Survey Research Methods (2024)
Vol. 18, No. 3, pp. 273-287
https://doi.org/10.18148/srm/2024.v18i3.8188
European Survey Research Association

An assessment of the utility of a Bayesian framework to improve response propensity modes in longitudinal data

Eliud Kibuchi[1] · Gabriele B. Durrant[2] · Olga Maslovskaya[2] ·
Patrick Sturgis[3]
[1]University of Glasgow
[2]University of Southampton
[3]The London School of Economic and Political Science

Response propensity (RP) models are widely used in survey research to analyse response processes. One application is to predict sample members who are likely to be survey nonrespondents. The potential nonrespondents can then be targeted using responsive and adaptive strategies with the aim of increasing response rates and reducing survey costs. Generally, however, RP models exhibit low predictive power, which limits their effective application in survey research to improve data collection. This paper explores whether the use of a Bayesian framework can improve the predictions of response propensity models in longitudinal data. In the Bayesian approach existing knowledge regarding model parameters is used to specify prior distributions. In this paper we apply this approach and analyse data from the UK household longitudinal study, Understanding Society (first five waves) and estimate informative priors from previous waves data. We use estimates from RP models fitted to response outcomes from earlier waves as our source for specifying prior distributions. Our findings indicate that conditioning on previous wave data leads to negligible improvement of the response propensity models' predictive power and discriminative ability.

Keywords: response propensity models; Bayesian; informative priors; nonresponse

1 Introduction

It has become more difficult in recent years to conduct social surveys because of an increase in nonresponse rates and survey costs (Carlson & Williams, 2001; de Leeuw & de Heer, 2002; Luiten, Hox and de Leeuw 2020). Increasing nonresponse rates reduce stakeholder confidence in the ability of surveys to inform public policy due to concerns about the representativeness of samples and the generalisability of findings to wider populations. Therefore, survey researchers are keen to understand and address the factors which influence nonresponse. Such factors include socioeconomic and demographic attributes of members of the public (Gjonça & Calderwood, 2004; Goldberg et al., 2001), salience of survey topics (Groves et al., 1992), and survey design characteristics (Fan & Yan, 2010; Moss, 1981).

An increase in nonresponse rates has resulted in interest among survey practitioners in developing improved understanding of nonresponse behaviour, chiefly with response propensity (RP) models (Särndal & Swensson, 1987). For example, Olson & Groves (2012) employed RP models to predict changes of individual response propensities under responsive and adaptive strategies. Durrant, et al. (2015) showed that the predictive power of RP models for final call outcome and length of call sequence improves when information from most recent calls is included as explanatory variables. However, the explanatory power of RP models in terms of pseudo R^2 tends to be low and ranges between 2 and 8 % (Fricker & Tourangeau, 2010; Kreuter & Olson, 2011; Olson & Groves, 2012). Therefore, ways of improving the predictive power of RP models is an active and important area of research in survey design.

Some of the steps taken to improve the predictive power of RP models in responsive and adaptive designs relate to collection and/or use of new auxiliary data such as paradata which are data about the survey process (Biemer et al., 2013; Durrant et al., 2015 and 2017; Kreuter, 2013; West, 2011). Alternatively, researchers can potentially improve prediction accuracy through implementation of more suit-

This article (https://doi.org/10.18148/srm/2024.v18i3.8188) contains supplementary material.

Corresponding author: Eliud Kibuchi, University of Glasgow, Glasgow, United Kingdom (Email: Eliud.Kibuchi@glasgow.ac.uk)

able statistical models. One possibility in the latter context is the use of a Bayesian approach which utilises informative priors to improve the prediction model (Coffey et al., 2020; Coffey & Elliott, 2023; Schouten et al., 2018; Wagner et al., 2023; West et al., 2023; Wu et al., 2022). For example, West et al. (2023) have shown that using historical data in a Bayesian setting can improve predictions of RP models in the early or middle periods of data collection in cross-sectional studies.

This paper aims to extend our understanding of the utility of the Bayesian framework in maximising prediction accuracy in RP models. We investigate whether using informative priors derived from the coefficients from models fitted to previous waves can improve the predictions of final call outcome in the current or subsequent waves of a longitudinal survey. In the context of longitudinal surveys, informative priors from previous waves can, in principle, pull the estimates of the current wave closer to the accurate values which may lead to improved bias in estimates (Fearn et al., 2004; Simon, 2009). Additionally, in a situation where both the informative prior and current data can accurately estimate the parameters of interest, their combination has the potential to reduce the mean squared error (MSE) due to reduction in the effects of sampling variation (i.e., shrinkage) (Fearn et al., 2004; Simon, 2009).

Our primary objective, then, is to assess whether the use of informative priors derived from models fitted to previous waves in a longitudinal survey can improve prediction accuracy in RP models. If this can be achieved, the model predictions can be used for better planning of fieldwork at the next wave, for example, by for example targeting monetary incentives, household communication strategies and sending more experienced interviewers to low propensity households. Model performance is assessed using a range of evaluation criteria such as Watanabe Akaike Information Criterion (WAIC), sensitivity, specificity, area under the receiver operating characteristic (ROC) curve, and positive and negative predicted values. Data from the UK Household Longitudinal Study Understanding Society Waves 1–5 are used for the analysis (University of Essex et al. 2016).

The remainder of this paper is structured as follows: We first provide the background and motivation for the study. We then describe the Understanding Society survey that form the basis of our analysis and the approach used to construct the analysis samples. This is followed by the description of the analysis methodology and presentation of our key findings. We conclude by a discussion of the implications of our findings for survey practice, acknowledgement of the limitations of our study and suggestions for future research.

2 Background and motivation

RP models were introduced by David, Little, Samuhel, & Triest (1983) who extended the propensity score theory of Rosenbaum & Rubin (1983) as a tool for evaluating nonresponse behaviour in surveys. RP models produce a single score which summarises the likelihood of a sample member responding to a survey request as a function of variables that are observed for both respondents and nonrespondents (Kalton & Flores-Cervantes, 2003). Traditionally, the method for estimating response propensities is a logistic regression model, where the outcome is a binary indicator of survey response. The fitted probabilities from such a model are the response propensities and these have been used for a variety of purposes, including obtaining a better understanding of nonresponse, and associated mechanisms (Durrant & Steele, 2009), developing nonresponse weights (Little, 1986), calculating representativeness indicators such as R-indicators and coefficients of variation (CVs) (Schouten & Cobben, 2007), predicting response outcomes either during or right at the end of data collection (Durrant et al., 2011, 2013, 2015, 2017), and providing information to target interventions for adaptive and responsive survey designs (Coffey & Elliott, 2023; Groves & Heeringa, 2006; Wagner et al., 2023).

