Statistical Science
2017, Vol. 32, No. 2, 293–312
DOI: 10.1214/16-STS584

Combining Survey Data with Other Data Sources

Sharon L. Lohr and Trivellore E. Raghunathan

Abstract. Collecting data using probability samples can be expensive, and response rates for many household surveys are decreasing. The increasing availability of large data sources opens new opportunities for statisticians to use the information in survey data more efficiently by combining survey data with information from these other sources. We review some of the work done to date on statistical methods for combining information from multiple data sources, discuss the limitations and challenges for different methods that have been proposed, and describe research that is needed for combining survey estimates.

Key words and phrases: Hierarchical models, imputation, multiple frame survey, probability sample, record linkage, small area estimation.

1. INTRODUCTION

How can we collect data that give accurate and timely estimates of quantities of interest, and assess the suitability of those estimates for answering research and policy questions? Probability sampling theory was developed beginning in the 1920s and 1930s (Neyman, 1934; Duncan and Shelton, 1992) to provide methods for collecting information efficiently and assessing the error arising from sampling. The early books and papers on probability sampling contrasted it with "judgment sampling," in which selection of units depends on an interviewer's or expert's judgment, and with "convenience sampling," in which the sample consists of whatever units are conveniently at hand. Many of the probability surveys in current use were launched decades ago, and when launched were often the only reliable source of information on the topic studied.

Probability samples are often tailored to answer the research and policy questions of interest but face a number of challenges. Response rates are decreasing worldwide and the response rate for a typical telephone survey is now less than 10% (Kohut et al., 2012)—far from the 95% response rate for mail surveys thought to be achievable by Deming (1950), page 35. Even high-quality face-to-face surveys such as the U.S. National Health Interview Survey (NHIS) have declining response rates, and the NHIS household response rate decreased from 92% in 1997 to 70% in 2015 (National Center for Health Statistics, 2016), with additional nonresponse occurring among individual persons within sampled households. Investigations to date have not found strong relationships between the response rate and bias, at least for some statistics (Groves, 2006), but the declining response rates have contributed to higher costs for data collection. The increased expense of conducting probability samples limits the sample sizes. Hence, reliable estimates for subpopulations of interest may require multiple years of data, if they can be calculated at all, and the estimates may be out of date when they are produced.

Parallel to these developments in the probability sampling arena, large amounts of data are now available in many forms. Traditional administrative sources such as the U.S. Decennial Census, tax records, or lists of recipients of social services continue to be available. Road cameras and satellites provide streams of information about traffic patterns and movements. Electronic health records contain the medical history and diagnosed conditions of large parts of the population. Police agencies post lists of crimes reported to them—sometimes within a day of the reporting. Social media

Sharon L. Lohr is Vice President, Westat, 1600 Research Boulevard, Rockville, Maryland 20850, USA (e-mail: sharonlohr@westat.com). Trivellore E. Raghunathan is Director of Survey Research Center, Institute for Social Research, and Professor of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48106, USA (e-mail: teraghu@umich.edu).

such as Facebook and Twitter capture expressed sentiments of the participants, and internet search engines track trending search items. Cellular telephone records provide locations of individuals and details of call locations and durations. Credit card records and shopper loyalty cards capture information on financial transactions. Web crawling software gathers information from web pages. Much of this information can be gathered faster and cheaper than data from a probability sample. The large sample sizes of these data sets can provide finer detail on subpopulations than a typical probability sample. Citro (2014) has emphasized the need to rely on multiple data sources—not just data from traditional probability samples—for producing statistics.

The field of statistics now faces opportunities (and, of course, challenges) in developing methods and frameworks to combine survey and nonsurvey data sources to produce estimates, while maintaining a probabilistic framework for drawing inferences of high quality and rigor. Such developments are important because the data sources differ in their quality and suitability for answering research questions, and many of the inexpensive data sources provide convenience samples. The set of income tax records gives a census of the entities filing taxes in a country; however, some entries in tax returns may be incorrect and the records do not include unreported income or nonfiling entities. The tax records also do not contain information on behavioral variables that may be of interest to researchers. Persons without health insurance are underrepresented in electronic health records. Social media capture the expressed views of persons who use the platform, but do not represent nonusers. Administrative records and large convenient data sets might not have the information needed for statistical purposes.

We review statistical methods that have been proposed for combining information from multiple probability samples and other sources to answer research and societal questions. All sources have advantages and deficiencies, and it is desired to leverage the advantages and reduce the deficiencies as much as possible. This goal accords with Deming's (1950), page 2, holistic view of sampling: "Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability." We summarize each method, highlight its potential gains and drawbacks, and assess the work done to date with respect to the goals of (1) increasing the precision, timeliness and granularity of estimates and (2) providing accurate estimates of uncertainty.

Probability samples have long used information from other sources whenever possible. Stratification and balanced sampling use auxiliary information in the design, while poststratification and regression estimation use auxiliary information to improve precision of estimates and to attempt to compensate for nonresponse and undercoverage. Section 2 briefly reviews these methods and establishes notation.

Sometimes the information from a survey can be augmented through linking individual records from the survey respondents with other data sets, as described in Section 3. Such linkage requires record identifiers that can be used to match records across sources. Record linkage can be thought of as imputing the auxiliary information from the linked records. Even when records are not linked, models developed on a high-quality data source can be used to impute information for responses of interest in other sources, and these methods are described in Section 4.

In other situations, individual records cannot or should not be linked because of insufficient information, privacy concerns, or lack of overlap among data sources. Many data sources report aggregate statistics and do not release individual records. In these situations, summary statistics can be calculated separately from each source and then combined, often by taking a weighted average of the summary statistics. Section 5 summarizes multiple frame survey methods used to aggregate estimates across data sources.

Sections 6 and 7 describe hierarchical models that can be used to combine estimates across studies. Small area (also called small domain) estimation methods borrow strength from administrative data to obtain estimates in subpopulations where the sample size from the probability sample is too small to produce reliable estimates. Many small area methods combine the data from the survey with predictions from a regression model using covariates from the administrative data, often using a hierarchical model in which the deviation of an area mean from the overall mean is represented by a random effect. Hierarchical models are also used for combining data sources, where the individual records from each data source are nested within data sources.

There are many potential advantages to using multiple data sources. These include being able to obtain information on more parts of the population with finer detail on subpopulations. Using administrative or sensor data can result in substantial cost savings. An additional advantage is being able to use multiple sources in survey design. Section 8 discusses using multiple data

sources to improve sampling frames and make the design of the entire data collection effort more efficient.

At the same time, there are also challenges for combining the information. Section 9 describes some of the statistical research needed for combining data sources. We recommend a modular approach, in which different methods may be used with different subpopulations, reflecting the availability of information.

2. MULTIPLE DATA SOURCES IN DESIGN AND CALIBRATION

Most probability samples use information from multiple data sources in the design and estimation as part of standard survey practice. The sampling frame may be constructed using information from a census, and variables in the frame can be used to stratify the sample and determine selection probabilities. A university conducting a survey of its students would have demographic information and information on major and academic performance for every student. Using the frame information in design allows better control of the sample, for example, by specifying a predetermined number of students from each academic division.

A probability sampling design assigns a probability $P(\mathcal{S})$ to each potential sample $\mathcal{S}$ that can be selected from the finite population, and these probabilities serve as the basis for inference. The probability that unit i is included in the sample is $\pi_i = P(i \in \mathcal{S})$, and the design weight is $d_i = 1/\pi_i$. Unit i in the sample is considered to represent d_i units in the population, so that the population total of a characteristic y can be estimated by $\sum_{i\in\mathcal{S}} d_i y_i$.

Calibration and poststratification, reviewed in Särndal (2007) and Brick (2013), use information from an external data source in the estimation. A vector of auxiliary variables $\mathbf{x}_i$ is known for each unit, i, in the sample, and the external data source is assumed to provide the exact value of the population totals for those variables, denoted $\mathbf{X}$. These control totals will be known if the sampling frame has the value of $\mathbf{x}_i$ for every unit in the population, as in a survey of university students, or may alternatively be obtained from an independent external source such as a population census. Calibration constructs adjusted weights w_i that satisfy the calibration constraints $\sum_{i\in\mathcal{S}} w_i \mathbf{x}_i = \mathbf{X}$ while minimizing a distance function between the adjusted weights w_i and the design weights d_i. Poststratification is a special case of calibration, in which the auxiliary variables are indicators for poststrata such as combinations of age, race, and sex. After the poststratification, the survey estimate for the number of persons in each age/race/sex cell is forced to agree with the control total for that cell.

Calibration, or other weight adjustment methods such as raking (Deville, Särndal and Sautory, 1993) or inverse propensity weighting (Rosenbaum and Rubin, 1983; Lee and Valliant, 2009; Valliant and Dever, 2011), are often used to adjust for nonresponse or undercoverage. The calibration constraints require that estimated population totals for the $\mathbf{x}$ variables, using the respondents to the survey, equal the external control totals $\mathbf{X}$: the calibration removes the bias in the calibration variables. It is hoped that the calibration will remove bias for other variables, too, but that hope is sometimes unfounded. Kohut et al. (2012), for example, found that estimates of civic engagement from low-response-rate surveys are higher than corresponding estimates from high-quality surveys, indicating that the weighting adjustments do not remove bias for these variables in the low-response-rate surveys. Calibration and other weighting adjustments are also sometimes used to attempt to adjust for bias from convenience samples (Baker et al., 2013). In this case, in the absence of known inclusion probabilities, the initial design weights are set to 1 and all weight variation comes from the calibration. Again, it is hoped that calibration removes the self-selection bias, although there is evidence that calibration may be less successful in reducing bias for nonprobability samples (Yeager et al., 2011).

