61.8)**	**321 (73.8)**	**161 (49.5)**	**49 (53.8)**	**78 (62.9)**
Other	**4 (5.0)**	**6 (2.6)**	**5 (3.4)**	**4 (11.8)**	**11 (2.5)**	**9 (2.8)**	**3 (3.3)**	**6 (4.8)**
Age	**58 (9.3)**	**56 (15.0)**	**66 (12.0)**	**56 (14.4)**	**59 (10.0)**	**66 (10.3)**	**66 (9.4)**	**59 (13.7)**
CCI	**3.80 (2.84)**	**2.70 (1.06)**	**3.24 (1.45)**	**2.47 (1.11)**	**2.72 (1.28)**	**3.42 (1.70)**	**3.36 (1.43)**	**2.77 (1.11)**

Summary statistics are given in Table 3. For gynecological cancer, there is no need to adjust for sex. We specify$N(0, 1)$ priors on all covariates except in the following instances: in the models for E/G, brain, anal, and rectum, we use tighter $N(0, .1)$ priors on the other race coefficient. Similarly, for the P/D/H model we use a $N(0, .1)$ prior on other insurance. The tight priors are to regularize coefficients that explode due too little variation in insurance status or race in a particular stratum. Non-bayesian analyses typically omit such variables (equivalent to a prior that the coefficient is exactly 0), but we choose to include them with a tight prior around 0 as a compromise. Note that the $N(0, 1)$ prior may seem overly informative, but on the log scale it is quite flat. It puts sufficient volume at incident rate ratios within $\exp(\pm 1.96)$ or within $(.14, 7.1)$.

For posterior sampling, we use hamiltonian monte carlo as implemented in Stan. We call Stan in R using the rstan package. For inference, we retain 10000 posterior draws after a 10000 burn-in. After obtaining these draws, we use HBB as described in Section 3.1.

For the BART model, we adjust for all of the same covariates. Draws of f_v under particular treatments were obtained using the BayesTree R package. We retain 1000 posterior draws for inference after discarding the first 1000 as burn-in. For the BART hyperpriors, we increase the power parameter from the default of 2 to 3. This is to favors more shallow trees which provides more regularization. After draws of f_v are obtained, we combine with HBB draws as described in Section 3.1.

Finally, we note that the effects in the gynecological cancer model, in particular, is highly variable. As there were only 4 subjects treated with proton therapy in this stratum and none of the four had events, this coefficient is not identifiable with data. This is manifest in the large interval in both the Poisson and BART models.

In order to more directly compare the results of the parametric Poisson model and the nonparametric BART model, we compute the implied probability of at least one adverse event under the Poisson model. Specifically, recall that the parametric Poisson model is fit to outcome Y, the count of adverse events. BART is fit to binary outcome $\tilde{Y} = I(Y > 0)$. Thus we can equivalently express the causal odds

$$\Psi(v) = \frac{E[\tilde{Y}^1 \mid V = v]/(1 - E[\tilde{Y}^1 \mid V = v])}{E[\tilde{Y}^0 \mid V = v]/(1 - E[\tilde{Y}^0 \mid V = v])} = \frac{P[Y^1 > 0 \mid V = v]/(1 - P[Y^1 > 0 \mid V = v])}{P[\tilde{Y}^0 > 0 \mid V = v]/(1 - P[\tilde{Y}^0 > 0 \mid V = v])}$$

So, it is possible to compare the results by post-processing the posterior draws of the Poisson model to compute the posterior probability of at least 1 event rather than the expected count - without need to refit the model. The results are displayed in Figure 2 - where the BART results from the main manuscript are reproduced in the right panel and the Poisson results are in the left panel.

