CREATING IMPROVED SURVEY DATA PRODUCTS USING LINKED ADMINISTRATIVE-SURVEY DATA[1]

[1] **Acknowledgement**

Any opinions and conclusions expressed here are those of the authors and do not necessarily represent the views of the New York Office of Temporary and Disability Assistance (OTDA) the U.S. Census Bureau or NORC at the University of Chicago. The ACS-OTDA data analysis was conducted at the Chicago Census Research Data Center by researchers with Special Sworn Status and the results were reviewed to prevent the disclosure of confidential information. We would like to thank Katharine Abraham for comments on an earlier draft. Meyer and Mittag would like to thank the Alfred P. Sloan Foundation for support. Mittag is thankful for financial support from the Czech Science Foundation (through grant no. 16-07603Y) and the Czech Academy of Sciences (through institutional support RVO 67985998).

ABSTRACT

Recent research linking administrative to survey data has laid the groundwork for improvements in survey data products. However, the opportunities have not been fully realized yet. In this paper our main objective is to use administrative-survey linked microdata to demonstrate the potential of data linkage to reduce survey error through model-based blended imputation methods. We use parametric models based on the linked data to create imputed values of Medicaid enrollment and food stamp (SNAP) receipt in public use data. This approach to blending data from surveys and administrative data through models is less likely to compromise confidentiality or violate the terms of the data sharing agreements among the agencies than releasing the linked microdata and we demonstrate that it can yield substantial improvements of estimate accuracy. Using the blended imputation approach reduces Root Mean Squared Error (RMSE) of estimates by 81 percent for state-level Medicaid enrollment and by 93 percent for sub-state area SNAP receipt compared to estimates based on the survey data alone. Given the high level of measurement error associated with these important programs in the United States, data producers should consider blended imputation methods like the ones we describe in this paper to create improved estimates for policy research.

1. INTRODUCTION

The goal of good survey design is to make conscious decisions about efforts and resources spent to reduce different types of survey error given the level of funding and the specific research question the survey sponsor would like the data to answer. Survey researchers catalogue the potential sources of survey errors that can influence survey-based estimates (Federal Committee on Statistical Methodology (FCSM) 2001) into five basic sources of error: (1) Sampling error (2) Sample coverage error (3) Non-response error –including both unit and item non-response (4) Measurement error and (5) Processing error (FCSM 2001). In this paper our main objective is to improve overall estimate accuracy by reducing measurement error in survey estimates of program participation by means of using models developed on linked administrative and survey data to impute program receipt status. We discuss methods to blend information from the linked data with the public use data that neither require access to the restricted linked data, nor compromise confidentiality in the public use data.[2] We show that estimates blending administrative and survey data substantially reduce error that has been observed in critical policy relevant survey estimates of Medicaid enrollment and Supplemental Nutrition Assistance Program (SNAP) receipt. We argue that such utilization of data linkage to improve estimates is a more cost-effective approach to increase survey accuracy than many other current practices used to reduce survey error.

Past research has demonstrated substantial measurement error and bias in survey estimates of Medicaid enrollment and SNAP receipt. Work using the Current Population Survey (CPS) found that 43% of those linked to administrative data with Medicaid coverage did not report having

[2] No additional data can be disclosed without affecting confidentiality at all, but the risk from disclosure of synthetic data or model parameters is minimal. See e.g. Reiter (2003) for a discussion.

coverage (false negatives). On the other hand only one percent of respondents in the CPS reported having Medicaid coverage that could not be confirmed through the linkage. This pattern results in a large net undercount (Davern et al 2009a).[3] Research on survey misreporting of SNAP has also found that a substantial share of recipients do not report receipt. For New York State, Celhay et al. (2015) found false negative rates of 42 and 26 percent in the CPS and American Community Survey (ACS) respectively. Meyer et al. (2014) found even higher rates in the same surveys for Illinois (48 and 32 percent) and Maryland (53 and 37 percent). On the other hand, the false positive rates (reported SNAP receipt that cannot be verified) are low at around one percent (e.g. 1 percent for the New York ACS), resulting in the substantial net underreporting of food assistance that is documented in Meyer et al. (2015a, b) and Meyer and Mittag (2015).

Survey error and the bias it causes for Medicaid and SNAP is a serious problem for the policy research community and the Federal Statistical system as these survey estimates are used for critical purposes. Medicaid and SNAP are two important noncash benefits provided by states and funded through a federal-state partnership. It is critical for surveys to measure them accurately as benefit recipients are better off than similar non-recipients, because they have more resources. For example when calculating the Supplemental Poverty Measure (which includes SNAP benefits, but not Medicaid) or related measures, having accurate benefit receipt information is critical to capture the full resources available to a person or family (U.S. Census 2015). The impact of these benefits on understanding poverty in the United States is large overall, and even larger for specific demographic groups (U.S. Census 2015). The adjustments for non-cash

[3] Note that the net undercount is smaller than the difference between the false negative and false positive rate, as the denominator of the false positive rate (true non-recipients) is much larger than the denominator of the false negative rate (true participants).

benefits often rely on survey responses for SNAP that are known to have significant measurement error and that undercount the participation in these programs (and as a result may over-count poverty). The errors also result in understatement of the poverty reduction of the various programs and the mistakes in the relative importance of the individual programs in poverty reduction.

In addition to measuring poverty these data are critical for (1) providing general knowledge and statistics on the programs (2) evaluating these programs to see whether specific policy objectives are met (3) aiding official Congressional Budget Office legislative “scoring” to provide cost estimates for critical legislative initiatives such as the Affordable Care Act (Congressional Budget Office 2007) as well as simulation models used by federal agencies such as the Urban TRIM model (Urban Institute 2015). They are also used for official purposes by agencies to develop important health expenditure estimates for the country and states (Cuckler et al. 2013). Given these important uses of the survey data and the evidence that these data have considerable measurement error and bias, improvements to the data could substantially aid policy-making.