The advantages of using external data sources in design and estimation are well known. Stratification almost always increases precision and allows better control of sample sizes for subpopulations. When the response rate is 100 percent, calibration also usually increases precision. When there is nonresponse, calibration and other weight adjustment methods remove or reduce bias in the $\mathbf{x}$ variables used in the calibration, and it is hoped that they reduce nonresponse bias in other variables as well.

These methods also have disadvantages if the external data sources have errors. A frame constructed from a data source that omits some of the population will have undercoverage. If that same frame is used to provide the control totals for the calibration, then the weight adjustments for the undercovered subpopulation will be too small. Control totals from independent sources may also have undercoverage or other errors. The NHIS, which asks respondents about their cellular and landline telephone usage, is often used to calibrate dual frame telephone surveys, discussed in Section 5. Yet the NHIS is itself a sample with sampling

error and potential nonresponse bias, and the errors in the calibration totals introduce additional uncertainty into the estimates. Renssen and Nieuwenbroek (1997) discussed calibrating two surveys to each other using variables common to both surveys.

When there is nonresponse, the properties of estimators calculated using the calibration weights depend on how well the calibration model captures the structure of the population or the response mechanism. Most published survey estimates report standard errors that are calculated under the very strong assumption that the calibration has removed all of the bias. If that assumption is wrong, then the standard errors understate the uncertainty of the estimates.

3. COMBINING INFORMATION FROM INDIVIDUAL RECORDS

In some cases, data records for individuals can be composited from different sources. This can be done to reduce burden for survey respondents, to fill data gaps, or to check accuracy of information. Record linkage, also known as data matching or entity resolution, merges records from different sources that are believed to belong to the same entity such as a person, household, or business. We give two recent examples.

The Canadian Income Survey informed respondents that Statistics Canada planned to combine the household's survey information with tax data (Statistics Canada, 2014). The questionnaire for the survey could therefore omit many of the income questions that had been in previous surveys, reducing the length of the questionnaire and allowing deeper exploration of other topics such as employment, housing, and disability. The information from tax returns was also used to adjust for nonresponse through calibration. This is an example of *exact* or *deterministic record linkage* (DRL), so called not because the method is always error-free but because the linked records agree on a set of characteristics (in this case, tax identification number) that is deemed to determine unique linkage.

Zolas et al. (2015) combined data from university administrative records on graduate students who received research funding with confidential survey information housed at the Census Bureau. Lacking a unique identifier across all sources, they used *probabilistic record linkage* (PRL, Fellegi and Sunter, 1969) to link persons in the university databases with Social Security Administration records and Census Bureau information by name, address and date of birth. This linkage then allowed the researchers to study the employment outcomes of the graduate students in the university databases. PRL methods typically calculate a similarity score for pairs of prospective matches using the pattern of agreements, disagreements, and near-agreements among the variables used in linking. A record from source A is linked with a record from source B if the similarity score exceeds a predetermined threshold. A comprehensive review of PRL methods is beyond the scope of this paper, and we refer the reader to the books of Herzog, Scheuren and Winkler (2007), Christen (2012), and Harron, Goldstein and Dibben (2016) for details of how similarity scores may be calculated.

False matches or missed matches can occur in either DRL or PRL when the linkage variables do not uniquely identify entities. Records may have typographical errors or variations (Robert may be the same person as Bob), be out of date, or have insufficient information for unique linkage (multiple persons may have the same name and date of birth, or date of birth information may be missing). Zolas et al. (2015) failed to match 20% of the doctoral recipients in the participating universities. Even small amounts of error in linkage can bias results (Bohensky et al., 2010): for example, graduate students who cannot be linked may be less likely to have found employment. Winkler (2014) reviewed recent research on accounting for linkage error in statistical analyses of linked data.

Bayesian record linkage methods calculate the posterior probability that two records match. The uncertainty about the linkage in the posterior distribution can then be propagated in other analyses. Steorts, Hall and Fienberg (2016) reviewed Bayesian linkage research and considered a formulation in which records from each data set are linked to latent "true" individuals. Under the assumption that the data sets are conditionally independent given the latent individuals, they calculated the posterior distributions of linkages with the latent individuals, which then allowed computation of linkage probabilities among the different data sets that preserve transitivity (i.e., if A matches B and B matches C, then A matches C).

Record linkage can be thought of as a form of imputation, in which the data fields from source B fill in those missing fields for the linked record in source A (Goldstein, Harron and Wade, 2012). In the Canadian Income Survey, the tax records supply the information on income that is no longer collected in the questionnaire.

Statistical matching, sometimes called data fusion, may be done when individual records cannot be linked.

Records, or groups of records, from source B are matched to similar records from source A using variables common to both sources such as demographic information. For example, source A might have information on heart disease for one set of persons and source B might have information on nutritional intake for a different set of persons, but both sources have information on each person's age, sex, race, ethnicity, and education. By matching records from source A with records from source B that have similar age, sex, race, ethnicity, and education, the analyst can explore relationships between nutritional intake and heart disease. Correlational relationships between the demographic variables and nutritional intake in source B, and between the demographic variables and heart disease in source A, are used to make inferences about the relationship between nutritional intake and health characteristics (Rodgers, 1984; Moriarity and Scheuren, 2001). Of course, such an analysis requires strong assumptions to be made about the comparability of the data sets and the nature of the relationship between nutritional intake and heart disease. A second type of data fusion involves using information from one source to impute variables into another source (Rässler, 2002) and this will be discussed in more detail in Section 4.

When records can be linked across sources with a high degree of accuracy, the linked data sets can provide information on many more variables than would be available from any of the data sources by themselves, and this allows researchers to explore multivariate relationships among these extra variables. Record linkage methods can also be used to augment the number of records in the combined data set, if records that cannot be linked are deemed to be separate entities.

However, it is often difficult to link records accurately, especially when there is little identifying information in the data files. The creation of linked databases also raises concerns about privacy and informed consent, and these issues will be discussed further in Section 9.

4. IMPUTATION

Combining information from multiple data sources naturally fits within the missing data framework given that not all variables are typically measured in every data set. Thus, a standard missing data pattern is obtained when the data sets are concatenated. In addition, many variables in each data set may also be subject to item missing data. Given this scenario, it is not surprising that an imputation-based approach offers a distinct advantage in creating estimates based on combining information from multiple data sources.

In this approach, variables that are missing from a data source are "filled in," or imputed. Many techniques are available for filling in the missing values (Durrant, 2009; Andridge and Little, 2010; Carpenter and Kenward, 2012), and the goal of all of these methods is to use information available in the survey and other sources to accurately predict missing items. Most of the applications of imputation for combining information across sources have relied on multivariate models to predict and then impute the missing items. Models developed on one data source may be used to impute missing variables in other sources. Alternatively, all records may be concatenated into one large data set and all missing items in the concatenated data may be imputed using one multivariate model or a sequence of regression models.

There are many advantages to being able to impute the missing items. The primary advantages of imputation are the abilities to augment the amount of information available for analysis, and to produce data sets without "holes" in them. Suppose that Survey A provides data on x and y, Study B provides data on y and z, and administrative data provide information about x and z. An imputation model making use of the bivariate relationships estimable from the individual sources can provide information about the relationship among all three variables. Clearly, combining data from these sources provides a means for inferring beyond the scope of each individual study.

Kim and Rao (2012) imputed a variable of interest y in a large survey that does not measure y directly but that does measure covariates $\mathbf{x}$. A second, smaller, survey measures both $\mathbf{x}$ and y, and a regression model predicting y as a function of $\mathbf{x}$ is fit to the data in this survey. That model is then applied to the $\mathbf{x}$ variables in the large survey to obtain imputed values for y. These imputed values are then used together with the weights from the large survey to estimate the population total for y; the standard error of the estimate depends on the sampling variability from the large survey and on the lack of fit of the regression model. This model has the strong assumptions that the $\mathbf{x}$ variables for the two surveys measure the same quantity (i.e., there is no measurement error due to mode effects or other sources of incomparability discussed below), and that the regression model developed on data from the small survey applies to the large survey.

Gelman, King and Liu (1998) used multiple imputation to combine information from a series of cross-sectional surveys where some questions are not asked

in some surveys. The particular problem involved combining data from 51 election polls conducted during the six months prior to the election. The goal was to assess the changes in vote intentions over time for different subgroups based on gender, age, party affiliations, etc. Using a hierarchical model to incorporate study differences, a fully Bayesian approach was used to draw values from the posterior predictive distribution of not asked or not answered items conditional on the observed data.

Raghunathan (2006) and Schenker, Raghunathan and Bondarenko (2010) used multiple imputation to correct for possible bias in self-reports of health conditions (such as diabetes, hypertension, or hyperlipidemia) in the NHIS using data from the National Health and Nutrition Examination Survey (NHANES) which collects data using both self-reports and clinical measures. An added advantage of this approach is that the national estimates of undiagnosed health conditions borrow strength from both surveys.