The effectiveness of RP models in helping survey researchers implement fieldwork decisions is, however, hindered by their generally low predictive power (Brick & Montaquila, 2009; Kreuter, Olson, et al., 2010; Olson et al., 2012; Olson & Groves, 2012). For example, a RP model developed by Olson & Groves (2012) had a pseudo R^2 of 2%. Olson, Smyth, & Wood (2012) investigated the effect of respondents' choice on their preferred survey mode using RP models and obtained pseudo R^2 ranging between 3.2 and 8%. The low predictive power of these models is primarily due to the use of auxiliary variables which are not strongly correlated with response outcomes because more strongly correlated variables are generally not available to be linked to the sampling frame (Brick & Montaquila, 2009; Kreuter, Olson, et al., 2010).

One of the strategies adopted by researchers to improve the fit of RP models involves the collection of new kinds of auxiliary variables and paradata to be used as predictors (Biemer et al., 2013; Blom, 2009; Peytcheva & Groves, 2009; Sinibaldi & Eckman, 2015; Sinibaldi, Trappmann, & Kreuter, 2014). For example, Durrant et al. (2015, and 2017) found that the inclusion of call record variables, especially from the most recent calls, improved the predictive power of RP models from 9 to 26%. Sinibaldi & Eckman (2015) used interviewer observations at call level and found an improvement of in terms of both pseudo R^2 and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curves. Likewise, Blom (2009) showed that the predictive power of RP models improved when demographic variables are combined with paradata using European Social Survey (ESS) data for nonresponse adjustment.

Historically, RP modelling has been implemented within a frequentist statistical framework (Durrant et al., 2015, 2017; Olson et al., 2012; Olson & Groves, 2012; Sinibaldi & Eckman, 2015). However, Bayesian models are increasingly being implemented and hold promise for improvements to prediction accuracy (Coffey et al., 2020; Coffey & Elliott, 2023; Schouten et al., 2018; Wagner et al., 2023; West et al., 2023; Wu et al., 2022). The main differences between frequentist and Bayesian frameworks are the treatment of the observed data and the interpretation of uncertainty (Fearn et al., 2004; Simon, 2009; Skorczynski, 2012; Wagenmakers et al., 2008). Statistical inferences based on the frequentist framework make probability statements about random events by assuming that parameters are fixed, and the data are random in which any uncertainty is due to randomness, while a Bayesian model treats all unknown quantities as random variables and represents uncertainty over those quantities using probability distributions (Fearn et al., 2004). In addition, Bayesian inferences are exact since they are conditioned on observed data satisfying the likelihood principle, unlike frequentist inferences that rely on asymptotic approximations (Steel, 2007).

The starting point of Bayesian analysis is expressing prior knowledge about unknown parameters in the form of a prior distribution (Zyphur & Oswald, 2015). The observed data are then combined with this prior distribution using Bayes' theorem to obtain an updated prior in the form of a posterior distribution (Fearn et al., 2004; Gill, 2014; Simon, 2009; Zyphur & Oswald, 2015). In many practical situations, there is little or no previous knowledge on the phenomenon of interest. This leads to the specification of 'vague' or 'noninformative' prior distributions that have minimal influence on the posterior estimates (Gill, 2014). However, when researchers have some pre-existing knowledge about the parameters of interest, it is possible to specify 'informative' prior distributions (Coffey et al., 2020; Coffey & Elliott, 2023; Gill, 2014; Schouten et al., 2018; Wagner et al., 2023; West et al., 2023; Wu et al., 2022). Information to specify informative priors can be derived from a range of sources including, but not limited to existing studies, expert opinions, and pilot studies (Gill, 2014; West et al., 2023; Wu et al., 2022).

In the context of a longitudinal survey, posterior distributions that summarise knowledge on the parameters at the current wave may also be used as the basis for deriving prior distributions for subsequent waves. This can, in principle at least, lead to estimates with reduced variance for accurate and stable predictions which can enhance optimal allocation of survey fieldwork efforts in the subsequent wave. This procedure is known as sequential Bayesian updating (SBU) (Lindley, 1972). SBU has been applied in fields such as survey methodology (West et al., 2023), traffic analysis (Yu & Abdel-Aty, 2013), big data applications using web sourced data (Oravecz et al., 2015), and in clinical trials (Viele et al., 2014) in which model parameters were updated as new data became available without the need to repeatedly compute the likelihood. It has also been found that Bayesian models with informative priors have increased power and reduced bias when implemented for datasets with small sample sizes (Schouten et al., 2018; van de Schoot et al., 2015).

The use of a Bayesian approach using informative priors has also attracted the attention of survey methodologists in recent years in the context of adaptive and responsive survey designs (Coffey & Elliott, 2023; Schouten et al., 2018; Wagner, 2016; Wagner et al., 2023; West et al., 2023). For instance, Schouten et al. (2018) found that a correctly specified Bayesian model leads to robust results compared to a non-Bayesian model especially when used for smaller sample sizes. West et al. (2023) showed that using quarterly historical data to define prior distributions can lead to higher quality predictions of RP models compared to standard approaches not using prior information. In the same vein, Wagner et al. (2023) and Coffey & Elliott (2023) showed that informative priors derived from historical data improved the evaluation of optimisation rules for data quality and costs in responsive and adaptive survey designs. In situations where historical data are not available, informative priors elicited from experts or derived from the existing literature have also been shown to be effective in improving prediction accuracy (Coffey et al., 2020; West et al., 2023; Wu et al., 2022). Our objective in this paper is to extend our understanding of whether and how informative priors can be used to improve RP prediction accuracy in a longitudinal survey context. Our rationale is that informative priors derived from the coefficients of the previous wave (t-1) RP model have the potential to produce MSE improvements in the estimates of the current wave (t) which in turn leads to stable and accurate predictions compared to a frequentist model. This can improve the adoption of adaptive survey designs aimed at reducing attrition among households with low response predictions in subsequent waves (t + 1) in longitudinal studies.

3 Data

Understanding Society is a large-scale household longitudinal survey which collects information on health, work, education, income, family and social life and aims to explain their stability and changes among individuals and households living in the UK (Buck & McFall, 2012; Knies, 2014). The survey uses a multi-stage sample design with clustering and stratification. Households are clustered within inter-

Table 1

Number of households in each wave linked to previous wave auxiliary data, missing cases and wave final sample size

		Missing cases		
Waves	Linked Households	*n*	%	*Final sample*
1 and 2	24,738	288	1.20	24,450
2 and 3	19,791	618	3.10	19,173
3 and 4	17,856	490	2.70	17,366
4 and 5	16,705	578	3.50	16,127

viewers and within the primary sampling units (PSU). The details of sample selection can be found in Lynn (2009). The study also collects call record data and interviewer observation variables (Knies, 2014). The survey aims to achieve interviews with all individuals in sampled households who are aged 16 years and above and young people aged 10–15. Data collection for each wave is scheduled across a 24-month period, with interviews taking place annually.

