Article

Field Methods
2022, Vol. 0(0) 1–13

DOI: 10.1177/1525822X221107053
journals.sagepub.com/home/fmx
SAGE

A Machine Learning Model Helps Process Interviewer Comments in Computer-assisted Personal Interview Instruments: A Case Study

Catherine Billington[1], Gonzalo Rivero[1], Andrew Jannett[1], and Jiating (Kristin) Chen[1]

Abstract
During data collection, field interviewers often append notes or comments to a case in open text fields to request updates to case-level data. Processing these comments can improve data quality, but many are non-actionable, and processing remains a costly manual task. This article presents a case study using a novel application of machine learning tools to assist in the evaluation of these comments. Using over 5,000 comments from the Medical Expenditure Panel Survey, we built features that were fed to a machine learning model to predict a grouping category for each comment as previously assigned by data technicians to expedite processing. The model achieved high top-3 accuracy and was incorporated into a production tool for editing. A qualitative evaluation of the tool also provided encouraging results. This application of machine learning tools allowed a small but worthwhile increase in processing efficiency, while maintaining exacting standards for data quality.

[1]Westat Inc., Rockville, MD, USA

Corresponding Author:
Catherine Billington, Westat Inc., 1600 Research Blvd., Rockville, MD 20850, USA.
Email: catherinebillington@westat.com

Introduction

In-person surveys sometimes plan for the collection and processing of interviewer requests for edits (Bricker et al. 2014; Kennickell 2007). When field interviewers transmit survey data, they may include comments requesting *post hoc* edits to be incorporated. For example, respondents sometimes offer updated or additional information after the interview has progressed too far to back up and make the correction. These interviewer comments contain information that alerts data technicians (DTs) to unusual responses or circumstances that can affect data quality (Athey and Kennickell 2005). Interviewer comments are more frequent in surveys that place a significant burden on respondents, such as asking them to recall information that is distant in the past or that spans a long reference period. Importantly, these comments are different from spontaneous notes left by respondents in self-administered surveys (see McClelland 2016) in that they are part of the collection protocol and may result in edits to the final dataset.

Comment review can provide valuable insights on imperfect question design, training gaps, or bias from an interviewer (Smith 2008 or Galasiński and Kozłowska 2010). At the same time, processing comments is time consuming and the comments are often superfluous or do not contain enough detail to be actionable (Kennickell 2007). The ability to reliably assess these comments and apply standardized data editing procedures in a timely manner is key to improving data quality and contributes to overall processing efficiency.

In this article, we discuss the use of natural language processing and machine learning as tools to reduce the effort required to address the requests left by field interviewers in comments. We address the key challenge raised in previous literature that evaluates the use of interviewer comments to monitor and improve data quality (Kennickell 2007; Windle 2015), namely, that the cost and speed of manual processing are the main factors influencing the scalability of comment review. We applied this approach to the Household Component of the Medical Expenditure Panel Survey (MEPS) sponsored by the Agency for Healthcare Research and Quality (AHRQ 2020), which uses a computer-assisted personal interview (CAPI) to collect data from respondents.

This article contributes to the growing literature on the application of natural language processing and machine learning to the analysis of text data in surveys. Previous research has focused on responses by study participants to open-ended questions, with the goal of upcoding them to discreet variables during data processing (He and Schonlau 2020, 2021; Schonlau et al. 2021) or using them to gain insights at the analysis stage (Pietsch and Lessmann 2018; Roberts et al. 2014). Here, we focus on ad hoc information entered by the interviewer that must be reviewed to determine whether an update to the survey data is merited. This article is also related to literature that uses interviewer observations to improve different steps of the data collection process (West 2013; West et al. 2014). The interviewer comments we study here are entered during instrument administration and can be interpreted as an additional method of communication that can help researchers improve the survey design.