Another example is given in He, Landrum and Zaslavsky (2014), where data from surveys, medical records, Medicare claim data, and cancer registries were combined to study hospice use in terminal cancer patients. All data sources had missing data and the multiple imputation relied on observed data from all sources.

There are number of challenges when implementing multiple imputation approach for combining information from multiple survey data sources. For example, surveys usually involve stratification, clustering and weighting for selection and nonresponse. Though each survey may represent the same or a similar population, the complex survey design differences have to be taken into consideration in deriving the combined estimates. The recent work of Dong, Elliott and Raghunathan (2014a, 2014b) proposed "uncomplexing" the survey data by simulating populations from each survey data and then combining using the superpopulation modeling framework. Zhou, Elliott and Raghunathan (2015) extended the approach when variables in each survey data are subject to item missing values.

Estimates based on combining information from multiple data sources are subject to errors due to incomparability as well as issues in modeling of those errors. Early references to address the issue of comparability in pooling data are Bancroft (1944) and Mosteller (1948). The latter is perhaps the first to discuss the bias-variance trade-off in pooling data and lay out conditions for deciding whether to pool or not. Here, we raise five potential sources of incomparability that need to be considered. These are raised in the context of imputation, but also apply to other methods for combining data sources.

A first source of potential incomparability is the differences in the types of respondents and the sources of respondent information. For example, in a household survey, the respondents may be interviewed face-to-face and report health conditions based on memory and recall. The data from the other source may be provided by physicians who may be consulting medical records to check for health conditions.

A second potential source of incomparability may arise due to mode of the interview. For example, one survey may be based on random digit dialing, the second survey may be based on face-to-face interviews, and the third survey may begin with a telephone mode but switch over to face-to-face interviews on a subset. Based on the effect of mode on the measurement of outcomes, pooling may introduce bias, if there is one preferred "gold standard" approach for collecting the information. In the absence of such a gold standard, combining data may be a better reflection of the population quantity since it accounts for differences among sources.

A third source of potential incomparability may arise due to survey contexts. For example, nationally representative data collected by a federal agency that is well known and well publicized may have different response error properties than a survey conducted through a reputable institution that is not as well known. The advance letter that is usually sent may also affect the measurement error properties.

A fourth source of potential incomparability may arise from differences in the survey design. For example, NHIS collects information in an interview setting whereas NHANES collects information in an interview setting but with an advance knowledge provided to the respondent that he/she may be selected for medical examination and specimen collection. Recalling abilities of the respondents may differ in these two survey settings.

A final source of potential incomparability may arise due to different wordings of the questions asking the same information. Other issues relate to placement of the questions in the survey instrument, protocol differences for the interviewer prompts, and additional questionnaire features.

These and other sources of incomparability affect combining information from multiple survey data sources. If nonsurvey data sources are also brought into

the mix, a lack of a probability survey framework to assess representativeness can be an additional source of incomparability.

The above discussion may appear to discourage combining information from multiple sources. On the contrary, the advantage of combining information is the ability to address analytic problems beyond the scope of any single survey, and imputation can provide a richness of data unavailable from any single source. Direct estimation techniques may not be applicable and some modeling approach may have to be used to properly harness and pool the information. There are no assumption-free approaches in statistics. The modeling framework provides a means for incorporating the study differences and one or more issues of incomparability in an explicit manner. An explicit modeling framework provides transparency. With limited data and complicated modeling, it is important to consider issues such as covariate selection, features to be incorporated, collection of auxiliary variables, and incorporation of model uncertainty.

5. MULTIPLE FRAME METHODS

In a multiple frame survey, samples are selected from each of F sampling frames and estimates from the samples are combined. A sample is selected from each of the frames, and the estimates from the different samples are combined. The different frames often include different subsets of the population. For example, frame A might cover the entire population of interest, such as the frame for the face-to-face NHIS; frame B might be a set of electronic medical records; frame C might consist of tax records. Some frames might not be well defined in advance, as would occur if the sample from frame D consists of volunteers responding to an internet survey. For some frames, such as electronic medical records or tax records, the frame itself may have the information of interest so that the entire data set may be used rather than sampling from it.

Multiple frame survey methods have several potential advantages. If each data source includes only a part of the population of interest, using multiple sources as frames can give better coverage of the population. Telephone surveys often take one sample from a frame of landline telephone numbers and an independent sample from a frame of cellular telephone numbers; using just the landline (or cellular) frame would exclude persons with exclusively cellular (or landline) telephone service from the survey. Multiple frame surveys can increase precision for little additional cost if data collection is inexpensive for some of the frames. This is particularly beneficial if the population being studied is a small component of the general population. Data collection has already been done for electronic medical records and tax records, and using them can increase the precision for the parts of the population they contain. The large sample size from these sources also provides more information on subpopulations such as persons with rare disorders or taxpayers who hold tax-exempt bonds. Lesser, Newton and Yang (2008) investigated use of lists of individuals belonging to disability organizations as sampling frames in their study of improving public transportation access for persons with disabilities. While these lists do not include all persons with disabilities, they reduce screening costs that are needed if respondents to a random digit dialing survey are asked questions to determine if they are in the population of interest. However, multiple frame surveys are more complicated than single frame surveys, and must be carefully analyzed to take advantage of the potential increased efficiency and to avoid bias.

The easiest way to use multiple frames, if feasible, is to create a single frame from the different sources before sampling by concatenating the frames and removing duplicates. This is not always possible, however: for a dual frame cellular/landline telephone survey, typically the sampling frames would consist of landline and cellular telephone numbers, and one would not know before sampling whether a person associated with a cellular telephone number also has access to landline service. If a single frame cannot be constructed using the frame information, then an alternative is to take independent samples from the different frames and then combine the data or estimates after sampling.

In multiple frame methods, the union of the frames is assumed to be the population of interest. The overlap of the frames creates overlap sets[1] consisting of the population units accessible through different combinations of the frames. Using the notation in Hartley (1974), overlap set a consists of the population units in frame A but none of the other frames; set ab consists of the units in frames A and B but not C or D; set $abcd$ consists of the units that could be accessed through each of the four frames. The overlap sets are disjoint and together comprise all of the population units that can be reached through at least one of the frames. If each of the F frames consists of the same set of population

[1]The multiple frame literature typically calls these *domains* rather than *overlap sets*; however, in this paper we use the term domain to denote a subpopulation of interest.

units, as would happen when all frames cover the entire population, there will be one overlap set. If all frames overlap but none has complete coverage, there can be up to $2^F - 1$ overlap sets. With $F = 2$ frames, there can be 3 overlap sets: units in frame A but not in B (set a), units in frame B but not in A (set b), and units in both (set ab).

Lohr and Rao (2006) and Metcalf and Scott (2009) summarized estimators for combining information from the F samples taken from the frames. The complications come in because units in more than one frame have multiple chances to be selected; in a dual frame survey, the units in overlap set ab can be sampled from either or both frames. If we simply concatenated the data sets without adjusting for the multiplicity, then the individuals in set ab would be overrepresented in the combined samples. The population total in overlap set k can be estimated by a weighted average of the estimated population totals from the individual frames that have observations in overlap set k:

$$\hat{Y}_k = \sum_{f \in k} \lambda_{kf} \hat{Y}_{kf}, \tag{5.1}$$

where $\sum_{f \in k} \lambda_{kf} = 1$. Then the overall population total is estimated by summing the pieces $\hat{Y}_k$ from the distinct overlap sets.

The λ_{kf}'s can be thought of as adjusting the respondents' weights for the multiplicity that occurs because units can appear in multiple frames. The estimate $\hat{Y}_{kf}$ from each source is assumed to be approximately unbiased after calibration has been performed. The relative importance λ_{kf} assigned to source f may be fixed in advance (it is common to use $\lambda = 1/2$ for the overlap set ab in dual frame surveys), based on the surveys' selection probabilities (Bankier, 1986; Kalton and Anderson, 1986), or determined so as to minimize the variance of the aggregated estimate (Hartley, 1962; Skinner and Rao, 1996). Chauvet and de Marsac (2014) studied estimators for two-stage dual frame surveys where the two surveys share some of the same primary sampling units.

To apply the appropriate weighting factor λ_{kf} to each sampled unit, one must be able to identify which overlap set it belongs to, or at the very least one must know how many sampling frames it belongs to (Mecatti, 2007; Rao and Wu, 2010). We know that a respondent to the NHIS is in the frame for that survey, but do we know whether he or she is represented in the set of electronic health records? In other words, how many times could the same person be represented in the combined data sets? We do not need to be able to link records across surveys, but we do need to know how many chances an individual has to be in the data set. The overlap sets for a multiple frame survey are often determined by asking respondents about their membership in other frames, and that sometimes introduces measurement error into the determination. For example, respondents to a dual frame cellular/landline telephone survey are usually asked about their relative usage of cellular and landline telephones to receive calls, but that determination may be imprecise. Lohr (2011) showed that dual frame methods can have less precision than using estimates from just one data source if individuals are misclassified in the wrong overlap set, and she and Stokes and Lin (2015) considered estimators that account for misclassification bias in dual frame surveys.

