Using leave-one-out cross validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale

Swen Kuh1,2 | Lauren Kennedy1,2 | Qixuan Chen3 | Andrew Gelman4

1 School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide, Australia
2 Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia
3 Department of Biostatistics, Columbia University, New York, New York, USA
4 Department of Statistics and Political Science, Columbia University, New York, New York, USA

Correspondence
Swen Kuh, School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide, Australia.
Email: swen.kuh@adelaide.edu.au

Funding information
National Institutes of Health,
Grant/Award Number: 5R01AG067149–02

In recent decades, multilevel regression and poststratification (MRP) has surged in popularity for population inference. However, the validity of the estimates can depend on details of the model, and there is currently little research on validation. We explore how leave-one-out cross validation (LOO) can be used to compare Bayesian models for MRP. We investigate two approximate calculations of LOO: Pareto smoothed importance sampling (PSIS-LOO) and a survey-weighted alternative (WTD-PSIS-LOO). Using two simulation designs, we examine how accurately these two criteria recover the correct ordering of model goodness at predicting population and small-area estimands. Focusing first on variable selection, we find that neither PSIS-LOO nor WTD-PSIS-LOO correctly recovers the models’ order for an MRP population estimand, although both criteria correctly identify the best and worst model. When considering small-area estimation, the best model differs for different small areas, highlighting the complexity of MRP validation. When considering different priors, the models’ order seems slightly better at smaller-area levels. These findings suggest that, while not terrible, PSIS-LOO-based ranking techniques may not be suitable to evaluate MRP as a method. We suggest this is due to the aggregation stage of MRP, where individual-level prediction errors average out. We validate these results by applying to the real world National Health and Nutrition Examination Survey (NHANES) data in the United States. Altogether, these results show that PSIS-LOO-based model validation tools need to be used with caution and might not convey the full story when validating MRP as a method.

KEYWORDS
LOO, model validation, MRP, population estimand, small-area estimation

Abbreviations: AR, autoregressive; elpd, expected log predictive density; LOO, leave-one-out cross validation; MRP, multilevel regression and poststratification; PSIS, Pareto smoothed importance sampling; SAE, small-area estimation.
1 | INTRODUCTION

Multilevel regression and poststratification (MRP) relies on models, so it is essential to have a method to validate our models. MRP has experienced a surge in popularity since its first proposal.1 This method is formed through two stages: the first stage requires a regularized model, which is used to directly model the outcome that we would like to estimate in the population by poststratification variables. The second uses this model to predict this outcome into the population and then aggregate over these predictions to make population and small-area estimates. As its popularity has grown, so too have the successful application areas.2-5 However, MRP as a method is beholden to the goodness of the model. We see this in the initial criticism by Si et al6 and subsequent reanalysis by Valliant7 that the goodness of the estimand is highly dependent on the goodness of the model. However, just how do we validate the goodness of the model, and which models should be included for comparison?

In this manuscript, we focus on two types of candidate models for comparisons: different variables and different priors. With the first model comparison, we consider which variables should be included in the model. We categorize variables into different types based on their impact on the estimand and work under the framework that a model designed for an MRP method should be concise (only include the most necessary variables). The second model comparison focuses on holding the variables constant but considering different forms of regularization instead.

To compare the goodness of the models, we focus on two methods. We consider first an approximation to LOO, Pareto smoothed importance sampling (PSIS-LOO)8 as implemented in the \texttt{loo} package in R, and second a weighted alternative of this approximation (WTD-PSIS-LOO). These methods were chosen for their practical speed (as they can be used without refitting the models), and their overall popularity (suggesting a high likelihood that they are already being used to validate MRP models). However, we demonstrate that this task is inherently difficult: while PSIS-LOO recovers the order of models well for individual-level estimates, at a wider population or small-area estimates, this ordering is not maintained. Although we find some reassurance that the best model is consistently identified, we suggest that challenges around ordering suggest researchers and practitioners should proceed with caution when using LOO for MRP.

1.1 What is MRP?

MRP1,2,9 was first proposed in a political science context.5,10 Since then, it has continued to be widely applied in political science,11-13 and expanded to other fields such as health sciences,5,14 survey research,6 and most recently in inferring the prevalence and transmission of the COVID-19 pandemic.15,16 This popularity brings two salient obstacles for practitioners who seek to use MRP credibly. First, MRP does not have a single goal but rather two: population-level and small-area estimations. Second, the diverse range of applications suggests that the models will differ based on context.

For this manuscript, we distinguish between the method of MRP and the model used in MRP. The method of MRP has two stages. First, a multilevel model (the model used in MRP) is fitted to a sample with an outcome of interest and predictive categorical variables. Then, in the second stage, we poststratify the posterior quantities from the first stage to the corresponding population, resulting in a posterior distribution for \( \theta \). Usually, the model is a multilevel regression (hence the namesake multilevel regression and poststratification), but other predictive models can also be used (eg, see Bisbee17; Liu and Chen18; Ornstein19; Gao et al20; Liu et al21).

We follow the following steps using notation similar to Gelman and Hill22 as cited in Gao et al23:

1. \textbf{Multilevel regression:}

   \[
   \Pr(y_i = 1) = \logit^{-1}\left( \sum_{k=1}^{K} \alpha_{\ell[i]}^{(k)} \right) \quad \text{for } i = 1, \ldots, n, \tag{1}
   \]
   \[
   \alpha_{\ell[i]}^{(k)} | \sigma^{(k)} \sim \text{normal}(0, \sigma^{(k)}) \quad \text{for } k = 1, \ldots, K; \ \epsilon^{(k)} = 1, \ldots, L^{(k)}, \tag{2}
   \]
   \[
   \sigma^{(k)} \sim \text{normal}^+(0, 1) \quad \text{for } k = 1, \ldots, K, \tag{3}
   \]

   where \( y_i \) is the outcome for each individual \( i \), \( k \) is the \( k \)-th predictor variable for the total \( K \) number of variables, \( \epsilon^{(k)} \) is the variable level and \( L^{(k)} \) is the total number of levels for the \( k \)-th variable. We use the symbol \( \ell \) for small-area level, not to be confused with \( j \) as the corresponding number of cells on the poststratification table.

2. \textbf{Poststratification:} The predictions are made for each combination of the levels in the variables.22 The variables often represent geographic and demographic descriptors, but not necessarily (eg, age groups, income level and residential
postcode, etc.). This matrix of all the combinations of levels is called a poststratification table. These predictions are then aggregated to obtain a population level estimand, $\theta$ by using known population sizes $N_j$ (usually from census or a large-scale survey at the national level) and the cell prediction $\theta_j$ for every $j$th cell in the poststratification matrix

$$\theta := \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}. \tag{4}$$

To extend this formula to small-area estimation, we define a set $\mathcal{S}$, which represents the cells of a particular category or level of the variable defined based on the poststratification table. In limit, $\mathcal{S}$, is the entire population. For population-level estimates, the sum of all cell sizes amounts to the population size $\sum_{j \in \mathcal{S}} N_j = N$, while for small-area estimation, the sum of all cell sizes would be the population size for the particular category of a variable of interest $\sum_{j \in \mathcal{S}} N_j = N_S$. The small area goal quantity $\theta_S$ would correspond to the proportion of individuals “in the sub-population that would respond yes to the survey question of interest”.

$$\theta_S := \frac{\sum_{j \in \mathcal{S}} N_j \theta_j}{\sum_{j \in \mathcal{S}} N_j}. \tag{5}$$

The two steps of MRP are often discussed separately, but they are not independent. To poststratify, we require the population distribution of the adjustment variables in the model to be known. If not, work is required to impute these variables in the population.\textsuperscript{6,24,25} This is not always possible, and where it requires substantial work to find data and build an imputation technique. If we could correctly identify the variables that are needed in the model for the MRP method, this would mean this work is only done where it is absolutely needed.

One of the strengths of MRP is the use of regularizing models. The regularizing model is particularly useful as it enables us to create estimates of groups with relatively small samples. Regularization is powerful in that it allows greater estimate stability and precision, but it also can create bias in estimands. Regularization is a feature of the model, but how can we identify the correct amount (tightness of prior) and type (e.g., a varying effect prior instead of an autoregressive prior)? If the regularization is too strong or inappropriate for the data and the sample size or size of a particular small area is small, the model will not correct this.

### 1.2 Challenges with validating MRP models

We are not the first to think of validating MRP model in terms of predictive ability. Wang and Gelman\textsuperscript{26} discuss the use of $k$-fold cross validation for multilevel models in a survey context. They note the difficulty of identifying where a partially-pooled model is best suited for the data, as well as noting the difficulty of identifying which folds to take in cross validation. Our work builds on this existing work by focusing on validating MRP the process (not just the model), as well as the use of leave-one-out cross validation; see Section 1.3.