Case Study: The Medical Expenditure Panel Survey

Health care in the United States is a uniquely complex system (see Jonas et al. 2007). Collecting data on the costs and use of that health care requires a similarly complex survey. The MEPS, sponsored by AHRQ, is “the most complete source of data on the cost and use of health care and health insurance coverage” in the United States. (AHRQ 2020). The study follows the civilian U.S. population to collect two years of data via five CAPI interviews. The household component is administered to a sample size of approximately 20,000 households interviewed each spring and 14,000 households each fall. This large survey includes 40 sections on topics such as medical expenses, access to care, child health, insurance, and income. The version of the CAPI instrument used for our analysis encompasses 1,267 possible questions and takes about 90 minutes to administer.

Respondents are asked to recall details about all medical events involving a health care provider during the previous three–six months on average. Follow-up questions ask about utilization and charge/payment information for each event, prescribed medicines and other medical expenses, employment information, health insurance information, and post-interview contact information for the household.

During administration, the CAPI system routes the interview to specific questions based on previous responses and preloaded household data. To avoid the time-consuming task of backing up to a previous section to correct a response, field interviewers can leave a comment at any question by pressing a short-cut key and entering information into an open text field. Field interviewers must select a single category for each new comment, indicating the broader topic of the comment itself, which is used by a team of quality control DTs to perform any necessary edits.

Interviewer Comments

The MEPS household interview asks respondents to recall details about a broad range of health-related events over the past few months. The survey is designed to facilitate recall through probes and by asking households to maintain diaries of their use of health care (Zuvekas 2011). However, respondents often recall additional information at later parts of the interview. To relieve respondent burden while preserving data quality, interviewers can enter this information as a comment at any point in the interview. This can be preferable to “backing up” through the interview to correct the data which can increase the duration of the interview (Olson and Smyth 2015) and affect the overall experience for the respondent (Tarnai and Paxson 2005).

Comments may include information about health care events (“R was diagnosed with diabetes at age 25 but does not have diabetes now”), about insurance providers (“PID also has AARP as a supplemental insurance”), about medications (“Ibuprofen 800 mg also prescribed for ID 104 on 17 June 2019”), or general information collected during the interview (“R’s employer should be spelled ‘Westat’”).

While these comments are often short, they take significant effort to process. Each comment must be reviewed and dispositioned manually as needing one or more data edits. In addition, comments can be difficult to interpret without the in-person context of the interview. Interviewers receive ongoing training to help them enter comments efficiently, but many comments are not actionable by DTs because they contain vague or superfluous information.

Improving Comment Quality from a Technological Perspective

The MEPS introduced a data quality control (DQC) system in the spring of 2018 with the goal of making data review and editing more efficient while preserving data quality. The new DQC system offers two main advantages. First, DQC is tightly integrated with the CAPI instrument. Editing in the data collection instrument allows instrument rules to be enforced in real time and shortens the time and effort required for quality edits and checks.

In addition, DQC provides users with a graphical interface for DTs, which facilitates the standardization of workflows and protocols. Field interviewers are required to select a category before entering each comment. The data review team also uses the comment categories to link the information from comments with the correct standardized data editing procedure. DTs follow step-by-step instructions, which are organized by comment category, to update the data during editing. This process ensures a standard approach for common data edits, provides documentation of decisions, increases efficiency, and allows for trackable metrics.

The motivation for incorporating machine learning in comment review is twofold. First, it can improve workflow efficiency and consistency across technicians. Even small gains on each item, due to the overall volume of comments for a long-running panel study, can amount to significant efficiency improvements over time. The eventual ability to identify comments that are not actionable earlier in the process is also worthwhile. Second, all editing procedures are standardized and indexed by comment category. Quickly identifying the correct category increases data reliability and overall quality.

Data, Methods, and Results

Data

For our analysis, we used 5,366 comments that had been processed by the DQC system along with the corresponding grouping category assigned by one of the DTs from a closed list of nine options plus one catch-all "Other comment." These categories, which cannot co-occur for a single comment, are listed in Table 1 along with their prevalence in the dataset. Note that the distribution is skewed; a few categories offer us less than 100 examples. This table tells us that amendments are often needed for categories that impose the highest burden on respondents, such as those involving the recall of health care events or payment details.