The additional complexity of multiple frame surveys has implications for nonresponse adjustments. Brick et al. (2011) discussed choosing the compositing factor λ to reduce nonresponse bias. It is often assumed that the weights from all samples are individually preadjusted for nonresponse using methods such as those described in Section 2; if desired, the weights can be calibrated again after the estimators are combined (Ranalli et al., 2016).

Much of the literature on multiple frame surveys assumes that the survey conducted from each frame asks the same questions, and that estimates from the overlap sets from different sources measure the same quantity. In a dual frame survey, this means that the expected value of the estimated population total from overlap set ab is the same for the estimator from frame A and the estimator from frame B. But the sources of survey incomparability discussed in Section 4 apply to multiple frame surveys as well. If the sample selected from frame A is collected using different survey questions, modes, or procedures than the sample from frame B, the estimated population totals in ab may differ because of the procedures or nonsampling error rather than because of sampling variability. These differences are of particular concern for the sources in which data collection is inexpensive, because the estimates may have different measurement error properties than the estimates from the expensive sources. It is important that these nonsampling errors be included in the measures of uncertainty about the survey estimates. Typically, it is recommended that the variance of the multiple-frame estimate for an overlap set k be estimated by summing the variances $\lambda_{kf}^2 V(\hat{Y}_{kf})$ for the components of the weighted sum in that set, but this formula accounts only for sampling error and does not

consider differences that may be due to different survey procedures or questions.

One method for evaluating potential bias is to use multiple frame methods on the different subpopulations, called domains, of the surveys. These domains can be distinct from the overlap sets. For example, in a dual frame telephone survey, the overlap sets would be persons with a landline phone only, persons with a cell phone only, and persons with both landline and cell phone. The domains studied could be different geographic regions or demographic subgroups. This allows the analyst to compare estimates from the different surveys in those domains. Merkouris (2004, 2010) used regression methods to adjust two surveys being combined, using the common variables from the surveys. Merkouris (2010) used regression estimators to combine information from multiple surveys and obtain small domain estimators. He considered the case in which there are multiple surveys of the same population, and calibrated the surveys to each other using variables that are common to both surveys.

Lohr and Brick (2012) considered dual frame estimation when one of the sources is considered to be unbiased but with small sample sizes in domains. The other data sources have larger sample sizes but potentially have differential bias across domains. The relative contributions of sources toward each domain estimator depend on the relative variances and the amount of differential bias. These methods allowed the differences among estimators that arose from nonsampling error to be included in the mean squared error estimates for the domains.

Multiple frame survey methods have great potential for combining information from data sources that are measuring the same quantities. As with all the other methods discussed in this paper, however, they have strong assumptions about the comparability of the data sources, and extending the methods to relax those assumptions is a promising area for research.

6. SMALL AREA ESTIMATION

A pressing need for many policy makers is obtaining estimates of important quantities at small geographic levels such as counties or states, or for a subgroup based on certain demographic characteristics (such as gender, age or race). Many national surveys are inadequate for constructing such estimates because the sample size in many domains of interest is too small, or may even be zero. Combining data from multiple sources provides the only meaningful way to develop estimates for domains, or areas, with small sample sizes.

Small area estimation methods combine information from a survey with auxiliary information from administrative data sources to calculate domain-level statistics. Fay and Herriot (1979) estimated the mean θ_d in domain d using a weighted average of estimates from two sources. The first estimate is $\bar{y}_d$, which is the estimated mean in domain d calculated directly from the survey. For many small domains, $\bar{y}_d$ is based on a small sample size and is imprecise; for some domains such as large states, however, $\bar{y}_d$ may have high precision. The second estimate uses a regression model predicting θ_d from domain-level covariates $\mathbf{x}_d$ that are available from an administrative data source to obtain prediction $\hat{\theta}_d$. The Fay–Herriot estimator of θ_d is $\lambda_d \bar{y}_d + (1 - \lambda_d)\hat{\theta}_d$, where $\lambda_d \in [0, 1]$ is larger if $V(\bar{y}_d)$ is small or if the regression model does not fit the data well (and, therefore, does not provide accurate predictions). The Fay–Herriot estimator is thus of the same form as (5.1), combining the direct estimate $\bar{y}_d$ from the survey with a regression prediction based on covariates from an administrative data source. A two-stage model underlies the Fay–Herriot estimator. First, the area-level means from the survey are assumed to follow a distribution with mean θ_d and sampling variance ψ_d, where ψ_d is estimated using the survey design and weights. The second stage relates the θ_d's to the external-source covariate information through a regression model, $\theta_d = \mathbf{x}_d'\boldsymbol{\beta} + v_d$, where v_d represents the error in prediction from using the regression model and is assumed to have mean 0.

The U.S. Small Area Income and Poverty Estimates (SAIPE) program (United States Census Bureau, 2016) uses a variant of this method to provide annual poverty statistics for states, counties and school districts. The direct estimates are one-year estimates from the American Community Survey (ACS), and the regression predictions use covariates from the Decennial Census, from tax records collected by the Internal Revenue Service, from the Supplemental Nutrition Assistance Program, and from population estimates. The use of the administrative data sources allows the U.S. Census Bureau to publish poverty statistics for every county and school district each year, even though the sample sizes for many of these areas are too small for the ACS estimate to be published.

When estimates are produced for nested areas of different sizes, it is often desirable to adjust estimates at finer levels of geography so that they aggregate to estimates at coarser levels of geography. In general, the

estimates for larger geographic areas are thought to be more reliable because they have a larger sample size from the survey and rely less on the model-based predictions which are based on model assumptions. The SAIPE program state-level estimates of the number of children in poverty are ratio-adjusted so that they sum to the national estimate of number of children in poverty that is calculated from the ACS. The county estimates within a state are also ratio-adjusted to sum to the state estimate, and the school district estimates are similarly benchmarked to the county estimates. In this way, the estimated counts of children in poverty are consistent across school districts, counties, and states, and the nation as a whole. Datta et al. (2011) reviewed benchmarking methods for small area estimates, and proposed a class of Bayesian small area estimators that constrain a weighted average of the posterior means to equal prespecified estimates. Pfeffermann and Tiller (2006) and Hyndman, Lee and Wang (2016) described methods that may be used to benchmark time series.

The Fay–Herriot model makes use of statistics computed for each area using the sampling weights from the survey, and uses individual records only through the area-level summaries. A unit-level small area model (Battese, Harter and Fuller, 1988) may be used when covariates are available for each population unit. A hierarchical model is used for the individual responses of survey participant j in area d:

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + v_d + e_{dj},$$

where the area-specific random effects v_d are assumed to have mean 0 and variance σ_v^2, and the individual-level errors e_{dj} are assumed to have mean 0 and variance σ_e^2. In this hierarchical model, individual respondents from an area are considered to be nested in that area. Rao and Molina (2015) provided a comprehensive description of models commonly used in small area estimation, including empirical Bayes, hierarchical Bayes, time series and spatial models. For most of these models, the $\mathbf{x}$ information is assumed to be measured exactly, and different methods are needed if the $\mathbf{x}$ information comes from another survey or a source with differential measurement error.

Ybarra and Lohr (2008) used a Fay–Herriot-type model, accounting for measurement error in the covariates, to estimate mean body mass index (BMI) for age/race/sex domains. NHANES calculates BMI from direct measurements of height and weight, and thus is thought to be more accurate than the measure of BMI from the larger NHIS that is calculated from self-reported weight and height. The measurement error models accounted for the sampling error in NHIS both in the calculation of λ_d (which was smaller if the NHIS estimate had higher variance) and in the mean squared error of the small area estimates. You, Datta and Maples (2014) used a bivariate Fay–Herriot model to incorporate the error from multiple sources when estimating disability.

Elliott and Davis (2005) and Raghunathan et al. (2007) further developed small area estimation by combining data from two surveys, NHIS and the Behavioral Risk Factor Surveillance System (BRFSS). Elliott and Davis (2005) used a model-assisted framework to match the respondents in the two surveys using propensity score methods and then used the combined data to develop Fay–Herriot-type estimates. On the other hand, Raghunathan et al. (2007) used an explicit Bayesian hierarchical model framework to model NHIS, which was assumed to provide unbiased estimates for telephone and nontelephone households, but for only a few counties. The NHIS was combined with BRFSS data, which provides biased estimates for the telephone households but for all the counties. Using auxiliary county and state level covariates, the estimands (the population-size-weighted county-level population means of telephone and nontelephone households) were simulated from their posterior distribution using Markov Chain Monte Carlo methods.

The arcsine square root transformation was applied to the county level direct estimates, in part to simplify the modelling by stabilizing the variances of the outcomes. However, the theory behind the variance stabilizing properties of the arcsine square root transformation is a large-sample theory, and thus the transformation might be less effective for some of the counties in the project that have sparse samples. To avoid making large sample approximations, the logit-normal model was also used which resulted in similar estimates but with an enormous increase in computational time and complexity. Current work is considering small area estimates by combining three different subpopulations within each area: households with a landline (with or without cell phones), nontelephone households, and cell-only households. Chen, Wakefield and Lumely (2014) reduced the computational complexity for Bayesian hierarchical small-area models by using an integrated nested Laplace approximation. Mercer et al. (2014) incorporated spatial random effects in models estimating smoking prevalence at the zip code level from BRFSS, and compared different model structures in a simulation study.

