sciendo

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 325–349, http://dx.doi.org/10.2478/JOS-2023-0015

Predicting Days to Respondent Contact in Cross-Sectional Surveys Using a Bayesian Approach

Stephanie Coffey[1] *and Michael R. Elliott*[2]

Surveys estimate and monitor a variety of data collection parameters, including response propensity, number of contacts, and data collection costs. These parameters can be used as inputs to a responsive/adaptive design or to monitor the progression of a data collection period against predefined expectations. Recently, Bayesian methods have emerged as a method for combining historical information or external data with data from the in-progress data collection period to improve prediction. We develop a Bayesian method for predicting a measure of case-level progress or productivity, the estimated time lag, in days, between first contact attempt and first respondent contact. We compare the quality of predictions from the Bayesian method to predictions generated from more commonly-used predictive methods that leverage data from only historical data collection periods or the in-progress round of data collection. Using prediction error and misclassification as short- or long- day lags, we demonstrate that the Bayesian method results in improved predictions close to the day of the first contact attempt, when these predictions may be most informative for interventions or interviewer feedback. This application adds to evidence that combining historical and current information about data collection, in a Bayesian framework, can improve predictions of data collection parameters.

Key words: Adaptive design; hurdle models; national health interview survey.

1. Introduction

Making contact with a survey respondent is an important part of the data collection process. Without a successful contact, it is not only impossible to obtain cooperation and a completed interview, possibly introducing nonresponse bias, but it may also be impossible to determine the eligibility of a case, affecting response rates, nonresponse adjustments, and ultimately, variances of estimates. Noncontact makes up a substantial portion of nonresponse in face-to-face surveys. Durrant and Steele (2009) discuss six large,

[1] U.S. Census Bureau, Research and Methodology, 4600 Silver Hill Road, Suitland, Maryland, 20746 U.S.A. Email: stephmcoffey@gmail.com
[2] University of Michigan, Department of Biostatistics, School of Public Health, University of Michigan, M4041 SPH II, 1420 Washington Heights, Ann Arbor, Michigan 48109, Michigan, 48105 U.S.A. Email: mrelliot@umich.edu

Acknowledgments: Research reported in this publication was supported by the National Center for Science and Engineering Statistics of the National Science Foundation under contract number 0099-FY16-IPA-0002, and the National Institute on Aging of the National Institutes of Health under Award Number R01AG058599. This study design and analysis was not preregistered. The opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the National Science Foundation, or the National Institutes of Health. All results have been reviewed to ensure that no confidential information is disclosed. The authors thank the journal reviewers as well as Stephen Blumberg, James Dahlhamer, James Wagner, Brady West, and Yan Li for their review and helpful comments. The data used to prepare this manuscript are restricted to uses approved by the National Center for Health Statistics (NCHS). As a result, these data cannot be included as part of supplementary materials due to privacy and confidentiality requirements. However, researchers can apply to access this data in the NCHS Research Data Center (RDC) by following the proposal process outlined at this location: https://www.cdc.gov/rdc/index.htm.

© Statistics Sweden

government-sponsored, face-to-face household surveys carried out in the UK which have noncontact rates ranging from 8% to 40% of sample cases. Williams and Brick (2018) report that for nine government-sponsored, primarily face-to-face household surveys in the United States, noncontact rates as a portion of nonresponse range from four percent to 40% for the most recent year included in the article, 2014. Through a trend analysis across national statistical institutes, Luiten et al. (2020) also found evidence that noncontact rates are increasing in the United States, though evidence was mixed in other countries.

Groves and Couper (1998) discuss methods for reducing nonresponse due to noncontact, including planning for numerous contact attempts in face-to-face surveys, and requiring interviewers to make attempts early enough in the data collection period to make contact with all sample units. Waiting too long to begin making attempts could lead to nonresponse due to noncontact rather than post-contact reluctance or refusal. Biffignandi and Bethlehem (2021) note that an effective call management system can mitigate this problem, particularly in a centralized telephone environment, however, that solution may not be applicable to decentralized interviewing environments, such as in-person interviewer-administered surveys.

In order to improve survey outcomes, survey managers might want to monitor the expected time lag, in days, between a first attempt and first contact with a sample member to better understand the expected progress of data collection operations. If, for example, these day lags are longer than expected, or are leading to nonresponse due to noncontact, survey managers could attempt to reduce this lag by intervening with respect to individual cases shortly after a first contact attempt is made. Those interventions could reflect different passive or active levels of management throughout data collection. Christy (2014) described an intervention where interviewers were provided the estimated response propensity of cases in their workload to help interviewers more efficiently plan their contact attempt strategies. Alternatively, an interviewer whose cases are approaching (or exceeding) expected lag times might prompt feedback from a supervisor, and the earlier that intervention can occur, the faster an improvement might be made (Edwards et al. 2020). Walsh and Coombs (2013) described a reassignment strategy based on a different parameter, the scaled evenness of finding attempts (SEFA), which they find was related to higher rates of contact and completion. To order to intervene during a data collection period, as the examples above suggest, the indicators or parameters of interest must be generated during the data collection period itself. However, the first step to being able to intervene to reduce these time lags is to estimate their expected length.

Other survey data collection parameters related to data quality, survey progress, and survey cost are commonly generated and monitored as part of data collection operations. These parameters may be defined at various levels of the data collection process, including the interviewer level (Mneimneh et al. 2018); the question-level (Mittereder and West 2021); the case-level (Wagner et al. 2012), and the survey level (LaFlamme and Karaganis 2010). Indicators of cost, progress, and quality have also served as inputs into responsive and adaptive designs (Groves and Heeringa 2006; Schouten et al. 2011) that allow for the balancing of cost and quality goals through centralized interventions including mode changes (Chesnut 2013; Coffey et al. 2019), case prioritization (Walejko and Wagner 2018; Tolliver et al. 2019; Peytchev et al. 2020) and stopping rules (Wagner and Ragunathan 2010; Peytchev 2014).

Of the varied data collection parameters that can be predicted, case-level response propensity is the most commonly estimated and used for interventions. Many implementations of response propensity prediction during survey data collection have used only the data available from the in-progress collection period to estimate coefficients in a logistic regression or probit model (Groves and Heeringa 2006; West and Groves 2013; Schouten et al. 2009, 2011; Chesnut 2013; Coffey et al. 2019). The data sources available for estimating response propensity in this setting include survey frame and sampling data and paradata (Couper 2000, 2017) for the in-progress period of data collection as it accumulates after each operation or contact attempt.

Other examples use only data from the past implementations of a survey to estimate coefficients in the response propensity model, which are then applied to covariates during the in-progress data collection period (Peytchev et al. 2010; Roberts et al. 2014; Calinescu et al. 2013; Jackson et al. 2020). Unlike the prior examples, these applications do not consider information learned during the in-progress data collection period.

More recent applications have incorporated both historical data and data from the in-progress data collection period in a Bayesian framework to improve predictions of response propensities (Schouten et al. 2018; West et al. 2021). Both studies found that Bayesian methods offered modest improvements in predictions of daily response propensities compared to predictions relying on either historical or current data alone, particularly when model misspecification was small.

For this application, we propose a model to predict our parameter of interest, the time lag between the first contact attempt and the first respondent contact in a face-to-face interviewer-administered survey. We use three sets of data for predicting lag: historical collection period data in a non-Bayesian prediction; current data from the in-progress data collection period data in a non-Bayesian prediction; and combining historical data and current data in a Bayesian prediction setting. The purpose of this research is to understand whether the Bayesian framework that led to improvements in predictions of response propensity can also lead to similar improvements in predictions of a time-to-event parameter, the estimated day lag between first contact attempt and first respondent contact. We apply our methodology to predict these contact lags in the National Health Interview Survey (NHIS), using monthly survey data from July 2014 through June 2016. We compare the case-level error in prediction of lags when using the Bayesian framework to methods that leverage data from either historical or in-progress data collection periods using maximum likelihood estimation. We also vary the strength of the priors put on the coefficients in the Bayesian framework to better understand how the strength of the priors affects prediction error. Additionally, we use the predictions generated from the different methods to illustrate how the different methods would perform when trying to identify cases with short or long lags between the first contact attempt and first respondent contact. Our goal is to contribute to the growing literature on the use of Bayesian methods for improving predictions of survey data collection parameters. After introducing the NHIS survey data (Section 2) we introduce a Weibull hurdle model which is used to predict the lag between first attempt and first contact and describe our method for prior construction (Section 3). We then present our analytic approach and evaluation methods (Section 4). Results are presented (Section 5) followed by a discussion (Section 6).

2. Description of Data

2.1. The National Health Interview Survey

The National Health Interview Survey (NHIS) is a monthly cross-sectional household interview survey that serves as the principal source of information on the health of the civilian noninstitutionalized population of the United States and is one of the major data collection programs of the National Center for Health Statistics (NCHS 2018).