Although Wang and Gelman\textsuperscript{26} do not explicitly discuss MRP, their focus on electoral survey predictions and simultaneous publication of work using MRP to make electoral predictions using highly unrepresentative data (e.g., the Xbox study of Wang et al\textsuperscript{4}) suggests that validating the MRP model was perhaps their underlying goal. In addition to the challenges of validating a multilevel model, we also argue there are MRP-specific challenges. We discuss three main challenges with validating MRP models.

When we began this work we believed that the biggest challenge would be differences between sample and population data. Naturally, if we are interested in applying an MRP method, there is an implicit assumption that the sample is not a random sample, and that the probability of being included in the sample is potentially related with the quantity we would like to estimate. In this work, we combine the computationally efficient PSIS-LOO method proposed by Vehtari et al\textsuperscript{8} with the weighted cross validation proposed by Lumley and Scott\textsuperscript{27} to address this issue.

However, later in the manuscript, we will argue that there is another challenge for the validation of the MRP method. That is, it is not individual predictions that need to be good, but rather the aggregations of these individual estimates. We argue that this means the cost (or utility) function (expected log predictive density (elpd)) employed by both Vehtari et al\textsuperscript{8} and Wang and Gelman\textsuperscript{26} is better used to evaluate the goodness of individual predictions than a method that requires aggregation like MRP.

The aggregation challenge is also seen in variable selection in causal inference. This should not be surprising, Mercer et al\textsuperscript{28} argue that there are several similarities between causal inference and survey adjustment. Of particular note is the distinction between bias variables (variables that if unadjusted for result in bias in estimates of the average treatment
effect) and precision variables (variables that if unadjusted for result in less precise estimates of the average treatment
effect) and precision variables (variables that if unadjusted for result in less precise estimates of the average treatment
effect). Similar variable distinctions are demonstrated in survey adjustment by Little and Vartivarian. While in causal inference, solutions like double-lasso effectively address this, this method is less suited to an MRP context due to the lack of information known about the population.

The final challenge is that the best model for MRP depends on what the goals of the method are (population versus small-area estimation). We demonstrate differences in the best model between different small areas. Although we do not claim a comprehensive solution to all three challenges, we demonstrate the efficacy of WTD-PSIS-LOO to solve the first challenge (the lack of random sampling), and argue that the underlying cause of the second (aggregation) is the use of elpd as a cost function (and discuss alternatives in the discussion) and demonstrate that clarity of the MRP goal is essential to correctly identify the best model.

1.3 Why LOO for MRP?

There are two primary drivers for focusing this work on leave-one-out (LOO) cross validation instead of k-fold cross validation. The first is entirely pragmatic. The availability of a fast and effective approximate estimate for leave-one-out cross validation (PSIS-LOO) proposed by Vehtari et al. and made accessible through the \texttt{loo} package. This makes PSIS-LOO practically appealing, especially given the lengthy computation time for complex MRP models. The second is the ability to easily incorporate survey weights into this approximation to adjust for sample representation (WTD-PSIS-LOO).

We investigate LOO in our MRP work as the first stage of an MRP method is usually a Bayesian multilevel model to predict at the population level. LOO is often used to compare and validate predictive Bayesian models, due to its attractiveness and simplicity. On top of that, Vehtari et al. propose PSIS-LOO as a faster approximation to LOO as fitting complex models such as a Bayesian multilevel model is computationally expensive. LOO and PSIS-LOO are essentially estimates to the elpd, as a measure of predictive accuracy of models. Vehtari et al. define the elpd for a new dataset as

$$\text{elpd} = \sum_{i=1}^{n} \int p(\hat{y}_i) \log p(\hat{y}_i|y)d\hat{y}_i,$$

where $p(\hat{y}_i)$ is the distribution of the true data-generating process for $\hat{y}_i$. While PSIS-LOO defined by Vehtari et al. takes the form of

$$\text{elpd}_{\text{PSIS-LOO}} = \sum_{i=1}^{n} \log \frac{\sum_{s=1}^{S} v_s p(y_i|\delta^S)}{\sum_{s=1}^{S} v_s f^*_s(y_i|x_i)},$$

where $v_s$ is the replaced notation for smoothed weights using the Pareto distribution for a long-tailed weight distribution, $\delta^S$ is the replaced notation for sampled posterior parameters $\delta$ for samples $s = 1, \ldots, S$. We replace some of the notations from Vehtari et al. to be consistent with the work in this manuscript. For more details, see Vehtari et al.

The model with the largest $\text{elpd}_{\text{PSIS-LOO}}$ is usually chosen to be the selected model, as a posterior predictive density indicates the new observation is well predicted by the model. The main goal of the MRP model is to make predictions at the population level, or sometimes at the small-area level. However, samples used in MRP are often unrepresentative of the population, so we propose a weighted alternative, as suggested in Lumley and Scott:

$$\text{elpd}_{\text{WTD-LOO}} = \frac{1}{n} \sum_{i=1}^{n} w_i \log f_\delta(y_i|x_i),$$

where $w_i$ is the survey weight for each individual $i$ and $f_\delta(y_i|x_i)$ is the conditional density of $y$ given $x$ with the parameters $\delta$.

Thus, our proposed weighted alternative WTD-PSIS-LOO takes the final form of

$$\text{elpd}_{\text{WTD-PSIS-LOO}} = \frac{1}{n} \sum_{i=1}^{n} w_i \log \frac{\sum_{s=1}^{S} v_s p(y_i|\delta^S)}{\sum_{s=1}^{S} v_s},$$

where $w_i$ is the survey weight for each individual $i$ and $f_\delta(y_i|x_i)$ is the conditional density of $y$ given $x$ with the parameters $\delta$.
where we replace \( f_g(y_i|x_i) \) in Equation (7) with the Pareto-weighted density \( (f_g^*(y_i|x_i)) \) in Equation (6). We calculate both WTD-PSIS-LOO and the unadjusted PSIS-LOO and compare their efficacy as a measure of predictive accuracy in our simulation studies.

1.4 How do we judge the MRP method?

So what is a good MRP model? As MRP methods consist of two stages, it is not as straightforward as simply finding a good model independent of the post-stratification step. Instead, we focus on the estimand. Existing MRP research has focused on identifying methods that produce population and sub-population estimates with high precision—narrow prediction intervals for the population estimates at the multi-level regression stage, and low bias—where the prediction for the population estimate is close to the assumed truth (e.g., see Buttice and Highton\(^2\); Downes et al\(^3\)).

This is beneficial because it allows the readers to separate the difference between methods that improve bias and methods that improve precision. In the context of survey work, it might be particularly important to avoid biased estimates in some applications. However, the calculation of bias requires knowledge of the underlying truth and so can only be used in simulation studies and some special contexts (e.g., political science applications such as post-election results as in Isakov and Kuriwaki\(^32\)). In this manuscript, our aim is to identify the usefulness of PSIS-LOO and WTD-PSIS-LOO, which do not require knowledge of the underlying truth and thus can be used to validate models in real-world settings. To facilitate easier comparison, we choose a single proper scoring rule that uses the underlying truth of the simulation to estimate the goodness of a predictive model. As we are comparing probabilistic estimates of a prediction, we choose the interval scoring rule in Gneiting and Raftery\(^33\):

\[
S_{int}^a(l, u; x) = (u - l) + \frac{2}{a}(l - x)\mathbb{I}\{x < l\} + \frac{2}{a}(x - u)\mathbb{I}\{x > u\},
\]

where \( a \) is the threshold for the \((1 - a) \times 100\%\) prediction interval \([u, l]\), \( u \) and \( l \) are the upper and lower bound of the prediction interval of the MRP estimates at both the population or the individual level and \( x \) is the population and individual truths from our simulations. The interval score rewards narrow prediction intervals and penalizes when an observation misses the interval. Hence, a lower value of the interval score indicates a good model fit.

Our argument proceeds as follows. In Section 2, we compare PSIS-LOO and WTD-PSIS-LOO in the context of population estimands predicted using different sets of variables. We demonstrate that PSIS-LOO and WTD-PSIS-LOO do not reflect the true model goodness at the population level. We continue our investigation by demonstrating that first, the best model for small-area estimations (SAE) depends on the small area being estimated, and second, PSIS-LOO and WTD-PSIS-LOO are not sensitive to that either. In Section 3, we compare two models with different regularizing priors in population and small-area estimands, and again demonstrate a lack of consistency in PSIS-LOO and WTD-PSIS-LOO when used to score our models. Together, these suggest that LOO-based model ranking techniques may not always be suitable to evaluate the goodness of MRP estimates. We discuss the implication of this in Section 5.