Small area methods use multiple sources to augment the information available at the domain level. As with imputation and multiple frame methods, this augmentation requires the use of model assumptions and we refer the reader to Rao and Molina (2015) for discussion of model misspecification in small area estimation.

7. HIERARCHICAL MODELS FOR COMBINING DATA SOURCES

In the hierarchical models used in Section 6 to obtain small area estimates, random effect terms are used to model the means of different domains. Hierarchical models can also be used to synthesize data from multiple sources: in this usage, random effect terms represent the means from different data sources, and individual data records from the studies (if available) are nested in the studies. The problem is structurally similar to that of random effects models used for meta-analysis (Sutton and Higgins, 2008), in which summary statistics from different studies are assumed to come from a normal distribution with mean θ, and a weighted average of the summary statistics from the different studies is used to estimate the underlying effect size of the treatment. The weights may be inversely proportional to variances, or experts' judgments may be used to assess the quality of the studies and downweight studies with lower quality (United States General Accounting Office, 1992; Turner et al., 2000; Greenland, 2005).

A number of models have been proposed that combine summary statistics—usually means—from different studies. Methods that rely upon summary statistics do not require access to the individual data records that comprise the studies, and thus can be used when access to the individual data records is restricted. The mean and its estimated variance for the subpopulation of interest is calculated separately for each data source using the design and the nonresponse-adjusted weights for that source.

Manzi et al. (2011) used a Bayesian analysis to estimate θ_d, the smoking prevalence in local area (domain) d, for each of 48 local areas in England. Prevalence estimates were available from seven different studies, but these studies differed in methodology and quality, and there was concern that estimates from some of the studies could be biased. The estimated prevalence for domain d from data source j, u_{dj}, is assumed to follow the model

$$u_{dj} = \theta_d + \delta_{dj} + e_{dj}, \tag{7.1}$$

where the bias δ_{dj} is assumed to follow a normal distribution with mean Δ_j and variance τ_j^2. The error term e_{dj} is the sampling error for the estimate u_{dj}, assumed to have mean 0 and variance σ_{dj}^2, where the variance is calculated from the sample design. Note that the model in (7.1) is similar to the Fay–Herriot model for small area estimation, with the additional feature that the model-based deviation component is allowed to have mean Δ_j rather than mean 0. Many of the properties of the estimates of θ_d depend on the constraints put on the mean bias Δ_j from source j. Manzi et al. (2011) adopted a vague prior distribution for the bias Δ_j, and constrained the mean prevalence over all domains to equal the prevalence estimate from the UK General Household Survey, which was considered to be highly accurate. Alternatively, one of the sources could be considered to be a gold standard with zero bias. Turner et al. (2009) discussed a framework for eliciting prior information on bias for multiple data sources.

Many hierarchical models that treat different studies as random effects incorporate the between-source variability into the measures of uncertainty. Thus, the estimate of smoking prevalence obtained from combining multiple studies may have larger standard error than an estimate constructed from one probability sample. The standard error of the estimate from a single probability sample includes the within-survey error, while the standard error of the estimate obtained by pooling surveys also includes the between-survey error.

Wang et al. (2012), Nandram, Berg and Barboza (2014), and Cruze (2015) discussed hierarchical Bayesian methods for combining information from multiple repeated surveys to obtain benchmarked estimates of state and regional crop yields in the United States. The quantity of interest is the true annual yield in year t, denoted as μ_t, and it is desired to estimate μ_t at different time points in the growing season. The first survey takes monthly field measurements, including acres planted, from a sample of sites in states that are the top producers for the crop being studied. The second survey is a monthly national interview survey asking farmers to estimate their expected yields for a range of crops. The measurements from the first and second surveys are taken throughout the growing season. Estimates of μ_t derived from these two sources tend to be biased; however, the biases are assumed to be consistent across years, and depend only on number of months before harvest. The third national survey occurs after harvest, and asks farmers about yield of

different commodities as well as other quantities. Because the third survey has large sample size and occurs after harvest, it is considered the gold standard for yield estimates—but it is not available for making preharvest estimates in the current year. The model uses the historical relationship between the gold standard estimate and the monthly estimates of crop yield from the other two surveys, so that the accruing information from the first two surveys can be used to update the forecast yield μ_t for the current year. The estimated crop yields for month m, year t, and survey j are assumed to be normally distributed with mean given by μ_t plus a bias term for the first two surveys that varies by survey and month. The bias term is assumed to be zero for the gold standard survey. The posterior distribution of μ_t for the current year is calculated by conditioning on the data available at the time of the forecast and including covariates such as weather information. This model uses the estimates of bias from the first and second surveys from previous years to adjust the current-year forecast for those biases. The posterior mean for crop yield is a weighted average of the bias-corrected estimates from the first and second surveys, the information from the third survey when available, and predictions using covariate information, with higher weights assigned to more precise sources. This methodology allows biased surveys to be used to produce more accurate and timely estimates of crop yield, along with measures of uncertainty.

The models discussed above combine summary statistics to improve the precision of estimates. Other studies have combined individual records with aggregated statistics through a hierarchical model. Wakefield and Salway (2001) presented a framework for using aggregated data, with attention to potential bias coming from variability of covariates in the different areas; Wakefield (2004) argued that sometimes informative priors are needed when fitting hierarchical models using aggregated data. Jackson, Best and Richardson (2008) used a hierarchical logistic model to study the risk of hospital admission for heart and circulatory disease. They had individual-level data on risk-behavior and socioeconomic covariates, and the outcome of hospital admission from the Health Survey for England; individual-level data on covariates from the UK census; and aggregate counts of hospitalization, and socioeconomic covariates, at the ward and district level. The individual-level logistic model had terms for area-level covariates; the model for aggregate-level data relied on the summary statistics for different areas as well as the within-area variability in covariates.

Finucane et al. (2014) combined information using a hierarchical Bayesian framework to estimate trends in mean systolic blood pressure for different countries. They had surveys and other data sources, of varying quality, from almost 200 countries. Some data sources contained individual records, while others only had summary statistics; some were rigorous national probability samples with high response rates, while others were less representative community studies. The hierarchical model used the estimated mean and standard deviation from each data source and year as input. Random effects terms captured the study-level heterogeneity. Finucane et al. (2014) used an informative prior distribution to account for the quality of the data sources; they constrained the variances of the different terms so that national probability samples were assumed to have lower model variance than regional studies, which in turn had lower model variance than community studies. This did not model the bias explicitly, but resulted in the community studies that were thought to be less reliable having less influence on the estimates of health characteristics. Finucane et al. (2015) used related methodology to estimate the distribution of child malnutrition for different countries.

The malaria atlas project (Bhatt et al., 2015) employed a Bayesian hierarchical model to study the infection prevalence of the malaria-causing parasite *Plasmodium falciparum* in sub-Saharan Africa from 2000 to 2015. Data sources included community-level measurements of the parasite rate from published literature (see http://www.map.ox.ac.uk/explorer/), national household surveys, and historical records that provided environmental covariates (such as temperature, surface wetness, and population) at a 5 km · 5 km spatial resolution. Spatial and temporal correlations were included in the model through a Matérn covariance function and first-order autoregressive terms. This model allowed the investigators to include uncertainty that arose from small sample sizes, information on observed clinical incidence rates, and the estimated parameters in the posterior distribution predicting prevalence.

The Global Burden of Disease Study used similar hierarchical Bayesian models to combine data from thousands of epidemiological sources as well as available national surveys in approximately 200 countries in order to study levels and trends of disease incidence, prevalence, and mortality (Vos et al., 2015; Wang et al., 2016). The hierarchical Bayesian population reconstruction method described by Wheldon et al. (2016) reconciled census counts with population

projections based on vital rates. The prior information came from expert opinion about the relative error of the data sources, and the methodology provided a mechanism for assessing biases in the different data sources.

Sweeting et al. (2008) used hierarchical models (see also Ades and Sutton, 2006) to evaluate the consistency of data sources (one capture-recapture study on intravenous drug users, four national household surveys asking about drug use, medical clinic data, blood donation records, and testing data) for estimating the prevalence of hepatitis C. The model included parameters for the bias from each sources. They re-estimated the model, leaving each data source out in turn, to investigate whether omitting sources changed the estimates.

Although it does not use an explicit hierarchical model, Brick's (2015) design-based framework for compositing multiple surveys is related to this work. Each source is weight-adjusted, using poststratification or inverse propensity weighting, and the variability among sources is used to estimate the variance of the mean estimated from the sources.

Strauss et al. (2001) studied hierarchical models to estimate the relationship between residential lead exposure and children's blood lead levels. NHANES provided information on blood lead concentrations, but not exposure; the U.S. Department of Housing and Urban Development (HUD) National Survey on Lead-Based Paint in Housing had exposure information but no information on lead levels in children. They used a third source that related lead exposure and blood lead concentration (but only for Rochester, New York). An additional complication occurred because the Rochester study measured lead exposure differently than the HUD study. The authors assumed that the true value of the lead exposure level was a latent variable, and modeled the Rochester and HUD lead exposure using covariates available in both sources. Adopting the strong assumption that the exposure/blood lead relationship found in Rochester held nationally, the model allowed the researchers to predict a national distribution of blood lead in children.