The sample design follows a multistage area probability design that permits the representative sampling of households and non-institutional group quarters (e.g., college dormitories). The entire United States is divided into approximately 1,700 primary sampling units (PSUs), which can consist of a county, a small group of adjacent counties, or a metropolitan statistical area. PSUs are selected, some with certainty (self-representing PSUs), and others with probability proportional to size (non-self-representing PSUs). Within those selected PSUs, clusters of households and non-institutional group quarters are selected. Additionally, an oversample is taken for geographies with higher expected populations of particular age and race/ethnicity groups. Interviewing for the NHIS is conducted continuously throughout each calendar year.

The U.S. Census Bureau is the data collection agent for the NHIS. Survey data are collected continuously throughout the year by Census interviewers. Cases released into the field are attempted for up to 30 days. The NHIS consists primarily of face-to-face interviews conducted in respondents' homes, but follow-ups to complete interviews may be conducted over the telephone. A telephone interview may also be conducted when the respondent requests a one or when road conditions or travel distances would make it difficult to schedule a personal visit before the required completion date (NCHS 2018). We used two years' worth of monthly data collection periods, spanning the months from July 2014 to June 2016 for this work.

2.1.1. Auxiliary Data Sources

Four data sources were used for this evaluation, in addition to the NHIS sample itself. The Census Bureau Planning Database (PDB) is a detailed dataset including sociodemographic information at the block group level that is produced annually using data from the five-year ACS estimates (Census 2008) and the Decennial Census. For this application, we used the 2016 release. For a full list of items on the PDB, see Census (2016). Second, we obtained a dataset of basic employment information about interviewers, including which regional office an interviewer belongs to and their experience level on the NHIS.

We also utilized two sources of paradata, the Neighborhood Observation Instrument (NOI) and the Contact History Instrument (CHI). In the NOI, interviewers are asked to record information about the housing unit and neighborhood from their own observations. These neighborhood observations require no contact with the respondent and should be recorded just prior to the first contact attempt. For the CHI, on the other hand, interviewers record the date, time and outcome of each contact attempt and information about interactions with sample persons. Additionally, the CHI includes information about the field management structure, including which interviewers are assigned to work each case and whether a case was reassigned during the field period. These data sources were selected for this application partially because of their availability, but also because they

Table 1. Auxiliary data sources, example data items and their availability for use in prediction of lag.

Data source	Available prior to start of the data collection period?	Fixed or time-varying	Example items
NHIS Sample File	Yes	Fixed	Sample unit case ID Census block-group-level ID Data collection sample month Census fieldwork regional office
Planning Database	Yes	Fixed	Census block group (BG) ID % HU in a BG that are mobile homes % Population in a BG that is urbanized % Pop in a BG with a college degree
Interviewer Information	Yes	Fixed	Interviewer ID New NHIS interviewer (< one year experience)
Interviewer Observations	No	Fixed	Sample unit case ID Day/time of contact attempt Interviewer ID Evidence of children at the housing unit (HU) Evidence of smoking at the HU Bars observed on windows of the HU
Contact history Instrument	No	Time-varying	Sample unit case ID Interviewer ID Day/time of contact attempt Outcome of contact attempt Activities completed on contact attempt Case reassignment indicator

include the types of data identified in Groves and Couper (1998) for predicting outcomes like household contactability, survey cooperation, and response propensity.

Table 1 lists the five data sources, whether the information contained in each source would be available prior to the start of data collection, whether the data items from the source are fixed or time-varying during the data collection period, and some example items from each source. For a full list of items included in the predictive models as part of this research, see the online supplementary materials.

Identifiers across the different data sources can be used to link the sources and generate a file at the contact attempt level for each sample unit in each data collection month. It is important to note that, until the interviewer makes their first contact attempt at a sample unit, the information from the NOI and the CHI is missing. Both of those data sources are only reported at or after the point-in-time of the first contact attempt, which restricts when they are available for prediction of the lag between first attempt and first contact.

2.2. Parameter of Interest: Lag Between First Contact Attempt and First Respondent Contact

For this study, we analyzed data from 24 months of the NHIS, ranging from July, 2014 through June, 2016. Our parameter of interest was the time lag, in days, between the first

contact attempt made on a sample unit, and the first respondent contact. We calculated this parameter at the case-level from the CHI data by subtracting the data collection day of the first attempt from the data collection day when the first contact with the respondent was reported. If respondent contact was made on the first contact attempt, the resulting lag was zero days. Otherwise, all lags are positive integers. Table 2 summarizes, by month, the approximate sample size in housing units of the NHIS, the percent of cases with a lag greater than zero, the mean and variance of the lags conditional on the lag being greater than zero, and the annual response rate for the NHIS. While the NHIS is conducted monthly, response rates are only reported yearly (NCHS 2018). Table 2 shows that the percentage of cases with a positive lag, and the means and variances of non-zero lags, are relatively consistent across the 24 months of data.

While the mean lag across months was relatively similar, the mean lag *during* any one month varies by the day of the data collection month when the first attempt is made. Again, the day of the first contact attempt was taken from the CHI data. Table 3 shows that, for example, cases first attempted during the first week of data collection have a mean lag of nearly nine days, while cases first attempted within the last week of the month have a mean lag of less than four days. This suggests that the day of the first attempt may be an

Table 2. Mean and variance of lag in days between 1st attempt and 1st contact by interview month.

Interview period	Number of cases	% Cases with lag > 0	Mean(lag) if lag > 0	Var(lag) if lag > 0	Response late (AAPOR RR6)
July 2014	5200	0.4171	8.890	48.67	
Aug 2014	5100	0.3990	8.216	41.90	
Sept 2014	4900	0.4131	7.941	42.54	73.8%
Oct 2014	5300	0.4236	8.009	43.89	annualized
Nov 2014	5000	0.3895	7.837	40.78	
Dec 2014	5100	0.4058	9.192	48.30	
Jan-2015	5600	0.3861	8.178	39.93	
Feb 2015	5400	0.4181	8.464	44.49	70.1%
Mar 2015	5400	0.3984	8.658	49.41	annualized
Apr 2015	5800	0.4165	8.396	41.84	
May 2015	5400	0.4026	8.424	45.74	
Jun 2015	5600	0.3842	8.072	43.88	
Jul 2015	5400	0.3938	8.589	40.90	
Aug 2015	5200	0.3929	8.385	46.09	
Sept 2015	5100	0.4003	8.449	43.12	
Oct 2015	5300	0.4015	8.847	49.48	
Nov 2015	5000	0.3754	8.062	42.60	
Dec 2015	4800	0.3915	9.684	59.51	
Jan-2016	5200	0.3876	7.797	35.65	
Feb 2016	6000	0.3952	8.657	42.47	
Mar 2016	6600	0.3893	8.861	51.00	67.9%
Apr 2016	6000	0.3947	8.665	43.58	annualized
May 2016	5700	0.3951	8.448	44.40	
Jun 2016	5800	0.4009	8.724	47.89	

Table 3. Mean and variance of lag in days between 1st attempt and 1st contact by week of 1st attempt, all months.

Week of data collection	Mean(lag) if lag > 0	Var(lag) if lag > 0
1	8.891	48.81
2	7.763	33.56
3	5.770	16.94
4	3.716	9.944

important predictor of the lag between first attempt and first contact. Lags in Table 3 were averaged over all 24 months of data collection used in our application.

These results make intuitive sense. As a fixed data collection period progresses, there are fewer remaining data collection days for cases to be contacted, and so interviewers may make contact attempts closer together, or at higher frequency to make contact with the sample member. Additionally, interviewers may have fewer remaining cases in their workloads, and so more attention can be paid to those remaining cases.

3. Modeling the Day Lag Between First Contact Attempt and First Respondent Contact

In order to model the day lag between first contact attempt and the first respondent contact, we needed to identify a model that can properly account for a lag of zero due to a contact being made on the first attempt. Additionally, our model needed to account for covariates that are available prior to the start of data collection as well as covariates available only after the first contact attempt is completed. As a result, we employed a hurdle model to combine two different processes – the likelihood of making contact on the first attempt, and a time-to-event, or survival, model to predict the length of the lag between the first attempt and first contact, given that contact was not made on the first attempt.

3.1. Hurdle Models

A hurdle model (Mullahy 1986) is useful when one believes there are two separate processes at work – one that generates zeros (in our case, "zero lags", where the interviewer makes contact on the first attempt) and the other that generates some non-zero value, in our case a positive lag (Ma et al. 2015). Hurdle models are similar to zero-inflated models (Rose et al. 2006) in the fact that they are useful for handling data with excess zeros. However, zeros are conceptualized differently between the two classes of models. Hurdle models are a mixture model where one component predicts whether a zero or non-zero status (the "hurdle"), and the second component estimates a positive value conditional the non-zero status. A zero-inflated model, on the other hand, allows zeros to appear in either component of the model. That is, even if a non-zero status is predicted, there is the chance that the second component of the model still results in a zero. Here, the hurdle model was conceptually intuitive. If the first contact attempt results in a noncontact, we would expect a nonzero lag for the case to result. The negative binomial model

construction would suggest that when there is a noncontact on the first attempt, there could still be a zero lag. See Feng (2021) for more discussion of zero-inflated and hurdle models.

