Multilevel regression and poststratification as a modelling approach for estimating population quantities in large population health studies: a simulation study

Marnie Downes[*,1,2], **and John B. Carlin**[1,2,3]

[1] Department of Paediatrics, The University of Melbourne, 50 Flemington Road, Parkville, 3052, Victoria, Australia.

[2] Murdoch Children's Research Institute, 50 Flemington Road, Parkville, 3052, Victoria, Australia.

[3] Centre of Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, 235 Bouverie Street, Carlton, 3053, Victoria, Australia.

Received zzz, revised zzz, accepted zzz

There are now a growing number of applications of multilevel regression and poststratification (MRP) in population health and epidemiological studies. MRP uses multilevel regression to model individual survey responses as a function of demographic and geographic covariates. Estimated mean outcome values for each demographic-geographic respondent subtype are then weighted by the proportions of each subtype in the population to produce an overall population-level estimate. We recently reported an extensive case study of a large nationwide survey and found that MRP performed favourably compared to conventional survey sampling weights for the estimation of population descriptive quantities in a highly selected sample. In this study, we aimed to evaluate, by way of a simulation experiment, both the accuracy and precision of MRP versus survey sampling weights in the context of large population health studies. While much of the research into MRP has been focused on U.S. political and social science, we considered an alternative population structure of smaller size and with notably fewer geographic subsets. We explored the impact on MRP performance of sample size, model misspecification, interactions, and the addition of a geographic-level covariate. MRP was found to achieve generally superior performance in both accuracy and precision at both the national and state levels. Results were generally robust to model misspecification and MRP performance was further improved by the inclusion of a geographic--level covariate. These findings offer further evidence that MRP provides a promising analytic approach for addressing participation bias in the estimation of population descriptive quantities from large scale health surveys and cohort studies.

Key words: Multilevel regression and poststratification; Participation bias; Selection bias; Simulation; Survey weighting.

[*] Corresponding author: Marnie Downes, e-mail: marnie.downes@mcri.edu.au, Phone: +61-03-9936-6049, Fax: +61-03-8341-6212

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX

1 Introduction

Multilevel regression and poststratification (MRP) was first described by Gelman and Little (1997) and Park, Gelman and Bafumi (2004), as a model-based approach to estimating a population parameter of interest, typically using data obtained from large-scale surveys. The method comprises two steps. Firstly, multilevel regression is used to model individual survey responses for the outcome measure of interest as a function of demographic and geographic covariates. In the second, poststratification step, the resulting outcome estimates for each demographic-geographic respondent subtype are weighted by the proportions of each subtype in the actual population to produce an overall population-level estimate.

The use of multilevel regression in the first step has two key advantages. First, the multilevel model allows us to consider many more respondent subtypes than would classical methods. This is enabled by the use of "random" or "modelled" effects, rather than "fixed" or "unmodelled" effects (Gelman and Hill, 2007), for all categorical covariates with more than two levels. That is, we assume the effects of the different levels of a covariate are related to each other by way of a common distribution. This ability to make use of more detailed demographic information as well as all information used in the sampling design leads to more accurate poststratification.

The second key advantage is that multilevel regression can be thought of as a method for compromising between the two extremes of excluding a categorial covariate from a model (complete pooling), or estimating target parameters separately within each level of the categorical covariate (no pooling) (Gelman and Hill, 2007). A complete pooling approach ignores variation between levels of a categorical covariate, while a no-pooling analysis overfits the data within each category. A multilevel model partially pools or "shrinks" categorical covariate parameter estimates towards their mean, with the degree of pooling determined from the data. Greater pooling occurs when the variance between categories is small and for categories with small sample sizes. With MRP, this partial pooling means estimates for relatively sparse respondent subtypes, also referred to as poststratification cells, can be improved through "borrowing strength" from demographically similar cells with richer data (Wang et al., 2015).

Poststratification-based estimation, the final step in the MRP approach, corrects for differential participation rates in the poststratification cells by weighting estimates across these cells so that they are more representative of the population of interest. This requires access to detailed data for the target population, specifically, population totals for the poststratification cells, which are defined by the cross-classification of all covariates included in the multilevel model. The use of predominantly demographic and geographic variables to define poststratification cells in MRP is largely driven by the limited overlap in information captured by surveys and publicly available population data.

1.1 Overview of MRP

Formally, the MRP approach defines, for a single outcome measure Y (which may be continuously-valued or binary), a multilevel regression model that specifies a linear predictor for the mean μ_j (or logit transform of the mean for a binary outcome) in poststratification cell *j*:

$$g(\mu_j) = g\left(\mathrm{E}\left[Y_{j[i]}\right]\right) = \beta_0 + \boldsymbol{X}_j^T\boldsymbol{\beta} + \sum_{k=1}^{K} a_{l[j]}^k$$

where $Y_{j[i]}$ is the outcome measurement for respondent i in cell j, β_0 is the fixed intercept, $\boldsymbol{X}_j$ is the unique covariate vector for cell $j = 1, \ldots, J$, $\boldsymbol{\beta}$ represents a vector of regression coefficients (fixed effects) and $a_{l[j]}^k$ the random (or modelled) effects where $l[j]$ maps the cell index j to the appropriate category l of variable k for $k = 1, \ldots, K$. All random effects are modelled using independent normal distributions: $a_l^k \sim \mathrm{N}(0, \sigma_k^2)$, $l = 1, \ldots, L_k$.

Consider a simple example where we have a dichotomous outcome measure and three categorical covariates: age-group (with 3 levels); education (with 4 levels): and region (with 5 levels). Together these covariates produce 3 x 4 x 5 = 60 unique respondent subtypes or poststratification cells. We could fit a multilevel regression model including all main effects and the interaction between age-group and education as follows:

$$\mu_j = \Pr\left(Y_{j[i]} = 1\right) = \mathrm{logit}^{-1}(\beta_0 + a_{l[i]}^{\mathrm{region}} + a_{m[i]}^{\mathrm{age}} + a_{o[i]}^{\mathrm{edu}} + a_{m[i],o[i]}^{\mathrm{age.edu}}),$$

The terms $a_{l[i]}^{\mathrm{region}}$, $a_{m[i]}^{\mathrm{age}}$, $a_{o[i]}^{\mathrm{edu}}$ and $a_{m[i],o[i]}^{\mathrm{age.edu}}$ are random (modelled) effects for the categorical covariates representing region, age, education and the age-by-education interaction. The subscripts indicate category membership of the i-th respondent, for example, $a_{l[i]}^{\mathrm{region}}$ would take values from the set $\{a_1^{\mathrm{region}}, a_2^{\mathrm{region}}, a_3^{\mathrm{region}}, a_4^{\mathrm{region}}, a_5^{\mathrm{region}}\}$. Each random effect is modelled using an independent normal distribution, for example, $a_l^{\mathrm{region}} \sim \mathrm{N}(0, \sigma_{\mathrm{region}}^2)$.