In this paper, we estimate the magnitude of data quality gains that would be possible if agencies or policy researchers began to routinely use imputations from models that rely on linked survey and administrative data. We model the relationship between “true” and reported receipt status in the linked data (using only public use file variables as predictors) and use this model to impute a more accurate receipt indicator in the public use data for all respondents. The models are estimated using the confidential linked data so that the parameters from the models can be released to the public. We demonstrate that such methods can substantially increase estimate quality in studies of program receipt, focusing on Medicaid and SNAP.

The methods could also be applied to other government programs important to understanding poverty on which administrative records exist, including Temporary Assistance for Needy Families (TANF), Supplemental Security Income and the Earned Income Tax Credit as well as other programs such as social security or unemployment insurance. We focus on binary variables here, but the approach also works for continuous variables known to be measured with error (e.g., self-reported height and weight data, housing prices, and employer characteristics). More generally, the methods can be used to amend or improve survey data whenever additional data can be linked to the survey, but making the linked data publicly available is infeasible, due to confidentiality concerns or data license restrictions.

2. LINKING DATA AS A WAY TO REDUCE MEASUREMENT ERROR IN ESTIMATES

One way to try to ameliorate the potential limitations of any data system is to combine it with other sources of data through linkage (National Academies of Sciences, Engineering and Medicine, 2017a,b). In this paper we do this by combining survey reported data with program administrative data and use the linked data to create models we can use to partially correct for measurement error in survey reported program participation. Previous studies have used linked data to: (1) examine sample coverage and accuracy (Bee, Gathright and Meyer 2015; Celhay, Meyer and Mittag, 2017), (2) impute variables (Davern et al. 2009a), (3) substitute administrative values for reported values (Nicholas and Wiseman,2010; Hokayem et al., 2015; Meyer and Mittag, 2015), (4) supplement survey reported data (Abowd et al. 2006), and (5) correct estimates (Davern et al 2009a; Schenker et al. 2010; Mittag 2017). These studies demonstrate that data linkage can improve survey accuracy by illuminating problems in survey design and methodology. Data linkage can also be used to improve survey accuracy more directly by combining information from the linked variables and the survey responses.

In this paper we explore one of the potential benefits of combining administrative data with survey data. Specifically, our central objective is to show the reduction in measurement error that results when a model-based blended imputation model is used to impute values to replace the fallible survey responses. We use reported survey data that have been linked to external administrative data to estimate the relationship between “true” and reported values of the variable of interest (e.g., being enrolled in Medicaid or receiving SNAP). Then, we use this model to impute improved receipt indicators. Using these imputed values instead of the misreported survey measures yields partially corrected estimates of the statistics of interest. An example of the method is Schenker et al. (2010) who start with a set of data from the National Health and Nutrition Examination Survey (NHANES) that has both the reported survey items and clinically measured items to diagnose hypertension, diabetes, and obesity. They then model the clinically diagnosed values using the survey reported values along with potential covariates of measurement error as predictors. Once the model is developed on the NHANES with both survey reported and clinically measured variables they use the model to multiply impute clinical outcomes in data for which they do not have the actual clinical measures in the National Health Interview Survey. For a Bayesian approach, see e.g. He and Zaslavsky (2009). Blackwell, Honaker and King (forthcoming) review such (multiple) imputation as a method to correct for measurement error. They propose a similar imputation strategy that relies on independence assumptions for situations in which validation data are not available.

Our analyses build on applications to Medicaid in Davern et al. (2009a) and SNAP in Mittag (2017). Davern et al. (2009a) link Medicaid administrative data to CPS data and then use the model developed on the linked data to impute an indicator of Medicaid enrollment in the CPS for subsequent years. Mittag (2017) employs a similar approach using SNAP administrative data linked to ACS data to correct estimates of food stamp receipt out of sample in the ACS. We discuss the advantages and disadvantages of these methods compared to other approaches such as direct substitution further below.

We use a Mean Squared Error (MSE) metric to measure the gains in estimate quality that can be realized by using linked data to create blended imputed estimates. The MSE is defined by:

MSE=Bias Squared + Variance. [Equation 1]

The MSE of an estimate is the expected value of the square of its deviation from the true parameter of interest, i.e. it combines estimator bias and variance. Thus, when evaluating the quality of different survey estimates, preference is given to the one with the smaller MSE. In our tables below we take the square root of the MSE or the Root Mean Squared Error (RMSE) in order to put the measure on the same scale as the original estimate. It is important to note that our RMSE is itself likely a biased estimate due to the fact that the measure of "truth" we use to compute it contains error as well. We discuss these points in the next section, when we review limitations of our approach.

Table 1 presents results from our application to Medicaid receipt. The first four columns of numbers in Table 1 are drawn from Davern et al. (2009a).[4] They used the 2001-2002 CPS linked to Medicaid Statistical Information System (MSIS) data from 2000-2002 to create a person level logistic regression model of Medicaid receipt (for detailed tables on the data linkage and the evaluation see US Census Bureau 2007; US Census Bureau 2008a; and US Census Bureau 2008b). Of the CPS respondents linked to MSIS, who show Medicaid enrollment in MSIS at some point during the reference period, roughly 43% do not report having Medicaid, resulting in a Medicaid undercount (Davern et al. 2009b). However, because 47% of those linked do correctly report, survey reported Medicaid enrollment is a critical predictor of true enrollment in the imputation model.