We used a logistic regression model to predict whether a noncontact would result on the first contact attempt, resulting in a non-zero lag. The covariates available for this portion of the hurdle model were restricted to those known prior to the first contact attempt, which are found in the sample file and the PDB, as shown in Table 1. For the second component of our model, which estimates the length of the non-zero lag in cases that have a first contact, we chose to use a survival model to allow us to retain right-censored cases in our analytic data sets. This means that cases for which the first contact attempt resulted in a noncontact, and for which contact with a respondent had not been made, could still be included in the set of cases used to estimate model coefficients. While only a small portion of the cases (less than 2% in any given month) were right-censored at the end of data collection, the use of a survival model is particularly salient when considering prediction methods using data from the in-progress data collection period, where many of the cases that will ultimately be contacted after nonzero lags have not yet had attempts that result in contact. If we chose an alternative count model, such as a negative binomial, only cases that had both a noncontact on the first attempt *and* an observed first contact would be included in the set of cases used to estimate model coefficients. For more information on how censored cases are still included in parameter estimates of survival models, see the online supplementary materials.

To determine which parametric survival distribution best fit our data, we compared the empirical distribution of positive lags (without controlling for any factors) in the NHIS data against Poisson, Weibull, Gamma, and Lognormal distributions both visually and with goodness-of-fit statistics, using the `fitdistrplus` package in R. Figure 1 below shows panels for each of the four distributions evaluated for one data collection period to illustrate how different time-to-event distributions compare with the empirical distribution for a given month.

The black bars and density curve display the empirical distribution of the actual lag between first attempt and first contact, based on the NHIS data, while the red density curve

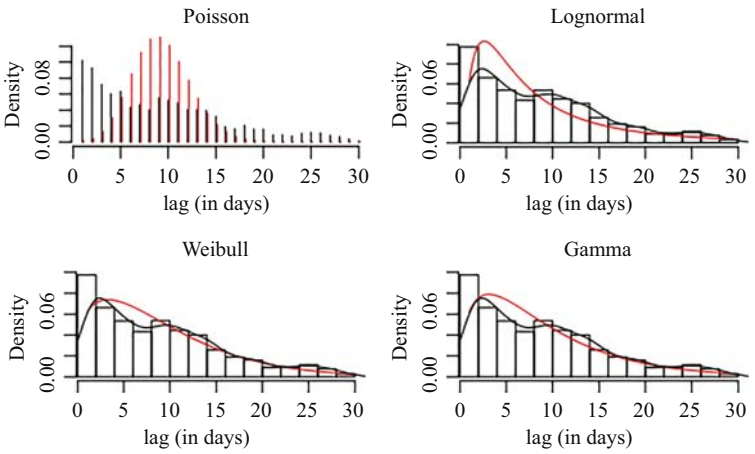

Fig. 1. Four parametric survival distributions compared to empirical distribution of lag (in days).

(or histogram, in the case of the Poisson distribution) represents the theoretical distribution when parameters are estimated from the NHIS data. The Poisson distribution provided the worst fit, followed by the lognormal distribution. The Weibull and Gamma distributions were similar in their fits, and so to decide, we examined Chi-Square tests, AIC, and BIC. For the month displayed in Figure 1, and across the 24 months of our data, the AIC and BIC scores of the Weibull distribution were smaller, and the Chi-Square test was larger, more often. As a result, we elected to use a Weibull distribution for model fitting and prediction.

3.2. The Weibull Hurdle Model

In this setting, the prediction of interest, y_i, is the length of time in days (lag) that will elapse between the first contact attempt and the first contact with a household sample member for the ith case. For each observation, with probability π_i, $y_i = 0$, and with probability $(1-\pi_i)$, $y_i > 0$. Thus the probability distribution function for an observation can be written as:

$$f(y_i;\beta,\ \sigma,\gamma) = \begin{cases} 0 & with\,probability\,\pi_i = \exp(\gamma' z_i)/(1+\exp(\gamma' z_i)) \\ g(y_i,\beta,\ \sigma) & with\,probability\,(1-\pi_i) = 1/((1+\exp(\gamma' z_i)) \end{cases}$$

where $g(y_i, \beta,\ \sigma)$ is the probability distribution function (p.d.f.) of the Weibull distribution and z_i are the covariates that govern the log-odds of the probability of a zero-lag contact. The likelihood function for the Weibull model in the presence of right censoring is:

$$L = \prod_{i=1}^{n} S(t_i)(\lambda(t_i))^{\delta_i},$$

where $S(t_i) = 1 - G(t_i;\ \beta,\ \sigma) = \exp(-exp(x_i'\beta/\sigma)t_i^{1/\sigma})$ is the survival function, $\lambda(t_i) = \frac{1}{\sigma} exp(x_i'\beta/\sigma)t_i^{((1/\sigma)-1)}$ is the hazard function, δ_i is the censoring indicator (equal to 1 if the failure time is observed and 0 if not), $t_i = \min(y_i, c_i)$ is the observed failure or censoring time, and x_i are the covariates that govern the log of the number of days until a non-zero lag contact. The full likelihood can then be written as follows:

$$L(\beta,\sigma,\gamma) = \prod_{i=1}^{n}\left[\left(\frac{exp(\gamma' z_i)}{1+exp(\gamma' z_i)}\right)^{1-d_i}\left(\left(\frac{1}{1+exp(\gamma' z_i)}\right)exp\left(-exp\left(\frac{x_i'\beta}{\sigma}\right)t_i^{\frac{1}{\sigma}}\right)\right.\right.$$
$$\left.\left.\left(\frac{1}{\sigma}exp\left(\frac{x_i'\beta}{\sigma}\right)t_i^{((1/\sigma)-1)}\right)^{\delta_i}\right)^{d_i}\right]$$

where d_i is an indicator specifying there *will be* a lag between the first attempt and first contact and now z_i are the covariates that govern the probability of a zero lag (note that x_i and z_i can be disjoint, partially overlapping, or equal to each other). The first term in the likelihood represents the probability that there will *not* be a lag between the first event and the first contact; that is, contact will be made on the first attempt. The second term represents the probability that there will be a positive lag, and the last term in the likelihood represents the survival portion of the model, accounting for censoring. (Note

that there are alternative parameterizations of the Weibull distribution that, e.g, replace $x_i'\beta$ with $\alpha_i = \exp\left(x_i'\beta\right)$ [scale parameter] and $\beta = \frac{1}{\sigma}$ [shape parameter].)

The likelihood includes parameters for both the binary process and the time-to-event process. Maximum likelihood estimation shows that the different contributing terms can be maximized separately allowing the model to be estimated in parts, computationally. Specifically, the log-likelihood can be written as:

$$l(\beta, \sigma, \gamma) = \sum_{i=1}^{n} \left[(1 - d_i)\gamma' z_i - \log(1 + \exp(\gamma' z_i)\right] + \sum_{i=1}^{n} d_i \left[\left(-\exp\left(\frac{x_i'\beta}{\sigma}\right) t_i^{1/\sigma}\right) + \delta_i \left(\frac{x_i'\beta}{\sigma} + \left(\frac{1}{\sigma} - 1\right)\log t_i - \log\sigma\right)\right]$$

As a result of this construction, the portion of the model that predicts the binary outcome (lag/no lag) can be estimated using logistic regression from all observations, and the portion of the model that predicts the lag length, can be estimated using Weibull regression among only those cases with a non-zero lag (Smithson and Merkle 2013, chap. 5). However, we refer to this as a single model, as the goal of this prediction is to determine the expected lag between first attempt and first contact, and both portions of this model are needed to arrive at that prediction.

3.3. *Prior Derivation for Bayesian Prediction*

We derive prior distributions for the coefficients in our Weibull hurdle model with the goal of improving the accuracy of predictions of those coefficients, in order to then improve the predictions of lag between first attempt and first contact during the data collection period. We assume that historical data will provide informative priors for each of our model coefficients. We accomplish this by first fitting a Weibull hurdle model using three months of historical data to obtain parameter estimates $\left(\hat{\gamma}, \hat{\beta}, \hat{\sigma}\right)$, and the associated estimated variances $\hat{V}\left(\hat{\gamma}, \hat{\beta}, \hat{\sigma}\right)$. These parameters capture the time-varying nature of some of the covariates within a single data collection period in a survival model framework.

Assuming approximate normality by the properties of maximum likelihood estimates, we form priors, $p\left(\gamma, \beta, \sigma\right) = N\left(\left(\hat{\gamma}, \hat{\beta}, \hat{\sigma}\right), c\hat{V}\left(\hat{\gamma}, \hat{\beta}, \hat{\sigma}\right)\right)$, where c is a constant that controls the degree to which the prior information is used in the daily estimation procedure. (A normally distributed prior distribution is a logical choice given the asymptotic normality of maximum likelihood estimates.) We assume that the estimates of the model coefficients will have similar precision each month, due to similar sample sizes. As a result, we would expect that the prior estimates for the coefficients, which are based on three months of data instead of just one, would be three times as precise as the coefficient estimates generated from just the in-progress data collection period. As part of a sensitivity analysis, we inflate the variance of each of the priors by a factor of c. By varying the inflation factor as 1/3, 1/2, 1, 2, and 3, we varied the precision of the prior coefficients to represent 27 months, 12 months, 3 months, 3/4th of a month, and 1/3rd of a month, respectively.

