**Non-response Bias Adjustments in the National Health Interview Survey (NHIS)**

Ronaldo Iachan, ICF

Matt Jans, ICF

Yangyang Deng, ICF

Adam Lee, ICF  
Stephen Blumberg, NCHS

Jim Dahlheimer, NCHS

*Introduction*

In long-running surveys, such as the National Health Interview Survey (NHIS), conducted by the National Center for Health Statistics (NCHS), sample and questionnaire redesigns offer an opportunity to assess and innovate weighting adjustments. NHIS has used a weighting class adjustment approach to correct for nonresponse bias, combined with post-stratification, to produce final weights since 2010. Advances in statistical modeling and nonresponse research over the past 20 years make this an opportune time for NCHS to reconsider the nonresponse weighting approaches used for NHIS.

The primary study goal was to assess the possibility of using nonresponse prediction to inform the nonresponse adjustment. The current NHIS nonresponse weighting approach only adjusts for the geographic area unit and does not account for other factors that predict nonresponse, such as neighborhood characteristics and contact attempt history. Using regression modeling and machine learning methods to predict nonresponse and creating weighting adjustment classes from those results allow NHIS to include more predictors of response in the adjustment. As nonresponse bias is most likely to occur when response propensity is correlated with a key health indicator (KHI), the study focused on variables that predict both response and KHIs.

The ICF team evaluated two regression approaches (single-level logistic regression and multilevel logistic regression) and two machine learning methods (random forest and least absolute shrinkage and selection operator [LASSO]) for nonresponse weighting. In addition, for calibration, the team compared the current NHIS post-stratification to age, sex, and race/ethnicity population control totals with two raking approaches: one used the same calibration dimensions, while the other added education, employment, and metropolitan statistical area (MSA).

This study was implemented in conjunction with the NHIS 2019 redesign, using data from the 2018 Quarter 4 Bridge Sample, in which only half of the NHIS sample received the redesigned protocols, and the 2019 Quarter 1 fully redesigned sample.

*Methods*

The NHIS data structure, including its contact history paradata and neighborhood observations, provides a rich, multilevel source of information that can be used to thoroughly understand nonresponse predictors and correlates of bias. While proximal measures of resistance derived from paradata tend to be the strongest predictors of unit nonresponse[[1]](#footnote-1),[[2]](#footnote-2), there is also evidence that community and neighborhood-level factors can influence participation[[3]](#footnote-3),4. Furthermore, repeated contacts to households during data collection create an additional level of nested information about the precursors of nonresponse[[4]](#footnote-4),[[5]](#footnote-5),[[6]](#footnote-6),[[7]](#footnote-7). These can be summarized to the sample unit (i.e., household level) or modeled explicitly, for example, with discrete time logistic regression models6.

The goals of this study were as follows:

1. Quantify the nonresponse bias present in the redesigned NHIS.
2. Evaluate current NHIS weighting compared with other weighting methods that take advantage of auxiliary data, paradata, and machine learning statistical methods, with a primary focus on bias reduction key health indicators (KHIs).
3. Provide NCHS with evidence and reasoning regarding whether to implement a new weighting approach for NHIS using these alternative approaches.

The study had three assessments stages:  
Stage 1: Assess which of the candidate regression models and machine learning methods should be used for nonresponse prediction. At this stage, the study focused on metrics that describe how well the models predict response (e.g., area under the curve [AUC] score and mean squared error [MSE]). This led to narrowing the focus to a multilevel model and a random forest method to carry forward into nonresponse weighting.

Stage 2: Assess the characteristics of the nonresponse adjustment weights and final calibrated weights. The NCHS-ICF team reviewed nonresponse and final (i.e., post-calibration) weight distributions, and design effects (DEFF) due to weighting.

Stage 3: Assess the impact of multilevel and random forest on both the point estimates and the standard errors at the nonresponse adjustment stage for key health indicators (KHI). Final weighted KHIs were also compared for weights using multilevel modeling at the nonresponse adjustment stage, and three types of calibration, which compared the current NHIS post-stratification to two iterative raking approaches.