This study uses the first five waves of the survey which were collected between January 2009 and December 2014. The General Population sample covering Great Britain (GB) only is used for the analysis, since the Northern Ireland (NI) sample does not contain call record data, which are required for the analysis, since previous wave call record data are incorporated in the models. The waves are linked pairwise (wave 1 to wave 2; wave 2 to wave 3, etc.) using unique personal identifiers. The analysis sample for each current wave consist of only those individuals who responded in the immediate previous wave as nonrespondents to that wave do not have observations on the variables included in the prediction model. This means there is a declining sample size in each subsequent wave due to attrition. Note also that each wave's sample size comprises both respondents and nonrespondents based on the final call outcome of that wave. For example, the sample for wave 2 consists of wave 1 respondents only. Details about the four pair-wise datasets across the five waves of interest are presented in Table 1.

As the Bayesian models with informative priors are likely to perform better when used with smaller sample sizes (Gill, 2014), we also derive informative priors for random subsamples of the full analysis samples. This is to investigate whether data used to derive informative priors have dominating effects on the posterior results. Therefore, random subsamples of 2%, 5% and 10% of the main sample are selected to obtain estimates for informative priors.

3.1 Outcome and explanatory variables

The outcome variable in the RP models is the final call outcome (of the current wave). This is coded (1) if at least one interview is conducted in a sequence of call attempts to a household, otherwise the response is recorded as unsuccessful (0). The choice of household level response as the variable of interest (as opposed to individual-level response) is motivated by the fact that this study aims to include variables from the call record data (paradata) RP models, which are only recorded at the household level, as is the case for most surveys. The definition of a call sequence in the response propensity models is informed by the definition presented in Durrant et al. (2015) and the distribution of the final call outcome (of the current wave) is presented in Table 2.

The explanatory variables included fall under the following four main groups:

1. Geographical and design variables: Government Office Regions (GORs), urban/rural indicator, and month and year of household issue.
2. Survey variables: lone parents, pensioners in household, employment status, number of cars, highest education qualification in household, household income, tenure, household size.
3. Interviewer observations: accommodation, relative condition of property, presence of unkempt garden in ad-

Table 2

Distribution of the final call outcome of the current wave in the final analysis sample

	At least one interview		*No interview*		
Waves	*n*	%	*n*	%	*Total*
2	18,928	77.40	5522	22.60	24,450
3	15,741	82.10	3432	17.90	19,173
4	15,016	86.50	2350	13.50	17,366
5	14,271	88.50	1856	11.50	16,127

dress, conditions of surrounding houses, presence of trash/litter/junk in street or road, heavy traffic on street or road, presence of car/van and children in household.

4. Call record data: length of call sequence, proportion of noncontacts, proportion of appointments, proportion of contacts, proportion of other call outcomes and proportion of interviews. The denominator used for all the proportions is the length of call sequence which is the total number of call attempts made to a household within one wave.

4 Methodology

The final call outcome (for the current wave) is modelled using binary logistic regression (Durrant et al., 2015; Hosmer & Lemeshow, 2000). Let the binary response of household i in current wave be denoted by $y_i, i = 1, \ldots$. The response variable for the final call outcome in current wave is given as:

$$y_i = \begin{cases} 1 & \text{successful final call outcome} \\ & \text{(at least one interview in a sequence)} \\ 0 & \text{unsuccessful final call outcome} \\ & \text{(no interview)} \end{cases} \tag{1}$$

for household i. The response probabilities for y_i are denoted as $\pi_i = Pr\,(y_i = 1)$ and $(1 - \pi_i) = Pr\,(y_i = 0)$. Observed responses y_i are proportions with the standard assumption that they are binomially distributed.

$$y_i \sim \text{Bin}\,(n_i, \pi_i) \tag{2}$$

where n_i is the current wave number of households. The logistic regression model, which predicts the response propensities in each current wave, is defined as

$$\text{logit}\,(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_J x_J = \boldsymbol{B}^T \boldsymbol{X}_j \tag{3}$$

where $\text{B} = (\beta_0, \beta_1, \ldots, \beta_J)$ is vector of current wave regression parameters and X_j is a vector of current wave explanatory variables at household level; with $j = (0, 1, \ldots, J)$ being an index for the number of explanatory variables.

The Bayesian logistic models are fitted using the *INLA* package (Fong, Rue, & Wakefield, 2010; Rue, Martino, & Chopin, 2009) in R version R-3.3.0. INLA produces fast and accurate approximations compared to Markov chain Monte Carlo (MCMC) alternatives for latent Gaussian models (Rue et al., 2016). INLA's Bayesian inferences are approximated deterministically, making it practically feasible to fit models which contain many regression parameters and complex structures (Rue et al., 2009). A detailed description of the INLA methodology can be found in Rue et al. (2009).

To complete the model described in Eq. 3, normal distributions denoted by $\beta_j \sim N\left(\mu_j, \sigma_j^2\right)$, $j = 1, \ldots j$, are specified as priors for regression parameters (Gelman et al., 2008). The normally distributed priors are not conjugate with the likelihood of the data, and they are incorporated in the model by altering the weighted least squares step of the algorithm and augmenting the approximate likelihood with the prior distribution (Gelman et al., 2008). The idea of conjugacy implies that prior-to-posterior updating yields a posterior that is also in the same distribution family. The analysis starts by specifying vague normal prior distributions denoted by $\beta_j \sim N\,(0, 10000)$ for regression parameters in the model predicting wave 2 final call outcome. Posterior summaries are then obtained from the INLA that summarise the knowledge on the parameters given the data. The posterior results are summarised in terms of the means that express the updated knowledge of the regression parameters and their variances. The estimated posterior means and variances are then used as informative priors for the subsequent wave analysis.

A frequently voiced concern in the use of Bayesian analysis is the 'subjectivity' associated with the choice of informative priors (Bijak & Bryant, 2016). Therefore, when informative priors are used, it is important to quantify prior impact under different specifications which involves fitting models with vague priors and altering the variance component of informative priors (Evans et al., 2011; Gill, 2014). That being said, it is important to note that the 'subjectivity' associated with informative priors is precisely what can improve model predictions when there is consistency in the data generating mechanism between posterior predictive distribution (i.e., likelihood of current data) and priors generated from previous waves. Since the normal distribution is a location-scale family distribution, altering the variance parameter provides the best way of assessing the sensitivity of the informative priors because the variance influences the posterior results' dispersion. Therefore, posterior sensitivity is assessed by multiplying the informative prior variance parameters by a factor of 0.1, 2.0, 5.0, 10.0, and 100.0 and by observing the effect on the resulting posterior distribution in terms of predictive and discriminative measures.