3.4. *Implementation*

The posterior distributions for coefficients of all covariates in the Weibull-hurdle model are estimated each day using the priors and all data through the most recent day from the in-progress data collection period. The `brms` package was then used to conduct

resampling and estimation for Bayesian inference of parameters and estimation of posterior predictions. The code for generating predictions via all methods for a given day and month are provided in the online supplementary materials.

4. Methods

4.1. Predicting Day Lag to First Contact: Four Methods

For this application, we used 24 months of data from the NHIS, covering the time period from July 2014 through June 2016. For each prediction data collection period, the prediction of interest was the length of time in days that would elapse between the first contact attempt and the first contact with a household sample member. In addition to the *Bayesian method* described in Section 3, we consider three alternative approaches for comparison:

1. The first method uses only available data from the in-progress data collection period to estimate coefficients for the covariates in a Weibull hurdle model, and then uses those coefficients to predict the expected lag for each open case in the current month. We refer to this as the *current method.*
2. The second method uses only historical data to estimate the mean expected lag. The expected lag for all cases in the current month, then, is just the overall average lag of the three prior months, ignoring any additional information. This leads to all sample units having the same predicted lag in a given month. We refer to this as the *mean method.*
3. The third method also uses historical data but estimates coefficients for the covariates in a Weibull hurdle model, and then uses the point estimates of those coefficients to predict the expected lag for each open case in in the current month. We refer to this as the *historical method.*

For the *mean, historical,* and *Bayesian* methods, which all leverage historical data, three consecutive months were used as the historical data, and the next month was considered the prediction period of interest. For example, if October 2014 was the predictive period of interest, July, August and September of 2014 would be used to generate the mean for the *mean* method, coefficients and standard errors for model covariates for the *historical* method, and point estimates and standard errors for use as priors in the *Bayesian* method. Table 4 summarizes which months of data are used for each method for a prediction period of interest.

We generate predictions at the case- and day- level for each of the 21 time periods of interest (October 2014–June 2016), giving us a set of daily predictions for each case that has not yet had a successful contact.

4.2. Evaluation Measures

The four methods will be compared primarily using measures of mean prediction bias, root mean squared error, and percent change between any two of the four methods. Cases in the in-progress month for which we are evaluating our predictions are included in these calculations if no contact with a sample member has been made on the day the prediction

Table 4. Summary of prediction methods by data types used.

Prediction method	Months used to estimate model coefficients	Coefficient parameters	Prediction period of interest
Current	October 2014 (in-progress)	Coefficient point estimates	October 2014
Mean	July 2014 August 2014 September 2014	Mean predicted lag	
Historical	July 2014 August 2014 September 2014	Point estimates for coefficients in Weibull hurdle model	
Bayesian	*Priors:* July 2014 August 2014 September 2014 (priors) *Likelihood:* October 2014 (in-progress)	Point estimates, standard errors for coefficients in Weibull hurdle model based on historical data; Posterior estimates and standard errors for coefficients in Weibull hurdle model based on priors plus October 2014 accumulated data	

has been made, and a first contact attempt was made prior to the day the prediction was made. Mean prediction bias (MPB) is expressed as:

$$MPB_{d_j}^{m} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}^{m}{}_{id_j} - y_i\right)$$

and root mean square prediction error (RMSE) is expressed as:

$$RMSE_{d_j}^{m} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}^{m}{}_{id_j} - y_i\right)^2}$$

where n is the number of cases eligible for inclusion in the error estimate (this varies throughout the prediction period of interest), $\hat{y}_i$ is the predicted value of the lag, y, for the ith case under the mth method on the dth day, for cases first attempted on day j, and y_i is the true value of the lag, y.

The percent change between the RMSE for the mth method and the method selected as the baseline, b, on the dth day for cases first attempted on day j is calculated as:

$$PCHG_{d_j}^{m_1 m_b} = 100*\left(\frac{RMSE_{d_j}^{m} - RMSE_{d_j}^{b}}{RMSE_{d_j}^{b}}\right)$$

Percent change in the MPB can be calculated similarly. Creating measures of MPB, RMSE, and PCHG based on d_j, the cutpoint by day the case was first attempted leads to a different number of summary error estimates on each day. For example, on day $d = 4$, for each of the methods, there would be three separate estimates of MPB, one each for $j = 1, 2, 3$, where j denotes the day the case was first attempted. One day $d = 25$, there would be up to 24 separate estimates for MPB. Days where no cases were attempted for the first time would not generate an estimate for MPB. This allows us to not only examine the distribution of MPB and RMSE at each cut point, but also examine how the distribution of error in predicted lag differs by how close to the day of first attempt the predictions are made.

The expectation is that the *Bayesian* method will produce predictions of the expected lag closer to that of the actual lag, resulting in a smaller overall MPB and RMSE, as the modeling procedure is effectively borrowing strength across the historical data and current data to make a prediction. For this evaluation, we do not use design-adjusted variance estimates.

4.3. Evaluation Methods

In order to compare the four discussed methods for predicting lag (*current (C), mean (M)*, *historical (H)*, and *Bayesian (B)*), we first compare the MPB and RMSE of the prediction of the lag for each of the four methods using boxplots that illustrate the mean and range of the errors in prediction of lag for eligible cases on each day. Predictions of the expected lag and the resulting MPB and RMSE of those predictions depend not only on the prediction method used, but also on when the prediction is made during data collection, which we refer to as the cut point. This is because each day, new cases are attempted for the first time, and contact is being made in other cases. Therefore, we are attempting to predict the expected lag for a pool of cases that is changing on a daily basis. In order to evaluate the quality of predictions, we evaluate MPB and RMSE at several cut points in data collection, after days 2, 4, 6, 8, 10, 15, 20 and 25.

Additionally, we are particularly interested in the quality of predictions of expected lag shortly after the first contact attempt is made. If we are able to correctly predict the lag close to the day of the first contact attempt, we maximize the window where we can implement an intervention to inform the interviewer or even try to reduce that expected lag. To see this, we recreate the initial plots comparing MPB and RMSE, but restrict them to cases who have had their first contact attempt occur within two days of the cut point. In other words, on day 6, these boxplots would only display the mean and range of errors in prediction of lag for cases first attempted on day 4 or 5. On day 10, these plots would be limited to cases first attempted on days 8 or 9.

Additionally, in the online supplementary materials, we include an example of how this prediction could be used during a data collection period to identify cases that are likely to have a long lag.

5. Results

5.1. Comparison of Four Predictive Methods (MPB and RMSE)

For each cut point during each of the 21 time periods, we used all open, uncontacted cases to generate estimates of MPB and RMSE. The boxplots were then generated using the $MPB^m_{d_j}$ and $MPB^m_{d_j}$ for all m time periods.

Figure 2 displays estimates of MPB for predictions of the lag and shows that all methods underestimate the actual lag between first attempt and first contact throughout the data collection process. Additionally, as data collection goes on, the underestimation MPB increases. It is also evident that, until day 15, all methods that use historical data in some form (*M, H, and B)* outperform the c*urrent* method with respect to prediction MPB. While the three methods that use historical data all perform similarly late in data collection, the *historical* and *Bayesian* methods perform better than the *mean* method until day 10. The *Bayesian* method appears to provide small improvements over the *historical* method over the 21 time periods, as evidenced by central tendencies of estimates of the MPB being closer to zero. The use of historical data is particularly helpful in reducing bias in the very early period, privileging the stronger priors; by day 4 however, the prior has little impact.

In addition to the overall predictive abilities of the four different methods, we are particularly interested in the temporal effectiveness of the predictions. In other words, we are interested in how well we can predict the expected lag near the day of the first attempt, when we are close to the entry point of a given case into the dataset. Just after the initial contact attempt, we have the least data and the most time to implement an intervention if needed. In order to explore this, we generated Figure 3, which is equivalent to Figure 2 but restricted the to include only measures of error based on cases worked within two days of the cut point. So, when $d = 4$, only cases that were first attempted on day 2 or 3 are included in the estimates of MPB. On day 25, only those cases first attempted on days 23 and 24 are included.

While Figure 3 follows the same general pattern as Figure 2, there are some differences. Most notable is that as time progresses through the data collection period, the MPB in predicting lag for recent cases is much smaller (the bias is closer to zero) for all cases, whereas in Figure 2, the under-estimation bias increases over time. This finding shows that the model is relatively accurate at predicting lag length close to the day of the first attempt, whether that lag is long or short. However, as time moves away from the initial contact attempt, the quality of predictions of the lag decreases for the outstanding cases. This suggests that cases where we underestimate lag, we underestimate it by a large amount. Again, the *current* data method performs the worst until late in data collection, but here, it never really outperforms the *historical* or *Bayesian* methods. Late in data collection, the *mean* method begins outperforming other methods, but this is not consistent throughout data collection. Again, we see that the central tendencies of the MPB in the *Bayesian* method are closer to zero than in the *historical* method, demonstrating a small improvement. Among the *Bayesian* methods, we again see an advantage for the strong prior at day 2, but little difference from day 4 on.