The multilevel regression estimate of a_l^{region}, the effect for a given region *l*, is a weighted average of the mean of the observations in the region and the mean over all the regions, where the weighted average reflects the relative amount of information available about the individual region and all the regions combined. For example, regions with smaller sample sizes carry less information so the weighting pulls the multilevel estimates closer to the overall region average, while regions with larger sample sizes carry more information so the corresponding multilevel estimates are closer to the individual regional averages (Gelman and Hill, 2007).

From the fitted multilevel regression model, outcome estimates, $\hat{\mu}_j$, for each poststratification cell *j*, are obtained. These are then weighted by the corresponding population totals, N_j, and summed to produce a poststratification estimate for the population parameter of interest as follows:

$$\hat{\mu}^{\text{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{\mu}_j}{\sum_{j=1}^{J} N_j}.$$

Similarly, an estimate at any subpopulation level s, can be derived by:

$$\hat{\mu}_s^{\text{PS}} = \frac{\sum_{j \in J_s} N_j \hat{\mu}_j}{\sum_{j \in J_s} N_j},$$

where J_s is the subset of all poststratification cells that comprise s.

1.2 Applications in opinion surveys

Early research into the application of MRP was almost exclusively performed in the context of presidential voting and social research in the United States (U.S.) with the primary aim of estimating state-level public opinion. It has since emerged as a “widely used gold standard for estimating preferences from national surveys” (Selb and Munzert, 2011). A particularly interesting application of MRP performed by Wang et al. (2015), showed the method to be effective in estimating the 2012 U.S. presidential election result using a highly non-representative sample of Xbox gaming users. Despite males aged 18–29 years being greatly overrepresented in the Xbox sample, MRP was able to produce estimates in line with leading traditional representative polls.

There have been several published studies in recent years, that have begun to examine the predictive accuracy of MRP with the intention of identifying the conditions under which it performs best. Two studies by Lax & Phillips (2009, 2013) found that MRP produced reasonably accurate state-level estimates of U.S. public opinion using sample sizes of approximately 1,000 – 1,500 respondents. It was found that the gains in MRP performance relative to survey disaggregation, the conventionally used alternative, were largely due to more accurate modelling of individual responses rather than the partial pooling of states towards the national mean. It was also shown that even fairly simple response models incorporating demographic and geographic covariates performed well.

An examination of a larger number of cases and a greater range of opinions led Buttice & Highton (2013) to draw less optimistic conclusions regarding the use of MRP for U.S. national surveys of typical size ($N \approx 1{,}500$). They found substantial variation in performance and argued that the conditions necessary for MRP to perform well will not always be met.

One of these conditions, widely agreed to be important by previous authors in the U.S context (Buttice and Highton, 2013; Lax and Phillips, 2009, 2013) is the inclusion of geographic-level covariates that account for a substantial amount of the geographic variation in the outcome. Buttice and Highton (2013) concluded that a strong geographic predictor emerges “as a necessary but not sufficient condition for MRP to perform well.”

1.3 Applications in population health and epidemiology

Historically, one of the most widely used methods for adjusting for known or expected discrepancies between sample and population when estimating population descriptive quantities in complex population health or epidemiological surveys has been the use of sampling weights. These weights are generally defined to reflect the number of alike individuals in the population represented by each respondent and are usually calculated as inverse probabilities of survey selection combined with sample-based adjustments for nonresponse. An important distinction between sampling weights and poststratification is that the former are known at the time the survey is designed whereas the latter can only be estimated after the data have been collected (Gelman and Carlin, 2002).

More recently, there has been somewhat of a shift away from this traditional, design-based paradigm in the analysis of surveys, with the current trend towards a combination of design- and model-based approaches(Keiding and Louis, 2016). MRP is one such model-assisted, design-based strategy as it brings together a model-based approach with the use of population information through poststratification.Previous applications of MRP in the context of population health and epidemiological studies have largely focused on producing small-area estimates of prevalence at state and local levels in the United States. Zhang et al. (2014) used MRP for estimation of the prevalence of chronic obstructive pulmonary disease using Behavioral Risk Factor Surveillance System 2011 data, and then further validated the MRP approach for the additional health indicators of current smoking, obesity, diabetes and lack of healthcare coverage (Zhang et al., 2015). Eke et al. (2016) applied MRP to predict rates of periodontitis from National Health and Nutrition Examination Survey (NHANES) 2009–2012 data and Lin et al. (2018) used MRP to estimate the prevalence of untreated dental caries among U.S. children aged 6–9 years from NHANES 2005–2010 data. In each case a fairly simple multilevel model was used including individual demographic covariates of age, gender and race/ethnicity as well as county-level poverty status. All four studies found MRP successful in producing small-area estimates that were reasonable and consistent with gold-standard direct estimates.

Most recently, we reported an extensive case study of a large nationwide longitudinal survey of Australian adult males and found that MRP provided a promising analytic approach to the estimation of population descriptive quantities from a highly selected sample (Downes et al., 2018). While the study employed a sound stratified, multistage cluster sampling design (Currier et al., 2016; Pirkis et al., 2017; Spittal et al., 2016), a participation fraction of 35% for the baseline wave implied considerable potential for non-representativeness or participation bias in the sample obtained. MRP showed greater consistency and precision across population subsets of varying sizes, when compared with using conventional survey weights.

This study aimed to evaluate, by way of a simulation experiment, both the accuracy and precision of MRP versus sampling weights in the context of large population health studies. We explored the impact on MRP performance of sample size, model misspecification, interactions, and the addition of a geographic-level covariate. This simulation study also aimed to extend MRP beyond the estimation of U.S. state-based public opinion by considering an alternative population of smaller size with notably fewer geographic subsets.