2 VARIABLE SELECTION

Our first challenge will consider the ability of PSIS-LOO and WTD-PSIS-LOO to identify models that are good at predicting population quantities (i.e., the population mean). We create a simulated scenario with two primary features. First, a sampling design where the probability of inclusion is strongly related to two of our predictors, and two are weakly related. Second, we create the relationship between the outcome \( y \) and four categorical predictors \( X \), two of which are strongly predictive of our outcome and two of which are weakly related. Together this forms a \( 2 \times 2 \) table for potential predictor variables shown in Table 1 below. According to Little and Vartivarian,\(^2\) we would expect a model with \( X_4 \)-omitted to

| Variable type               | Weakly predictive of inclusion | Strongly predictive of inclusion |
|-----------------------------|--------------------------------|---------------------------------|
| Weakly predictive of the outcome | \( X_1 \) (ignorable) | \( X_3 \) (inconsequential) |
| Strongly predictive of the outcome | \( X_2 \) (precision) | \( X_4 \) (bias) |
produce biased MRP estimands and a model with $X_2$-omitted to produce less precise MRP estimates. Hence, for simplicity, we name the variables $X_4$ and $X_2$ as bias and precision variables accordingly. The remaining two are denoted *ignorable* (not strongly related to the probability of inclusion or outcome) and *inconsequential* (strongly related to sample inclusion but not the outcome). We name the variables accordingly for ease of reading.

### 2.1 Simulation design I

We use 100 simulated iterations. For each, we create a new population of size $N = 10,000$. We define the population consisting of four variables in the set $K = (X_1, X_2, X_3, X_4)$ with $k$ denoting the $k$th member of the set. First, we generate four continuous variables from a normal distribution with a mean zero and a standard deviation of two. Then, the probability of outcome and inclusion probability were set to have an inverse-logit relationship with the sum of the continuous variables:

\[
\text{Probability of outcome: } \Pr(y = 1) = \text{logit}^{-1}(0.1X_1 + 1X_2 + 0.1X_3 + 1X_4),
\]

\[
\text{Inclusion probability: } \Pr(I = 1) = \text{logit}^{-1}(0.1X_1 + 0.1X_2 + 1X_3 + 1X_4).
\]

In our simulation, we denote 0.1 as the coefficient for a weak relationship and 1 for a strong relationship with the outcome and the inclusion probability.

We create a binary outcome variable $y$ through the binomial formulation using $\Pr(y)$ obtained in Equation (10). The inclusion probabilities $\Pr(I)$ are generated as in Equation (11). Then, each $k$th continuous $X$ variable is discretized into $L^{(k)} = 5$ groups of equal range. The relationship between the outcome and inclusion probability were set up with the continuous variables instead of the discrete variables to allow stronger trends in their relationships. We depict the relationships between the discretized covariates and the inclusion probability as well as the probability of outcome in Figure 1. We use only the discrete variables and binary outcome in all of our models.

Now, we have generated the population with a binary outcome, discrete variables and continuous inclusion probability. Overall, our simulation setup has 100 different populations of size $N = 10,000$ and samples of size $n = 1000$. To do this, we take a sample of size 980 according to the inclusion probability, and the remaining $L^{(k)} \times K = 20$ are sampled by enforcing each of the five levels for the four variables are included in the sample. We do this to avoid running into issues when constructing weights for a level that does not exist in the sample. We repeat the steps above for a 100 times to construct different populations and samples each time. This is considered the superpopulation approach, as we generate a different population each time. We employ this approach to test the efficacy of PSIS-LOO and WTD-PSIS-LOO for different simulated populations.

For each sample, we then fit an MRP model as in Equations (1) to (3). We fit the Bayesian multilevel logistic model using the probabilistic programming language Stan\textsuperscript{34,35} as called from R. A total of 15 different models were fitted by regressing the binary outcome on all combinations of the $X$ variables ($2^4 - 1$, excluding the null model).* We then poststratify the predicted population quantities from the models using formula (4).

To assess the performance of PSIS-LOO and WTD-PSIS-LOO in different MRP settings, for each sample, we calculate PSIS-LOO values for each model using the `loo` package in R. The WTD-PSIS-LOO is calculated by using raked weights through the population margins for the set of variables $K = (X_1, X_2, X_3, X_4)$ using the `rake` function from the `survey` package in R.†

### 2.2 Results

To investigate variable selection between different groups of models, we examine the PSIS-LOO-based model score values against the interval score, the selected goodness measure of a model. We study these on three instances, at the individual level, MRP population estimand level and the small-area estimand levels. We describe and discuss the findings in the following sections.

#### 2.2.1 Individual level

We first consider for each model, how well PSIS-LOO and WTD-PSIS-LOO accord with the individual-level prediction accuracy in the sample (top row of Figure 2) as measured by the interval score (averaged over 100 iterations). If the
PSIS-LOO approximately represents the true individual prediction goodness, we should see a negative and roughly linear relationship. As we see in the top left panel of Figure 2, this broadly is true. We see a clustering in model goodness, which will be a salient component of our results. The bias-precision models (blue-toned dots) are preferred by both scores, followed by the precision(-variable)-only models (green-toned dots), bias(-variable)-only models (red-toned dots) and the “irrelevant” variable models (gray-toned dots).

When we move to WTD-PSIS-LOO (top right column of Figure 2), the correlation between the two ranking criteria decreases slightly (from −0.98 to −0.95), and the bias- and precision-only variable models (red- and green-toned dots) are ranked similarly by WTD-PSIS-LOO, but not by the interval score. In other words, based on WTD-PSIS-LOO, both bias- and precision-only models are equally preferred. This might seem counterintuitive at first, but consider that we calculated the mean individual interval score in the sample, not the population. The sampling design has changed the relative proportions of different $X_4$ (bias) levels in the sample (which led to the precision-only models being favoured over the bias-only models in the top left panel), but in the population $X_2$ and $X_4$ had similar distributions by creation.

If we look at the mean individual interval score in the population (bottom row of Figure 2), we see that at the population level, the mean interval score for individuals are equally preferred for the bias- and precision-only models. As mentioned above, $X_2$ and $X_4$ in the population have similar distributions and this is reflected by the interval score measure, but not the unadjusted PSIS-LOO. When we adjust to WTD-PSIS-LOO using weights, both the interval score and WTD-PSIS-LOO agrees with the bias- and precision-only models, resulting in a huge overlap in the bottom right panel of Figure 2 and a slight increase in correlation between the two ranking criteria (from −0.96 to −0.97).
FIGURE 2  Interval score for average individual predictions (y-axis) and PSIS-LOO-based model scores (x-axis) in the sample and the population for 100 iterations. The interval score for the population individuals is calculated by multiplying individual raked weights by the mean interval score at the sample level. The four groups of colour represent different models (blue tone: models with both “bias & precision”; red tone: models with “bias-only”; green tone: models with “precision-only”; and gray tone: models with the “irrelevant” variables in them.) The x-axis on the left column panel represents PSIS-LOO as per loo package and the right column panel shows our proposed alternative WTD-PSIS-LOO. Ideally, the ranking of the models should be similar on both panels (which would be represented by a linear relationship between the two model scores).

2.2.2  |  MRP estimand

Keeping in mind what the relationship should look like (top left of Figure 2), we now turn our attention to comparing PSIS-LOO and WTD-PSIS-LOO to the population interval score in Figure 3. On the x-axis, the full model (blue-toned dots) are preferred based on all three measures: PSIS-LOO, WTD-PSIS-LOO and interval score. Unfortunately, we see a large deviation between the ordering of the precision-only and bias-only models. While PSIS-LOO slightly prefers the precision-only model, the interval score reveals that the bias-only model is almost as good as the full model. We see a similar problem for the WTD-PSIS-LOO method, with a slight reordering of the bias- and precision-only models (correlation from −0.53 to −0.69), but not sufficiently to properly capture the true ordering. These bring suspicions that PSIS-LOO-based scores concern only about the strength of relationship of the variable with the outcome, but not with the inclusion probability. We conduct extra simulations to confirm this and included them in the Appendix (Figure A1).