This spectrum of mis-specified priors gives the relative weighting of the variance for the likelihood function from highly to less informative priors. For example, the estimated variance multiplied by 0.1 and 2.0 implies that the new variance is based on the effective sample sizes that are 10 and 0.5 times larger than the original sample sizes re-

Table 3

Prior distributions used for Bayesian response propensity models in each wave

Prior type	*Specification of prior distribution for regression parameters in wave n*	*Model name*	*Effective sample size relative to original*
Vague	$\beta_k^n \sim N\,(0{,}0.001)$	M1	–
Informative	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2\right)$	M2	1
	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1} \times 0.1\right)^2\right)$	M3	10
	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1} \times 2\right)^2\right)$	M4	0.50
	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1} \times 5\right)^2\right)$	M5	0.20
	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1} \times 10\right)^2\right)$	M6	0.10
	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1} \times 100\right)^2\right)$	M7	0.01

spectively. As an uncertainty measure, the variance works well for determining prior impact, where a higher variance 'flatten' out the informative prior, making it less informative (Gill, 2014). The different prior specifications employed in this study are presented in Table 3 where models M1 and M3–M7 makes the prior wave sample smaller, while M3 makes it larger.

The posterior results from the best fitting model in each wave are then used to specify the informative priors for the subsequent wave model. We do not consider correlation structures among the regression parameters due to the large number of explanatory variables used in the models, which make it computationally demanding. The model parameters for frequentist models are fitted using Maximum Likelihood Estimation (MLE) in *Stats* Package in R for comparison purposes (R Core Team, 2015).

4.1 Model Selection

The variables included in the final models were selected in a two-step process for both the frequentist and Bayesian models and using the wave 2 outcome, with models for subsequent waves employing the same set of explanatory variables for comparability. The explanatory variables with p values < 0.05 for frequentist models and 95% credible intervals that do not cover zero for Bayesian models were selected for inclusion in the propensity models based on bivariate comparisons. Contingency tables with zero or low cells that may cause numerical problems in models are grouped (i.e., categorical levels that have few cases are combined into one group). The strength of association between each of the explanatory variables and the final call outcomes were assessed using Cramer's V which is a measure of correlation for categorical variables (Liebetrau, 1983). Cramer's V ranges from 0 to 1 where they indicate no and strong associations between two variables respectively.

The second step involved iterative refitting of frequentist and Bayesian models using a forward selection approach (Hosmer & Lemeshow, 2000). The explanatory variables that were significant in both the frequentist and Bayesian models were retained in the final model. This ensures that explanatory variables selected for inclusion in the final frequentist and Bayesian models are equivalent for comparison purposes. For the frequentist and Bayesian models, the Akaike Information Criterion (AIC) and WAIC measures were used for selecting the final models (Freese & Long, 2006; Gelman et al., 2013). AIC is calculated using the maximum likelihood estimate, while WAIC is computed using log pointwise predictive density and both adjust for the effective number of parameters. The model with the lowest AIC and WAIC value when compared with alternative models is considered to have the best fit to the data. WAIC, as an out of sample predictive measure of the estimated model, is also used to evaluate whether use of informative priors leads to an improvement in the MSE. In addition, the proportion of variance in the final call outcome accounted for by the explanatory variables in the frequentist models is assessed using a Nagelkerke pseudo R^2 (Nagelkerke, 1991).

Although the AIC, WAIC, and pseudo R^2 are useful for evaluating model adequacy, they cannot assess the accuracy of the model predictions in terms of correctly classifying nonrespondents and respondents (Plewis et al., 2012). In addition, using WAIC and AIC makes it difficult to compare the predictive performance of frequentist and Bayesian models directly. We therefore supplement our diagnostics with measures for classification, discrimination (sensitivity and specificity), prediction (positive and negative predicted values), and AUC of the ROC which help to deal with the issues of arbitrary cut-off values in discrimination and

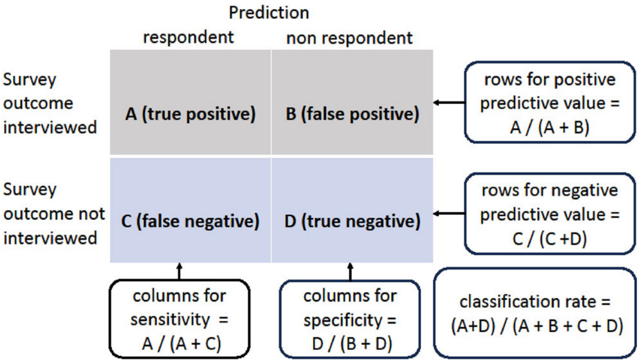
Prediction
respondent
non respondent
Survey outcome interviewed
A (true positive)
B (false positive)
rows for positive predictive value = A / (A + B)
Survey outcome not interviewed
C (false negative)
D (true negative)
rows for negative predictive value = C / (C +D)
columns for sensitivity = A / (A + C)
columns for specificity = D / (B + D)
classification rate = (A+D) / (A + B + C + D)

Fig. 1

Graphical representation of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and classification rate derivation

prediction (Durrant et al., 2017; Pepe, 2003; Plewis et al., 2012).

An overall summary of the predictive power of the model is the proportion of the correct classifications referred to as the classification rate, which measures the proportion of households that would be correctly classified by the model. Sensitivity is the proportion of households that experience no interview and are correctly predicted as such, while specificity is the proportion of households which are correctly predicted as providing at least one interview (Agresti, 2013; Durrant et al., 2015; Plewis et al., 2012). The positive predictive value (PPV) is the probability that a household is indeed a nonresponse given that it is predicted as nonresponse, while the negative predicted value (NPV) is the probability that a household is indeed a response given that it is predicted as a response (Agresti, 2013; Durrant et al.,

Table 4

Evaluation criteria for frequentist models using Akaike Information Criteria (AIC), Nagelkerke's pseudo R2 and Watanabe Akaike Information Criteria (WAIC) for Bayesian models

Waves	*Model*	*AIC*	*Nagelkerke R^2 (%)*	*WAIC*
1 and 2	Frequentist	12,559	7	–
	M1	–	–	12,561
2 and 3	Frequentist	8701	5	–
	M1	–	–	8701
	M2	–	–	8704
	M3	–	–	8857
	M4	–	–	8693
	M5	–	–	8696
	M6	–	–	8699
	M7	–	–	8701
3 and 4	Frequentist	6865	6	–
	M1	–	–	6865
	M2	–	–	6868
	M3	–	–	6997
	M4	–	–	6854
	M5	–	–	6858
	M6	–	–	6862
	M7	–	–	6871
4 and 5	Frequentist	5593	6	–
	M1	–	–	5594
	M2	–	–	5810
	M3	–	–	5627
	M4	–	–	5604
	M5	–	–	5593
	M6	–	–	5592
	M7	–	–	5594

2015; Plewis et al., 2012). A graphical representation of sensitivity, specificity, NPV and PPV, including their derivations, are given in Fig. 1. The R package *epiR* is used to evaluate classification rate, sensitivity, specificity, NPV and PPV (Mark et al., 2016).

The AUC of the ROC curve measures the model's ability to discriminate between households which did not have interviews and those which had at least one interview (Plewis et al., 2012). The AUC represents an overall accuracy of model predictions and has a range of 0.5 to 1.0. A value of 0.5 means the model predictions are no better than random guessing, while a value of 1.0 represents perfect discrimination between households that experience at least one interview and those which do not. The ROC curves are implemented in the R *pROC* package, a tool for visualising, smoothing, and comparing ROC curves (Robin et al., 2011).