*Results*

Table 1 shows the various data sources used in the models.

Table 1. NHIS Data, Paradata, and Auxiliary Data Sources

| Data Source | Level | Potential Use in  Predicting Response |
| --- | --- | --- |
| **Census Planning Database** | Community/ Neighborhood | * Community-level resistance predictors (i.e., low response score) |
| **Neighborhood Observation Instrument (NOI)** | Neighborhood | * Neighborhood predictors of nonresponse and poor health outcomes |
| **Unit Control File** | Sample unit (address) | * Geographic nonresponse predictors * Methodological controls (e.g., whether the housing unit is a screening or interview unit) |
| **Survey Data** | Sample unit (respondents only) | * Nonresponse bias metrics |
| **Contact History Instrument (CHI)** | Contact attempt | * Interim and final dispositions (i.e., refusals, interview completion) * Methodological controls (e.g., whether contact was made by phone) |

Unit response predictor variable selection was conducted in two stages. First, bivariate analyses helped identify an initial list of potential predictors nonresponse predictors variables that had a significant relationship with household response, adult response, and child response (*α =* 0.1). Second, potential predictors that demonstrated a statistically significant association with response were assessed for their association with each of the KHIs. This two-step approach was used because nonresponse bias occurs when a characteristic of sampled units (i.e., household or people) is associated with both response and a substantive variable of interest. Thus, statistically significant predictors of response that are also associated with KHIs should be most effective at reducing nonresponse bias. The number of significant associations between each predictor and each KHI was then used to select final predictors for the nonresponse bias models.

The predicted response probabilities from the multilevel model and random forest method were divided into quintiles with group 1 representing the lowest propensity to respond and group 5 representing the highest propensity to respond. Due to a small sample size in the first quintile, and the related risk of creating very large weights, response quintiles 1 and 2 were combined into a single low response propensity group. Quintiles 3 through 5 were combined into a high response propensity group to facilitate nonresponse bias identification using the methods previously used in NHIS.[[8]](#footnote-8) The method assumes that low response propensity respondents are similar to nonrespondents for the KHIs used in this study, and that the factors that predict response propensity and bias among respondents extend to nonrespondents.

The multilevel model identified potential bias in 12 of the 15 KHIs, while the random forest method only identified bias in 10. The amount of nonresponse bias identified was very similar, on average, across KHIs between the multilevel model and random forest method (about 2–7 on the *t-value scale* for each prediction approach). However, the KHIs on which bias was identified differed by prediction method. Both methods identified bias in the following KHIs: public insurance (+), private insurance (-), uninsured (-), doctor visit (+), influenza vaccination (+), disability (+), having a prescription for hypertension (+), excellent or very good health (-), usual source of care (+),

Multilevel regression uniquely identified bias in emergency department visits in the past year (+), mental health counseling (+), and skipping medication due to cost (-). The random forest method uniquely identified bias in obesity (+). Neither model identified statistically significant bias in past-year asthma episodes or cigarette smoking. The greatest amount of bias identified in both models was for having public health insurance, receiving an influenza vaccination, and having hypertension.

*Conclusions*

Compared with the current NHIS nonresponse weighting method, regression and machine learning methods both adjusted estimates more strongly (i.e., moved estimates further from the base weights). However, these adjustments reduced effective sample size due to high variability in these weights. This design effect due to weighting could be mitigated by trimming or capping the nonresponse weights in a method similar to the current NHIS method.

The tested raking calibration methods, when compared to the current NHIS post-stratification, show that raking to the same demographic dimensions produced very little change in final estimates. Adding education, employment, and MSA moved point estimates further from the post-stratified estimates (but never more than a percentage point over all estimates assessed) and half a percentage point for low prevalence estimates.