2 Methods

This simulation study was informed by an extensive case study focusing on the baseline wave of the *Ten to Men* nationwide survey of Australian males aged 10–55 years (Downes et al., 2018).

2.1 The population

The population of interest for the simulation study mirrored that of the case study. Relevant data were obtained from the 2011 Australian Census of Population and Housing (Australian Bureau of Statistics, 2011) in the form of cross-classification frequency tables, based on the place of usual residence of all persons enumerated. These data were exported from the Australian Bureau of Statistics website and its online tool, TableBuilder.

The following variables were extracted for use as potential geographic and demographic covariates in multilevel modelling: remoteness classification (major cities, inner regional, outer regional; remote and very remote areas were excluded); Australian state or territory (New South Wales: NSW, Victoria: VIC, Queensland: QLD, Western Australia: WA, South Australia: SA, Tasmania: TAS, Australian Capital Territory: ACT, Northern Territory: NT); age-group (18–19 years, 20–24 years, 25–29 years, 30–34 years, 35–39 years, 40–44 years, 45–49 years, 50–55 years); Indigenous status (no, yes); and socio-economic index (based on education and occupation; area-based deciles 1–10). This set of covariates represented a slightly simplified version of that used in the final model for the primary outcome measure reported in the published case study (Downes et al., 2018).

Together, these five variables led, in principle, to 3,840 (3 x 8 x 8 x 2 x 10) unique poststratification cells. However, some of these cells do not exist for structural reasons (e.g. there are no major cities defined in TAS or NT and no outer regional areas in ACT) (Australian Bureau of Statistics, 2013), so the actual number of poststratification cells was 3,200. Of these, approximately 12% were empty in the population. The population size of Australian males aged 18–55 years was 5,090,397. To provide further context for the population of interest, Australia is a country in the Southern Hemisphere occupying land area of similar size to the U.S. Australia is comprised of eight states and territories, which vary greatly in terms of land area and population density (Table 1).

In order to create non-representative samples in which simple estimation methods would produce biased estimates, we needed to define two data generation models, the first to generate the outcome measure and the second to generate survey participation probabilities. Both models were conditional on the same set of covariates outlined above and are described in turn below.

2.2 Generation of population outcome data

The primary outcome used in this simulation study was designed to replicate that from the informing case study, a dichotomous outcome defined as participation in physical activity at levels sufficient to confer a health benefit (Australian Institute of Health and Welfare, 2003). The population prevalence was set at approximately 65%.

Outcome data were generated for all individuals in the population using a multilevel logistic regression model including all the geographic and demographic covariates listed above. This required full specification of all model parameters including: the fixed intercept, fixed linear trends for age-group and socio-economic index, fixed coefficients for the effects of Indigenous status and remoteness classification, and random or modelled effects drawn from normal distributions with mean zero and specified standard deviations for state, as well as for age-group and socio-economic index (representing departures from the linear trends in each case). Values for these model parameters were taken from estimates from the informing case study and in some cases, magnified slightly in order to represent plausible and practically meaningful effect sizes. A listing of all model parameter values used for data generation is provided in Supporting Information Table 1.

The multilevel model produced a probability of outcome "success", p_j, for each unique poststratification cell, j, where all individuals belonging to the same poststratification cell had an equal probability of success. The number of successes (and consequently, failures) in each poststratification cell, j, was therefore generated by taking a random observation from the Binomial distribution $Bin(N_j,p_j)$ where N_j was the known population frequency in poststratification cell, j.

2.3 Simulation and generation of survey participation probabilitiesRandom samples of n individuals were then selected from the population, stratified according to remoteness classification, with oversampling of the inner and outer regional areas (major cities 55%, inner regional 25%, outer regional 20%, compared to major cities 74%, inner regional 17%, outer regional 9% in the population).

A second multilevel logistic regression model, again including all geographic and demographic covariates, was used to simulate survey participation probabilities. All sampled individuals belonging to the same poststratification cell, j, had equal probability of survey participation, q_j and each sampled individual's participation indicator was generated using $Bin(1,q_j)$. Participation probabilities were re-calculated for each simulated dataset.

The magnitudes of the parameters in this model were informed by an analysis of responders versus non-responders in the informing study, where the overall survey participation rate was about one-third. Covariates were limited to geographic variables and socio-economic index. Model parameters were specified such that the direction of all participation biases resulted in underestimation of outcome prevalence. This allowed us to see clearly the impact of these biases on the performance of the various estimation methods rather than the biases operating in different directions and possibly cancelling each other out.

2.4 Analysis of simulated samples

For each scenario under investigation, $S = 500$ simulated samples were created. We estimated the prevalence of the dichotomous outcome measure for the Australian male population aged 18–55 years as a whole and for each Australian state and territory using MRP and also using unweighted and weighted approaches for comparison. The unweighted prevalence estimate was simply the mean of the raw data. The weighted approach incorporated conventional survey weights, calculated as inverse probabilities of survey selection adjusted for the observed (simulated) participation. Consistent with the informing case study, only geographic covariates were used to construct the sampling weights and as a result, there were 20 unique weighting values, one corresponding to each state-by-remoteness classification combination.

Multilevel models for the MRP approach were fitted using approximate marginal maximum likelihood as implemented in the `glmer()` function of the `lme4` package (Bates, Maechler and Bolker, 2013) in `R`. Ideally, we would have performed a full Bayesian analysis in `RStan` (Stan Development Team, 2016), but this was not feasible due to the excessive time required to obtain complex posterior distributions over many simulated samples. As a result, it was not possible to calculate estimated standard errors or 95% credible intervals for MRP national and state-level prevalence estimates. The empirical standard error (SE) of the simulated distribution of prevalence estimates was therefore used as a basis for comparing the three estimation methods in terms of precision. Bias was considered the primary performance measure of interest and was estimated using the average difference of prevalence estimates from known true population values.

This simulation framework allowed us to investigate the influence of several factors on MRP performance for estimating population descriptive quantities. We considered three sample sizes: $n =$ 30,000, 15,000 and 4,500, which, after a participation fraction of approximately one-third, resulted in target sample sizes of 10,000, 5,000 and 1,500 respectively. The largest target sample size of 10,000 was comparable to the informing case study while the smallest target sample size of 1,500 reflected the typical size of surveys of public opinion (Buttice and Highton, 2013; Lax and Phillips, 2009, 2013).