These measures are evaluated using out-of-sample predictions of test data as this approach is less sensitive to outliers and overfitting (Hastie et al., 2009). This is done by partitioning the analysis samples into training and testing subsets which are used for model fitting and evaluation respectively (Hastie et al., 2009). We use 50% of the sample for an out-of-sample prediction. The training and testing subsets are obtained by randomly splitting the given wave data using the R *caret* package (Kuhn et al., 2016). Cross-validation was done by splitting each dataset twice into a training dataset and a validation dataset.

Table 5

Results of classification table and AUC of ROC curves, sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) for the final call outcome

Wave	*Modelling approach*	*Classification (%)*	*AUC (%)*	*Sensitivity (%)*	*Specificity (%)*	*PPV (%)*	*NPV (%)*
1 and 2	Frequentist	78	64	52	78	3	99
	M1	78	64	53	78	3	99
2 and 3	Frequentist	82	62	25	82	0	100
	M1	82	62	27	82	0	100
	M2	82	57	–[a]	82	0	100
	M3	82	62	50	82	0	100
	M4	82	62	40	82	0	100
	M5	82	62	30	82	0	100
	M6	82	62	27	82	0	100
	M7	82	62	27	82	0	100
3 and 4	Frequentist	87	63	40	87	0	100
	M1	87	63	40	87	0	100
	M2	87	58	–[a]	87	0	100
	M3	87	62	50	87	0	100
	M3	87	58	–[a]	87	0	100
	M5	87	63	25	87	0	100
	M6	87	63	40	87	0	100
	M7	87	63	40	87	0	100
4 and 5	Frequentist	89	64	45	89	1	100
	M1	89	64	45	89	1	100
	M2	89	52	–[a]	89	0	100
	M3	89	63	43	89	0	100
	M4	89	64	38	89	0	100
	M5	89	64	50	89	1	100
	M6	89	64	56	89	1	100
	M7	89	64	45	89	1	100

[a] Division by zero

In summary, if using previous wave RP model coefficients to specify informative priors for the subsequent wave model is effective it should lead to a reduction in WAIC values (improved MSE) and an increase in sensitivity, specificity, NPV and PPV values (reduced bias). In a situation where WAIC is lower without a change in sensitivity, specificity, NPV and PPV values, it shows an improvement in MSE without a loss in bias. Our analysis proceeds in two stages: 1) identify variables using wave 2 to be included in the predictions of the final call outcome, and 2) evaluate whether inclusion of regression coefficients for these variables from the RP model from the previous wave as informative priors can improve predictions of final call in the subsequent waves. This is important since it provides a mechanism which enables the prediction of final call outcome for waves 2, 3, 4, and 5.

5 Results

Our results consist of 23 models for final call outcome across Waves 2, 3, 4, and 5, respectively. In Wave 2, only two models are presented: a frequentist model and a Bayesian model with vague priors. A total of 9 models (i.e., a frequentist, a Bayesian model with vague prior, and 7 Bayesian models with different specifications of informative priors) are fitted for the final call outcome at the subsequent Waves (Waves 3, 4, and 5, respectively). The posterior summaries from the Bayesian model with the lowest WAIC among alternative models in the previous wave is used to specify the informative priors for the current wave analyses. At each wave, a model with vague priors is used as the reference when comparing the predictive performance of informative prior models. Cramer's *V* values obtained for the final call outcome ranged between 0.02 to 0.26 indicating a weak bivariate correlation between outcome and explanatory variables used in all models.

Table 4 presents pseudo R^2 coefficients and the values for AIC and WAIC for the 23 models in Waves 2, 3, 4 and 5. Table 4 shows that in Wave 3, all models with different specifications of informative priors have lower WAIC values compared to a model with vague priors except models M2 and M3, which have higher WAIC values indicating a poor model fit. The inclusion of previous wave data via informative priors in Wave 4 produces mixed results. Models M2, M3 and M7 have higher WAIC values compared to model M1 with vague prior indicating a poor model fit. However, there is a slight decrease in WAIC values for models M4, M5 and M6 compared to model M1 showing an improvement in model fit. In Wave 5, models with informative priors have higher WAIC values compared to a model with vague priors, while models M5, M6, and M7 have WAIC values similar to the ones obtained in vague priors' model. This indicates that our strategy of using previous wave data in the final call models via informative priors does not improve the model fit especially when the prior wave sample is larger. However, reduction of the effective prior sample by half (M4) yields a somewhat smaller WAIC, indicating an improved MSE.

Table 4 also shows that the Nagelkerke pseudo R^2 values for the frequentist models are between 4.9 and 7% for the final call outcome, which are similar to pseudo R^2 values of nonresponse models reported in previous studies (Olson et al., 2012; Olson & Groves, 2012). These results indicate that the use of informative priors leads to a slight improvement in model fit in the earlier waves of the survey relative to models with vague priors which is consistent with the findings of West et al. (2023). However, the performance of the Bayesian models is worse at later waves. This difference between earlier and later waves could be due to substantive changes in the survey fieldwork, such as the introduction of phone interviews for some households in wave 3. In addition, this may also be explained by the reduction in strength of borrowed information in later waves due a longer period which renders informative priors from previous wave less consistent with current wave data. Therefore, using the immediate previous wave is expected to be more informative about households in comparison to the informative priors derived from the full follow up data (e.g. priors derived from combined waves 2, 3 and 4 on wave 5 outcome). This is because household characteristics are more likely to be stable in the short to medium term. The RP models with informative priors with larger variances (standard deviation multiplied by a factor of 10 and 100) have WAIC values similar to those for the models with vague priors.

Table 5 presents the classification tables and AUC values for ROC curves based on 50% out-of-sample predictions. For classification tables, 50% of the cases for the final call outcomes are classified correctly by chance, with values above 50% indicating more predictive powers. Here, the observed classification values are 82%, 87% and 88% in Waves 3, 4 and 5 respectively. These values are similar to the proportion of households which had at least one interview at each wave because classification rates are sensitive to the largest category of the response variable (Agresti, 2013). These classification values for the final call outcomes, then, show that the models do not perform better than the observed distribution.

The AUC values of ROC curve greater than 50% indicate that any discrimination for the outcome is not due to random variation, with values above 70% considered to represent better discrimination (Hosmer & Lemeshow, 2000). For the final call outcomes, Table 5 shows the AUC values obtained in all waves range between 62 and 64%, indicating a minimal discrimination. In all waves, the differences in AUC values for models with informative and vague priors ranged between ±0.0% and ±0.03%, which are negligible. Although there are slightly higher AUC values for RP models with informative priors, they are not statistically significant. Overall, the results show that the use of informative priors does not lead to improvements in the predictive power of the models.