Stratifying on survey reported Medicaid status, Davern et al. (2009a) estimated two models to partially correct for survey measurement error. The first model used a logistic regression to

[4] The standard errors for the imputed Medicaid Enrollment estimates in Davern et al. (2009a) were incorrect and did not appropriately adjust for the design effect of the CPS complex sample design. The standard errors in Table 1 of this paper for imputed Medicaid by state have been adjusted for the design effect of the CPS survey.

predict whether a person received Medicaid in MSIS given that they did not report having Medicaid in the survey (i.e., a false-negative model). The second model predicted whether a person received Medicaid given that they had reported Medicaid coverage in the survey (a true positive model). Both models condition on key covariates such as age, sex and income and include state fixed effects. The models also include reported Medicaid enrollment, which is a key predictor of actual enrollment as pointed out above. Fitting a well-specified model is crucial for the accuracy of the correction, but a standard specification question that is beyond the scope of this paper. Davern et al. (2009a) discuss estimation of the Medicaid models we use here further, see Appendix A of this paper for the estimated model parameters.

We use the coefficients from these two logistic regression models from 2000 and 2001 to predict each person's probability of being enrolled in Medicaid in the 2007 and 2008 CPS (covering calendar years 2006 and 2007) given their self-reported coverage and other key co-variates such as age, sex, income and state of residence. We then use these person level predicted probabilities to estimate the number of people having Medicaid by state, which are reported in Table 1.

The point of this paper is to add the evaluation of estimate accuracy in the last four columns of Table 1 below. The bias is estimated as the difference between the state estimate of enrollment in 2006-2007 and the Medicaid enrollment numbers found on Kaiser State Health facts (Kaiser State Health Facts 2015). The Kaiser number is likely imperfect as well. Bias can vary from state to state given that each state has different ways they compile the data for Kaiser. In addition, concept alignment between the CPS measure and the Kaiser measure is not perfect: The Kaiser measure is an average monthly enrollment and the CPS is a measure of Medicaid enrollment "at any point in the last year." Thus, the CPS number should be higher and include more enrollees who churn on and off the program throughout the year. In general, this would mean the administrative data counts of Medicaid enrollment should be even higher than the Kaiser counts for the “any time in the past year” enrollment number. Finally, universes between CPS and Kaiser are not the same. Kaiser includes people in group quarters and who may have died during the year who would not be counted in CPS. The impacts of these universe adjustments are important, but will not significantly impact the findings of the paper (see US Census Bureau, 2008 to better understand the magnitude). Nevertheless, the Kaiser numbers are an independent estimate of enrollment in those years for comparison purposes. The first column of RMSEs are for the unadjusted CPS, i.e. what you would get if you simply tabulated the CPS public use file for those two years and created a two-year average. The second column of RMSEs compares the Kaiser rate to the CPS imputation based on the individual level predicted probabilities. The final column is the percent reduction (negative numbers are the percent increase) between the two RMSEs for any given state. The Medicaid example does not account for variance added by imputation so that the MSEs for the imputed model are too small.[5] However, the additional variance due to imputation modeling will likely be small relative to the reduction in bias (as we find in the SNAP example below).

--Insert Table 1 About Here --

For the U.S. as a whole the RMSE for the model-based blended imputed estimate is 81% lower than the RMSE for the direct CPS estimate. This is a substantial reduction in RMSE which is

[5] We no longer have access to the Medicaid microdata to correct the variance estimates.

mainly due to bias being reduced. The direct CPS estimate of the Medicaid coverage rate for the U.S. is 11.4% and the imputed estimate is 13.8% which is much closer to the 14.3% in the Kaiser State Health Facts. In most states the RMSE decreased between the CPS direct survey estimate and imputed estimate. There are, however, 14 states that saw an increase in bias. One would expect estimates from some states to be closer to truth by chance due to the substantial sampling error in the CPS. Due to the extrapolation, we cannot hold sampling error fixed here. The fact that RMSE increases for a smaller fraction of areas when we hold sampling error fixed in the second application suggests that this explains some, but not all of these increases. Increases in RMSE may also stem from unobserved differences in reporting rates, which may lead us to overstate receipt in states with the most accurate reporting. The largest increases were in Utah, Arizona, North Dakota, North Carolina, Oregon, Iowa, Montana, Kansas and Missouri. These are states in which the CPS fares particularly well: In these 9 states the difference between the CPS direct survey estimate and the Kaiser rate was only .8% on average whereas in the 37 states with positive reductions in RMSE the average difference was 3.2%. When known, differences in reporting rates can be incorporated in the imputation model, making it desirable for future research to look for potential reasons and to attempt to improve on the fit of the model for these states.[6]

Our second illustration of how blending based on models from linking data can improve survey estimates examines SNAP receipt for small geographic areas[7] in New York State. The results in

[6] For the state of Montana, the increase in the bias in the modeled results derives from the fact that over half of those on Medicaid were missing the linking information. Thus in Montana's case, too few people are imputed to have Medicaid as over half the enrollees were not linkable to the CPS (US Census Bureau 2008a).

[7] We use the counties that can be identified in the public use ACS data and pool counties that cannot be separated in the public use data.

Table 2 are similar to those for Medicaid in Table 1. They are based on the model and results in Mittag (2017), which uses administrative SNAP records linked to the ACS to develop a method of correcting survey estimates for measurement error. The validation data were created by linking administrative records on monthly SNAP payments for all recipients in New York State from the New York State Office of Temporary and Disability Assistance (OTDA) to the 2010 ACS survey data. The administrative records are based on actual, validated receipt and the two data sources are linked with a high match rate at the household level. Thus, even though they are not free of error, the linked data appear accurate enough that we consider them to be the assumed best or unbiased measure of receipt. For further descriptions of data linkage and accuracy, see e.g. Celhay et al. (2015), Cerf Harris (2014), Mittag (2017) and Scherpf et al. (2014). As Celhay et al. (2015) show, the linked data reveal substantial error in reported SNAP receipt and amounts. For example, 26 percent of administrative data recipient households do not report SNAP receipt in the ACS (false negatives). On the other hand, the false positive rate (true non-recipients reporting SNAP receipt) is low at 1.2 percent, resulting in the substantial net underreporting of government transfers that is documented in Meyer et al. (2015a,b) and Meyer and Mittag (2015).