At the analysis stage, we first considered the "correct" specification of the multilevel model, that is, the multilevel model including the full set of demographic and geographic covariates used to generate the outcome and survey participation indicators. We then investigated the impact of model misspecification by omitting the age-group covariate when fitting MRP to the simulated datasets.

We explored the inclusion of a random effect representing an interaction between state and remoteness classification into the outcome generation and sampling processes and then the impact of including and excluding this interaction in the MRP analysis. We first defined a standard deviation parameter for this state-by-remoteness interaction equal to that observed in the informing case study and then a second scenario where this standard deviation was doubled.

Finally, we also examined the potential role of a state-level covariate. Additional between-state variation was first introduced by specifying a large between-state standard deviation in both the outcome generation and survey participation models. Since this represented random between-state variation, models including and excluding a state-level covariate were expected to perform equally well. Next, we retained a small between-state standard deviation but introduced additional between-state variation by specifying, in both outcome and survey participation data generation models, a state-level covariate effect. The choice of state-level covariate was not limited to data collected from the survey. Given the outcome measure of interest aimed to replicate participation in sufficient physical activity, we hypothesised that an appropriate state-level covariate might be mean global solar exposure (MJ/m^2) obtained for the capital city of each state and territory (Australian Government Bureau of Meteorology, 2018). As Table 1 shows, Australia is a vast country with minimum and maximum latitude values of –43.64° and –10.69° respectively (Australian Government Geoscience Australia, 2019). As a result, different parts of the country experience very diverse weather conditions. We therefore deemed it appropriate to define an association between the state-level covariate of global solar exposure and participation in sufficient physical activity, specifically, a quadratic association, where both low and high levels of solar exposure were associated with reduced participation (see Supporting Information Table 1 for model parameter values used).

3 Results

Across the set of five simulation scenarios, the true national population prevalence ranged from 63.4% to 65.6% with an overall average of 64.6%. True prevalence at the state-level ranged from 51.6% to 76.7%. All results reported and plotted below represent bias, in absolute percentage points, from these true prevalence values.

The first scenario considered in the simulation study involved a large target sample size of 10,000 and a correct MRP model fit including all the geographic and demographic covariates of remoteness classification, age-group, state or territory, Indigenous status and socio-economic index. The mean sample size by state ranged from 95 in ACT and 160 in NT to 3,208 in NSW.

Estimated bias (%) and the associated standard error for national and state-level prevalence estimates using the three estimation approaches for this scenario are shown in Figure 1A. At the national level, the correct MRP model resulted in the smallest bias (–0.09%, SE=0.51%), followed by the weighted approach (–2.5%, SE=0.54%) and then the unweighted approach (–6.6%, SE=0.50%). The same pattern of results was observed at the state level, particularly for the larger states of NSW, VIC, QLD, WA and SA. For the smaller states and territories (TAS, ACT and NT), the biases of the unweighted and weighted approaches were more similar to each other due to these states containing only a subset of the three remoteness strata. For NT, the unweighted and weighted results were in fact identical since this state consisted of only outer regional areas (no major cities or inner regional areas) so all sampled individuals from this state received the same sampling weight. Nonetheless, for each of TAS, ACT and NT, the correct MRP model was still superior to both weighted and unweighted approaches. The estimated bias of the correct MRP model was at most 0.62% across all states and territories.

Figure 1A also shows that MRP achieved greater precision (smaller standard errors) than the unweighted and weighted approaches, with this gain becoming more pronounced as the state-level sample size decreases.

3.1 Varying sample size

Figures 1B and 1C show results for the smaller target sample sizes of 5,000 and 1,500 respectively. For these two scenarios, the mean sample size by state ranged from 48 and 15, respectively, in ACT, to 1,604 and 482 in NSW. At the national level, while the bias was reasonably consistent across the three sample sizes for both the unweighted and weighted approaches (approximately –6.6% and –2.5% respectively), the MRP approach showed slight increases in bias with decreasing sample size suggesting that performance is improved with larger sample sizes. The bias for the correct MRP model at the national level was –0.51% (SD=1.4%) for a target sample size of 1,500 and –0.17% (SD=0.71%) for a target sample size of 5,000 compared to –0.09% (SD=0.51%)

for a target sample size of 10,000. As expected, all methods showed reduced precision for smaller sample sizes.

At the state level, in almost all cases, MRP demonstrated superior performance in terms of both accuracy and precision relative to the unweighted and weighted approaches at each sample size. There was, however, some evidence of too much shrinkage toward the mean, or over-pooling, by the multilevel regression model for the smaller states and territories at the smallest sample size. For NT and TAS, regions with the lowest true prevalence, MRP resulted in a considerable positive bias (1.8%, SE=2.9% and 1.0%, SE=2.1% respectively) suggesting too much partial-pooling or shrinkage towards the national estimate and the other state estimates with larger sample sizes.

3.2 Model misspecification

Figures 1A, 1B and 1C also show the bias and associated standard error for national and state-level prevalence estimates by sample size when the MRP model was misspecified by omitting the age-group effect. For the largest sample size of 10,000, the bias at the national level (–1.1%, SE=0.53%) was larger than when the correct model was fitted (–0.09%, SE=0.51%), but still considerably less than the unweighted and weighted approaches. Very similar results were observed at the state level with the correct MRP model consistently superior to the misspecified MRP model. Figure 2 shows the bias and standard error by levels of age-group, the variable excluded from the misspecified MRP model. By not allowing for the strong negative linear association between age-group and outcome, the misspecified MRP model significantly underestimated the outcome prevalence for the younger age-groups and overestimated for the older age-groups.

3.3 Interactions

The results of including a state-by-remoteness interaction as a random effect in both the true outcome and participation models are summarised in Figure 3. There were minimal differences in bias between the MRP models including and excluding the interaction at the national level, for both interaction effect sizes and target sample sizes of 10,000 and 1,500. The two models were also reasonably consistent across the states and territories, but there was some evidence of improved performance when including the interaction term for the largest sample size.

Standard errors for MRP were slightly larger when the interaction term was included, reflecting a reduction in the amount of partial pooling or “borrowing strength” between states. Results (not shown) by age-group, remoteness classification and socio-economic index also indicated that the inclusion of an interaction term did not greatly affect MRP performance for estimating outcome prevalence for the population as a whole and by levels of main effects.