Overall, sensitivity values range from 25 to 56%, indicating moderate success in the proportion of households correctly predicted as not being interviewed (Table 5). However, the low PPV values (ranging from 0 to 3%) suggest that none of the households which were not interviewed were predicted as nonrespondents. Considering the analysis sample at each wave consisted of all respondents in the immediately previous wave, this may have contributed to the low proportion of households that were nonrespondents and, therefore, a low PPV. The specificity values range from 78 to 89% suggesting that most households are correctly predicted as being interviewed. The higher NPC values (99–100%) show that most households, which were indeed interviewed, are correctly predicted as respondents.

Table 5 also shows improvement of sensitivity values for the final call outcome model (M3) in waves 3 and 4 relative to model (M1). The sensitivity values for model (M2) give a non-numeric value (Nan—which occurs when fraction's numerator is zero) in all waves indicating that a tight informative prior does not correctly predict any households that were not interviewed. The reason for this is that the informative prior specified is very strong, since it puts most of its mass on parameter values that are large in absolute value and, therefore, strongly influences the posterior inference. Considering the percentage of households that were not interviewed was lower (i.e. ranged between 11.5 to 23%) compared to those interviewed (77.4 to 89%), conditioning on a tight informative prior may have contributed to the prediction that none of the households were correctly classified as nonrespondents. The mis-specified informative prior models with larger variances have similar sensitivity values as vague priors, except in Wave 5 which has slightly improved values. In addition, the specificity values for models with informative and vague prior models are similar in each wave. Sensitivity and specificity results show that the use of previous wave information does not improve the discrimination power of the models. Table 5 also shows that the positive and negative predictive values for final call outcome models with informative priors and vague priors in waves 3, 4, and 5 are similar, with very small differences of±1%. Note that sensitivity, specificity, PPV and NPV values in Table 5 are integers because of the nature of data used. The coefficient estimates for RP models (i.e., M1, M2, M3 and M7) for waves 3, 5 and 4 are provided in Tables A3, A4 and A5 in the Supplementary material.

The additional analysis (Tables A6 and A7 in the Supplementary material) based on a random subsample of 10% of the main sample had similar results in terms of discrimination and prediction power. Thus, a smaller sample size does not have an impact on the discrimination and prediction power in response propensity models. The results for sensitivity analyses using 2% and 5% subsamples were similar to those obtained for 10%. Analyses using the length of the call sequence as the outcome also produced similar results. This confirms the finding that the use of informative priors based on previous waves does not seem to lead to significant improvements of the predictive ability of the models (at least not for the example analysed here).

We also investigated whether inclusion of explanatory variables in RP models that are highly correlated with the final call outcome (i.e. income and employment variables) only influenced the strength of informative priors in improving the predictive power (results are shown in Tables A8 and A7 in the Supplementary material). The results demonstrate that also these models' predictive and discrimination abilities are not very different from those obtained in the main analysis. This indicate that the strength of correlation between variables in the data used for this analysis does not influence the effectiveness of the borrowed information.

6 Discussion

To better understand survey nonresponse and to counter the effects of rising nonresponse rates, there is a need to improve the generally low predictive power of RP models. Our goal in this paper has been to explore whether gains in predictive power can be obtained by using previous wave information to specify informative priors in a Bayesian analysis at the subsequent wave. The utility of such an approach has been demonstrated in cross-sectional survey contexts (Schouten et al., 2018; Wagner et al., 2023; West et al., 2023). In principle such an approach is potentially very attractive because it would enable predictions of whether a sample unit will respond or not by updating current wave data with information contained in regression coefficient estimates of previous waves through the specification of informative priors, and thus generating cumulative predictions. Our rationale is that Bayesian estimation with informative priors can potentially achieve bias-reduction when the observed data are in some sense misleading as the prior can pull the estimates closer to the truth. Additionally, a Bayesian framework might also offer efficiency improvement; if both the prior and the data provide correct estimates, their combination implies more information about the quantity of interest, in effect, a larger sample.

Our findings are not encouraging. The RP models with informative priors are not a significant improvement compared to models with vague priors. Models with informative priors with half the effective sample size of the current wave (i.e., M4) have a slightly better predictive accuracy indicating an improvement in MSE without losses in bias. However, their specificity values are similar in each wave, indicating no improvement in the predictive power. Some small improvements in sensitivity values for models with informative priors were observed in earlier wave. This was as expected since they pull the estimates closer to the true values, but this effect diminishes and then reverses in later waves. We speculate that this is because at later waves informative priors were misleading when combined with the current wave data due to the introduction of a mixed-mode design during the last two quarters of Wave 3 in this survey. This makes earlier information about the correlates of response from earlier waves less relevant. These findings are consistent with Schouten et al. (2018) and West et al. (2023) who find informative priors derived from more recent data to be better at improving the predictive power of RP models (i.e., improved bias) compared to using earlier data. The high specificity and NPV values obtained in our study implies this approach can be adopted in early stages of data collection to predict sample units that are likely respond leading to reduced survey costs and improved response rate.

Altering the variance component of the informative prior did not produce notable changes in the range of the predictive and discrimination measures. This also suggests that informative priors were not effective in improving bias. The discrimination values indicate that models with better fit in terms of WAIC do not produce better discriminative power. Also, the AUC values alongside the positive and negative predicted values from models with informative priors show no prediction improvements compared to models with vague priors. Discriminative and predictive results obtained from smaller subsamples were not appreciably different to those for the main analysis.

An important assumption in this study involves specifying no correlations among regression parameters, which is informed by weak correlations between explanatory variables and the complexity involved in trying to incorporate covariance structure with higher dimensionality (due to many explanatory variables) into the model. The length of call sequence as response outcome was also analysed. For this analysis, the discriminative and predictive results were similar to the results reported in this paper. It observed that using different samples with small sizes leads to similar conclusions as the main sample. However, note that subsamples were obtained randomly from the main data, and it is probable that using a different survey with a small sample size might lead to different results.

Previous studies suggest that available paradata and auxiliary data are not sufficiently correlated with the response outcomes for effective predictive accuracy in household survey responses (Kreuter, Couper, et al., 2010; Kreuter, Olson, et al., 2010; Olson & Groves, 2012). Our findings suggest that borrowing this weakly predictive information from previous waves does not improve predictive accuracy of subsequent waves, since such informative priors do not bring any additional information (Kaplan et al., 2023). However, they may lead to stable estimates over time (improved MSE) especially when data generating mechanisms for data used to derive informative priors and current data are consistent. That is, model prediction accuracy (MSE), and not power (i.e. bias) of the final call outcome, improves because informative priors from previous waves can help provide a better representation of data patterns and relationships in the current wave. It is also important to note that, while informative priors can improve model predictions, this approach can also increase bias when informative priors are not consistent with the likelihood (current wave) and should therefore only be used with careful evaluation of model performance.