Our model used these previously coded comments to predict the correct category for new comments arriving into the system. We used two types of features: the section and question number at which the comment was entered in the CAPI, and the text of the comment itself. In particular, we built 12 features that captured attributes that are relevant for identifying categories. The first nine features assessed the *presence of a date, dollar amount, zip code, telephone number, a verb referring to an edit, a respondent, age of a person, any city or state, or any personal name.* The logic behind these features is straightforward; a comment indicating the respondent used a specific pharmacy to fill a prescription, for instance, may include the pharmacy's phone number or address. Similarly, comments about changes to the payment details are likely to include dollar amounts and perhaps a date. We captured this information using regular expressions and a named entity recognition (NER) engine (Honnibal and Montani 2017).

The remaining three variables indicated *whether the comment mentions prescription drugs, health insurers, or medical providers.* To build these variables, we used reference databases maintained by the study that listed all

Table 1. Counts and Percentage of the Observation Data in 10 Categories.

Category	*N*	%	Category	*N*	%
Health care events	2,611	48.66	Employment	267	4.98
Health insurance	663	12.36	RU member refusal	129	2.39
Prescribed medicines	646	12.04	Condition	124	2.31
Other	434	8.09	Glasses/Contact lenses	81	1.51
RU/ RU member	335	6.24	Other medical expenses	76	1.42

prescription drugs and insurers and providers in the geographical area of interest. To retrieve this information, we pre-processed the comments using an NER engine to identify likely term candidates in addition to a custom dictionary that included informal references such as "rx," "meds," or "tabs."

The identification of names of persons and organizations presented additional challenges. Because the text is entered manually, an exact match between the identifier used in the comment and the reference database is unlikely. For instance, the interviewer may have captured the name of "Jane Q. Doe" as "Doe, J.," "Jayne Doe," or "Dr. Jane Doe, MD." To address this, we used a combination of token-level and character level string distances to compute a similarity score between the name mentioned in the comment and the one in the reference database. We also used a full-text search engine (Kononenko et al. 2014) to minimize computation time.

All remaining tokens were used as a "bag of words," a popular technique that transforms texts into a matrix that indicates the presence or absence of a given word in a given sentence.

To illustrate the process, consider the following comment as an example.

> *Richard Roe PID 101, medical visit on 6/20/2018 to Dr Jane Doe, at 1600 Research Blvd, Rockville, MD 20850, 301-251-1500. Pt went in after having extreme puffy eyes, reaction to foliage Co-pay $50, don't know total charge. New meds recorded.*

Our process first flagged the tokens that were identified as a reference to the respondent, a date, or an address, and transformed them into custom indicators. Then, all other tokens were standardized to limit the number of variations. For instance, we transformed them to lower case, removed punctuation and common stop words (e.g., "the," "a," "do," "I," "on," etc.), and finally each word. As a result, the pre-processed sample comment would look like this:

> -NAME- -RESPONDENT- medical visit -DATE- -NAME- research blvd –ADDRESS- -ZIPCODE- -PHONE- -RESPONDENT- extreme puffy eye reaction foliage pay -MONEY- know total charge new med record

which was then fed into a machine learning model as a vector of presence–absence of a given token in a sentence.

Model Selection and Performance

We evaluated several alternative classification models.[1] As our final model, we selected a one-versus-all elastic net classification model (Zou and Hastie 2005), which offers the advantage of being simple and fast to train, easy to interpret, and has been shown to produce high performance in a variety of tasks. In our case, an elastic net showed performance comparable to more complicated black-box models. We discuss model performance in the next section.

The hyper-parameters of the model (the size of the penalization factor and the mixing parameter between L1 and L2 penalization factors) were selected via fivefold cross-validation in a simple random sample of 80% of the data. We probed a fixed grid of values for the hyper-parameters and the final option of each category was selected based on the highest testing accuracy.

The performance of the models reported below are based on the 20% of the data that were not used for training. The category-level performance of the model is summarized in Table 2. The model has an overall accuracy of 88.36%, precision of 86.48%, recall of 76.52%, and F1 score of 81.21%. However, this number hides some degree of heterogeneity across the categories, which partially reflects the imbalance in the number of cases available in training. The category Health Insurance, for instance, is correctly predicted 98.79% of the time; the category Health Care Events, meanwhile, has a slightly lower rate of correct prediction of 93.02%.