There has been a great deal of work in biostatistics on pooling information from different studies. Pocock (1976), Raghunathan (1991), and Prentice et al. (1992) pooled information from a randomized trial with retrospective data from historical controls, using Bayesian methods to model potential bias in the historical controls. Stuart (2010) reviewed research on methods that may be used to match treatment and control groups across multiple sources. Dugoff, Schuler and Stuart (2014) extended propensity-based matching methods to complex surveys.

One challenge when combining individual records from different sources is how to treat the survey weights from individual sources (Rao et al., 2008). When summary statistics are combined, the individual-source survey weights are used to calculate the means for each domain and then a weighted average is taken of these means. When combining individual records across data sources, two sets of weights are used: (1) the weights used to generalize each survey to its population, described in Section 2, which are based on the inverse of the selection probabilities with adjustments for nonresponse and calibration; and (2) the relative contribution of each individual source toward the combined estimate. In multiple frame methods, the weights within each overlap set are multiplied by the value of λ for that frame and overlap set to account for the multiplicity. A similar method could be used with hierarchical models, but more research on this topic is needed. Korn and Graubard (1999), Chapter 8, discussed the calculation of weights when pooling multiple surveys.

Hierarchical models have many advantages for combining data. As with imputation methods, they provide a transparent model framework for combining the information with explicit assumptions. The model assumptions can be strong, however, and the measures of uncertainty, while accounting for variability among sources, often do not account for potential model deficiencies.

8. DESIGNING STUDIES TO LEVERAGE MULTIPLE DATA SOURCES

The increasing availability of multiple data sources opens up new options for survey design. As described in Section 5, data sources may have information on different but overlapping parts of the population. Electronic medical records might provide information on persons who have used certain medical services, but other sources are needed to provide information about the health characteristics of persons who have not used those medical services. A data source may provide accurate information for some populations, but may be thought to have bias for other subpopulations.

With multiple sources available, the goal of the design is to leverage the strengths of each source to provide an accurate picture of the population and of subpopulations of interest. In this section, we consider the situation in which administrative data sources are available for some subpopulations and it is desired to use

those sources when designing a probability sample that will (1) provide a check on the accuracy of the other sources for variables of interest and (2) provide accurate information on subpopulations that are underrepresented in the administrative sources. There is a danger that all administrative sources may undercover the same subpopulations: for example, persons without health insurance may be missing from electronic health records and from insurance records. The survey design needs to capture the subpopulations underrepresented in other sources.

The administrative data sources may be used in several ways during the design process. First, they may be used when constructing the frame. Section 5 discussed combining estimates from multiple frame surveys when the information could not be consolidated prior to sampling. But of course in some situations, the information from the sources can be linked and consolidated to form a better sampling frame with rich auxiliary information. This auxiliary information may be used to improve the efficiency of the stratification of the sample, or may be used in conjunction with balanced sampling (Valliant, Dorfman and Royall, 2000). This also provides higher quality information for surveys of particular subpopulations. If it is desired to take a survey of persons with asthma, the data sources may provide better information for screening eligibility of the sample.

A second use of the information from other sources is to provide contextual variables for the survey. Nachman and Parker (2012) linked respondents from the NHIS to information from the U.S. Environmental Protection Agency AirData system to study the relationship between exposure to pollutants and outcomes such as asthma and bronchitis. They linked the latitude and longitude of the survey respondent to the kriged prediction of fine particulate matter at that latitude and longitude. This linkage provided important contextual variables for interpreting the NHIS data.

Third, the administrative data may provide information for dealing with nonresponse in the survey. If survey records can be linked, the administrative data may be used to impute information for nonrespondents. Tax records, for example, could be used to impute missing income information for nonrespondents to the survey.

The information from the administrative sources may also provide valuable information for nonresponse assessment and follow-up in surveys. Adaptive (also called responsive) survey design often uses information from multiple sources (Groves and Heeringa, 2006; Wagner and Raghunathan, 2007; Wagner et al., 2012; Tourangeau et al., 2017) to modify the protocol for survey data collection while in the field. These methods often use paradata—data about the process of collecting the survey data, such as number of contact attempts or neighborhood observations—to adapt the survey design. Data from external sources such as sensor data could also be used for these design modifications.

Smith (2011) reported the recommendations of an international workshop on using auxiliary data to detect and adjust for nonresponse bias in surveys. Auxiliary data from population registers, linked databases, the sampling frame, or paradata can provide case-level information for assessing potential nonresponse bias, while independent population estimates from censuses or high-quality surveys such as the ACS can be compared with survey estimates. The report noted, however, that adding more auxiliary data "increases the likelihood of deductive disclosure and thus potentially undermines confidentiality" (Smith, 2011, page 395).

Fourth, the entire data collection can be designed to make use of the multiple sources of data. If the records from different sources can be linked and merged before sampling to construct a rich sampling frame, then the sample can be allocated optimally using stratification or balanced sampling. Thus, if frame A is nearly complete but expensive to sample, frame B is incomplete but less expensive, and the frames can be combined before sampling, then the design can specify obtaining the information from overlap sets b and ab from frame B, and only using the expensive frame A to collect information on overlap set a.

If the population source information is unknown before sampling, however, the design needs to ensure that all parts of the population are represented in the sample. Hartley (1962) derived the optimal sample sizes along with the optimal compositing factor λ for dual frame designs where overlap-set membership is unknown in advance. When frame A is nearly complete but expensive to sample, and frame B is incomplete, a larger sample size should be taken from frame A when: (1) the cost per unit is higher in frame B, or (2) a larger proportion of the population is in overlap set a, and thus cannot be sampled from frame B. Lohr and Brick (2014) found that for many cost structures it made economic sense to use a two-phase screening survey for the expensive frame A, where the interview was terminated after determining that the unit was also in the less expensive frame B. This is of course less efficient than if the frame membership is known before sampling, because extra effort must be expended to obtain

screening interviews for persons sampled from frame A whose data are not used in the estimation.

We recommend a modular approach to survey design, in which the design makes use of the different information sources available for different parts of the population. With data collection planned to take advantage of administrative sources, the survey design can concentrate on parts of the population less represented in other sources.

9. OPPORTUNITIES

All of the methods for combining information reviewed in this paper have strengths and shortcomings. Linking records allows the most efficient use of information, but accurate linkage is not always possible and linkage can raise privacy concerns. Imputation can allow use of data sources that contain some but not all of the variables of interest by imputing the missing variables through multivariate relationships determined from sources that have the other variables, but the imputation models are strong: if the imputation model is developed on a source that has different relationships than the source where the imputation is applied, then the imputed values may be misleading. Multiple frame methods allow information from many sources to be composited, but require accurate information about the frame membership of the sampled units. Hierarchical models are powerful tools for combining information from surveys, but a big challenge with these methods is accounting for bias from different sources. All of the methods other than deterministic record linkage rely on models, and the results need to be investigated for sensitivity to those models.

The survey designer and analyst may wish to use different methods for different subpopulations, reflecting the availability and quality of sources available for those subpopulations. In the United States, most persons aged 65 and over are on Medicare, and so are represented in the records of the Centers for Medicare and Medicaid Services (CMS). It may be possible to link those records with records from electronic health records to obtain more detailed information about that subpopulation. Younger persons, however, are less likely to be in the CMS records and for those subpopulations hierarchical modeling or multiple frame methods may be needed. Thus, we see the problem of combining information from different sources as a mosaic, where different sources contribute to constructing the entire picture.

Much of the literature on meta-analysis and on combining surveys discusses having higher reliance on "high-quality" data sources when they are available, and downweighting the contributions of low-quality data sources. This raises the question of how to determine the quality of a data source. Berlin and Rennie (1999) listed qualities of well-designed, high-quality clinical trials. Citro and Straf (2013) and the American Association of Public Opinion Research (2015) gave characteristics of high-quality surveys but did not provide metrics for quantifying survey quality. The development of metrics for the quality of estimates from different data sources—going beyond sampling variability to consider measurement error, nonresponse bias, and stability over time—is a crucial area for research.

Estimates calculated from different sources are often further apart than can be explained by the sampling error of the respective sources. These extra differences are often due to nonsampling errors such as undercoverage, nonresponse, different question wording or modes, and measurement errors, as discussed in Section 4. Making use of estimates from different sources, then, can be used to provide a measure of uncertainty about estimates that includes some of the nonsampling error. Some of the hierarchical models discussed in Section 7 incorporate the estimated bias from sources into the posterior uncertainty about the parameters. These do not always capture all of the potential sources of bias, however, and more research is needed in this area.

Another area for research is the use of multiple sources to improve nonresponse bias assessment and adjustment. The standard practices of calibration and poststratification make use of a single external source, considered to be a gold standard, to adjust weights of respondents so that survey estimates conform to the external population totals. In the absence of a single gold standard, however, it may be possible to use the information from different sources to calibrate survey data. In related work, it may be possible to use multiple sources of data in adaptive design or to assign protocols dynamically.