The fifth column of Table 2 provides estimates of receipt rates that we consider to be unbiased from the linked data. We estimate these rates for the 39 county groups that can be identified in the ACS public use data. Comparing these receipt rates to the survey based estimates in the first two columns underlines that there is net underreporting in all but one area, and that reporting rates vary between these areas. Cerf Harris (2014) examines reporting rates at the county level in detail.

The main objective of this paper is to assess how the survey estimates compare to the results in columns three and four, which contain estimates of the receipt rate using an imputation model to create blended imputed estimates that partially correct the survey reports. The imputations are based on the method in Mittag (2017), who uses the linked 2010 ACS data to estimate the conditional distribution of administrative SNAP receipt and amounts received given reported receipt and other covariates. The conditional distribution of SNAP amounts can be seen as a continuous distribution with a mass point at 0. However, we are only concerned with receipt and not with amounts received here. Therefore, we only use the estimate of the binary part of the distribution, which is a standard probit model. In addition to reported SNAP receipt, the model conditions on a large set of demographic and economic variables, including household composition, age, education and income. The model does not condition on any geographic information, so that the variation between counties we examine here is only captured by the covariates. This makes the reduction in RMSE particularly noteworthy, because accuracy could still be improved by incorporating geographic information. As above, specification of the imputation model is crucial for the accuracy of the correction, but beyond the scope of this paper. It is discussed further in Mittag (2017). Appendix Table A2 of this paper contains the estimated parameters of the conditional distribution.

We use the parameters of this model to predict a probability of SNAP receipt for each household as with Medicaid above. We then generate a receipt variable by taking 20 random draws from a Bernoulli distribution with the predicted probability for every household in the 2010 New York ACS sample. Taking multiple draws makes simulation error negligible and thus reduces SEs and avoids having to correct the SEs for simulation error. [8]

[8] A key difference between multiple imputation and the approach we take here is that we estimate the statistic of interest from the multiple stacked imputations, rather than averaging estimates from repeated single imputations. For the subgroup means we estimate here, the two approaches are equivalent, but estimates and SEs differ in general. As discussed in Mittag (2017), correlations and model parameters as in Schenker et al. (2010) may be inconsistent under single and standard multiple imputation, but the methods discussed here yields consistent estimates.

--Insert Table 2 About Here --

The last three columns of Table 2 contain RMSE defined the same way as for Medicaid above. We compute the bias in the survey and imputation based estimates as the difference in the numbers from the linked data in the fifth column. The population totals to which we compare our estimates are affected by errors in the administrative data and linkage errors. However, these errors should be small and outweighed by the benefit that using the linked data ensures that the numbers are for the same population as our improved survey estimates (which exclude group quarters and the homeless) and that subject definitions are comparable. Contrary to the Medicaid application, the imputation model is estimated using the same sample. Mittag (2017) further discusses extrapolation across time and geography. We are mainly interested in the percent reduction in RMSE when replacing the survey reports by the imputations in the last column, i.e. by how much the imputations reduce error compared to uncorrected survey based estimates. The numbers for the entire state of New York in the last row show that the blended imputed estimates reduces RMSE by an impressive 93 percent. This is similar in magnitude to the reduction in RMSE for Medicaid and again driven by the reduction in bias. The standard errors are slightly larger than in the survey, but they are small in both cases due to the large sample. Thus, bias is the main determinant of RMSE. The reduction in bias more than makes up for the increase in standard errors. The survey understates receipt by 25 percent, while the imputations fall short of the actual share of recipients by 1 percent only.

This pattern also drives the results at the local level. The survey numbers underestimate receipt rates in all but one county, while the imputation based numbers do not seem to be systematically biased. They are larger than the assumed “true” numbers in 21 out of 39 areas and smaller in 18 areas. The imputation based rates are more accurate than the survey in terms of estimated RMSE in 31 out of 39 areas. The reductions in RMSE are substantial: In 29 of these 31 areas, RMSE is reduced by 25 percent or more, and in 15 areas the imputation based measure cuts the error by more than half. However, RMSE of the imputed receipt rate is larger than the survey RMSE in 8 of the 39 areas. As with Medicaid, this result is primarily due to the fact that the survey closely replicates the numbers from the linked data for these 8 areas, i.e. it is mainly driven by the good performance of the survey in these counties.

3. DISCUSSION

Recent federal data initiatives emphasize linking and combining data as a promising way to improve data for policy purposes. For example, key recommendations in the report of the Commission on Evidence-Based Policymaking (Commission on Evidence-Based Policymaking 2017) call for producing higher quality data by linking and combing data. After the commission report was released, the U.S. Office of Management and Budget (OMB) put out a request for information to help improve federal statistics stating that “a priority has been placed on using new techniques and methodologies based on combining data from multiple sources” (Federal Register, January 12, 2018). We believe our paper demonstrates an operationally achievable way of accomplishing the goal of combining data from multiple sources to improve data quality and ensure data can be widely disseminated for evidence based policy making. This section first briefly discusses why data linkage is a cost-effective way to reduce MSE in surveys compared to other common approaches. We then illustrate how model-based blended imputations compare to two key alternatives: the status quo and directly replacing survey reports with the linked administrative values. We compare the strengths and weaknesses of these three approaches using the data quality criteria of the FCSM (2001).