We see the same pattern in the RMSE of the lag predictions in Figure 4. The *current* method performs worse than the other methods until mid-way through the data collection period, and the other three methods are competitive with each other. Again, the *Bayesian* method appears to provide improvements over the *historical* method until late in the data collection period, as the central tendencies and intraquartile ranges are smaller in the *Bayesian* method than in the *historical* method. As with bias, there is a tendency for the stronger prior to perform better with respect to RMSE at day 2, but then little difference thereafter.

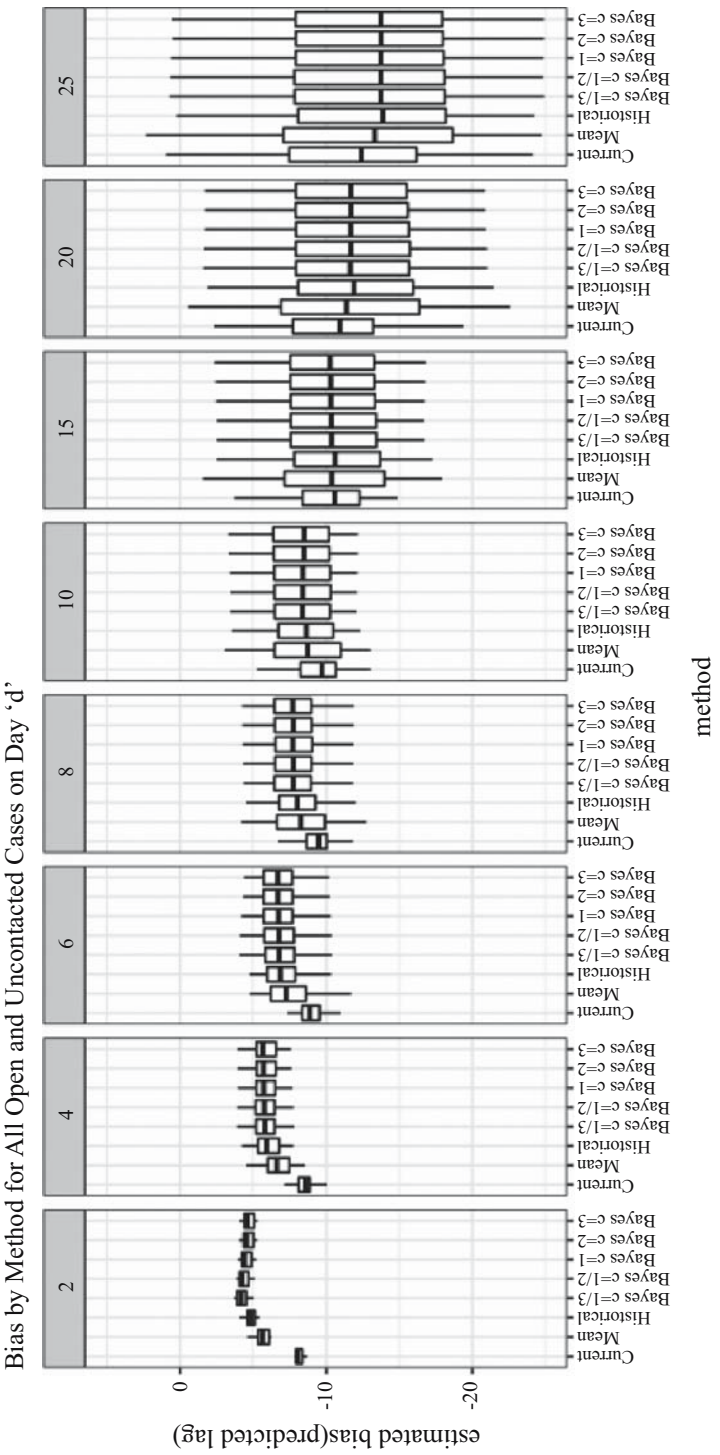

Fig. 2. Bias by prediction method for all open and uncontacted cases on cut point day "d".

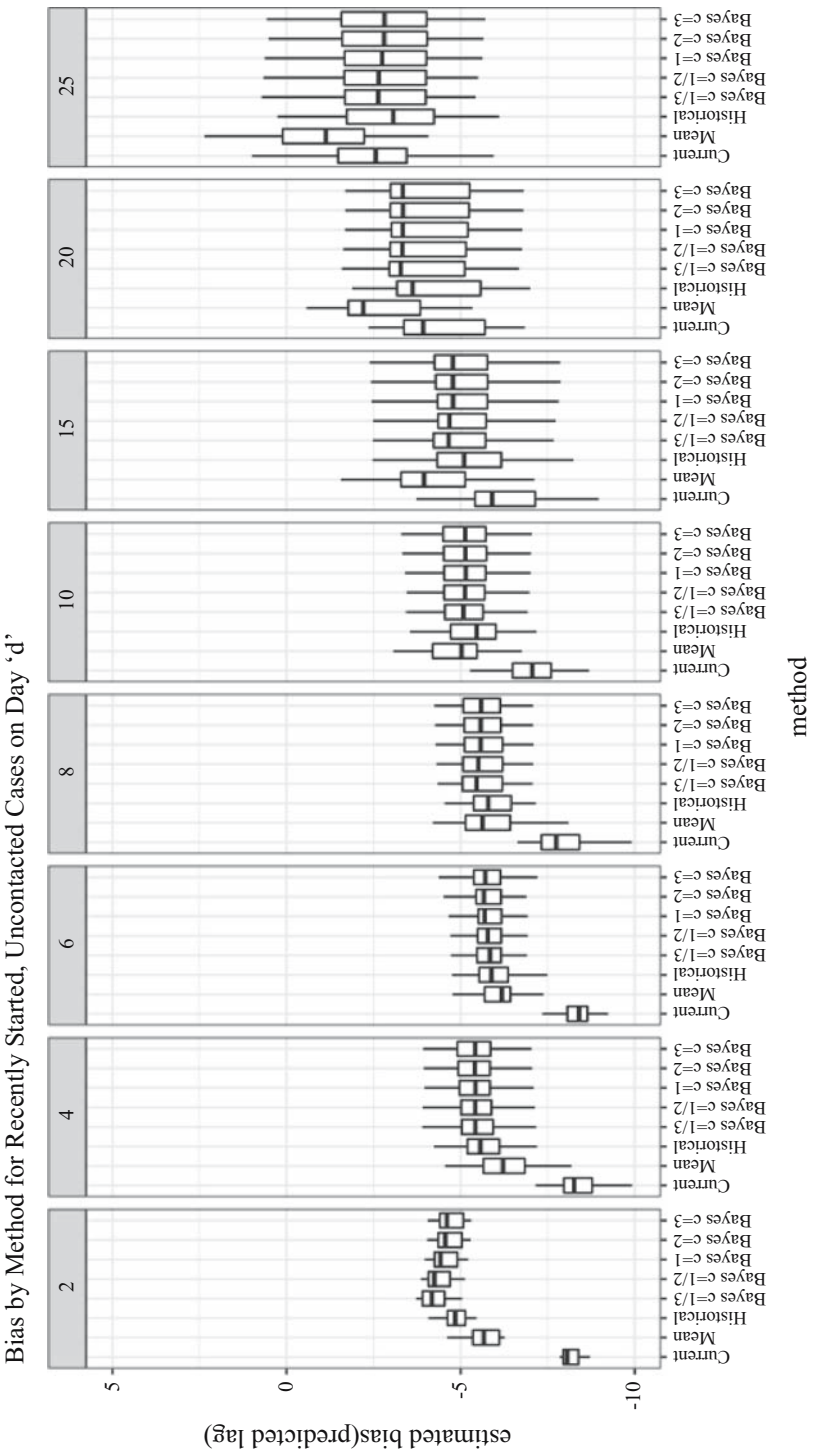

Fig. 3. Bias by prediction method for recent open and uncontacted cases on cut point day "d".