The analyses have a few limitations. First, the weights used in this study were not trimmed or capped as typically done with NHIS weights, limiting comparisons between the current NHIS weighting method and the candidate methods. Trimming or capping high weight values to match the NHIS protocols would reduce the design effects due to weighting and could potentially change the effects that the candidate methods have on point estimates and standard errors. Second, these results only reflect the influence of weighting method on NHIS for one time point and one quarter of data. To understand the effect of these methods on trends over time, additional quarters should be added, and estimates should be assessed for any other temporal aggregations that are publicly reported (i.e., year). Third, we assessed the effect of the candidate weighting methods on a selected set of health estimates.

We identified several other avenues for future NHIS weighting and recommendations as follows.

* Prediction methods, whether regression or machine learning, will likely improve nonresponse bias correction, but may increase overall variance, thus reducing effective sample size.
* Nonresponse weights based on multilevel logistic regression models, on average, perform better than the random forest machine learning method, and have a lower design effect. Combined with transparency and ease of implementation, multilevel regression models are recommended for further nonresponse weighting exploration.
* An extended raking procedure should be considered. This includes conducting iterative raking instead of single post-stratification and expanding the used demographic control totals from only age, sex, and race/ethnicity to also include education, employment, and MSA to provide further bias reduction and correct any overcorrection introduced at the nonresponse adjustment stage.

While the results did not conclusively point to one method over another, and the study did not assess the impact of the weighting methods on trends, the results provide support for significant changes in NHIS weighting procedures. This study discusses the logistical challenges in implementing these methods and suggests additional research that would help clarify the differential impacts of various weighting methods and provide more confidence for NCHS moving forward.

1. Dahlhamer, J. M., & Simile, C. M. (2009). Subunit nonresponse in the National Health Interview Survey (NHIS): An exploration using paradata. In Proceedings of the Survey Research Methods Section, American Statistical Association (pp. 262–276). Retrieved from <http://www.amstat.org/Sections/Srms/Proceedings/y2009/Files/302933.pdf> [↑](#footnote-ref-1)
2. Jans, M., McLaughlin, K., Viana, J., Grant, D., Park, R., & Ponce, N. A. (2018, in press). Geographic correlates of nonresponse in California: A cultural ecosystems perspective. In T. J. Dorer, B. Pennell, I. Stoop, & B. Dorer (Eds.), Advances in Comparative Survey Methodology. Hoboken, NJ: Wiley. [↑](#footnote-ref-2)
3. Casas-Cordero, C., Kreuter, F., Wang, Y., & Babey, S. (2013). Assessing the measurement error properties of interviewer observations of neighbourhood characteristics. Journal of the Royal Statistical Society. Series A, (Statistics in Society), 176(1), 227–249. <https://doi.org/10.1111/j.1467-985X.2012.01065.x> [↑](#footnote-ref-3)
4. Dahlhamer, J. M., & Jans, M. (2011). With an eye toward responsive design: The development of response propensity models with the National Health Interview Survey. Presented at the 22nd International Workshop on Household Survey Nonresponse, Bilbao, Spain. [↑](#footnote-ref-4)
5. Durrant, G. B., & D’Arrigo, J. (2014). Doorstep interactions and interviewer effects on the process leading to cooperation or refusal. Sociological Methods & Research, 43(3), 490–518. <https://doi.org/10.1177/0049124114521148> [↑](#footnote-ref-5)
6. Durrant, G. B., D’Arrigo, J., & Müller, G. (2013). Modeling call record data: Examples from cross-sectional and longitudinal surveys. In F. Kreuter (Ed.), Improving Surveys with Paradata (pp. 281–308). John Wiley & Sons, Inc. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9781118596869.ch12/summary> [↑](#footnote-ref-6)
7. Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and noncontact 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> [↑](#footnote-ref-7)
8. Dahlhamer, J. M., & Simile, C. M. (2009). Subunit nonresponse in the National Health Interview Survey (NHIS): An exploration using paradata. In Proceedings of the Survey Research Methods Section, American Statistical Association (pp. 262–276). Retrieved from <http://www.amstat.org/Sections/Srms/Proceedings/y2009/Files/302933.pdf> [↑](#footnote-ref-8)