A unique feature of longitudinal studies such as Understanding Society is that responsive and adaptive strategies are adopted as the survey progresses, which may lead to changes in the auxiliary data compositions across waves for effective borrowing of previous wave's data via Bayesian sequential updating (Gill, 2014; Plewis et al., 2012; Schouten et al., 2018). According to Gill (2014), the use of informative priors derived from previous data can be uncertain if the data generating mechanism keeps changing over time relative to the data used for estimating the first posterior estimates. It has also been shown that the robustness of informative priors depends on the time difference between the historical and current data (Schouten et al., 2018; West et al., 2023). Bayesian sequential updating restricts inclusion of additional variables in the RP models as the survey progresses since the selection of the explanatory variables is done during the initial wave (Oravecz et al., 2015).

We also found that the data forming the likelihood component from the current wave dominates the posterior estimates, rendering information borrowed from previous waves as informative priors less relevant. Usually, the likelihood component depends on the sample size, which implies that the influence of an informative prior from previous waves decreases in longitudinal studies with large samples (Lynch, 2007; Schouten et al., 2018). However, the dominating effect of the likelihood in this context is not always dependent on the sample size but also how strongly the data contribute to the posterior. The results from the informative priors estimated using subsamples showed that previous wave data had a dominating effect on the posterior results irrespective of the specification of the priors. The results from mis-specified informative priors show robustness in the model specification since alterations in the variance component do not lead to large changes in the ranges of the predictive and discrimination measures. Although variance as an uncertainty measure works well for determining prior impact, when altered, it is a poor detector of any prior and likelihood conflict which occurs when the prior puts all its mass in the tails of the likelihood. The prior and likelihood conflict may be detected using prior to posterior divergences measures. These measures were not considered in this study but could be considered in future analyses.

Our findings are consistent with those reported by Durrant et al. (2017), suggesting it is difficult to predict nonresponse with the sorts of variables that are typically available for this task, possibly exacerbated by the potential that nonresponse reflects a quite random process. That being said, it may be fruitful to explore different sources of prior information in the longitudinal survey context, such as qualitative analysis of reasons of why households refuse to participate in surveys in the first place, or use of expert surveyors. However, caution is needed since informative priors derived using this process may potentially end up masking the true estimates of the real data if they have different data distributions. Those results might point survey methodologists to the direction of where the efforts should be put in order to improve survey efficiencies.

While our largely null findings are disappointing, our findings contribute to a better understanding of how previous wave data can be leveraged to improve predictions of RP models in longitudinal settings. While our analysis examples showed only slight or no real improvements in response predictions, the procedures and principles developed here will, we hope, help to establish a new framework for the exploration of other sources of informative priors under different study settings. We encourage future researchers in this area to apply and extend the approach we have implemented here to other surveys and country context.

Acknowledgements This work was supported by the Economic and Social Research Council (ESRC) as part of Work Package 1 of the ESRC National Centre for Research Methods (2014–2019), grant number ES/L008351/1; ESRC National Centre for Research Methods (2020–2025), grant number ES/T000066/1; and the Economic and Social Research Council PhD Studentship grant number ES/J500161/1. Eliud Kibuchi is also funded by the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17).

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd edn.). John Wiley& Sons.

Biemer, P.P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 147–168. https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, *4728*, 1–19. https://doi.org/10.1080/00324728.2015.1122826.

Blom, A.G. (2009). Nonresponse bias adjustments: what can process data contribute ? *Social Sciences*.

Brick, M.J., & Montaquila, J.M. (2009). Nonresponse and weighting. https://doi.org/10.1016/S0169-7161(08)00008-4.

Buck, N., & McFall, S. (2012). Understanding Society: design overview. *Longitudinal and Life Course Studies*, *3*(1), 5–17. https://doi.org/10.14301/llcs.v3i1.159.

Carlson, B.L., & Williams, S. (2001). *A comparison of two methods to adjust weights for non-response: propensity modeling and weighting class adjustments*. Proceedings of the Annual Meeting of the American Statistical Association, August 5–9.

Coffey, S.M., & Elliott, M.R. (2023). Optimizing data collection interventions to balance cost and quality in a sequential multimode survey. *Journal of Survey Statistics and Methodology*.

Coffey, S.M., West, B.T., Wagner, J., & Elliott, M.R. (2020). What do you think? Using expert opinion to improve predictions of response propensity under a bayesian framework. *Methoden, Daten, Analysen*, *14*(2).

David, M., Little, R.J.A., Samuhel, M.E., & Triest, R.K. (1983). *Nonrandom nonresponse models based on the propensity to respond*. Proceedings of the Business and Economic Statistics Section. (pp. 168–173). American Statistical Association.

Durrant, G.B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK government surveys. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *172*(2), 361–381. https://doi.org/10.1111/j.1467-985X.2008.00565.x.

Durrant, G.B., D'Arrigo, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*(4), 1029–1049. https://doi.org/10.1111/j.1467-985X.2011.00715.x.

Durrant, G.B., D'Arrigo, J., & Müller, G. (2013). Modeling call record data: examples from cross-sectional and longitudinal surveys. In *Improving surveys with paradata: analytic uses of process information* (pp. 281–308).

Durrant, G.B., Maslovskaya, O., & Smith, P.W.F. (2015). Modeling final outcome and length of call sequence to improve efficiency in interviewer call scheduling.

Journal of Survey Statistics and Methodology, *3*(3), 397–424. https://doi.org/10.1093/jssam/smv008.

Durrant, G. B., Maslovskaya, O., & Smith, P. W. F. (2017). Using prior wave information and paradata: Can they help to predict response outcomes and call sequence length in a longitudinal study? *Journal of Official Statistics*, *33*(3), 801–833. https://doi.org/10.1515/jos-2017-0037.

Evans, M., Jang, G. H., & Jan, M. E. (2011). Weak Informativity and the Information in One Prior Relative to Another. *26*(3), 423–439. https://doi.org/10.1214/11-STS357.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: a systematic review. *Computers in Human Behavior*, *26*(2), 132–139. https://doi.org/10.1016/j.chb.2009.10.015.

Fearn, T., Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., & Fearn, T. (2004). Bayesian data analysis. *Biometrics*, *52*(3), 696. https://doi.org/10.1007/s13398-014-0173-7.2.

Fong, Y., Rue, H., & Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, *11*(3), 397–412. https://doi.org/10.1093/biostatistics/kxp053.

Freese, J., & Long, J. S. (2006). *Regression models for categorical dependent variables using stata* (3rd edn.). College Station: Stata Pres.

Fricker, S., & Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, *74*(5), 934–955. https://doi.org/10.1093/poq/nfq064.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. https://doi.org/10.1007/s11222-013-9416-2.

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (20th edn.). CRC press.

Gjonça, E., & Calderwood, L. (2004). 2. Socio-demographic characteristics. http://www.elsa-project.ac.uk/uploads/elsa/report03/ch2.pdf

Goldberg, M., Chastang, J. F., Leclerc, A., Zins, M., Bonenfant, S., Kaniewski, N., Schmaus, A., Niedhammer, I., Piciotti, M., Chevalier, A., Godard, C., & Imbernon, E. (2001). Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term Epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. *American Journal of Epidemiology*, *154*(4).