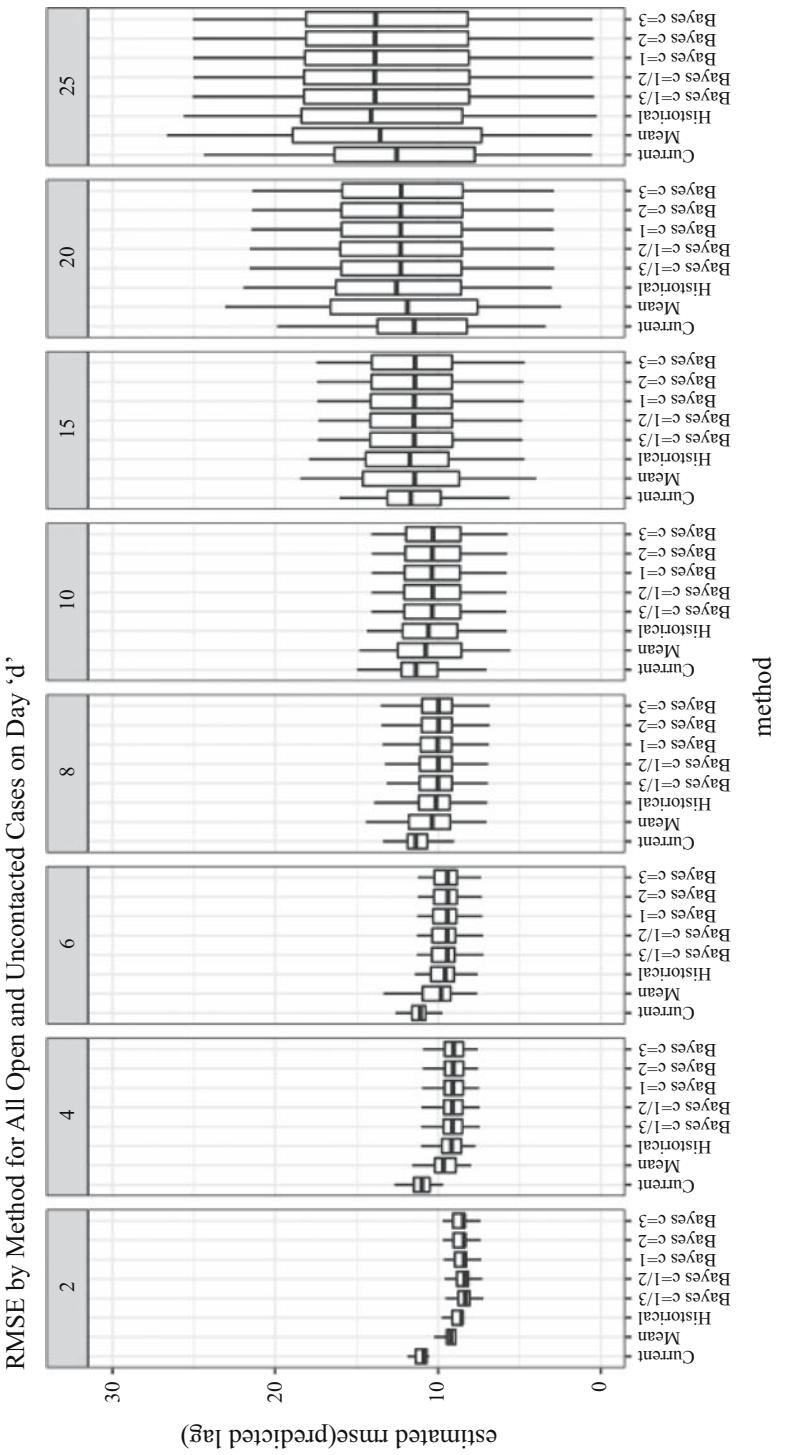

Fig. 4. RMSE by prediction method for all open and uncontacted cases on cut point day "d".

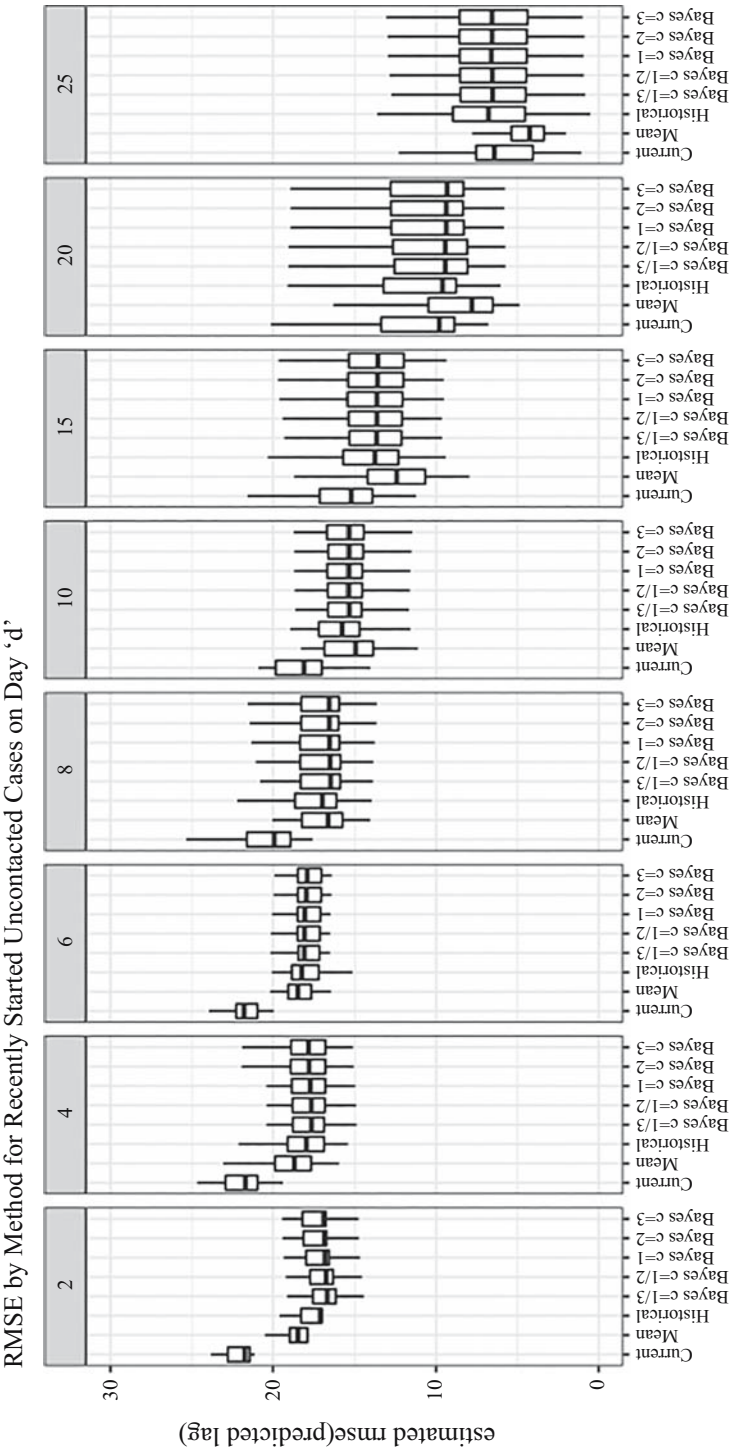

Fig. 5. RMSE by prediction method for recent open and uncontacted cases on cut point Day "d".

Figure 5 displays the RMSE of predictions of lag for recent cases and continues to demonstrate that the use of historical information is generally helpful for improving predictions. Additionally, throughout data collection, the *Bayesian* method provides small improvements in the central tendencies of RMSE over the *historical* method. Again. there is a tendency for the stronger prior to do better with RMSE at Day 2, but then little difference thereafter.

Figure 6 plots the percent change in RMSE and MPB of the *Bayesian* method where the prior is equivalent to 1/3 of a month from the *historical* method. Each point on the scatter plot represents a cut point (prediction day) by day of first attempt. Cases receiving their first contact attempt in the prior two days and cases attempted more than two days prior to the cut point are plotted separately. For example, one data point in the plot of recent cases would be for the measures of MPB and RMSE on Day 4, for cases first attempted on Day 2, in a given data collection period.

Most of the data points fall in the southwest quadrant, representing a reduction of both RMSE and MPB in the *Bayesian* method when compared to the *historical* method. This is true for both recent and older cases. Additionally, when there are improvements in MPB and RMSE, those improvements have a larger range (reductions in the bias and RMSE reaching 20% and 10% respectively) than the situations where the bias or RMSE increase when the *Bayesian* method is used (increases are generally limited to 10% increases in MPB and 5% increases in RMSE). These results suggest the *Bayesian* approach can provide modest improvements to our predictions.

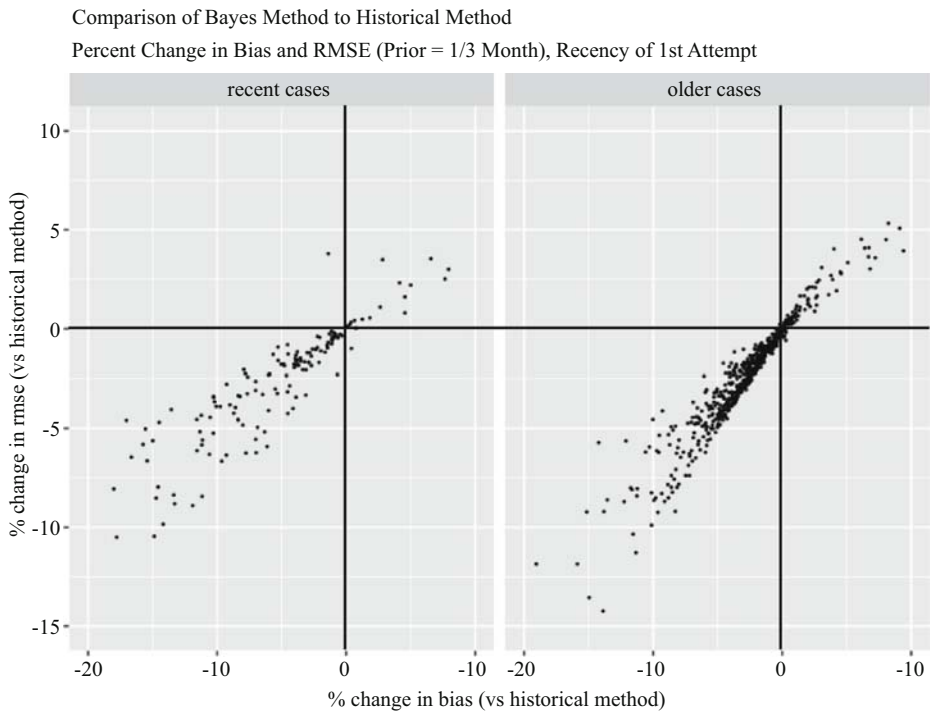

Fig. 6. Percent change in RMSE and bias of bayesian method versus. historical method by recency of first attempt.