3.4 State-level covariates

Figures 4A and 4B compare the performance of MRP conducted with and without the inclusion of a state-level covariate, monthly mean solar exposure, in a scenario that introduced a large amount of random between-state variation for target sample sizes of 10,000 and 1,500 respectively. As expected, there was very little difference, particularly at the national level and for the larger states.

In the final scenario, where greater between-state variation was introduced by way of a quadratic association with global solar exposure, the inclusion of this (correctly specified) state-level covariate in the MRP model was found to considerably improve the accuracy of prevalence estimates for the smaller states of TAS, ACT and NT (Figures 4C and 4D for target sample sizes of 10,000 and 1,500 respectively) by reducing the amount of over-pooling. For example, with a sample size of 10,000, bias reduced in magnitude from 2.2% (SE=2.0%) to 0.06% (SE=1.7%) for TAS, from –1.4% (SE=1.2%) to –0.03% (SE=0.9%) for ACT and from 4.1% (SE=2.0%) to –0.23% (SE=2.6%) for NT. The inclusion of the state-level covariate had little impact on MRP performance at the national level and for the larger states, mostly due to there being little room for improvement beyond already very accurate estimates.

3.5 Monte Carlo standard errors

Monte Carlo standard errors (MCSEs) for the primary performance measure of bias (Morris, White and Crowther, 2019) were calculated to be no more than 0.07% across all prevalence estimates at the national level. At the state level, MCSEs for the MRP approach were less than 0.15% for the larger states and no more than 0.35% for the smaller states, while for the weighted and unweighted approaches, MCSEs were less than 0.25% for the larger states and no greater than 0.65% for the smaller states. This was considered an acceptable level of precision. Uncertainty in the outcome data generated for the population, namely the randomly chosen values for random effects in the outcome generation multilevel model, was not accounted for in the simulations as our intention was to characterise inference from a fixed or finite population. However, a small number of repeated simulation runs with different random number generator seeds and hence, different random effect values, confirmed consistent overall conclusions.

4 Discussion

We conducted a simulation study to evaluate the accuracy and precision of multilevel regression and poststratification (MRP) compared with survey weighting for the estimation of population descriptive quantities in the presence of participation bias within large population health studies. We explored the impact on MRP performance of sample size, model misspecification, interactions, and the addition of a geographic-level covariate. MRP was found to achieve generally superior performance in both accuracy and precision at both the national and state levels, followed by weighted and then unweighted approaches. MRP was generally robust to model misspecification, but had a tendency to over-pool between-state variation in the outcome, particularly for smaller states and when sample sizes were small.

Not surprisingly, MRP performed better when the multilevel model was correctly specified, particularly when obtaining outcome estimates for levels of a covariate incorrectly excluded from the model. General advice on variable selection for model-based estimation approaches is often to include all variables thought to have an important impact on sampling and nonresponse, if they are also potentially predictive of the outcome of interest (Gelman, 2007). While it is unlikely that one could ever specify the true correct model, MRP should make it possible to get closer to this by allowing the inclusion of a larger number of covariates as modelled rather than fixed effects. When determining which variables to include in a MRP model, we recommend, as common sense suggests, to include or at least investigate all variables for which outcome estimates by each level separately are desired. Failure to do so could result in wildly inaccurate estimates, as was found when outcome estimates were obtained by age-group when it had been excluded from the MRP model.

While small improvements in MRP performance were seen with larger sample sizes, very accurate results were achieved at the national level and for the larger states with the smallest target sample size of 1,500. Outcome prevalence estimates at the national level were very robust to model misspecification due to the exclusion of important main effects, interactions and geographic-level covariates. Studies where it is of primary interest to estimate descriptive quantities for the target population as a whole, can therefore expect to achieve accurate results with MRP using fairly simple demographic-geographic response models. For studies where population subsets such as geographic regions are also of interest, more detailed consideration is recommended in determining the demographic and geographic covariates for inclusion into the MRP model.

Lax and Phillips (2013) found that the inclusion of interactions between individual-level demographic variables was not necessary for small samples and the results of this study support this finding. However, we did find some benefit from including interactions for larger sample sizes, where the interaction effects are estimated more precisely with more data available.

In the original analysis of the motivating case study, there was very little variation observed between states and territories after the influence of individual-level demographic variables was accounted for. It was, therefore, assumed that the addition of a state-level covariate would not have made much difference to the results (Downes et al., 2018). When additional between-state variation was introduced into the simulated data through a strong state-level covariate, however, improved MRP performance was found, particularly for the smaller states where the inclusion of the state-level covariate appeared to lessen the extent of the over-pooling or "borrowing strength" between states. This reduction in partial pooling between states was also evident in the slight increase in the standard errors of the prevalence estimates when the state-level covariate was included in the model.

One key difference between MRP and the use of conventional survey weights is that a single set of weights is generally constructed and applied to all analyses, while the application of MRP can be tailored specifically for each outcome measure. While being potentially time consuming and more demanding to apply in practice, this approach offers great flexibility when multiple outcomes are of interest. Warshaw and Rodden (2012) advocated for "optimising an MRP model for a particular research question" and this was strongly endorsed by Lax and Phillips (2013) particularly with regard to selection of geographic-level covariates, where covariates that work well for one outcome may not work well for another and vice versa. The choice of geographic-level predictors is also made difficult by the fact that they are not limited to data collected from the survey: they often come from external sources, which opens up a very large number of possibilities. While this can make model selection challenging, it also provides great potential to improve outcome estimates by explaining geographic-level variation using informative covariates not necessarily considered at the time of study design.

This simulation study has not considered any clustering within the survey sampling process. Multi-stage surveys, particularly those with many clusters each containing only a small number of participants appear to be a challenge for the MRP approach. Model coefficients representing cluster random effects would be estimated based on small numbers of observations and it is unclear how best to generate cluster random effect coefficients for unsampled clusters in the poststratification step. While the informing case study employed a multi-stage sampling design including a large number of clusters representing small geographical areas, implementation of MRP ignoring this clustering still produced acceptably accurate and precise population estimates. More work is required, however, to fully understand the value of MRP in the presence of clustering.