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *169*(3), 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *56*, 475–495.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd edn.). Wiley.

Kalton, G., & Flores-cervantes, I. (2003). Weighting Methods. *19*(2), 81–97.

Kaplan, D., Chen, J., Lyu, W., & Yavuz, S. (2023). Bayesian historical borrowing with longitudinal large-scale assessments. *Large-Scale Assessments in Education*, *11*(1), 2.

Knies, G. (2014). Understanding society-the UK household longitudinal study waves 1–5 user manual. *Innovation*. https://doi.org/10.2307/3348243.

Kreuter, F. (2013). *Improving surveys with Paradata: analytic uses of process information*. John Wiley & Sons.

Kreuter, F., & Olson, K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods & Research*, *40*(2), 311–332. https://doi.org/10.1177/0049124111400042.

Kreuter, F., Couper, M. P., & Lyberg, L. E. (2010). The use of paradata to monitor and manage survey data collection. *Measurement*, 282–296.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-cordero, C., Lemay, M., Peytchev, A., Groves, R. M., & Raghunathan, T. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(2), 389–407. https://doi.org/10.1111/j.1467-985X.2009.00621.x.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Candan, Y., & Can, T. (2016). caret: classification and regression training. R package version 6.0-68. https://cran.r-project.org/package=caret

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: a longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J.L. Eltinge & R.J.A. Little (Eds.), *Survey nonresponse* (pp. 41–54). Wiley.

Liebetrau, A.M. (1983). *Measures of association*. SAGE.

Lindley, D. (1972). *Bayesian statistics: a review*. Society for Industrial and Applied Mathematics.

Little, R. (1986). Survey nonresponse adjustments. *International Statistical Review*, *54*, 139–157. https://doi.org/10.2307/1403140.

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469–487.

Lynch, S.M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer.

Lynn, P. (2009). Sample design for understanding society. Understanding society working paper series, 2009–01. http://research.understandingsociety.org.uk/publications/working-paper/2009-01

Mark, S., Telmo, N., Cord, H., Jonathon, M., Javier, S., Ron, T., Jeno, R., Jim, R.-C., Paola, S., Peter, S., Kazuki, Y., Geoff, J., Sarah, P., Simon, F., Ryan, K. (2016). epiR: tools for the analysis of epidemiological data. R package version 0.9-74. https://cran.r-project.org/package=epiR

Moss, G.M. (1981). Factors affecting response rate and response speed in a mail survey of part-time university students. *XI*(3).

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691.

Olson, K., Groves, R.M. (2012). An examination of within-person variation in response propensity over the data collection field period. *Journal of Official Statistics*, *28*(1), 29–51.

Olson, K., Smyth, J.D., Wood, H.M. (2012). Does giving people their preferred survey mode actually increase survey participation rates? An experimental examination. *Public Opinion Quarterly*, *76*(4), 611–635. https://doi.org/10.1093/poq/nfs024.

Oravecz, Z., Huentelman, M., Vandekerckhove, J. (2015). Sequential Bayesian updating for big data. In *Big data in cognitive science: from methods to insights* (pp. 100–150).

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Peytcheva, E., Groves, R.M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, *25*(2), 193.

Plewis, I., Ketende, S., Calderwood, L. (2012). Assessing the accuracy of response propensity models in longitudinal studies. *Survey Methodology*, *38*(2), 167–171.

R Core Team (2015). *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77.

Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rue, H., Martino, S., Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.pdf. *Journal of the Royal Statistical Society, Series B*, (71), 319–392.

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K. (2016). Bayesian computing with INLA: a review. 1–26. http://arxiv.org/abs/1604.00860

Särndal, C.-E., Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. In *International Statistical Review/Revue Internationale de Statistique* (pp. 279–294).

van de Schoot, R., Broere, J.J., Perryck, K.H., Zondervan-Zwijnenburg, M., van Loey, N.E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*(1), 25216. https://doi.org/10.3402/ejpt.v6.25216.

Schouten, B., Cobben, F. (2007). R-indexes for the comparison of different fieldword strategies and data collection modes. http://hummedia.manchester.ac.uk/institutes/cmist/risq/schouten-cobben-2007-a.pdf

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., Wagner, J. (2018). A Bayesian analysis of design parameters in survey data collection. *Journal of Survey Statistics and Methodology*, *6*(4), 431–464.

Simon, J. (2009). *Bayesian analysis for the social sciences*. Wiley.

Sinibaldi, J., Eckman, S. (2015). Using call-level interviewer observations to improve response propensity models. *Public Opinion Quarterly*, *79*(4), 976–993. https://doi.org/10.1093/poq/nfv035.

Sinibaldi, J., Trappmann, M., Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: purchasing commercial auxiliary da ta or collecting interviewer observations? *Public Opinion Quarterly*, *78*(2), 440–473. https://doi.org/10.1093/poq/nfu003.

Skorczynski, M.L. (2012). *Bayesian statistics in practice*

Steel, D. (2007). Bayesian confirmation theory and the likelihood principle. *Synthese*, *156*(1), 53–77. https://doi.org/10.1007/s11229-005-3492-6.

University of Essex. Institute for Social and Economic Research, NatCen Social Research, Kantar Public. (2016). Understanding Society: Waves 1–6, 2009–2015. [Data collection]. 8th Edition. UK Data Service. SN: 6614, https://doi.org/10.5255/UKDA-SN-6614-9

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J.G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*, *13*(1), 41–54. https://doi.org/10.1002/pst.1589.Use.

Wagenmakers, E., Morey, R.D., Lee, M.D. (2008). Bayesian benefits for the pragmatic researcher. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). Springer.

Wagner, J. (2016). Using Bayesian methods to estimate response propensity models during data collection. http://hummedia.manchester.ac.uk/institutes/cmist/BADEN/workshop-2016/AAPOR_James.pdf

Wagner, J., Zhang, X., Elliott, M.R., West, B.T., Coffey, S.M. (2023). An experimental evaluation of a stopping rule aimed at maximizing cost-quality tradeoffs in surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*.

West, B.T. (2011). Paradata in survey research. *Survey Practice*, *4*(4), 1–8.

West, B.T., Wagner, J., Coffey, S., Elliott, M.R. (2023). Deriving priors for Bayesian prediction of daily response propensity in responsive survey design: historical data analysis versus literature review. *Journal of Survey Statistics and Methodology*, *11*(2), 367–392.

Wu, S., Schouten, B., Meijers, R., Moerbeek, M. (2022). Data collection expert prior elicitation in survey design: two case studies. *Journal of Official Statistics*, *38*(2), 637–662.

Yu, R., Abdel-Aty, M. (2013). Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. *Accident Analysis and Prevention*, *56*, 51–58. https://doi.org/10.1016/j.aap.2013.03.023.

Zyphur, M.J., Oswald, F.L. (2015). Bayesian estimation and inference: a user's guide. *Journal of Management*. https://doi.org/10.1177/0149206313501200.