How can we interpret these findings? First, there is good news. PSIS-LOO and WTD-PSIS-LOO do correctly suggest that the full model is the most appropriate for an MRP model in our simulation study. However, beyond this, the ordering is not correct. This is an issue for two reasons. One, there is no guarantee that we would always have the true model in the pool of models, and hence we are somewhat reliant on the ordering being fully correct. Second, there is a disadvantage to MRP models that is larger than they need to be, particularly if some of the variables are noncensus variables and therefore require substantial work to be imputed in the population. In the case of our bias-only model, it is almost as good as the full model, but with one less variable. In some contexts, this model could be highly desirable. Our hypothesis of why this occurs lies with the aggregation stage of MRP, where errors can be averaged out like a beneficial Simpson’s paradox. However, as our populations tend to the smaller samples (as in small-area estimation), will PSIS-LOO get closer
FIGURE 3  MRP population estimand interval score (y-axis) vs. PSIS-LOO-based model score (x-axis) in 100 iterations. The four groups of colour represent different models (blue tone: models with both “bias-precision” in it; red tone: models with “bias-only”; green tone: models with “precision-only”; and gray tone: models with the “irrelevant” variables in them.) The x-axis on the left panel represents PSIS-LOO as per `loo` package and the right panel shows the weighted alternative WTD-PSIS-LOO. Ideally, the ordering should be similar on both panels.

2.2.3 Small-area estimands

MRP is a dual-purpose method. In addition to being interested in the population estimand, many MRP applications also aim to make small-area estimates as well. Using the simulation setup we described in Section 2.1, we examine the interval score for small-area estimates in each level of the variable. As WTD-PSIS-LOO slightly corrects for our representation issue, we focus on WTD-PSIS-LOO in this section.

To examine small-area estimation, we poststratify each model to each level of the variable instead of to the population. Let $K$ be the set of variables we wish to do small-area estimations on, such that $K = \{X_1, X_2, X_3, X_4\}$. Denote $k$th member of this set as $X^{(k)}$. Then, we aim to identify whether WTD-PSIS-LOO correctly preserves the ordering of models for each $k$th variable. To do this, we need to modify the calculations of WTD-PSIS-LOO so it only represents the prediction error for the level $\ell$ (eg, 18 to 25-years-old) for the $k$th variable (eg, age group) in a way that is comparable across simulation iterations (remembering that by chance each sample will likely have a different sample size in each cell).

We do this by taking the mean of the individual elpd values obtained through WTD-PSIS-LOO, but now only where $i \in S$ where $S$ is the set of individuals that are in our target small-area level $\ell$ for each $k$th variable in $K$. Now, our $\theta_3$ in Equation (5) corresponds to the estimand for each level of the four variables.

To examine each of the model validation scores, we evaluate WTD-PSIS-LOO and interval score at each of the variable levels. We take the mean of WTD-PSIS-LOO values at each small-area instead of the sum due to the varying sample sizes in each group. Let $\xi_{\ell}$ be the set of individuals where $X_k \in \ell$ and $\ell$ be the $\ell$th level for each covariate $X_k$, for $\ell = 1, \ldots, L^{(k)}$, and $L$ be the total number of levels where $L = \sum_k L^{(k)}$. The mean of WTD-PSIS-LOO values for each small-area is then calculated by
Small-area estimations of interval score (y-axis) vs. mean WTD-PSIS-LOO scores (x-axis) of the 5th level ($\ell = 5$) for each of the variables $X_1, X_2, X_3,$ and $X_4$ for all 15 models in 100 iterations. The mean WTD-PSIS-LOO values of each level of the variables are as calculated through the \texttt{loo} package. The four groups of colour represent different models (blue tone: models with both “precision & bias” in it; red tone: models with “bias-only”; green tone: models with “precision-only”; and gray tone: models with the “irrelevant” in them. The horizontal panels show each variable and the vertical panels represent the levels of the variables.

\begin{equation}
\widehat{\text{elpd}_{\text{WTD-PSIS-LOO}}}[\ell] = \sum_{i \in \ell} \frac{\overline{\text{elpd}_{\text{WTD-PSIS-LOO}}}[i]}{|x|} \quad \text{for } i = 1, \ldots, n, \ell = 1, \ldots, L^{(k)},
\end{equation}

where $|x|$ is the length of vector for all observations in $x$. We make similar small-area calculations for the interval score by replacing $\text{elpd}_{\text{WTD-PSIS-LOO}}$ with interval score in Equation (12).

With the calculated small-area model scores, we plot the small-area interval score and mean WTD-PSIS-LOO values for each variable when level $\ell' = 5$ in Figure 4. We present the figure for all variable levels in the Appendix (Figures A3-A4). In Figure 4, we see that the best model for small-area estimation differs based on the area being estimated. This suggests that MRP model validation might depend not only on the outcome being estimated, but also on the target estimand, and raises questions about how practitioners who aim to estimate multiple small areas, as well as an overall population should correctly identify the “best” model. In terms of which model is preferred, we see in the top right of Figure 4 that the models without the precision variable (“bias-only” and “irrelevant” models) have substantially worse interval scores for
small-area estimations of the precision variable (remember at the population this was not true). Conversely on the bottom right, for our bias variable, we see that models without the bias variable in them (“precision-only” and “irrelevant”) are substantially worse in terms of interval score. If we look at the x-axis only, WTD-PSIS-LOO does not seem prefer any of the models for our ignorable and inconsequential variables. When we look at the precision variable (top right plot), in terms of WTD-PSIS-LOO, although there are some preference for the models with the precision variable in them, the distinction across all models is less clear. This pattern is similar for the bias variable.

We also see a similar problem as we saw in the population estimates when it comes to the usefulness of WTD-PSIS-LOO. While the best model differs for different variables, the actual ordering does not relate to the mean WTD-PSIS-LOO at all (for a particularly alarming example see the bottom-left figure). This suggests that a good MRP model for SAE differs based on the small area being estimated. WTD-PSIS-LOO, however, is unable to distinguish between the different models with distinct inclusion of the variables at the small-area level, as shown in all of the panels in Figure 4. In the next section, we investigate if this challenge is specific to variable selection or if WTD-PSIS-LOO would be suitable when comparing models with the same variables but different priors.

3 | PRIOR SELECTION

To investigate the efficacy of PSIS-LOO-based scores on using different priors in our model, specifically in estimating small areas, we adapt simulation design I in Section 2.1. The main difference from the previous simulation design is that we follow the setup closely of Gao et al. 23 and apply an autoregressive (AR) prior (also named structured prior) on our bias variable $X_4$. We further describe the change in simulation design in the following section.

3.1 | Simulation design II

The idea of this simulation is similar to the design in Section 2.1 but with some exceptions. Instead of generating our continuous variables from a normal distribution, we now simulate $X$’s from a uniform distribution on $(-3, 3)$. When discretizing $X_1, X_2$ and $X_3$ into $L(k) = 5$ groups, for $k = 1, 2, 3$. For $X_4$, we discretize it into $L(k) = 12$ groups, for $k = 4$ as in Gao et al. 23 to allow the investigation of MRP prediction in smaller group sizes. We do this to test the efficacy of different priors and their effect on PSIS-LOO-based scores in SAE. As in Section 2.1, we now have four variables, one strongly related to the probability of inclusion and outcome, the bias variable ($X_4$), one strongly related to the outcome, the precision variable ($X_5$) and two that are ignorable ($X_1$) and inconsequential ($X_3$) to our MRP predictions. The relationship between the variables with inclusion probability and outcome respectively are depicted in Figure 5. In estimating small-area outcomes using MRP, we want to generate a stronger relationship between the variable and outcome, to investigate the effectiveness of the use of prior. Therefore, instead of creating coefficients for $X_4$ as a random draw from a zero-mean normal distribution with a standard deviation of 2, we simulate a smooth relationship as the levels increase, mimicking the data simulation as in Gao et al. 23 to investigate the improvements of the use of a structured prior:

\[
\text{Probability of outcome: } Pr(y_i = 1) = \logit^{-1} (-1 + 0.05X_1 + 0.5X_2 + 0.05X_3 + 1.5f(X_4)), \tag{13}
\]

\[
f(x) = 0.7 - 3.0 \exp\left(-\frac{x}{0.2}\right), \quad x \in (-3, 3), \tag{14}
\]

where $f(x)$ in Equation (14) is the increasing-shaped preference for the outcome as in Gao et al. 23.