Our findings were largely consistent with the existing literature in political science, that respectable estimates could be achieved by MRP with a sample size as small as 1,500 (Buttice and Highton, 2013; Lax and Phillips, 2009, 2013) One important difference in our simulations, was that a sample size of 1,500 was distributed across only eight Australian states and territories rather than 50 U.S. states. This same sample size dispersed across fewer geographic regions implies greater precision of state-level estimates and it is possible that reasonable estimates could be achieved in the Australian context with an even smaller sample size. While it was originally hypothesised that the addition of a geographic-level covariate would be of lesser value with only eight states and less between-state variation to explain, this was found not to be the case, leading us to conclude that MRP performance may be largely comparable when applied to these two differing population contexts.

Results of this simulation study indicate that MRP provides generally superior performance in both accuracy and precision relative to the use of conventional survey weights for addressing potential participation bias in the estimation of population descriptive quantities from large scale health surveys. Future research may involve performing similar simulations for populations with differing geographical structures or developing some user-friendly software tools to facilitate more widespread usage of this method. We may also consider the application of MRP to more complex problems such as estimating changes in prevalence over time in a longitudinal study.

Figures

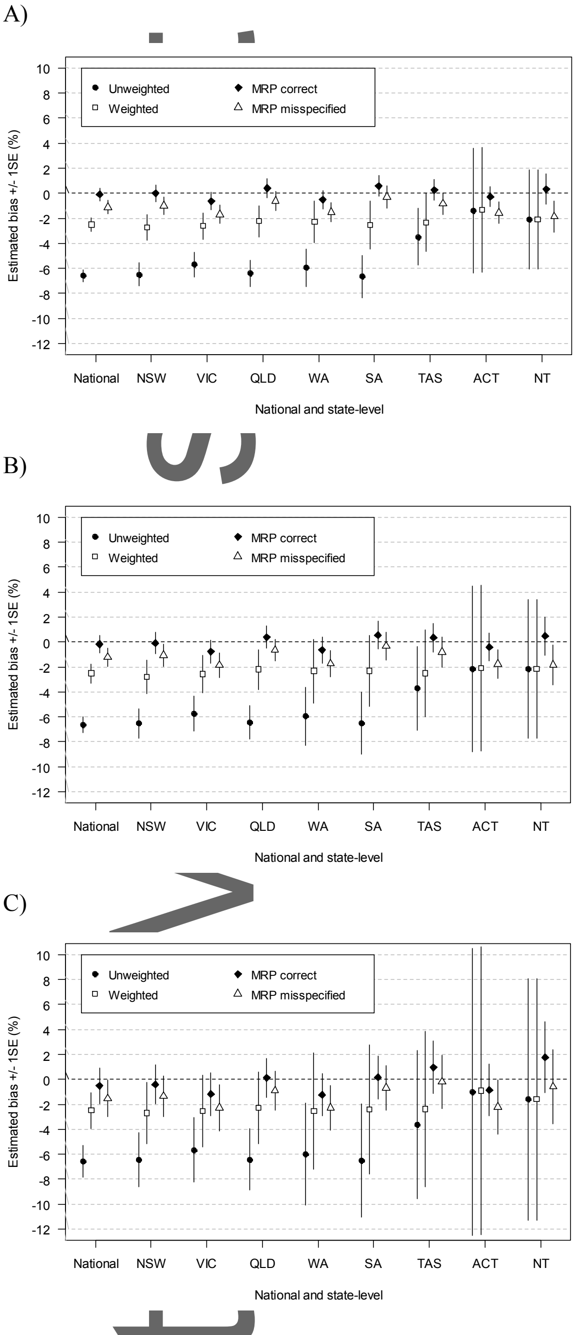

Figure 1 Estimated bias (%) and associated empirical standard error (SE) for national and state-level prevalence estimates using: i) unweighted estimation; ii) survey sampling weights; iii) the correct MRP model; and iv) a misspecified MRP model (excluding the age-group main effect) for the following target sample sizes: A) n =10,000; B) n =5,000; C) n =1,500.

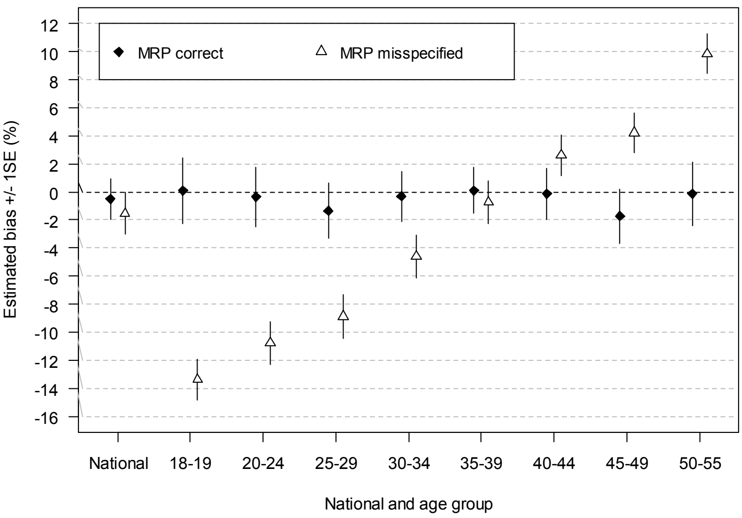

Figure 2 Estimated bias (%) and associated empirical standard error (SE) for prevalence estimates by age-group, using: i) the correct MRP model (including the age-group main effect); and ii) a misspecified MRP model (excluding the age-group main effect) for a target sample size of n =1,500.