3.1. Comparisons to Other MSE Reduction Approaches

Model-based blended imputation may not always yield the largest feasible error reduction, but we argue that it offers a less cost effective way to lower the MSE of survey statistics than other commonly employed approaches on which large amounts of money are spent. Survey researchers often use tools such as larger sample sizes and/or reducing non-response as ways to reduce MSE in surveys. Larger sample sizes reduce MSE but this option is both expensive and grows less effective at reducing variance (and MSE) with each additional case that is added to the sample (and it also adds respondent burden as more sample is added). Another common strategy to reduce bias is to reduce survey nonresponse. Reducing survey non-response through additional effort (more telephone calls, more in person attempts to recruit a household, more mailings, etc.) and the use of incentives is costly and there is little evidence it improves data quality. Research has shown that spending considerable funds on strategies aimed at increasing response rates can indeed increase response rates. However, survey research is concerned with response bias and not response rates per se. But nonresponse has been shown to have little impact on the bias survey estimates (Groves 2006; Groves et al. 2008). Also, linkages to administrative data have demonstrated that nonresponse bias is small for key policy relevant variables such as income (Bee, Gathright and Meyer, 2015; Celhay, Meyer and Mittag, 2017). However, as we show in our paper these same linkage studies often show significant amounts of measurement error in survey responses that often lead to sizeable bias in survey estimates. Thus, we believe that expensive attempts to reduce MSE (such as increasing sample sizes or increasing effort to convert non-respondents) should be evaluated to make sure that they are cost effective ways of reducing MSE relative to other alternatives such as data linkage.[9]

3.2. Data Quality of Model-Based Blended Imputations Estimates Versus Alternatives

To compare model-based blended imputations to key alternative approaches on criteria that are relevant to statistical agencies, we use the data quality framework developed in FCSM (2001). The FCSM identifies the four key elements of data quality as (1) accuracy, (2) relevance, (3)

[9] Data linkage to administrative data can also facilitate other survey improvements besides reducing measurement error. For example, there is strong evidence that linking the sample frame to other sources of data can help surveys more efficiently allocate resources used in household listing (Montaquila 2011).

timeliness and (4) accessibility. The two alternatives we explore are, first, not making any changes (i.e., having the agencies maintain current practice); and second, direct substitution (i.e., having the agencies link the data and directly replace the survey report by the actual value from the administrative data -- rather than imputing a value as was done in this paper).

If the statistical agencies do not make any changes from current practice we believe that data quality will continue to be a problem. A mounting literature demonstrates that the current approach is not *accurate* by showing that estimates of key policy importance are biased by the substantial amount of measurement error (e.g. Davern et al. 2009; Meyer and Mittag, *forthcoming*). The results in this paper show that the blended imputation approach would substantially improve estimate *accuracy* over the current practice. In addition to *accuracy* the imputed estimates could also bring gains in *relevancy* in that the survey could be enhanced with additional information from the administrative data. For example, one could use the blended imputations to develop monthly enrollment flags instead of indicators of ever being enrolled during a 12-month reference period. Such additional programmatic detail has the potential to improve the policy *relevance* of the blended imputations for policy research purposes.

On the other hand, the blended imputation approach cannot improve over current practice in terms of *timeliness* and *accessibility*. Our approach may slightly reduce timeliness if creating the imputations delays data release or if the imputations are produced after data release. In order to mitigate the impact *timeliness* and *accessibility*, the model coefficients could be created by the statistical agency themselves (similar to what appears in Appendix A of this paper) and distributed separately so as to not interrupt current data processing. Another approach could be to have the statistical agency grant access to the linked data to a third party using the Research Data Center (RDC) network. Interested third parties could include those working on micro simulation models that rely on the survey data such as the Urban Institute's TRIM (Urban Institute 2015), Congressional Budget Office simulation models (Congressional Budget Office 2007), or RAND's COMPARE (Eibner et.al. 2010) simulation models as well as those groups that disseminate the survey microdata such as IPUMS, ICPSR or NBER. If granted access these third parties could estimate imputation model coefficients and/or the imputations themselves and distribute them through the current dissemination channels such as IPUMS, ICPSR, or NBER. This would impinge on *accessibility* of the imputed data, because the imputations are not provided with the core data product, but this downside could be mitigated if data disseminations channels such as IPUMS, ICPSR or NBER would include the imputed values in the version of the data they distribute.

The second alternative approach to model-based blended imputation approach is direct substitution of administrative data for survey data. Direct substitution is more *accurate.* In terms of *relevance,* direct substitution is better on some dimensions, but worse on others. And finally, the direct substitution approach is likely to be less *accessible* and *timely* than the blended imputation. To elaborate on these points, the blended imputation approach is less *accurate* than the direct substitution approach for two main reasons. First, the model based blending approach adds variance from estimated parameters of the model and imputation. The added variance from estimated parameters decreases the effective sample size of the linked data.[10] In addition to variance, the direct substitution method is more *accurate* because it does not rely on a potentially misspecified model. Most specification questions can be assessed with standard tests (see Davern

[10] Standard errors need to be adjusted for this additional variance, which makes the model-based imputation approach less convenient. Methods to do so are well developed but depend on details of the implementation. If one uses the imputation model to create multiple imputations or synthetic data, SEs can be estimated as discussed in Rubin 1996, Raghunathan, Reiter and Rubin (2003) and Reiter (2003). When using the imputation model to integrate out the error free measure as in Mittag (2017), SEs can be corrected for simulation error as discussed in McFadden (1989) and for estimated first stage parameters as described in Newey and McFadden (1994). If one is willing to specify prior distributions, Bayesian survey inference provides a compelling way to measure uncertainty, see Little (2012) for a discussion.

et al. (2009a) and Mittag (2017) for discussions of the models we use here). For imputation models, particular attention should be paid to the choice of conditioning variables. As discussed in Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006), the imputation model should condition on all covariates in the outcome model. The goal of the blended imputations is to reproduce the distribution of the accurately measured variable or its joint distribution with the relevant covariates. Researchers developing the imputation models have access to the linked data, so how closely a given model reproduces these distributions can be measured and tested using Kolmogorov- or Cramer-von Mises-type statistics. Most household surveys are used for a wide range of purposes, so that the ideal imputation model may depend on the statistic of interest. This can be addressed by producing different models for different purposes, and this again presents a slight downside in terms of convenience compared to direct substitution. Also, as programs and reporting errors may change over time, the imputation models should constantly be evaluated and improved. And it is likely that one model may not be appropriate for all use cases and the development of additional models for specific use cases is recommended.