We discussed linking records among sources that have identifying information. Such linkage raises concerns about the confidentiality of respondents' data. The information contained in a single data source might be insufficient to identify an individual, but the extra variables contained in the linked sources may increase the chances of disclosure. Fellegi (1999), page 6, described record linkage as "intrinsically intrusive of privacy." Daas et al. (2015) discussed concerns about privacy that can arise when using large

nonsurvey data sources for official statistics. Sometimes, privacy concerns can be lessened if aggregated statistics are combined instead of linking individual records, although these methods too can compromise the confidentiality of individuals' or subpopulations' information. Duncan, Jabine and de Wolf (1993) provided guiding principles for balancing the needs of data access with the need to protect the confidentiality of survey respondents' information. Many of the statistical techniques for reducing disclosure risk in that report, however, were conceived in an era in which fewer data and less sophisticated identification techniques were available. More research is needed on privacy-preserving methods for releasing data; the differential privacy framework of Dwork (2011) can provide a mathematical foundation for such work (Machanavajjhala and Kifer, 2015). One possible area for research is on use of hierarchical models to obtain aggregate statistics that protect privacy.

In this article, we have concentrated on statistical methods for combining information. An important factor not discussed here is the issue of obtaining consent from participants to have their information combined with information from other sources. Several of the studies we cited (e.g., Nachman and Parker, 2012; Zolas et al., 2015) linked records across multiple sources. When should consent be obtained from survey respondents for their survey-provided information to be linked with other sources? Even if the information released from the analysis is in the form of aggregated statistics, the linkage creates a database that could potentially be obtained by hackers.

Record linkage and other methods for combining information across sources also raise questions about data ownership. Does a college student own the data about her test scores, class attendance, analytics from online classes, library usage, and cafeteria purchases, or do those belong to the educational institution (Jones, Thomson and Arnold, 2014)? How should society balance patients' ownership of their electronic health records, fitness tracking data, and genetic information with potential benefits that could arise from sharing data (Kish and Topol, 2015; Kostkova et al., 2016)? Hurst (2015) discussed data ownership issues in his testimony to the U.S. House of Representatives Committee on Agriculture, and proposed a "Transparency Evaluator" that would accompany data collection. Farmers providing data would be told who controls their data, who can access them, and how the data will be used, along with other information about the data curation.

Increasingly, rich administrative data sources such as credit card transactions, electronic health records and social media are owned and harnessed by private companies. At the same time, increasing costs, decreasing budgets, and lower cooperation of the public in providing data for federal and state surveys are threatening the federal statistical system. Thus, for the methods reviewed in this paper to be useful, a framework of a private-public partnership will need to be forged to use all available data for the benefit of society.

Many of the probability sampling designs in current use were developed at a time when other sources of information were not available. If these data collections were designed starting over, it is likely that the designs would make use of the wealth of information now available from multiple data sources. The availability of multiple data sources opens multiple opportunities for research on designing the data collection using a systems-based approach; on linking records; on developing imputation, multiple frame, and hierarchical models for combining data; on developing measures of uncertainty that reflect the nonsampling errors from various data sources; and on preserving privacy for individuals who contribute their data. The use of multiple data sources has great potential for capturing more of the population, saving resources by making use of cheaper sources of information, obtaining more information on subpopulations, and improving the temporal and spatial granularity of information used for research and public policy.

ACKNOWLEDGMENTS

The authors thank the reviewers for helpful comments and suggestions for additional references.

REFERENCES

ADES, A. E. and SUTTON, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. *J. Roy. Statist. Soc. Ser. A* **169** 5–35. MR2222010

AMERICAN ASSOCIATION OF PUBLIC OPINION RESEARCH (2015). Code of Professional Ethics and Practices. Available at https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx.

ANDRIDGE, R. R. and LITTLE, R. J. A. (2010). A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* **78** 40–64.

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A., GILE, K. J. and TOURANGEAU, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1** 90–143.

BANCROFT, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Stat.* **15** 190–204. MR0010373

BANKIER, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *J. Amer. Statist. Assoc.* **81** 1074–1079.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey snd satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.

BERLIN, J. A. and RENNIE, D. (1999). Measuring the quality of trials: The quality of quality scales. *J. Amer. Med. Assoc.* **282** 1083–1085.

BHATT, S., WEISS, D. J., CAMERON, E., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., BATTLE, K. E., MOYES, C. L., HENRY, A., ECKHOFF, P. A. et al. (2015). The effect of Malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526** 207–211.

BOHENSKY, M. A., JOLLEY, D., SUNDARARAJAN, V., EVANS, S., PILCHER, D. V., SCOTT, I. and BRAND, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Serv. Res.* **10** 1–7.

BRICK, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *J. Off. Stat.* **29** 329–353.

BRICK, J. M. (2015). Compositional model inference. In *Proceedings of the Survey Research Methods Section* 299–307. Amer. Statist. Assoc., Alexandria, VA.

BRICK, J. M., CERVANTES, I. F., LEE, S. and NORMAN, G. (2011). Nonsampling errors in dual frame telephone surveys. *Surv. Methodol.* **37** 1–12.

CARPENTER, J. and KENWARD, M. (2012). *Multiple Imputation and Its Application*. Wiley, Hoboken, NJ.

CHAUVET, G. and DE MARSAC, G. T. (2014). Estimation methods on multiple sampling frames in two-stage sampling designs. *Surv. Methodol.* **40** 335–346.

CHEN, C., WAKEFIELD, J. and LUMELY, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat. Spatiotemporal Epidemiol.* **11** 33–43.

CHRISTEN, P. (2012). *Data Matching*: *Concepts and Techniques for Record Linkage*, *Entity Resolution*, *and Duplicate Detection*. Springer Science & Business Media, New York.

CITRO, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Surv. Methodol.* **40** 137–161.

CITRO, C. F. and STRAF, M. L., EDS. (2013). *Principles and Practices for a Federal Statistical Agency*, 5th ed. National Academies Press, Washington, DC.

CRUZE, N. (2015). Integrating survey data with auxiliary sources of information to estimate crop yields. In *Proceedings of the Survey Research Methods Section* 565–578. Amer. Statist. Assoc., Alexandria, VA.

DAAS, P. J. H., PUTS, M. J., BUELENS, B. and VAN DEN HURK, P. A. (2015). Big data as a source for official statistics. *J. Off. Stat.* **31** 249–262.

DATTA, G. S., GHOSH, M., STEORTS, R. and MAPLES, J. (2011). Bayesian benchmarking with applications to small area estimation. *TEST* **20** 574–588.

DEMING, W. E. (1950). *Some Theory of Sampling*. Wiley, New York.

DEVILLE, J.-C., SÄRNDAL, C.-E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.* **88** 1013–1020.

DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014a). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **40** 29–46.

DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014b). Combining information from multiple complex surveys. *Surv. Methodol.* **40** 347–354.

DUGOFF, E. H., SCHULER, M. and STUART, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Serv. Res.* **49** 284–303.

DUNCAN, G. T., JABINE, T. B. and DE WOLF, V. A. (1993). *Private Lives and Public Policies*: *Confidentiality and Accessibility of Government Statistics*. National Academies Press, Washington, DC.

DUNCAN, J. W. and SHELTON, W. C. (1992). U.S. Government contributions to probability sampling and statistical analysis. *Statist. Sci.* **7** 320–338. MR1181415

DURRANT, G. B. (2009). Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. *International Journal of Social Research Methodology* **12** 293–304.

DWORK, C. (2011). A firm foundation for private data analysis. *Commun. ACM* **54** 86–95.

ELLIOTT, M. R. and DAVIS, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *J. Roy. Statist. Soc. Ser. C* **54** 595–609. MR2137256

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

FELLEGI, I. P. (1999). Record linkage and public policy: A dynamic evolution. In *Record Linkage Techniques*—1997: *Proceedings of an International Workshop and Exposition* 1–12. National Academy Press, Washington, DC.

FELLEGI, I. P. and SUNTER, A. B. (1969). A theory of record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

FINUCANE, M. M., PACIOREK, C. J., DANAEI, G. and EZZATI, M. (2014). Bayesian estimation of population-level trends in measures of health status. *Statist. Sci.* **29** 18–25. MR3201842

FINUCANE, M. M., PACIOREK, C. J., STEVENS, G. A. and EZZATI, M. (2015). Semiparametric Bayesian density estimation with disparate data sources: A meta-analysis of global childhood undernutrition. *J. Amer. Statist. Assoc.* **110** 889–901. MR3420668

GELMAN, A., KING, G. and LIU, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *J. Amer. Statist. Assoc.* **93** 846–857.

GOLDSTEIN, H., HARRON, K. and WADE, A. (2012). The analysis of record-linked data using multiple imputation with data value priors. *Stat. Med.* **31** 3481–3493. MR3041825

GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data. *J. Roy. Statist. Soc. Ser. A* **168** 267–306. MR2119402

GROVES, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Q.* **70** 646–675.

GROVES, R. M. and HEERINGA, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *J. Roy. Statist. Soc. Ser. A* **169** 439–457. MR2236915

HARRON, K., GOLDSTEIN, H. and DIBBEN, C. (2016). *Methodological Developments in Data Linkage*. Wiley, Hoboken, NJ.

HARTLEY, H. O. (1962). Multiple Frame Surveys. In *Proceedings of the Social Statistics Section*, *American Statistical Association* 203–206. Amer. Statist. Assoc., Alexandria, VA.

HARTLEY, H. O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, *Ser. C* **36** 99–118.