6. Discussion and Future Work

This article discussed several methods for generating predictions for the estimated lag (in days) between first attempt and first contact for a case in the NHIS. The results showed that using model-based methods that leverage data external to the current data collection, in this case historical survey data, can be useful for improving predictions in the current month.

Because of the time-varying nature of the "day of the first attempt covariate", the *historical method* and the *Bayesian method* were able to capture that variability, resulting in better predictions, particularly near the point of first attempt. Further, the model-based methods improved estimates with relatively few covariates, including aggregate geographic information, basic interviewer information, and the day of first attempt. Improvements could be more significant with more information, either for individual cases in a longitudinal setting, or a richer auxiliary frame data in a cross-sectional setting.

The *Bayesian* method provided modest improvements over the *historical* method, with the largest gains visible early in the data collection period. Additionally, the *Bayesian* method provided small improvements in the identification of cases with excessive lags at a higher sensitivity than the other methods, particularly when weak priors were used. When sensitivity and specificity were considered equally in an application of the use of predictions, the *Bayesian* method performed slightly better early in data collection, though the *historical* method was competitive. In general, early in data collection the benefit of the Bayesian framework is that external data can be incorporated into the prediction process as priors. Those priors can help improve the stability of predictions when working with the partially accumulated data that exist early in the data collection period. Later in data collection, the likelihood (data from the in-progress period) takes over, reducing bias that might have been introduced by considering historical data only, although the prior continues to reduce the variance of the predictive coefficients, stabilizing the posterior predictions of expected lag. The *Bayesian* method also offers flexibility in how much influence the prior has on the posterior prediction, in the form of the constant, c.

There are limitations to the model-based predictions. From a data quality perspective, the NOI and the CHI data are self-reported by interviewers, who are expected to record outcomes of contact attempts immediately after they occur. West and Kreuter (2013), Biemer et al. (2013), and Bates et al. (2010) have discussed potential quality issues in self-reported interviewer paradata, due to under-reporting of undesirable behaviors, or the error inherent in guesswork. Additionally, as shown in the comparison of Figure 2 and Figure 3, predictions can be more inaccurate the farther in time away from the first attempt for a case, meaning the covariates currently in the model underestimate the length of the lag for cases with long lags. This may be for one of two reasons. First, in the hurdle portion of the model, estimated by logistic regression, the discrimination between sample units with zero lag and those with positive lags is not perfect. As a result, some units for which we predict a zero lag with high confidence may, in fact, have a long lag, and the high probability of a zero lag would cause our predictions to underestimate the expected lag. Then, in the Weibull portion of the model, which estimates lag length for non-zero lags, the estimate of the shape parameter may also underestimate the actual lag.

Future work could extend into several different areas. First, we used a standard parametric Weibull model to predict time (days) to contact. Models for integer-coarsened

data, perhaps as a special case of interval-censored data (Sun 2006), could be considered, although we believe the impact of this level of refinement would be relatively minor. More work could be done to append auxiliary variables onto the survey data, and potentially include a flexible set of model covariates over time, in order to improve model fit and subsequent predictions. Household-level data may be more predictive than aggregate geographic information from the PDB. Additionally, this method should be replicated for different estimators in different data collection settings. Selected data collection parameters and associated models would likely be different across surveys with different data collection periods, modes, or data collection features. In situations where time-to-event modeling is useful or appropriate for estimating selected data collection parameters, it may be useful to explore additional tools for identifying the appropriate survival model, including non-parametric methods, such as Kaplan-Meier curves, in addition to the goodness-of-fit tests used in Subsection 3.1.

More benefit could be extracted from the Bayesian framework, as well. Here, we leveraged the Bayesian framework primarily to combine both historical and current accumulating data to maximize the information used in our predictions of posterior coefficient estimates. But after that, we only used the point estimate of the predicted lag for evaluating our predictions or identifying short and long lags in the online supplementary application. We could also leverage the credible interval around predictions for evaluation or application. In the evaluation phase, we could evaluate whether the predicted credible interval for a case covers its actual lag, as opposed to only looking at the distance between the predicted point estimate and actual lags. Similarly, we could rank cases that have minimally overlapping credible intervals by predicted lags and evaluate whether those rankings match the rankings in the actual lags. This would help us understand whether our predictions of lag were biased downward but ordered correctly. Additionally, we could evaluate the accuracy of predictions for cases with narrow credible intervals versus wide intervals, to better understand if there are domains or types of cases for which our predictions are more accurate than others.

In an application setting, the credible interval could be used to help identify cases for intervention. In the case where a threshold is used (e.g., lags of at least a certain number of days are considered long), the credible interval could be used to determine whether a case met that threshold. For example, in the case where a point estimate met the threshold, but a large portion of the credible interval did not meet the threshold, we may not consider that case to have a long lag. The additional information provided by the credible interval would help identify the cases likely to have long lags in a more complete way. One could also consider the distribution of rank order, for example focusing on cases that we find to be likely to have the longest times to contact for intervention if resources are limited for such intrerventions. The strategy for assigning a value to c is also important. For this application, we varied the value of c, allowing the priors to represent a range of 1/3 of a month of data to three months of data for the results in order to demonstrate the effect of prior data strength on the posterior predictions of lag. However, the Bayesian framework could be extended by putting a prior on the value for c using power prior methodology (Ibrahim and Chen 2000; Hobbs et al. 2011). Power priors provide a statistical framework for weighting priors in a more sophisticated way than we did for this work.

Finally, it is important to consider how to utilize these improved predictions during data collection. Although this article has focused on the technical issue of predicting the day lag

between first attempt and first contact, there is a broader issue of exactly how these results could be used to affect action in the field. There are at least two major categories of actions that could be taken. The first is load leveling. Given the extra effort that a successful contact will entail, one could ensure that interviewers are assigned a reasonable mix of cases that will allow them to get through a workday without leaving a large number of uncontacted cases because they had too many likely responders on a given day or end up short of cases to work because they had too many unlikely responders. The second category follows adaptive or responsive design interventions. For example, if time-to-contact was uncorrelated with response propensity or outcomes of interest, then at some point during data collection, cases could be dropped from operations if the estimated time-to-contact was too long. Alternatively, if time-to-contact was negatively correlated with response propensity, or correlated in any way with outcomes of interest, cases with longer predicted time-to-contact could be assigned high priority early in the collection period so that interviewers spent more time on those cases, or have alternative data collection strategies applied, such as assigning the case to multiple interviewers, leaving materials that might improve the chance for future contact, and so forth.

7. References

Bates, N., J. Dahlhamer, P. Phipps, A, Safir, and L. Tan. 2010. "Assessing Contact History Paradata Quality Across Several Federal Surveys," In Proceedings of the American Statistical Association 2010 Joint Statistical Meeting, Vancouver, Canada. Available at: http://www.asasrms.org/Proceedings/v2010f.html (accessed July 2017).

Biemer, P., P. Chen, and K. Wang. 2013. "Using Level-Of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society* A176: 147–168. DOI: https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Biffignandi, S., and J. Bethlehem. 2021. Web Surveys and Other Modes of Data Collection. In *Handbook of Web Surveys*. DOI: https://doi.org/10.1002/9781119371717.ch6.

Calinescu, M., S. Bhulai, and B. Schouten. 2013. "Optimal Resource Allocation in Survey Designs," *European Journal of Operations Research* 226: 115–121. DOI: https://doi.org/10.1016/j.ejor.2012.10.046.

Census Bureau. 2008. *A Compass for Understanding and Using American Community Survey Data*. Available at: https://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf (accessed October 2017).

Census Bureau. 2016. *Planning Database Documentation.* Available at: https://www.census.gov/research/data/planning_database/2016/docs/PDB_Block_Group_2016-07-28a.pdf (accessed July 2017).

Chesnut, J. 2013. "Model-based mode of data collection: switching from internet to mail in the American Communities Survey." *2013 American Communities Survey Research and Evaluation Report Memo Series* #ACS-13RER, 18, 1–17. Available at: https://www.census.gov/library/working-papers/2013/acs/2013_Chesnut_01.html (accessed July 2017).