Then we generate a binary outcome $y$ using the new probability of outcome in Equation (13). Now, we have a binary outcome, three variables $X_1, X_2, X_3$ with 5 levels and a bias variable $X_4$ with 12 levels. To motivate small-area estimations, we use a sample size of $n = 500$ to investigate the efficacy of PSIS-LOO-based scores on SAE of MRP predictions. As in the earlier simulation design, the sample is generated by taking the size of $(n - L)$ according to each of their inclusion probability. The remaining $L$ samples are simulated by making sure each of the variable levels is included in the sample. We demonstrate and discuss the results of sample size $n = 1000$ in the Appendix. Using Stan, we fit the multilevel model similar to Equation (1) but for each small area $l$ for the $k$th variable in the set of variables $K$:

\[
Pr(y_{i} = 1) = \logit^{-1} \left( \beta_0 + \sum_{k \in K} \sum_{l \in l_{k}} a^{(k)}_{l_{k}} \right) \quad \text{for } i = 1, \ldots, n; \quad \ell^{(k)} = 1, \ldots, \ell^{(k)}. \tag{15}
\]
Instead of the base normal prior distribution used in MRP as in simulation I, we assume an AR prior distribution for our bias variable $X_4$:

$$a_1^{(X_4)} | \rho, \sigma^{(X_4)} \sim \text{normal}(0, \frac{1}{\sqrt{1 - \rho^2}} \sigma^{(X_4)}) $$  \hspace{1cm} (16)

$$a_\ell^{(X_4)} | a_{\ell-1}^{(X_4)}, \ldots, a_1^{(X_4)}, \rho, \sigma^{(X_4)} \sim \text{normal}(\rho a_{\ell-1}^{(X_4)}, \sigma^{(X_4)}) $$  \hspace{1cm} (17)

To consider the use of AR priors in SAEs, we compare the PSIS-LOO and WTD-PSIS-LOO preferred model from the previous section—the “full model” that uses all four variables as the predictor, with and without the structured prior on the bias variable $X_4$ in the following section.

### 3.2 Results

To investigate the use of AR priors, we compare the mean interval score and mean PSIS-LOO-based model scores for the full model with all variables, particularly with AR prior and with non-AR prior (base normal prior) on the bias variable. We first compute the difference in mean interval score and mean PSIS-LOO-based model scores (PSIS-LOO and WTD-PSIS-LOO) for all of the population individuals from simulation design II. We plot the difference in model scores for the population individual, MRP estimand and small-area estimand level. We calculate the difference in both scoring criteria to see if the model with AR prior performs better than the non-AR (base normal) prior, and if the interval score and PSIS-LOO agree on the differences.
3.2.1 | Individual level

We take the individual predictions at the population level and calculate the difference in the mean interval score and PSIS-LOO-based model scores in Figure 6. The difference in mean interval score is calculated using \( \text{diff}_{\text{IS}} = \text{mean interval score (AR prior)} - \text{mean interval score (non-AR prior)} \), where IS stands for interval score. We take the negative value of the difference as a smaller value of interval score indicates a better predictive model. We calculate the difference in mean PSIS-LOO-based models score using \( \text{diff}_{\text{LOO-based}} = \text{LOO score (AR prior)} - \text{LOO score (non-AR prior)} \), where LOO score indicates PSIS-LOO or WTD-PSIS-LOO values.

The green quadrant signifies when PSIS-LOO corresponds to the perceived true goodness fit of the model where AR prior is working better. The red quadrants are where the model scores disagree in preference for the use of AR priors. At the individual level, we see that both scores perceive a similar preference for the AR prior model, where most points are scattered on the top right green quadrant. The similarity holds true for both PSIS-LOO and WTD-PSIS-LOO.

3.2.2 | MRP estimand

For Figure 7, we take the MRP population estimand predictions and calculate the difference in interval score and difference in PSIS-LOO-based scores. The difference in mean for the models using AR and non-AR priors on the bias variable is calculated similarly as before. We see that at the MRP population estimand level, the AR prior
model is only preferred by the true interval score in some of the cases, while for most iterations it signifies no difference in the use of AR prior. However, the PSIS-LOO-based model scores still suggest a difference between the models.

We then calculate the small-area estimates of interval score and PSIS-LOO-based model score difference in each of the levels of the bias variable $X_4$. As both WTD-PSIS-LOO and PSIS-LOO show similar results, we focus on WTD-PSIS-LOO for clarity. We plot the score difference in each small area of the bias variable $X_4$ between the two models with different priors in Figure 8 below.

### 3.2.3 Small-area estimands

In Figure 8, we see that the small-area difference in PSIS-LOO values corresponds well to the interval score in levels 1, 2, and 3 of the bias variable $X_4$, as most of the points are in the green quadrant. This is where the group sample size is the smallest, averaging about 18 units in the sample. The figure shows that as the variable level increases (so does the sample size due to the simulation setup), the points cluster around the origin. This signifies no difference in the use of AR prior on the higher variable level and increased sample sizes. This is coherent with the finding in Gao et al\textsuperscript{23} where the structured prior works better for smaller group sizes than in bigger groups which has more stratification of levels in the variable. We find the structured prior is less useful when the total sample size increases to 1000 instead of 500, and display the results in the Appendix (Figure A5).
4 | NHANES DATA APPLICATION

We replicate our simulation studies with the National Health and Nutrition Examination Survey (NHANES) data collected by the National Center for Health Statistics. The NHANES data contains a number of subsamples, which we will use to estimate a quantity known in the full sample. In particular, we use two subsamples that exist in the NHANES data and generate another artificial sample based on the NHANES variables.

The NHANES data was collected to assess the health and nutritional status of adults and children in the United States and are released to the public every two years. However, data collection in the NHANES 2019 to 2020 cycle was suspended in March 2020 due to the coronavirus disease 2019 (COVID-19) pandemic. Thus, we use the NHANES 2017—March 2020 prepandemic data, which combines data from 2019 to March 2020 with data from the NHANES 2017 to 2018 cycle to form a nationally representative sample.38

For illustration purposes, we select high blood pressure (yes/no) as our outcome variable and include a total of 15 demographic and health-related variables collected in the NHANES that are potentially predictive of high blood pressure or to the subsample inclusion.39-41 To create a synthetic population, we keep only the individuals with the
full set of these 15 predictor variables in the NHANES data \((N = 6686)\). To create different samples, we select two subsamples that already exist in the NHANES data, the fasting subsample and dietary 2-day sample.\(^{38}\) We include all individuals in these two subsamples with pre-existing weights and randomly select 1500 individuals for each of the samples \((n = 1500)\). Since those existing weights were created for inference of the general population of the United States but not our synthetic population, we discard those weights and create weights to adjust to the full sample in our analyses. In addition, we generate a third sample by simulating a stronger relationship between a selected set of variables and the inclusion probability. The equation of inclusion probability for the generated sample is included in the Appendix. This is to ensure we have a more balanced number of variables in each of the categories as in Table 1.

Together we have a synthetic population and three samples from the NHANES data. To create survey weights for our samples to make them representative of the synthetic population, we use Bayesian additive regression trees (BART) to predict sample inclusion using 15 predictor variables.\(^{42}\) We first fit BART on the synthetic population data with the sample inclusion indicator (1 for sampled and 0 for nonsampled) as the binary outcome against all the predictor variables using the \(\text{dbart}\) package.\(^{43}\) Then we predict the probability of inclusion for all sampled individuals. The inverse of the predicted probabilities are our respective sample weights for the individuals. We repeated the steps to create a set of weights for each of the three samples.

![Mean of correctly predicting individual outcomes](image)

**Figure 9** Comparison of PSIS-LOO-based model scores with the average individual population goodness. On the y-axis we represent one minus the mean of prediction of outcome for average individual predictions, and on the x-axis the PSIS-LOO-based model scores (x-axis) in the sample and the population for three NHANES (sub)sample. The mean of prediction of outcome is computed by calculating the proportion of correctly predicted outcomes from our models. The six symbols represent the models with six different sets of variables (circle: model with all the variables; square: model with both bias and precision variables; triangle: model with bias variables only; diamond: model with precision variables only; inverted triangle: model with inconsequential variables only; and cross: model with ignorable variables only). The three groups of color represent three different samples: dietary, fasting and generated. The x-axis on the left column panel represents PSIS-LOO as per \(\text{loo}\) package and the right column panel shows our proposed alternative WTD-PSIS-LOO. Ideally, the ranking of the models should be similar on all the panels (which would be represented by a linear relationship between the two model scores).
To align with the earlier simulations, we need to identify variables that are bias, precision, inconsequential, and ignorable variables. We use lasso regressions from the glmnet package\(^4^4\) on the NHANES synthetic population to select predictors that are predictive of the high blood pressure outcome and each of the sample inclusion indicator. We categorise the variables in each sample accordingly as documented in Tables A4 and A5 and plot the probability of high blood pressure (outcome variable) in each of the variable categories in the population and the ratio of the proportion of respondents in sample to population in each of the variable categories in samples in Figures A6 and A7 in the Appendix.

To investigate the impact of the types of variables on the MRP estimates, we fit six MRP models for each sample using the different sets of predictor variables. The six models are (1) model with all the variables, (2) model with both bias and precision variables only, (3) model with bias variables only, (4) model with precision variables only, (5) model with inconsequential variables only, and (6) model with ignorable variables only. As in the simulation studies, we calculate the respective PSIS-LOO and WTD-PSIS-LOO values for each model and compare them with the calculated interval score and the proportion of correctly predicting the outcome for the individuals as in Figures 9 and 10.