When comparing the policy *relevance* of the direct substitution method to the model-based imputation method there are pluses and minuses for each. For direct substitution (as with the model-based imputation method) relevant details from the administrative data can be included other than just enrollment or receipt. These details include the months the person was enrolled, the basis of eligibility, the exact program of enrollment (e.g., State Children's Health Insurance Program, limited benefits Medicaid program versus a full benefits Medicaid program), the services received, and the amount of benefits (among many others). The advantage of using direct substitution for these extra policy relevant details is that this information is measured more *accurately* than in the blended imputation approach.

The policy *relevance* advantage for the blended imputation approach is that high quality administrative data are often not available for some geographic areas or time periods, or some households cannot be linked (e.g., survey respondents opt out of linkage or survey/administrative data are missing identifiers used for linkage). In such cases, model-based blended imputation can use geographic areas and time periods with linked data to develop models and then extrapolate. This approach, although susceptible to model variance and misspecification, can still lead to significant reductions in MSE. Mittag (2017) discusses the required conditions and finds substantial improvements in accuracy even though the assumptions are at best approximations in his application. In the Medicaid empirical example we included in this paper, the model was created using MSIS data linked to 2000-2002 CPS data and was applied to microdata from the 2007-2008 CPS. There is likely to be some extrapolation error in this case since several states experienced changes in their Medicaid program over this time span. However, as we showed in the analysis presented in this paper the reductions in MSE are substantial nonetheless. A final benefit of the blended imputations is that the imputation model can be extended to impute true receipt for households that the agency is unable to link (e.g., the respondent opts out of linkage). Such extensions could also address the consequences or linkage errors (incorrect or incomplete linking identifiers on the survey or administrative data) into account. Linking data on a regular basis will improve our understanding of the conditions under which extrapolation works and thereby help to validate and improve the imputation models and the policy *relevance* of the data that result. This can make the blended imputations more policy *relevant* than direct substitution in cases where linkage is not possible or imperfect.

The possibility to extrapolate also gives the blended approach an advantage over direct substitution in terms of *timeliness*. The administrative data needed for direct substitution may sometimes not be ready for linking in a timely manner to allow for direct substitution. Using previous years of linked data for modeling, while potentially less accurate, will result in the production of more *timely* estimates for use in policy research.

The final and likely most important advantage of the blended imputation over direct substitution is *accessibility*. As pointed out by Bound, Brown and Mathiowetz (2001), linked data or validation data more generally are usually only accessible to a small group of researchers. The main reason for this situation is that making the linked data publicly available makes it easier to identify individuals and thereby carries a confidentiality risk for both the survey and the administrative data. In the past this risk has been deemed to be too large to allow for the release of directly substituted data. This situation may change in the future, possibly by adding noise or coarsening the linked variables or by creating an infrastructure for secure data access. But neither the model coefficients nor the model based imputed values present as much de-identification risk as direct substitution as long as there is an imperfect model fit (although models also carry disclosure risk (Reiter and Mitra 2009)). As Appendix A shows, statistical agencies are willing to release the required parameters, so the blended imputations approach is already feasible. Thus, the blended imputation approach has an advantage for public *accessibility* in that the model error may be large enough to pass data producing organization’s confidentiality review.

4. CONCLUSION

All data (including survey and administrative data) have errors. However, it is critical that we move beyond acknowledging data limitations and create new data products that blend the strengths of each data system to reduce known errors. Such innovative methods to mitigate the flaws in any one data system have the potential to improve public policy decisions. From our two analyses of Medicaid and Food Stamps we argue that in the realm of survey errors that (a) we can address and (b) have a measurable impact on data quality, reducing measurement error through linkage of administrative data to survey data is a way to achieve substantial MSE reductions. Current practice does not incorporate the results from linkage studies into the most widely used and circulated data products from data producers such as the U.S. Census Bureau.[11] We believe that they can and should do more to correct for known survey measurement error. At a minimum, data producers (potentially in collaboration with the broader research community) should create alternatives to their standard data products that are known to have pronounced measurement error in policy relevant variables.

We have demonstrated one approach for creating products that allow analysts to partially correct known measurement errors. The examples of Medicaid and SNAP receipt underline that the resulting improvements can be substantial as they reduced RMSE by 81 and 93 percent compared to the survey estimates for the geographic areas we examined. The model-based blended imputation approach has been found to work well for a wide range of use cases. It extends to multivariate analyses and more complex estimators. Schenker, Raghunathan, and Bondarenko

[11] Although we note a good recent example is in Motro and Roth (2017).

(2010) and Mittag (2017) impute both binary and continuous variables and find the blended imputation to work well for multivariate and non-linear models.