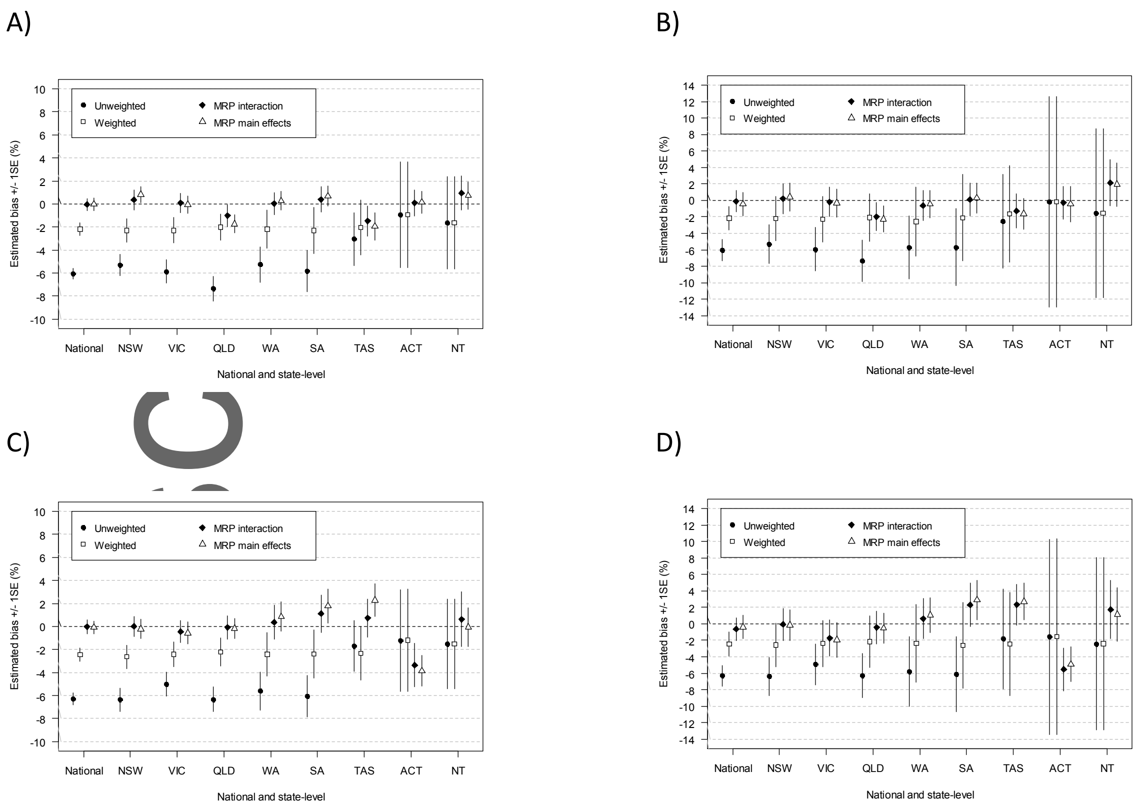

Figure 3 Estimated bias (%) and associated empirical standard error (SE) for national and state-level prevalence estimates using: i) unweighted estimation; ii) survey sampling weights; iii) the correct MRP model including a state-by-remoteness interaction effect ($\sigma_{\text{state.remote}}$ = 0.08 (size 1×TTM; panels A and B), $\sigma_{\text{state.remote}}$ = 0.15 (size 2×TTM; panels C and D)); and iv) a misspecified MRP model (main effects only, no interaction) for target sample sizes of n =10,000 (panels A and C) and n =1,500 (panels B and D).

]

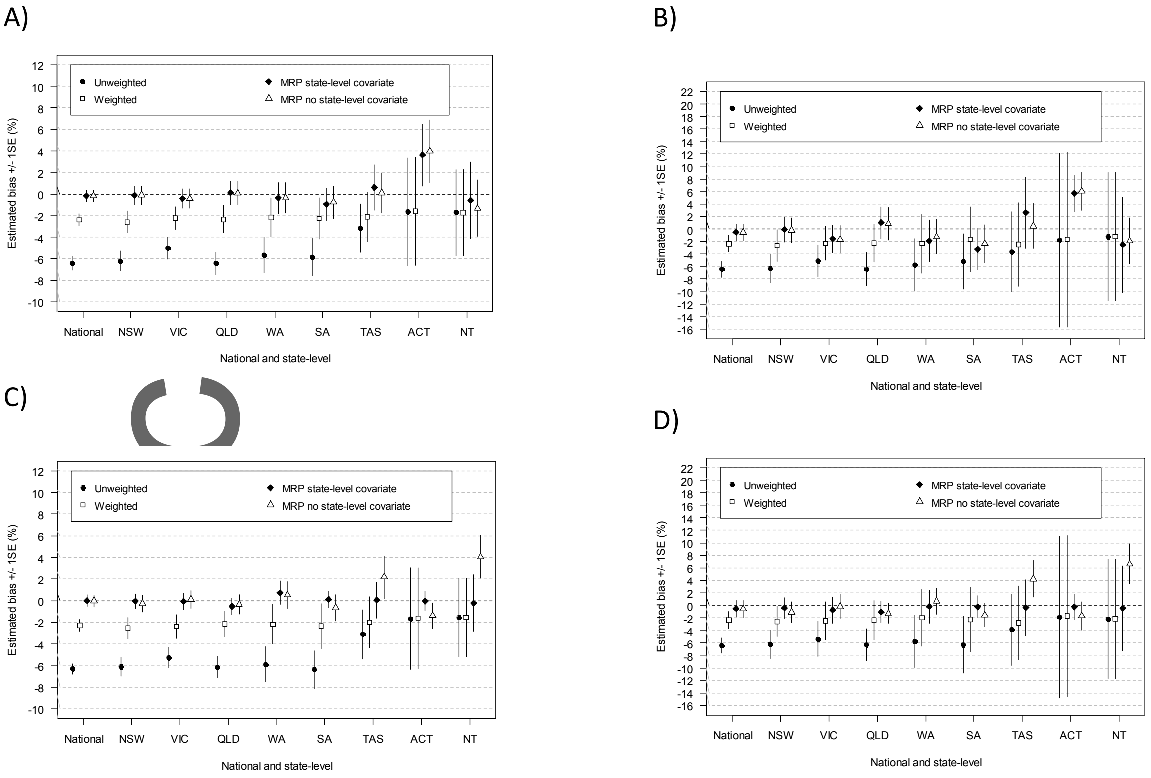

Figure 4 Estimated bias (%) and associated empirical standard error (SE) for national and state-level prevalence estimates using: i) unweighted estimation; ii) survey sampling weights; iii) the correct MRP model including a state-level covariate; and iv) a misspecified MRP model excluding a state-level covariate, for two scenarios: increased random between-state variation (panels A and B) and a quadratic association with a state-level covariate (panels C and D) for target sample sizes of n =10,000 (panels A and C) and n =1,500 (panels B and D).

Tables

Table 1 Background information for each Australian state and territory including: population size (males aged 18–55 years); population size (total); land area (km^2); latitude of capital city (degrees, ^{o}); % of state population living in the capital city; and solar exposure of capital city (MJ/m^2).