Christy, J. 2014. "Use of Response Propensity Scores to Direct CAPI Field Activity." In Proceedings of the American Association for Public Opinion Research Annual Conference, May 15–18, Anaheim, CA, USA. DOI: https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Coffey, S., B. Reist, and P.V. Miller. 2019. "Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates," *Journal of Survey Statistics and Methodology* 8: 726–747. DOI: https://doi.org/10.1093/jssam/smz026.

Couper, M.P. 2000. "Usability Evaluation of Computer-Assisted Survey Instruments," S*ocial Science Computer Review* 18: 384–396. DOI: https://doi.org/10.1177/089443930061800402.

Couper, M.P. 2017. "Birth and Diffusion of the Concept of Paradata " *Advances in Social Research* 18: 14–26. Available at: https://jasr.or.jp/english/JASR_Birth%20and%20-Diffusion%20of%20the%20Concept%20of%20Paradata.pdf (accessed October 2020).

Durrant, G.B., and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys," *Journal of the Royal Statistical Society* A172: 361–381. DOI: https://doi.org/10.1111/j.1467-985X.2008.00565.x.

Edwards, B., H. Sun, and R. Hubbard. 2020. "Behavior change techniques for reducing interviewer contributions to total survey error." In *Interviewer Effects from a Total Survey Error Perspective*: 77–90. Chapman and Hall/CRC.

Feng, C.X. 2021. "A comparison of zero-inflated and hurdle models for modeling zero-inflated count data," *Journal of Statistical Distributions and Applications* 8: 8. DOI: https://doi.org/10.1186/s40488-021-00121-4.

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*, New York: Wiley.

Groves, R.M., and S.G. Heeringa. 2006. "Responsive Designing for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistical Society* A169: 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Hobbs, B.P., B.P. Carlin, S.J. Mandrekar, and D.J. Sargent. 2011. "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials." *Biometrics* 67: 1047–1056. DOI: https://doi.org/10.1111/j.1541-0420.2011.01564.x.

Ibrahim, J.G., and M.H. Chen. 2000. "Power prior distributions for regression models." *Statistical Science* 15: 46–60. DOI: https://doi.org/10.1214/ss/1009212673.

Jackson, M.T., C.B. McPhee, and P.J. Lavrakas. 2020. "Using Response Propensity Modeling to Allocate Noncontingent Incentives in an Address-Based Sample: Evidence from a National Experiment," *Journal of Survey Statistics and Methodology* 8: 385–411, DOI: https://doi.org/10.1093/jssam/smz007.

Laflamme, F., and M. Karaganis. 2010. "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." In Proceedings of the Q2010, 3–6 May, Helsinki, Finland. Available at: https://www.researchgate.net/profile/Francois-Laflamme-3/publication/228583181_Implementation_of_Responsive_Collection_Design_for_CATI_Surveys_at_Statistics_Canada/links/5436b2120cf2dc341db46d70/Implementationof-Responsive-Collection-Design-for-CATI-Surveys-at-Statistics-Canada.pdf (accessed July 2017).

Luiten, A., J. Hox, and E. de Leeuw. 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. DOI: https://doi.org/10.2478/-JOS-2020-0025.

Ma, L., M. Yan, and J. Weng. 2015. "Modeling Traffic Crash Rates of Road Segments through a Lognormal Hurdle Framework with Flexible Scale Parameter," *Journal of Advanced Transportation* 49: 928–940.

Mittereder, F., and B.T. West. 2021. "A Dynamic Survival Modeling Apporach to the Prediction of Web Survey Breakoff." *Journal of Survey Statistics and Methodology* 10: 945–978. DOI: https://doi.org/10.1093/jssam/smab015.

Mneimneh, Z., L. Lyberg, S. Sharma, M. Vyas, D.B. Sathe, F. Malter, and Y. Altwaijri. 2018. "Case Studies on Monitoring Interviewer Behavior in International and Multinational Surveys." In *Advances in Comparative Survey Methods*. John Wiley & Sons. DOI: https://doi.org/10.1002/9781118884997.ch35.

Mullahy, J. 1986. "Specification and Testing of some Modified Count Data Models," *Journal of Econometrics* 33: 341–365. DOI: https://doi.org/10.1016/0304-4076(86)90002-3.

NCHS, National Center for Health Statistics. 2018. *National Health Interview Survey: Survey Description*. Technical Report. Available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2017/srvydesc.pdf (accessed July 2017).

Peytchev, A., J. Rosen, S. Riley, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias through Case Prioritization." *Survey Research Methods* 4: 21–29. DOI: https://doi.org/10.18148/srm/2010.v4i1.3037.

Peytchev A. 2014. *Models and Interventions in adaptive and responsive survey designs*. DC-AAPOR Panel on Adaptive Survey Design. Washington D.C. Available at: http://dc-aapor.org/ModelsInterventionsPeytchev.pdf (accessed July 2017).

Peytchev, A., D. Pratt, and M. Duprey. 2020. "Responsive and Adaptive Survey Design: Use of Bias Propensity During Data Collection to Reduce Nonresponse Bias," *Journal of Survey Statistics and Methodology* 10: 131–148. DOI: https://doi.org/10.1093/jssam/smaa013.

Roberts, C., C. Vandenplas, and M.E. Stahli. 2014. "Evaluating the Impact of Response Enhancement Methods on the Risk of Nonresponse Bias and Survey Costs." *Survey Research Methods* 8: 67–80. DOI: https://doi.org/10.18148/srm/2014.v8i2.5459.

Rose, C., S. Martin, K. Wannemuehler, and B. Plikaytis. 2006. "On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data," *Journal of Biopharmaceutical Statistics* 16: 463–481. DOI: https://doi.org/10.1080/10543400600719384.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for Representativeness of Survey Response." *Survey Methodology* 39: 29–58. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf?st=Ox4bRwqM.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Survey Response," *Journal of Official Statistics* 27(2): 231–253. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf.

Schouten, B., N. Mushkudiani, N. Shlomo, G. Durrant, P. Lundquist, and J. Wagner. 2018. "A Bayesian Analysis of Design Parameters in Survey Data Collection," *Journal of*

Survey Statistics and Methodology 6: 431–464: DOI: https://doi.org/10.1093/jssam/smy012.

Smithson, M., and E. Merkle. 2013. "Chapter 5: Count Variables," In *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. New York: Chapman and Hall/CRC. Print.

Sun, J. 2006. *The Statistical Analysis of Interval-Censored Failure Time Data* New York: Springer.

Tolliver, K., J. Fields, S. Coffey, and A. Nagle. 2019. "Combating Attrition Bias Using Case Prioritization in the Survey of Income and Program Participation." In Proceedings from the 2019 AAPOR Conference, Toronto, Ontario, USA. Available at: http://www.asasrms.org/Proceedings/y2019/files/1199523.pdf (accessed July 2017).

Wagner, J., and T.E. Raghunathan. 2010. "A New Stopping Rule for Surveys." *Statistics in Medicine* 29: 1014–1024. DOI: https://doi.org/10.1002/sim.3834.

Wagner, J., and F. Hubbard. 2014. "Producing Unbiased Estimates of Propensity Models during Data Collection," *Journal of Survey Statistics and Methodology* 2: 323–342. DOI: https://doi.org/10.1093/jssam/smu009.

Wagner, J., B.T. West, N. Kirgis, J.M. Lepkowski, W.G. Axinn, and S.K. Ndiaye. 2012. "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection." *Journal of Official Statistics* 28: 477–499. Available at: https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/use-of-paradata-in-a-responsive-design-framework-to-manage-a-field-data-collection.pdf (accessed July 2017).

Walejko, G., and J. Wagner. 2018. "A Study of Interviewer Compliance in 2013 and 2014 Census Test Adaptive Designs" *Journal of Official Statistics* 34(3): 649–670. DOI: https://doi.org/10.2478/jos-2018-0031.

Walsh, R., and J. Coombs. 2013. "Case Reassignment: When making contact is a two-person job." In Proceedings of the Joint Statistical Meetings, Boston, MA, USA: 3690–3700. Available at: https://www.researchgate.net/publication/296847275_Case_Reassignment_When_making_contact_is_a_two-person_job (accessed July 2017).

West, B.T., and R.M. Groves. 2013. "The PAIP Score: A Propensity-Adjusted Interviewer Performance Indicator," *Public Opinion Quarterly* 77: 352–374. DOI: https://doi.org/10.1093/poq/nft002.

West, B.T., and F. Kreuter. 2013. "Factors Affecting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth (NSFG)." *Public Opinion Quarterly* 77: 522–548. DOI: https://doi.org/10.1093/poq/nft016.

West, B.T., J. Wagner, S. Coffey, and M.R. Elliott. 2021. "Deriving Priors for Bayesian Prediction of Daily Response Propensity in Responsive Survey Design: Historical Data Analysis vs. Literature Review." *Journal of Survey Statistics and Methodology* 11: 367–392. DOI: https://doi.org/10.1093/jssam/smab036.

Williams, D., and J.M. Brick. 2018. "Trends in U.S Face-to-Face Household Survey Nonresponse and Level of Effort." *Journal of Survey Statistics and Methodology* 6: 186–211. DOI: https://doi.org/10.1093/jssam/smx019.

Received April 2022
Revised October 2022
Accepted January 2023