We use the FCSM (2001) elements of data quality (accuracy, relevance, timeliness and access) to evaluate the blended imputations. We provide evidence that a key advantage of model-based imputation is its improvement of *accuracy* compared to current practice. We argue that the improvement in *accuracy* is large enough to outweigh the disadvantages in *timeliness* and *access*. Direct substitution would be more accurate than the blended imputations (and have similar *relevance*). If linked data can be made *accessible* to a wider audience in a *timely* fashion, then advantages of direct substitution may easily make this approach preferable. However, *access* to directly substituted data may not be able to be made public making this route more difficult. And, therefore, given the current state of affairs the blended imputation appears preferable on grounds of *accessibility* and *timeliness* despite the loss off *accuracy*. And the method we propose does not pose as great a risk to data confidentiality and the privacy of respondents nor does sharing it publicly violate the terms of some of data sharing agreements between agencies.

Additional data products that improve accuracy through model-based imputations could be created based on existing data linkage projects. These additions to current data products could consist of (1) a set of models complete with coefficients like the ones we generated in this paper for Medicaid and SNAP -- so that users of the data could use them to create imputations themselves, or (2) a separate imputed variable for all survey persons/households using models like the ones we have used in this paper that is included in future data products. Little (2012) discusses the advantages of these two options further.

The reasons why it is now imperative to use linked data in the creation of official statistics, reports and data products are: (1) the foundational research for use of linked administrative data and survey data has been conducted for several potential sources (2) there is clear evidence from these research projects that the amount of bias due to measurement error in the survey data could be significantly reduced (3) the necessary infrastructure for sharing data among federal agencies is in place and directives have been supplied by the Office of Management and Budget (Burwell 2014; O'Hara 2016). Now is the time to start building the data products that use blended survey and administrative data in production as it will improve official statistics, reports and data products. While not all linked administrative data and survey data are ready for production we believe that there are substantive areas of policy research (such as Medicaid enrollment, Medicare enrollment, SNAP and other program receipt, and uninsurance calculations) that have the needed agreements in place and ongoing linkage projects. These projects can be leveraged to improve our ability to make policy relevant estimates to evaluate and cost out policy proposals for use by organizations such as the Congressional Budget Office, the Congressional Research Service and the Office of the Actuary at the Centers for Medicare & Medicaid Services.

REFERENCES

Abowd, J. M., Stinson, M., and Benedetto, G. (2006). Final report to the social security administration on the SIPP/SSA/IRS public use file project. *U.S. Census Bureau Working Paper*.

Alexander, J. Trent, Michael Davern and Betsey Stevenson. *(2010)* "Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications." *Public Opinion Quarterly.* 74 *(3):* 551-569.

Bee, C. Adam, Graton Gathright, and Bruce D. Meyer 2015, "Bias from Unit Non-Response in the Measurement of Income in Household Surveys." University of Chicago working paper.

Blackwell, Matthew, James Honaker, and Gary King. forthcoming. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods & Research*.

Bollinger, Christopher R., and Barry T. Hirsch. 2006. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics,* 24(3): 483-519.

Burwell, Sylvia. 2014. "M-14-06: MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES: Guidance for Providing and Using Administrative Data for Statistical Purposes." Office of Management and Budget. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

Congressional Budget Office. 2007. “Background Paper: CBO’s Health Insurance Simulation Model a Technical Description.” Congressional Budget Office. October 2007. Washington DC. https://www.cbo.gov/publication/19224?index=8712

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2015. “Measurement Error in Program Participation.” Unpublished paper.

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2017. “An Empirical Total Survey Error Decomposition Using Data Combination.” Unpublished paper.

Cerf Harris, Ben. (2014). Within and Across County Variation in SNAP Misreporting: Evidence from Linked ACS and Administrative Records. *CARRA Working Paper #2014-05.* U.S. Census Bureau.

Commission on Evidence-Based Policymaking. 2017. *The Promise of Evidence-Based Policy*. https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

Cuckler, Gigi, and Andrea Sisko. 2013. “Modeling Per Capita State Health Expenditure Variation: State-Level Characteristics Matter.” *Medicare and Medicaid Research and Review.* 3(4):E1-E24. https://www.cms.gov/mmrr/Downloads/MMRR2013_003_04_a03.pdf

Davern, Michael, Holly Rodin, Timothy J. Beebe, and Kathleen Thiede Call. 2005 “The Effect of Income Question Design in Health Surveys on Family Income, Poverty and Eligibility Estimates.” *Health Services Research.* 40(5):1534-1552.

Davern, Michael, Holly Rodin, Kathleen Thiede Call, and Lynn A. Blewett. 2007. “Are the CPS Uninsurance Estimates Too High? An Examination of Imputation.” *Health Services Research*. 42(5): 2038-2055.

Davern. Michael, Jacob Klerman, Jeanette Ziegenfuss, Victoria Lynch, and George Greenberg. 2009a “A Partially Corrected Estimate of Medicaid Enrollment and Uninsurance: Results from an Imputational Model Developed off Linked Survey and Administrative Data.” *Journal of Economic and Social Measurement.* 34(4):219-240.

Davern, Michael, Jacob Alex Klerman, David Baugh, Kathleen Call, and George Greenberg. 2009b “An Examination of the Medicaid Undercount in the Current Population Survey (CPS): Preliminary Results from Record Linking.” *Health Services Research.* 44(23) 965-87.

Davern, Michael, Donna McAlpine, Timothy J. Beebe, Jeanette Ziegenfuss, Todd Rockwood and Kathleen Thiede Call. 2010 “Are Lower Response Rates Hazardous to Your Health Survey? An Analysis of Three State Health Surveys.” *Health Services Research.* 45 (5): 1324–1344.

Davern, Michael. 2013. “Nonresponse Rates are a Problematic Indicator of Nonresponse Bias in Survey Research.” *Health Services Research*: 48(3):905-912.

Eibner, Christine, Federico Girosi, Carter C. Price, Amado Cordova, Peter S. Hussey, Alice Beckman, Elizabeth A. McGlynn. 2010. “Establishing State Health Insurance Exchanges Implications for Health Insurance Enrollment, Spending, and Small Businesses.” RAND Corporation. Santa Monica, CA.