Table 3 shows a category-level evaluation of the model through accuracy, precision, and recall. The *recall* of the model indicates how many of the comments belonging to a given category the model predicted correctly. For instance, out of the 23 comments belonging to the category "Glasses/Contact Lenses" in the test set, 11 (47.8%) are correctly classified. The *Precision* value captures how many of the comments predicted as belonging to a given category correspond to that category. For instance, in Table 3, out of the 151 cases that were predicted to be in the category "Prescribed medicines," 133 (88.1%) of them were actually coded by a human as such. As anticipated, performance depends heavily on the data available for training. However, the model performs sufficiently well in most categories with precision and recall

Table 2. Model Performance in the Testing Dataset.

Category	Precision	Recall	*N*	Category	Precision	Recall	*N*
Health care events	0.897	0.963	507	Employment	0.902	0.899	62
Health insurance	0.905	0.884	129	RU member refusal	0.941	0.842	19
Prescribed medicines	0.881	0.887	148	Condition	0.929	0.433	30
RU/RU member	0.882	0.788	85	Glasses/Contact lenses	0.846	0.478	23
Other	0.717	0.729	59	Other medical expenses	0.750	0.750	12

Table 3. Precision, Recall and NDCG for Top K Categories.

K	Precision *m*	Precision *M*	Recall *m*	Recall *M*	NDCG
1	0.883	0.865	0.883	0.765	0.883
2	0.475	0.425	0.951	0.877	0.926
3	0.323	0.275	0.971	0.911	0.936
4	0.244	0.215	0.979	0.931	0.939
5	0.197	0.180	0.985	0.950	0.942

above 90%. There are some clear exceptions. Most notable among them is the case of "Other," the catch-all category for changes that do not fit any other category. Additionally, the model has very low recall for the categories related to medical conditions and glasses and contact lenses, which indicates that there may be some pattern in the data of comments that belong to these categories that we do not have sufficient information to capture.

As discussed above, the goal of the model is to reduce the burden for DTs by presenting them with an appropriate subset of all the options that contains the true category for each comment. To address this, we evaluated the model using its performance in the top K categories. We show in Table 3 the micro (m) and macro (M) precision@K and recall@K (Manning et al. 2008:161) for the top K subset of categories. The table also shows the net discounted cumulative gain or NDCG (Wang et al. 2013), which measures ability of the model to produce a correct *ranking* of categories. The NDCG penalizes incorrect rankings (whenever the true category is not among the top K predictions) by a factor inversely proportional to the logarithm of the position in which the correct category is placed. That is, a model that places the correct category in position $K+1$ (in the order of predicted probabilities) will be worse than one that places it in position K by a factor that grows with K.

Table 3 shows that precision drops most significantly (second decimal) after we include it in the final set $K = 3$ predicted categories while the recall is not greatly affected once we start including more than one category. The increase in the NDCG is at the third decimal place once we include more than $K = 3$ categories. As a result, we decided to show the DTs only the top three predicted categories, sorted by their predicted probability, as the combination with the best trade-off between precision and recall.

An Evaluation of the Model in Production

We added a user interface (UI) to display the results of our machine learning model to the existing DQC UI so that it could function seamlessly with the current data editing workflow. Exhibit 1 shows the UI section where DTs can review and edit the comment category. The field category displays in the