Similar to our simulation studies, we examine the performance of PSIS-LOO-based model scores on both the sample individual and at the population level. For visualization purposes, we scale the model scores plotted on the x-axis so as to be consistent across the samples. On the y-axis, we calculate one minus the mean of correctly predicting the binary outcome on the individual level.\(^5^\) We separated out the model with all the variables in them and “bias-precision” variables only.\(^6^\) We see in Figure 9 that both PSIS-LOO and WTD-PSIS-LOO accord with the individual-level predictive accuracy and rank the models similarly. In particular, PSIS-LOO-based scores slightly prefer the “bias-precision” models over the “all variables” models, which is similar to the individual goodness averaged over the whole population (y-axis). The weighted version WTD-PSIS-LOO did not seem to affect the model ranking in each sample.

---

**FIGURE 10** Scaled interval score for MRP population estimands (y-axis) and PSIS-LOO-based model scores (x-axis) for three NHANES (sub)samples. The six symbols represent the models with six different sets of variables (circle: model with all the variables; square: model with both bias and precision variables; up-pointing triangle: model with bias variables only; diamond: model with precision variables only; down-pointing triangle: model with inconsequential variables only; and cross: model with ignorable variables only). The three groups of colour represent the three different samples: dietary, fasting and generated. The x-axis on the left column panel represents PSIS-LOO as per loo package and the right column panel shows our proposed alternative WTD-PSIS-LOO. Ideally, the ranking of the models should be similar on both panels (which would be represented by a linear relationship between the two model scores).
From this we can determine that PSIS-LOO can correctly identify the best model for individual predictions. We now turn our attention to the goodness of models for population estimates. We compare the scaled PSIS-LOO and WTD-PSIS-LOO to the scaled population interval score in Figure 10. Again, the scores are scaled within each sample to allow consistent comparisons. The figure shows the ranking of the models is inconsistent. On the x-axis, PSIS-LOO slightly prefer the model with “bias-precision” (square dots) over model with all the variables (round dots) in them but the interval score disagrees especially for the dietary (orange-toned) and fasting (green-toned) samples. For the generated sample (pink-toned), the interval score suggests the bias-only model (up-pointing triangle) is the most preferred. The weighted version WTD-PSIS-LOO only slightly adjusts but did not change the overall ranking.

In our simulation studies, PSIS-LOO-based scores generally suggest the full model (all the variables included) as the most appropriate, and was unable to distinguish between the models that have bias- or precision-only variables. In this real data application, at the individual level, both PSIS-LOO-based scores generally agrees with the interval score. There are no distinct differences when using either the PSIS-LOO or WTD-PSIS-LOO as the model scoring criteria. However, when making a population MRP-based estimate, the model rankings are different across samples and the six models with different sets of variables. Notably, the ranking of bias- and precision-only models are not consistent across samples. Regardless, in no sample did the PSIS-LOO scores reflect the population MRP goodness score, as noted in our simulation studies.

5 | DISCUSSION

Validating MRP is not straightforward. We focused here on using leave-one-out cross validation and demonstrated three unique challenges.

The first challenge is *representation*: MRP is typically used in settings where the sample is not representative of the broader population, but the leave-one-out cross validation assumes that it is. In our simulation studies, this turned out to be the least impactful challenge. More alarming is the impact that aggregation has on the usefulness of PSIS-LOO-based measures. While in MRP, individual prediction error can average out in group averages, when using a cost function like expected log predictive density, these errors simply grow. We suspect this causes the inconsistencies in ordering that we observed in this work. Lastly, we observed that the best model depended on the target estimand (particularly population versus small-area).

This all suggests that PSIS-LOO should be used with caution, if at all, to validate MRP models. The first challenge of representation, could be solved with a weighted alternative of PSIS-LOO. We had initially included an MRP version of this where weights are not available. However, it is eventually decided that the adjustment for representation of the sample was a relatively small contributor to the model validation issues as represented by the score studies in this work. When using PSIS-LOO for MRP model validation, the weighted version helps slightly to rebalance the ordering, but as we found that is not the biggest challenge.

The next major challenge is *aggregation*. As we saw in our results, using PSIS-LOO had the potential to suggest models that had more variables than needed for a reliable MRP estimate. In particular, PSIS-LOO was able to distinguish models that have *both* variables that are strongly predictive of the outcome and inclusion. However, it does not differentiate between models that only have one of the variables. This has implications for practitioners. If we rely only on PSIS-LOO, a model with noncensus variables might get selected, which can be time consuming if not impossible to create in the population. Our working hypothesis for this challenge is the default use of elpd as a cost metric in PSIS-LOO. Further work would be needed to confirm this, but if this is the case, it raises challenges for other elpd-based tools used with MRP, such as stacking.

The final challenge is *multiple goals*, which have been hinted at in previous work (eg, Gao et al\(^ {23} \)) but are highlighted when we consider the best model for different small areas. The MRP model best suited to the job depends on the goal. That MRP is a multi-use tool, for both population and small-area estimation, has always been a valuable feature. However, balancing multiple uses of this tool when it comes to validating and selecting the model has been challenging and requires further work.

An advantage of using the MRP method is that the multilevel regression allows adjustment to be made for multiple background variables, including those used to construct weights, so that the regression and poststratification can be performed unweighted. Comparison of different MRP models thus corresponds in some sense to comparison of weighted and unweighted estimates, with simpler models that adjust for fewer factors being analogous to unweighted or partially weighted fits, and saturated models that adjust for all weighting factors corresponding to fully weighted fits. However if reliable weights exist, we could compare models that produce weighted and unweighted estimates, but there is no guarantee that PSIS-LOO will reflect the correct ordering of models when comparing different models consisting different set of variables, as discussed at length in this paper.
Overall, we have illustrated a cautionary tale. In our studies, PSIS-LOO and WTD-PSIS-LOO typically prefer a model that would give good MRP estimands. However, the ordering of all other models implies that our confidence in this depends on which models were included. We suggest potential reasons for this, but until these have been investigated further we suggest caution in interpreting PSIS-LOO-based model validation tools. Other strategies, such as comparisons to high-quality auxiliary surveys (where available) and known truths might still be our best tools when evaluating MRP-derived estimates.

ACKNOWLEDGEMENTS
We thank Dan Simpson for advice and comments. Open access publishing facilitated by The University of Adelaide, as part of the Wiley - The University of Adelaide agreement via the Council of Australian University Librarians.

FUNDING INFORMATION
We gratefully acknowledge support from National Institutes of Health grants 5R01AG067149–02.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study will be available in the LOO_MRP repository at https://github.com/lauken13. Part of this data were derived from the NHANES available in the public domain at https://www.cdc.gov/nchs/nhanes/, modifications available from the corresponding author upon request.

ENDNOTES
*A summary of all the models is provided in the Appendix.
†A function loo_wtd is written to calculate the WTD-PSIS-LOO by using raked weights and svytotal function from the survey package.27
‡The mean sample size of each variable level for all 100 iterations are documented in the Appendix (Table A1).
§The average group sizes of each variable level for 100 iterations are documented in the Appendix (Table A2).
¶To accord with Figure 2 for ease of comparison.
§The ‘all variables’ model was categorised under the ‘bias-precision’ models in the simulation studies.