The estimates are generally consistent between the models, except the parametric model tends to yield narrower intervals due to the smoother model. BART tends to have slightly wider intervals (e.g. in the p/d/h stratum) - perhaps consistent with the bias-variance trade off that comes with parametric and

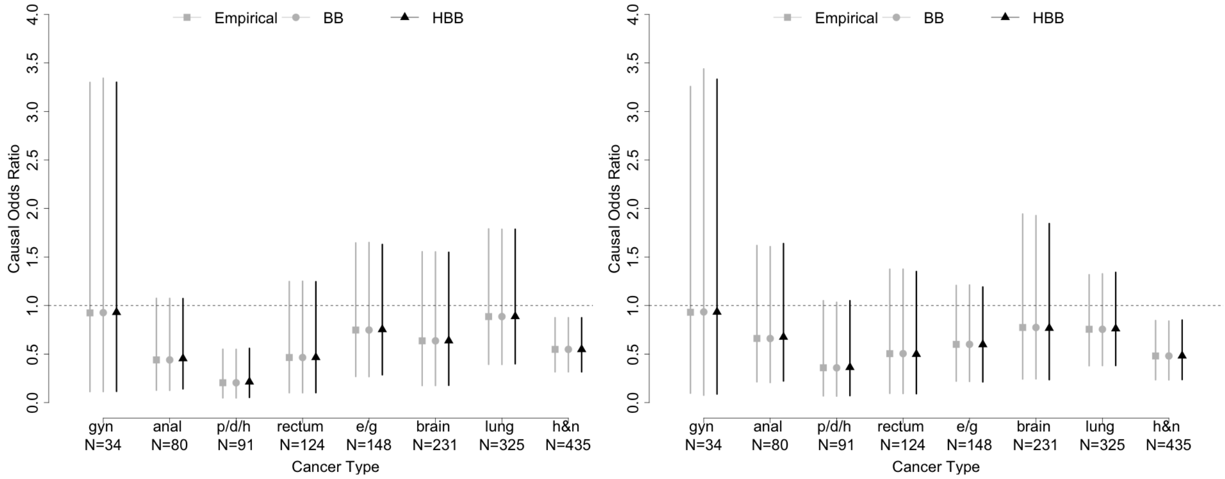

Fig. 2: Posterior mean and 95% credible interval estimates of stratum-specific causal odds ratio of at least one adverse event under Poisson model (left) and BART (right).

nonparametric models in general.

As an additional sensitivity, we have also repeated the Poisson analysis for several values of $M = 0$ to $M = 435$ (the size of the largest observed stratum). The idea behind this stems from the fact that the HBB is a compromise between two extremes. With $M = 0$, the HBB reduces to the posterior mean of the stratum-specific BB where no information is borrowed. On the other extreme, as M gets large, we shrink completely to the overall empirical distribution across all strata. So, for the same outcome model, the range of possible results as we toggle M is determined by the discrepancy between these two distributions - by construction the HBB will yield an answer "between" these two.

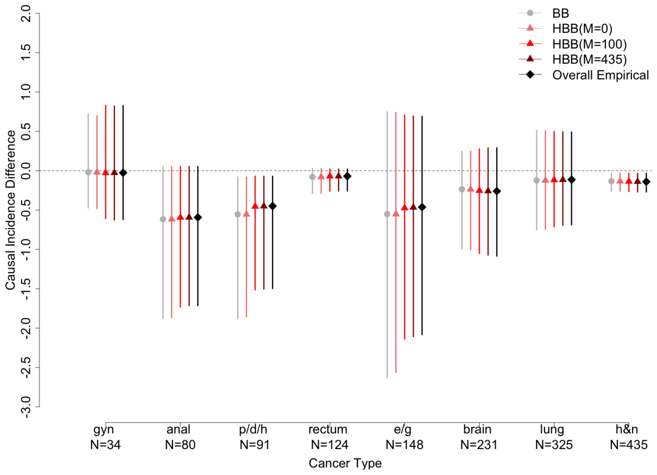

Fig. 3: Posterior mean results for the Poisson data analysis for a range of M.

A sensitivity analysis could be done by computing the causal effect posterior obtained under both the BB and the overall empirical distribution. If they very different, then changing M won't change results too much. If they differ greatly, then increasing M can be impactful.

In Figure 3 we have plotted the results of the Poisson model analysis of the photon-proton data in Figure 3 of the manuscript. In the paper we reported results with HBB setting $M = 100$ - but here we show results of the HBB under different M. Note that as we increase M, results move from the BB estimate at one extreme to the overall empirical at the other. These two extremes are not that different across strata and so we don't expect M to make a huge difference in this particular analysis. At the same time, this also acts as a sensitivity for the usual BB: if the BB results were very different from results based on the overall empirical, then the analyst may want to consider whether some partial pooling is necessary and may want to justify why they made the very informative prior decision to not pool if it would have made a difference. On the other hand, if the difference is small, then they may feel more comfortable with the BB.