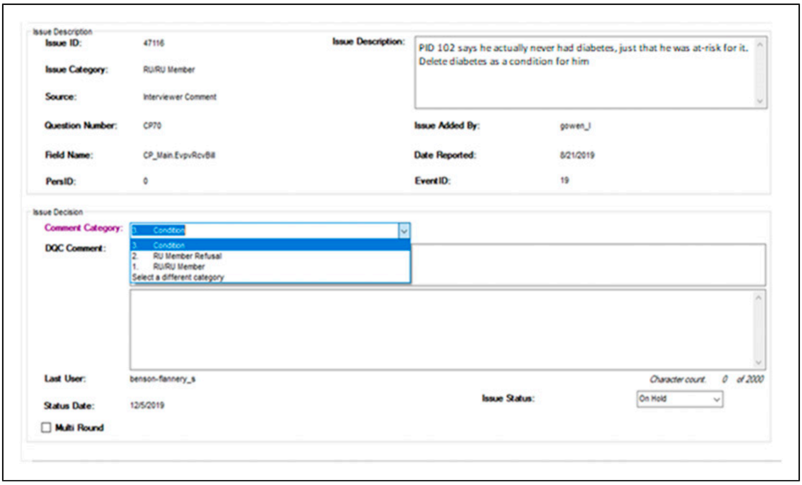

Exhibit 1. User interface in data quality control integrating the model predictions.

window labeled "Comment Category." Immediately below, the top three selections by the model are displayed in descending order. While we decided to display the categories in order based on probability (high to low), we did not display the percent probability to avoid influencing the selection by the DT. In our design, if the DT does not agree with any of the three categories, they can select the fourth (unnumbered) option, "Select a different category." This activates a drop-down display of all 10 categories.

Approximately four months after the introduction of NLP to the DQC UI, 2,900 comments had been processed through the NLP and evaluated by the DTs. For just under 90% of comments, the category selected by NLP as the most likely was confirmed by the DT. When including comments where the correct category according to the DT was among the top three most likely categories according to NLP, our success rate increased to 97.52%.

The performance of the model is dependent on the category it is predicting as the most likely one. Table 4 shows the number of comments within each category selected by the model as the most likely one, along with the frequency with which the DTs agreed with that assessment. Top category success rates vary across categories, from 72.25% among comments categorized by the model as RU/RU Member, to 95.74% among comments categorized by the model as belonging to the Employment category.

DTs added qualitative support for these findings during a debriefing session. Overall, the 9 DTs felt that the model worked well. They found it easy to use ("I don't know how it could possibly be easier to use.") and confirmed that the model selected the correct category most of the time. One suggestion

Table 4. Model Performance by Category.

Category	N	%	Category	N	%
Employment	188	95.74	RU member refusal	38	89.47
Health insurance	355	94.65	Prescribed medicines	397	88.41
Glasses/Contact lenses	14	92.86	Other medical expenses	32	87.50
Condition	69	91.30	Other	182	82.97
Health care events	1452	90.29	RU/RU member	173	72.25

for improvement requested that the model present categories for new comments or edits. In the current design, the model does not present a category for data checks or edits that the DTs add, such as updated information received through the Help Desk.

The consensus is that the machine learning model provides inappreciable gains in time or workflow efficiency. This is expected. We anticipated DTs would save only a very small amount of time for each case overall. However, due to the size and longevity of the study as well as the volume of comments, even small gains will add up. A similar experience with the large impact in efficiency due to small improvements to data quality control is reported by Windle (2015).

Conclusions

This article presents a case study for an innovative approach to using machine learning tools to assist with the time-consuming evaluation of interviewer comments on a large CAPI panel study. We used a standard machine learning model with manually built features and natural processing language to predict the most likely grouping category as needed for processing each comment. The model resulted in high statistical performance as compared to hand-coded data from previous waves. We then integrated the model in a data QC application, which showed the top three predicted categories but allowed technicians to make the final decision. In production, the model had a success rate of 97.52% and DTs found the application easy to use. Although time savings per comment are very small, we believe that the use of the application over time with a large volume of comments will be worthwhile. Among other things, using a machine learning model helped standardize our approach to data editing, which can increase data reliability and overall quality.

In future research, we plan to leverage newer developments in natural language processing that can power models pretrained on different but related problems (Vaswani et al. 2017) to try and programmatically identify non-actionable comments (comments that are too vague to justify a data edit). These comments account for roughly half of comments received and are costly to review. We will also review the impact of using the tool on the coders' decisions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. All models were fit using the scikit-learn library (Pedregosa et al. 2011).

References

AHRQ. 2020. *Medical expenditure panel survey*. North Bethesda, MD. May 29, 2020. https://meps.ahrq.gov/mepsweb/ (accessed May 29, 2020).