Federal Committee on Statistical Methodology (FCSM). 2001. "Measuring and Reporting Sources of Error in Surveys." Washington DC: Statistical Policy Office, Office of the Management and Budget. http://www.fcsm.gov/01papers/SPWP31_final.pdf

Federal Register. 2018. Office of Management and Budget. 83 FR 1634, P. 1634-35. https://www.federalregister.gov/documents/2018/01/12/2018-00400/request-for-information

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (4): 646-75.

Groves, R.M., E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly*. 72: 167-189.

He, Yulei, and Alan M. Zaslavsky. 2009. "Combining information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies." Biometrics. 65(3): 946-952.

Hirsch, Barry T., and Edward J. Schumacher. 2004. "Match Bias in Wage Gap Estimates Due to Earnings Imputation." *Journal of Labor Economics*. 22(3): 689-722.

Hokayem, Charles, Bollinger, Christopher R., and Ziliak, James P. 2015. The role of CPS nonresponse in the measurement of poverty. *Journal of the American Statistical Association*, *110*(511), 935-945.

Kaiser State Health Facts. 2015. “2006-2007 total monthly Medicaid enrollment December avg.” Downloaded from Kaiser state health facts 9/20/2015. http://kff.org/medicaid/state-indicator/monthly-medicaid-enrollment-in-thousands/ for notes and sources.

Kish, Leslie. 1965. Survey Sampling. Wiley and Sons. New York; New York.

Little, Roderick J. A. 2012. “Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics.” *Journal of Official Statistics*, 28(3): 309-334.

Meyer, Bruce D., Robert Goerge, and Nikolas Mittag**.** 2014. “Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation.” Unpublished paper.

Meyer, Bruce D. and Nikolas Mittag. 2015. “Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness and Holes in the Safety Net,” NBER Working Paper 21676, October.

Meyer, Bruce D., Mok, W.K.C. and Sullivan, J.X. 2015a. The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences. *Harris School of Public Policy Studies, University of Chicago Working Paper.*

Meyer, Bruce D., Mok, W.K.C. and Sullivan, J.X. 2015b. Household Surveys in Crisis. *Journal of Economic Perspectives*, 29(4): 199-226.

Mittag, Nikolas. 2017. “Correcting for Misreporting of Government Benefits.” Working Paper.

Montaquila, Jill, Hsu, Valerie, and Brick, J. Michael. 2011. Using a match rate model to predict areas where USPS-Based address lists may be used in place of traditional listing. Public Opinion Quarterly, 75, 317-335.

Motro, Joanna and Veronica Roth. 2017. "Using Administrative Records and Parametric Models in 2014 SIPP Imputations" Working Paper, U.S. Census Bureau.

National Academies of Sciences, Engineering and Medicine. 2017a. "Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy." Washington, D.C: The National Academies Press.

National Academies of Sciences, Engineering and Medicine. 2017b. "Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps." Washington, D.C: The National Academies Press.

Nicholas, J. and Wiseman, M. 2010 Elderly Poverty and Supplemental Security Income, 2002-2005. Social Security Bulletin, Vol. 70(2).

O'Hara, Amy. 2016. "Use of Administrative Records to Reduce Burden and Improve Quality." Committee on National Statistics Workshop on Respondent Burden March 8, 2016. Washington DC.

Plotzke, Michael, Jacob Alex Klerman and Michael Davern. *2011*. "How Similar Are Different Sources of CHIP Enrollment Data?" *Journal of Economic and Social Measurement*, 36(3): 213 – 25.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. 2003. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*. 19: 1-16.

Reiter, J. P. 2003. Inference for partially synthetic, public use microdata sets. *Survey Methodology*. 29(2): 181-188.

Reiter, Jerome P. and Mitra, Robin (2009) "Estimating Risks of Identification Disclosure in Partially Synthetic Data," *Journal of Privacy and Confidentiality*: Vol. 1 : Iss. 1 , Article 6. Available at: http://repository.cmu.edu/jpc/vol1/iss1/6

Rubin, Donald B. 1996. "Multiple Imputation after 18+ years." *Journal of the American Statistical Association*. 9(434):473-89.

Schenker, N., Raghunathan, T. E. and Bondarenko, I. 2010, Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. Statist. Med., 29: 533–545. doi: 10.1002/sim.3809

Scherpf, E., Newman, C., and Prell, M. (2014), "Targeting of Supplemental Nutrition Assistance Program Benefits: Evidence from the ACS and NY SNAP Administrative Records". *Working Paper*.

Urban Institute. 2015. TRIM3 project website, trim3.urban.org, downloaded on November 13 2015.

U.S. Census Bureau. 2015. The Supplemental Poverty Measure: 2014. P60-254, September 2015.

U.S. Census Bureau. 2007, Phase I research results: Overview of National Medicare and Medicaid Files. Report of the research project to understand the Medicaid undercount: The University of Minnesota's State Health Access Data Assistance Center, the Centers for Medicare and Medicaid Services, the Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, and the U.S. Census Bureau. Washington DC: U.S. Census Bureau.

US Census Bureau. 2008a. "Phase II Research Results: Examining Discrepancies between the National Medicaid Statistical Information System (MSIS) and the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)." US Census Bureau: Washington DC. https://www.census.gov/did/www/snacc/docs/SNACC_Phase_II_Full_Report.pdf

US Census Bureau. 2008b. "Phase III Research Results: Refinement in the Analysis of Examining Discrepancies between the National Medicaid Statistical Information System (MSIS) and the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)." US Census Bureau: Washington DC. https://www.census.gov/did/www/snacc/docs/SNACC_Phase_III_Executive_Summary.pdf