5 Hyperparameter Updating

In the manuscript, we propose setting α_v empirically. Here we illustrate how one could take a fully Bayesian approach and update $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_v)$ under some prior $g_\alpha(\alpha; \nu)$, where ν are the prior hyperparameters. The goal is not to sample from the joint posterior

$$p(\pi^1, \pi^2, \dots \pi^K, \pi, \alpha \mid W, V) = p(\pi^1, \pi^2, \dots \pi^K, \pi \mid \alpha, W, V) \cdot p(\alpha \mid W, V)$$

Thus provided, we can find the marginal $p(\alpha \mid W, V)$, then we can sample in two steps. Update α by drawing from the marginal posterior. Then, conditional on draws of α and the data we draw from the conditional of the weights as illustrated in the main manuscript. Under the HBB, the data model is $P_v(W \mid \pi^v) = \sum_{i=1}^n \pi_i^v \cdot \delta_{W_i}(W)$, where $\pi^v \mid \pi, \alpha_v \sim Dir(\alpha_v \pi)$ and $\pi \mid\sim Dir(0_n)$ and $\alpha \sim g_\alpha(\alpha; \nu)$. So, the marginal posterior is obtained by integrating the weights over the $n-$dimensional simplex Π^v,

$$\begin{aligned}
p(\alpha \mid W, V) &= \int\limits_{\Pi} \int\limits_{\Pi^1} \int\limits_{\Pi^2} \cdots \int\limits_{\Pi^K} p(\pi^1, \pi^2, \dots \pi^K, \pi, \alpha \mid W, V) d\pi^K \dots, d\pi^2, d\pi^1 d\pi \\
&\propto \int\limits_{\Pi} \Big\{ \prod_{v=1}^K \int\limits_{\Pi^v} \frac{\Gamma(\sum_{i=1}^n \alpha_v \pi_i)}{\prod_{i=1}^n \Gamma(\alpha_v \pi_i)} \prod_{i=1}^n (\pi_i^v)^{\alpha_v \pi_i + \delta_v(V_i) - 1} d\pi^v \Big\} \prod_{i=1}^n \pi_i^{-1} d\pi \; g_\alpha(\alpha; \nu) \\
&\propto \int\limits_{\Pi} \Big\{ \prod_{v=1}^K \frac{\Gamma(\alpha_v)}{\prod_{i=1}^n \Gamma(\alpha_v \pi_i)} \frac{\prod_{i \in S_v}^n \Gamma(\alpha_v \pi_i + 1) \prod_{i \notin S_v}^n \Gamma(\alpha_v \pi_i)}{\Gamma(\alpha_v + n_v)} \Big\} \prod_{i=1}^n \pi_i^{-1} d\pi \; g_\alpha(\alpha; \nu)
\end{aligned}$$

This second line follows from apply Bayes' rule and substituting the forms of the models and the third line follows because each integral is over the kernel of a Dirichlet distribution, with concentration parameter vector comprised of the $\alpha_v \pi_i + \delta_v(V_i)$. When further simplified we have,

$$\begin{aligned}
p(\alpha \mid W, V) &\propto \int\limits_{\Pi} \Big\{ \prod_{v=1}^K \frac{\Gamma(\alpha_v) \alpha_v^{n_v}}{\Gamma(\alpha_v + n_v)} \Big\} (\prod_{i=1}^n \pi_i) \prod_{i=1}^n \pi_i^{-1} d\pi \; g_\alpha(\alpha; \nu) \\
&\propto \Big\{ \prod_{v=1}^K \frac{\Gamma(\alpha_v) \alpha_v^{n_v}}{\Gamma(\alpha_v + n_v)} \Big\} g_\alpha(\alpha; \nu)
\end{aligned}$$

This result also appears in Equation 10 of Escobar and West (1995) in the context of the Dirichlet Processes. One can obtain posterior samples from this distribution using the data augmentation scheme discussed by Escobar and West if g_α is a product of independent gamma densities (and ν being the collection of shapes and scales) or a Metropolis-Hastings step otherwise. However, note that the marginal posterior only depends on the data through the sample size n_v - which we use to set α_v empirically.