ORCID
Swen Kuh https://orcid.org/0000-0002-7293-4094
Andrew Gelman https://orcid.org/0000-0002-6975-2601

REFERENCES
1. Gelman A, Little TC. Poststratification into many categories using hierarchical logistic regression. Surv Methodol. 1997;23:2127-2136.
2. Park DK, Gelman A, Bafumi J. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. Polit Anal. 2004;12(4):375-385.
3. Lax JR, Phillips JH. Gay rights in the states: Public opinion and policy responsiveness. Am Polit Sci Rev. 2009a;103(3):367-386.
4. Wang W, Rothschild D, Goel S, Gelman A. Forecasting elections with non-representative polls. Int J Forecast. 2015;31(3):980-991.
5. Downes M, Gurrin LC, English DR, et al. Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. Am J Epidemiol. 2018;187(8):1780-1790.
6. Si Y, Trangucci R, Gabry JS, Gelman A. Bayesian hierarchical weighting adjustment and survey inference. arXiv:1707.08220 2017.
7. Valliant R. Comparing alternatives for estimation from nonprobability samples. J Survey Stat Methodol. 2020;8(2):231-263.
8. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017;27(5):1413-1432.
9. Park DK, Gelman A, Bafumi J. State-level opinions from national surveys: Poststratification using multilevel logistic regression. In: Cohen JE, ed. Public Opinion in State Politics. Stanford, CA: Stanford University Press; 2006:209-228.
10. Warshaw C, Rodden J. How should we measure district-level public opinion on individual issues? J Polit. 2012;74(1):203-219.
11. Lax JR, Phillips JH. How should we estimate public opinion in the states? Am J Polit Sci. 2009b;53(1):107-121.
12. Buttice MK, Highton B. How does multilevel regression and poststratification perform with conventional national surveys? Polit Anal. 2013;21(4):449-467.
13. Ghitza Y, Gelman A. Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. Am J Polit Sci. 2013;57(3):762-776.
14. Zhang X, Holt JB, Yun S, Lu H, Greenland KJ, Croft JB. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. Am J Epidemiol. 2015;182(2):127-137.
15. Covello L, Gelman A, Si Y, Wang S. Routine hospital-based SARS-CoV-2 testing outperforms state-based data in predicting clinical burden. Epidemiology. 2021;32(6):792.
16. Machalek DA, Vette KM, Downes M, et al. Serological testing of blood donors to characterise the impact of COVID-19 in Melbourne, Australia, 2020. PLoS One. 2022;17(7):e0265858. doi:10.1371/journal.pone.0265858
17. Bisbee J. BARP: Improving Mister P using Bayesian additive regression trees. Am Polit Sci Rev. 2019;113(4):1060-1065.
18. Liu Y, Chen Q. Bayesian inference of finite population quantiles for skewed survey data using skew-normal penalized spline regression. J Survey Stat Methodol. 2020;8(4):792-816.
19. Ornstein JT. Stacked regression and poststratification. Polit Anal. 2020;28(2):293-301.
20. Gao Y, Kennedy L, Simpson D. Treatment effect estimation with Multilevel Regression and Poststratification. arXiv preprint, arXiv:2102.10003 2021.
21. Liu Y, Gelman A, Chen Q. Inference from non-random samples using Bayesian machine learning. J Survey Stat Methodol. 2022;11(2):433-455.
22. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press; 2007.
23. Gao Y, Kennedy L, Simpson D, Gelman A. Improving multilevel regression and poststratification with structured priors. Bayesian Anal. 2021;16(3):719-744.
24. Kastellec JP, Lax JR, Malecki M, Phillips JH. Polarizing the electoral connection: partisan representation in Supreme Court confirmation politics. J Polit. 2015;77(3):787-804.
25. Kuriwaki S, Ansolabehere S, Dagonel A, Yamauchi S. The geography of racially polarized voting: calibrating surveys at the district level. Am Polit Sci Rev. 2023;1-18. doi:10.1017/S0003055423000436
26. Wang W, Gelman A. Difficulty of selecting among multilevel models using predictive accuracy. Stat Interface. 2015;8(2):153-160.
27. Lumley T, Scott A. AIC and BIC for modeling with complex survey data. J Survey Stat Methodol. 2015;3(1):1-18.
28. Mercer AW, Kreuter F, Keeter S, Stuart EA. Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. Public Opin Q. 2017;81(S1):250-271.
29. Little RJ, Vartivarian S. On weighting the rates in non-response weights. Stat Med. 2003;22(9):1589-1599.
30. Urminska O, Hansen C, Chernozhukov V. Using double-lasso regression for principled variable selection. Available at SSRN 2733374 2016.
31. Vehtari A, Gelman A, Gabry J, Yao Y. LOO: Efficient leave-one-out cross-validation and waic for Bayesian models. R package version 2.6.0. 2023. https://mc-stan.org/loo/
32. Isakov M, Kuriwaki S. Towards principled unskewing: Viewing 2020 election polls through a correct lens from 2016. Harvard Data Sci Rev. 2020;2(4):69.
33. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102(477):359-378.
34. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. J Stat Softw. 2017;76(1):76.
35. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. 2019. https://mc-stan.org/users/documentation/
36. Kuh S, Kennedy L, Chen Q, Gelman A. Using leave-one-out cross validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale. Statistics in Medicine. 2024;43(5):953-982. doi: 10.1002/sim.9964

APPENDIX

Types of models fitted in simulation design I and II
In simulation I, we expect $2^4 - 1 = 15$ different models for the combination of the four simulated variables. The 15 models are, in R programming syntax:
Bias-precision
1) $y \sim (1|X_1) + (1|X_4)$
2) $y \sim (1|X_1) + (1|X_2) + (1|X_4)$
3) $y \sim (1|X_2) + (1|X_3) + (1|X_4)$
4) $y \sim (1|X_1) + (1|X_2) + (1|X_3) + (1|X_4)$

Bias-only
5) $y \sim (1|X_4)$
6) $y \sim (1|X_1) + (1|X_4)$
7) $y \sim (1|X_3) + (1|X_4)$
8) $y \sim (1|X_2) + (1|X_3) + (1|X_4)$

Precision-only
9) $y \sim (1|X_2)$
10) $y \sim (1|X_1) + (1|X_2)$
11) $y \sim (1|X_2) + (1|X_1)$
12) $y \sim (1|X_1) + (1|X_2) + (1|X_3)$

Irrelevant
13) $y \sim (1|X_1)$
14) $y \sim (1|X_3)$
15) $y \sim (1|X_1) + (1|X_3)$

Average sample sizes in simulation studies

TABLE A1  Average group sizes of our 100 simulated samples of size $n = 1000$ for each of the variables $X_1, X_2, X_3,$ and $X_4$ in simulation I.

| Variable | 1  | 2  | 3  | 4  | 5  |
|----------|----|----|----|----|----|
| $X_1$    | 13 |    | 211| 549| 214|
| $X_2$    | 12 |    | 202| 546| 226|
| $X_3$    |  2 |    |  89| 544| 340|
| $X_4$    |  2 |    |  86| 548| 339|

TABLE A2  Average group sizes of our 100 simulated samples of size $n = 500$ for each of the variables $X_1, X_2, X_3,$ and $X_4$ in simulation II.

| Variable | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $X_1$    | 95  | 98  | 99  | 103 | 105 | NA  | NA  | NA  | NA  | NA  | NA  | NA  |
| $X_2$    | 95  | 96  | 101 | 100 | 107 | NA  | NA  | NA  | NA  | NA  | NA  | NA  |
| $X_3$    | 39  | 68  | 100 | 132 | 161 | NA  | NA  | NA  | NA  | NA  | NA  | NA  |
| $X_4$    | 14  | 18  | 22  | 27  | 33  | 39  | 44  | 50  | 55  | 61  | 66  | 71  |

TABLE A3  Average group sizes of our 100 simulated samples of size $n = 1000$ for each of the variables $X_1, X_2, X_3,$ and $X_4$ in simulation II.

| Variable | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $X_1$    | 189 | 194 | 202 | 206 | 209 | NA  | NA  | NA  | NA  | NA  | NA  | NA  |
| $X_2$    | 188 | 193 | 202 | 204 | 212 | NA  | NA  | NA  | NA  | NA  | NA  | NA  |
| $X_3$    |  76 | 134 | 200 | 264 | 325 | NA  | NA  | NA  | NA  | NA  | NA  | NA  |
| $X_4$    |  26 |  34 |  44 |  54 |  66 |  78 |  89 | 101 | 112 | 123 | 132 | 141 |
Change in strength of relationship of $X_2$ and $X_4$ in simulation I

To examine if PSIS-LOO-based model scores mainly concerns about the relationship of a variable with the outcome but not the inclusion, we modify Equation (10) in simulation I. We change the strength of relationship between $X_2$ (precision) and $X_4$ (bias) with the outcome to 1.5 instead of 1, respectively, and keep the rest of the simulation design unchanged. We do this to evaluate if PSIS-LOO-based scores are sensitive to a stronger relationship between the variables and outcome. As a result, we see on the top panel that when the strength of relationship of $X_2$ and the outcome is stronger, PSIS-LOO model scores rank the (green-toned) precision-only models second, right after the (blue-toned) full model with both bias and precision variables in them. On the bottom panel, when we increase the strength of relationship of $X_4$ and the outcome, we see that the (red-toned) bias-only models are ranked second. The interval score on the y-axis consistently prefers the models with both bias-precision and bias-only models. This confirms our hypothesis of PSIS-LOO-based scores rank mainly based on the strength of relationship with the outcome but not the inclusion, which introduces specific challenges when validating MRP models, discussed at length in this paper.
Small-area estimation in simulation I: Mean PSIS-LOO scores