Athey, L., and A. B. Kennickell. 2005. Managing data quality on the 2004 Survey of Consumer Finances. Paper presented at the Annual Meetings of AAPOR, Miami. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.149&rep=rep1&type=pdf (accessed August 1, 2020).

Bricker, J., K. Moore, and R. Windle. 2014. Examining interviewer–respondent interactions in the Survey of Consumer Finances (SCF). In *Proceedings of the 2014 Survey Research Methods Section of ASA*, 2162–68. http://www.asasrms.org/Proceedings/y2014/files/312089_88725.pdf (accessed August 1, 2020).

Galasiński, D., and O. Kozłowska. 2010. Questionnaires and lived experience: Strategies of coping with the quantitative frame. *Qualitative Inquiry* 16:271–84.

He, Z., and M. Schonlau. 2020. Automatic coding of text answers to open-ended questions: Should you double code the training data? *Social Science Computer Review* 38:754–65.

He, Z., and M. Schonlau. 2021. Coding text answers to open-ended questions: Human coders and statistical learning algorithms make similar mistakes. *Methods, Data, Analyses* 15:103–20.

Honnibal, M., and I. Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/ (accessed August 1, 2020).

Jonas, S., R. Goldsteen, and K. Goldsteen. 2007. *An introduction to the US health care system*. New York: Springer Publishing Company.

Kennickell, A. B. 2007. Look and listen, but don't stop: Interviewers and data quality in the 2007 SCF. Paper presented at the Joint Statistical Meetings of the American

Statistical Association, Salt Lake City. https://www.federalreserve.gov/econresdata/scf/files/asa20072.pdf (accessed August 1, 2020).

Kononenko, O., O. Baysal, R. Holmes, and M. W. Godfrey. 2014. Mining modern repositories with ElasticSearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, 328–31. https://dl.acm.org/doi/abs/10.1145/2597073.2597091 (accessed August 1, 2020).

Manning, C., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval.* Cambridge: Cambridge University Press.

McClelland, S. I. 2016. Speaking back from the margins: Participant marginalia in survey and interview research. *Qualitative Psychology* 3:159–65.

Olson, K., and J. D. Smyth. 2015. The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology* 3: 361–96.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–30.

Pietsch, A.-S., and S. Lessmann. 2018. Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics* 1:93–116.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. Kushner Gadarian, B. Albertson, and D. G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58:1064–82.

Schonlau, M., H. Gweon, and M. Wenemark. 2021. Automatic classification of open-ended questions: Check-all-that-apply questions. *Social Science Computer Review* 39:562–72.

Smith, M. V. 2008. Pain experience and the imagined researcher. *Sociology of Health & Illness* 30:992–1006.

Tarnai, J., and M. C. Paxson. 2005. Interviewer judgments about the quality of telephone interviews. In *Proceedings of the 2005 Survey Research Methods Section of the American Statistical Association*, 3988–94. http://www.asasrms.org/Proceedings/y2005/files/JSM2005-000464.pdf (accessed August 1, 2020).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS 2017)* 30: 5998–6008.

Wang, Y., L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu. 2013. A theoretical analysis of NDCG ranking measures. *Proceedings of the 26th Annual Conference on Learning Theory, PMLR* 30:25–54.

West, B. T. 2013. An examination of the quality and utility of interviewer observations in the national survey of family growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176:211–25.

West, B. T., F. Kreuter, and M. Trappmann. 2014. Is the collection of interviewer observations worthwhile in an economic panel survey? New evidence from the

German Labor Market and Social Security (PASS) Study. *Journal of Survey Statistics and Methodology* 2:159–81.

Windle, R. 2015. Improving editing efficiency: How a comprehensive program interface reduces the time cost of the comment review process. In *Proceedings of the 2015 Survey Research Methods Section of the American Statistical Association*, 3621–29. http://www.asasrms.org/Proceedings/y2015/files/234196.pdf (accessed August 1, 2020).

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: 301–20.

Zuvekas, S. H. 2011. The effects of recall length and reporting aids on household reporting of health care events in the medical expenditure panel survey. *Journal of Economic and Social Measurement* 36:321–43.