	Population size[1,2]	Population size[1]	Land area[3]	Latitude of capital city[4]	% state population living in capital city[5]	Solar exposure of capital city[6,7]
Australian state/territory	(Males aged 18-55)	(Total)	(km^2)	Degrees (^{o})		(MJ/m^2)
New South Wales	1,641,497	6,917,656	801,150	−33.86	63.5%	16.4
Victoria	1,323,496	5,354,039	227,444	−37.81	74.7%	15.1
Queensland	1,005,253	4,332,737	1,729,742	−27.47	47.7%	18.5
Western Australia	514,458	2,239,171	2,527,013	−31.95	77.2%	19.3
South Australia	370,425	1,596,569	984,321	−34.93	76.7%	17.3
Tasmania	109,454	495,351	68,401	−42.88	42.7%	13.6
Australian Capital Territory	94,310	357,218	2,425	−35.28	99.8%	17.2
Northern Territory	31,504	211,943	1,347,791	−12.46	56.9%	21.2
Australia (Total)	5,090,397	21,504,684	7,688,287			

[1] Australian Bureau of Statistics, 2011 Australian Census, Census TableBuilder online.

[2] Including major cities, inner regional, outer regional areas only, and non-missing values for age, indigenous status and SEIFA index

[3] Australian Government Geoscience Australia. https://www.ga.gov.au/scientific-topics/national-location-information/dimensions/area-of-australia-states-and-territories. Accessed March 17, 2019.

[4] simplemaps Geographic Data Products. https://simplemaps.com/data/au-cities. Accessed March 17, 2019.

[5] Australian Bureau of Statistics, 2011 Census QuickStats. http://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2011/quickstat/0. Accessed March 17, 2019.

[6] Monthly mean global solar exposure, 1990–2018.

[7] Australian Government Bureau of Meteorology, Climate Data Online. http://www.bom.gov.au/climate/data/index.shtml. Accessed July 4, 2018.

Conflict of Interest

The authors have declared no conflict of interest.

References

Australian Bureau of Statistics (2011). How Australia takes a Census 2011. Canberra, ACT: Australian Bureau of Statistics; 2011 Apr. Report No.: 2903.0.

Australian Bureau of Statistics (2013). Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure, July 2011. Canberra, ACT: Australian Bureau of Statistics; 2013 Jan. Report No.: 1270.0.55.005.

Australian Government Bureau of Meteorology (2018). Climate Data Online. http://www.bom.gov.au/climate/data/index.shtml. Accessed July 4, 2018.

Australian Government Geoscience Australia (2019). https://www.ga.gov.au/scientific-topics/national-location-information/dimensions/area-of-australia-states-and-territories. Accessed March 17, 2019.

Australian Institute of Health and Welfare (2003). The Active Australia Survey: a guide and manual for implementation, analysis and reporting. 2003, Canberra, Australia: Australian Institute of Health and Welfare.

Bates, D., Maechler, M., and Bolker, B. (2013). lme4: Linear mixed-effects models using S4 classes. 2013. URL: http://CRAN.R-project.org/package=lme4. R package version 0.999999-2.

Buttice, M. K., and Highton, B. (2013). How does Multilevel Regression and Poststratification perform with conventional national surveys? *Political Analysis* **21**, 449-467.

Currier, D., Pirkis, J., Carlin, J., and al., e. (2016). The Australian Longitudinal Study on Male Health - Methods *BMC Public Health* **16**(Suppl 3:1043), 6-13.

Downes, M., Gurrin, L., English, D*., et al.* (2018). Multilevel Regression and Poststratification: A Modelling Approach to Estimating Population Quantities From Highly Selected Survey Samples. *American Journal of Epidemiology* **187**, 1780-1790.

Eke, P. I., Zhang, X., Lu, H*., et al.* (2016). Predicting periodontitis at state and local levels in the United States. *Journal of Dental Research* **95**, 515-522.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science* **22**, 153-164.

Gelman, A., and Carlin, J. (2002). Poststratification and weighting adjustments. In: Survey Nonresponse (R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little, eds.) 289-302. Wiley, New York. .

Gelman, A., and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge ; New York: Cambridge University Press.

Gelman, A., and Little, T., C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23**, 127-135.

Keiding, N., and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society Series a-Statistics in Society* **179**, 319-376.

Lax, J. R., and Phillips, J. H. (2009). How should we estimate public opinion in The States? *American Journal of Political Science* **53**, 107-121.

Lax, J. R., and Phillips, J. H. (2013). How should we estimate subnational opinion Using MRP? Preliminary findings and recommendations. Presented at Midwest Political Science Association. (2013) http://www.columbia.edu/~jrl2124/mrp2.pdf

Lin, M., Zhang, X., Holt, J. B., Robison, V., Li, C. H., and Griffin, S. O. (2018). Multilevel model to estimate county-level untreated dental caries among US children aged 6-9years using the National Health and Nutrition Examination Survey. *Preventive Medicine* **111**, 291-298.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 1-29.

Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* **12**, 375-385.

Pirkis, J., Currier, D., Carlin, J., and al., e. (2017). Cohort profile: Ten to Men (The Australian Longitudinal Study on Male Health). *International Journal of Epidemiology* **46**, 793-794i.

Selb, P., and Munzert, S. (2011). Estimating constituency preferences from sparse survey data using auxillary geographic information *Political Analysis* **19**, 455-470.

Spittal, M., Carlin, J., Currier, D., *et al.* (2016). The Australian Longitudinal Study on Male Health sampling design and survey weighting: Implications for analysis and interpretation of clustered data. *BMC Public Health* **16**(Suppl 3:1043), 15-22.

Stan Development Team (2016). RStan: the R interface to Stan, Version 2.9.0. 2016. URL: http://mc-stan.org.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting* **31**, 980-991.

Warshaw, C., and Rodden, J. (2012). How should we measure district-level public opinion on individual issues? *The Journal of Politics* **74**, 203-219.

Zhang, X., Holt, J. B., Lu, H., *et al.* (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *American Journal of Epidemiology* **179**, 1025-1033.

Zhang, X., Holt, J. B., Yun, S., Lu, H., Greenlund, K. J., and Croft, J. B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology* **182**, 127-137.