**FIGURE A2** Small-area estimates of interval score (y-axis) vs. mean PSIS-LOO scores (x-axis) of each level of the variables $X_1, X_2, X_3$, and $X_4$ for all 15 models in 100 iterations. The mean PSIS-LOO values of each level of the variables are as calculated through the `loo` package. The four groups of colour represent different models (blue tone: models with both “precision & bias” in it; red tone: models with “bias-only”; green tone: models with “precision-only”; and gray tone: models with the “irrelevant” in them. The horizontal panels show each variable and the vertical panels represent the levels of the variables. We see that the preferred model for small-area estimation differs based on the area being estimated.
Small-area estimation in simulation I: Mean WTD-PSIS-LOO scores

**FIGURE A3** Small-area estimates of interval score (y-axis) vs. mean WTD-PSIS-LOO scores (x-axis) of each level of the variables $X_1$, $X_2$, $X_3$, and $X_4$ for all 15 models in 100 iterations. The mean WTD-PSIS-LOO values of each level of the variables are as calculated through the `loo` package. The four groups of colour represent different models (blue tone: models with both “precision & bias” in it; red tone: models with “bias-only”; green tone: models with “precision-only”; and gray tone: models with the “irrelevant” in them. The horizontal panels show each variable and the vertical panels represent the levels of the variables. We see a similar pattern to the PSIS-LOO scores where the best model for small-area estimation vary based on the area being estimated.
Interval score, bias, and precision of MRP estimates in simulation I

**Figure A4** Comparison of the interval score, bias, and precision of the MRP estimates for the different models. The left panel shows the interval scores for the MRP estimates. The middle panel shows the difference between MRP prediction and the population truth (bias of MRP estimates). The points are the difference between the mean and ground truth and error bars are the difference between the 90% prediction interval and the truth. The right panel shows the violin plot for the width of the 90% prediction interval. Each point is calculated by taking the difference between 95% and 5% prediction interval value.
FIGURE A5  Similar to Figure 8 but using a size of 1000 for each sample. The average group sizes of each variable level for 100 iterations are documented in Table A3. The results appear to be less conclusive, where the use of AR prior does not make a huge difference, which is to be expected in larger sample size, as MRP is found to be the most useful in small-area estimation.
**NHANES data: Coefficients from glmnet output**

| Coefficient | High blood pressure | Dietary sample | Fasting sample | Generated sample |
|-------------|---------------------|----------------|----------------|------------------|
| (Intercept) | 3.81e−01            | −1.13e+00      | −1.25e+00      | −9.55e−01        |
| age18-29   | −9.90e−01           | −1.33e−02      | .              | .                |
| age30-39   | −4.78e−01           | 5.28e−02       | .              | .                |
| age40-49   | −1.23e−01           | .              | .              | −1.35e−01        |
| age50-59   | .                   | .              | .              | .                |
| age60-69   | 4.25e−01            | .              | .              | .                |
| age70-80   | 7.91e−01            | .              | 5.05e−02       | .                |
| ethnicityMexAme | −3.20e−02 | .              | 7.30e−02       | −4.93e−01        |
| ethnicityOtherHisp | . | .              | .              | .                |
| ethnicityNonHispWhite | . | .              | .              | −8.89e−02        |
| ethnicityNonHispBlack | 2.78e−01 | 8.11e−03 | .              | 2.82e−01         |
| ethnicityNonHispAsian | . | .              | .              | .                |
| ethnicityOther | . | .              | .              | .                |
| gendermale | .                   | .              | 7.38e−15       | .                |
| genderfemale | . | .              | .              | .                |
| educless9th | . | .              | .              | .                |
| educ9-11thgrade | . | −4.52e−02 | .              | .                |
| educHSgrad | .                   | .              | .              | .                |
| educSomecollegeAA | . | .              | .              | .                |
| educCollegegrad | . | 4.00e−02     | .              | .                |
| marital_statusmarried/living tgt | . | .              | .              | .                |
| marital_statuswidowed/separated | 1.70e−02 | .              | .              | .                |
| marital_statusnever married | . | .              | .              | .                |
| phys_actno | 5.77e−02            | 4.78e−02       | .              | .                |
| phys_actyes | −1.23e−13 | −7.94e−15 | .              | .                |
| overweightno | −5.00e−01 | −8.16e−02 | −1.18e−02 | −6.11e−02        |
| overweightyes | 2.25e−04 | .              | 6.66e−15       | .                |
| diabetesno | −6.43e−01           | .              | .              | .                |
| diabetesy | 1.37e−01            | 9.37e−03       | .              | .                |
| diabetesborderline | . | −2.81e−01 | .              | .                |
| trb_sleepno | −3.93e−01 | .              | .              | .                |
| trb_sleepyes | 1.24e−14 | .              | .              | .                |
| pov_levellow income | . | −9.32e−03 | .              | .                |
| pov_levelmiddle income | . | .              | .              | −4.83e−01        |
| pov_levelhigh income | . | 1.04e−02 | .              | 4.87e−02         |
| pov_leverfused | . | .              | .              | .                |
| pov_leveldont know | . | .              | .              | 3.38e−01         |
### TABLE A4 (Continued)

|                | High blood pressure | Dietary sample | Fasting sample | Generated sample |
|----------------|---------------------|----------------|----------------|------------------|
| smk_tobcigno   | .                   | .              | .              | .                |
| smk_tobcigyes  | .                   | .              | .              | .                |
| sodium_lvl_low | .                   | −1.09e+00      | .              | .                |
| sodium_lvl_intermediate | . | . | . | −3.15e–01 |
| sodium_lvl_high | . | −3.41e−02 | . | . |
| elst_status_complete | . | 3.53e−02 | . | . |
| elst_status_partial | . | 9.96e−02 | −1.62e−01 | 8.85e−03 |
| elst_status_ineligible | . | . | . | . |
| hiv_test_yes | .                   | .              | .              | .                |
| hiv_test_no | .                   | −8.94e−02      | .              | .                |
| hiv_test_dont_know | . | 1.86e−01 | . | . |
| urn_vollow | .                   | .              | 1.38e−01       | .                |
| urn_vol_med | .                   | .              | 8.94e−04       | .                |
| urn_vol_high | .                   | −6.53e−02      | −3.56e−02      | .                |
| urn_vol_very_high | . | . | −4.52e−02 | . |

### NHANES data: Bias, precision and irrelevant variables

#### TABLE A5
Variables categorised according to Table 1 chosen from lasso fitted on the NHANES synthetic population using blood pressure and each of the sample indicator as the outcome variable.

| Variable type/sample | Bias                                    | Precision                  | Inconsequential                                      | Ignorable                                      |
|----------------------|-----------------------------------------|----------------------------|-------------------------------------------------------|-----------------------------------------------|
| Dietary              | Age, diabetes, overweight, ethnicity, physical activity | Trouble sleeping, marital status | Elastography status, education, ever tested for HIV, poverty level, sodium intake level, urine volume collected | Gender, smokes tobacco or cigarette |
| Fasting              | Overweight, ethnicity                   | Age, diabetes, trouble sleeping, physical activity, marital status | Elastography status, urine volume collected | Education, ever tested for HIV, gender, poverty level, smokes tobacco or cigarette, sodium intake level |
| Generated            | Age, overweight, ethnicity              | Diabetes, trouble sleeping, physical activity, marital status | Elastography status, gender, poverty level, sodium intake level | Education, ever tested for HIV, smokes tobacco or cigarette, urine volume collected |

### NHANES data: Inclusion probability for the generated sample

We create a stronger relationship of the selected variables with the inclusion probability for the generated sample. The selected variables are chosen by selecting half of the variables from the set that are predictive of the outcome and half from the rest of variables.

\[
\text{inclusion probability} = \log^{-1}(0.3, -0.5, -1.5, 0.1, 0.2, 1.3) \times \text{age} + [-0.8, 0.3] \times \text{overweight} \\
+ [-1.8, 0.4, -0.6, 1.7, 0.8, 1.5] \times \text{ethnicity} + [-0.8, 1.4, 2.2, 1.9] \times \text{elastography status}
\]
\[ + [-0.5, 1.3] \times \text{gender} + [0.1, -1.5, 0.7, 1.8, 2.1] \times \text{poverty level} \]
\[ + [0.9, -0.5, 1.4, 1.5] \times \text{sodium intake level} \]

NHANES data: Probability of outcome and proportion of respondents in sample to proportion

**FIGURE A6** Probability of high blood pressure (outcome variable) in each of the variable categories in the population and the ratio of proportion of respondents in sample to population in each of the variable categories in samples.
FIGURE A7  Probability of high blood pressure (outcome variable) in each of the variable categories in the population and the ratio of proportion of respondents in sample to population in each of the variable categories in samples.