HE, Y., LANDRUM, M. B. and ZASLAVSKY, A. M. (2014). Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: A multiple imputation approach. *Stat. Med.* **33** 3710–3724.

HERZOG, T. N., SCHEUREN, F. J. and WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer Science & Business Media, New York.

HURST, B. (2015). Big Data and Agriculture: Innovations and Implications. Statement of the American Farm Bureau Federation to the House Committee on Agriculture, available at http://agriculture.house.gov/uploadedfiles/10.28.15_hurst_testimony.pdf.

HYNDMAN, R. J., LEE, A. J. and WANG, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Comput. Statist. Data Anal.* **97** 16–32.

JACKSON, C., BEST, N. and RICHARDSON, S. (2008). Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J. Roy. Statist. Soc. Ser. A* **171** 159–178. MR2412651

JONES, K. M., THOMSON, J. C. and ARNOLD, K. (2014). Questions of data ownership on campus. *EDUCAUSE Review*, *August* 1–10.

KALTON, G. and ANDERSON, D. W. (1986). Sampling rare populations. *J. Roy. Statist. Soc. Ser. A* **149** 65–82.

KIM, J. K. and RAO, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika* **99** 85–100.

KISH, L. J. and TOPOL, E. J. (2015). Unpatients—Why patients should own their medical data. *Nat. Biotechnol.* **33** 921–924.

KOHUT, A., KEETER, S., DOHERTY, C., DIMOCK, M. and CHRISTIAN, L. (2012). *Assessing the Representativeness of Public Opinion Surveys*. Pew Research Center, Washington DC. Available at http://www.people-press.org/files/legacy-pdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf.

KORN, E. L. and GRAUBARD, B. I. (1999). *Analysis of Health Surveys*. Wiley, New York.

KOSTKOVA, P., BREWER, H., DE LUSIGNAN, S., FOTTRELL, E., GOLDACRE, B., HART, G., KOCZAN, P., KNIGHT, P., MARSOLIER, C., MCKENDRY, R. A. et al. (2016). Who owns the data? Open data for healthcare. *Frontiers in Public Health* **4** 1–6.

LEE, S. and VALLIANT, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **37** 319–343.

LESSER, V. M., NEWTON, L. and YANG, D. (2008). Evaluating Frames and Modes of Contact in a Study of Individuals with Disabilities. Paper presented at the Joint Statistical Meetings, Denver, Colorado.

LOHR, S. L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Surv. Methodol.* **37** 197–213.

LOHR, S. L. and BRICK, J. M. (2012). Blending domain estimates from two victimization surveys with possible bias. *Canad. J. Statist.* **40** 679–696. MR2998856

LOHR, S. L. and BRICK, J. M. (2014). Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology* **2** 388–409.

LOHR, S. L. and RAO, J. N. K. (2006). Estimation in multiple-frame surveys. *J. Amer. Statist. Assoc.* **101** 1019–1030. MR2324141

MACHANAVAJJHALA, A. and KIFER, D. (2015). Designing statistical privacy for your data. *Commun. ACM* **58** 58–67.

MANZI, G., SPIEGELHALTER, D. J., TURNER, R. M., FLOWERS, J. and THOMPSON, S. G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *J. Roy. Statist. Soc. Ser. A* **174** 31–50. MR2758280

MECATTI, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Surv. Methodol.* **33** 151–157.

MERCER, L., WAKEFIELD, J., CHEN, C. and LUMLEY, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spat. Stat.* **8** 69–85. MR3326822

MERKOURIS, T. (2004). Combining independent regression estimators from multiple surveys. *J. Amer. Statist. Assoc.* **99** 1131–1139. MR2109501

MERKOURIS, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 27–48. MR2751242

METCALF, P. and SCOTT, A. (2009). Using multiple frames in health surveys. *Stat. Med.* **28** 1512–1523. MR2649709

MORIARITY, C. and SCHEUREN, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *J. Off. Stat.* **17** 407–422.

MOSTELLER, F. (1948). On pooling data. *J. Amer. Statist. Assoc.* **43** 231–242.

NACHMAN, K. E. and PARKER, J. D. (2012). Exposures to fine particulate air pollution and respiratory outcomes in adults using two national datasets: A cross-sectional study. *Environ. Health* **11** 1–12.

NANDRAM, B., BERG, E. and BARBOZA, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environ. Ecol. Stat.* **21** 507–530. MR3248537

NATIONAL CENTER FOR HEALTH STATISTICS (2016). Survey Description, National Health Interview Survey, 2014. Centers for Disease Control and Prevention, Hyattsville, MD. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2015/srvydesc.pdf.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97** 558–625. MR0121942

PFEFFERMANN, D. and TILLER, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *J. Amer. Statist. Assoc.* **101** 1387–1397. MR2307572

POCOCK, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic. Dis.* **29** 175–188.

PRENTICE, R. L., SMYTHE, R. T., KREWSKI, D. and MASON, M. (1992). On the use of historical control data to estimate dose response trends in quantal bioassay. *Biometrics* **48** 459–478. MR1173492

RAGHUNATHAN, T. E. (1991). Pooling controls from different studies. *Stat. Med.* **10** 1417–1426.

RAGHUNATHAN, T. E. (2006). Combining information from multiple surveys for assessing health disparities. *Allg. Stat. Arch.* **90** 515–526.

RAGHUNATHAN, T. E., XIE, D., SCHENKER, N., PARSONS, V. L., DAVIS, W. W., DODD, K. W. and FEUER, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *J. Amer. Statist. Assoc.* **102** 474–486. MR2370848

RANALLI, M. G., ARCOS, A., RUEDA, M. D. M. and TEODORO, A. (2016). Calibration estimation in dual-frame surveys. *Stat. Methods Appl.* **25** 321–349. MR3539496

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken, NJ. MR3380626

RAO, J. N. K. and WU, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *J. Amer. Statist. Assoc.* **105** 1494–1503. MR2796566

RAO, S. R., GRAUBARD, B. I., SCHMID, C. H., MORTON, S. C., LOUIS, T. A., ZASLAVSKY, A. M. and FINKELSTEIN, D. M. (2008). Meta-analysis of survey data: Application to health services research. *Health Serv. Outcomes Res. Methodol.* **8** 98–114.

RÄSSLER, S. (2002). *Statistical Matching*: *A Frequentist Theory*, *Practical Applications*, *and Alternative Bayesian Approaches*. *Lecture Notes in Statistics* **168**. Springer, New York. MR1996879

RENSSEN, R. H. and NIEUWENBROEK, N. J. (1997). Aligning estimates for common variables in two or more sample surveys. *J. Amer. Statist. Assoc.* **92** 368–374. MR1436124

RODGERS, W. L. (1984). An evaluation of statistical matching. *J. Bus. Econom. Statist.* **2** 91–102.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

SÄRNDAL, C.-E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.* **33** 99–119.

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. *Stat. Med.* **29** 533–545. MR2758451

SKINNER, C. J. and RAO, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *J. Amer. Statist. Assoc.* **91** 349–356. MR1394091

SMITH, T. W. (2011). The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. *Int. J. Public Opin. Res.* **23** 389–402.

STATISTICS CANADA (2014). Note to Users of Data from the 2012 Canadian Income Survey, available at http://www.statcan.gc.ca/pub/75-513-x/75-513-x2014001-eng.htm.

STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2016). A Bayesian approach to graphical record linkage and de-duplication. *J. Amer. Statist. Assoc.* **111** 1660–1672. MR3601725

STOKES, L. and LIN, D. (2015). Measurement error in dual frame designs. Paper presented at the Joint Statistical Meetings, Seattle WA.

STRAUSS, W. J., CARROLL, R. J., BORTNICK, S. M., MENKEDICK, J. R. and SCHULTZ, B. D. (2001). Combining datasets to predict the effects of regulation of environmental lead exposure in housing stock. *Biometrics* **57** 203–210. MR1833308

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

SUTTON, A. J. and HIGGINS, J. (2008). Recent developments in meta-analysis. *Stat. Med.* **27** 625–650. MR2418504

SWEETING, M. J., DE ANGELIS, D., HICKMAN, M. and ADES, A. E. (2008). Estimating hepatitis C prevalence in England and Wales by synthesizing evidence from multiple data sources. Assessing data conflict and model fit. *Biostatistics* **9** 715–734.

TOURANGEAU, R., BRICK, J. M., LOHR, S. and LI, J. (2017). Adaptive and responsive survey designs: A review and assessment. *J. Roy. Statist. Soc. Ser. A.* **180** 203–223. MR3600507

TURNER, R. M., OMAR, R. Z., YANG, M., GOLDSTEIN, H. and THOMPSON, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat. Med.* **19** 3417–3432.

TURNER, R. M., SPIEGELHALTER, D. J., SMITH, G. C. S. and THOMPSON, S. G. (2009). Bias modelling in evidence synthesis. *J. Roy. Statist. Soc. Ser. A* **172** 21–47. MR2655603

UNITED STATES CENSUS BUREAU (2016). Model-Based Small Area Income & Poverty Estimates (SAIPE) for School Districts, Counties, and States. Available at http://www.census.gov/did/www/saipe/.

UNITED STATES GENERAL ACCOUNTING OFFICE (1992). Cross-Design Synthesis: A New Strategy for Medical Effectiveness Research. U.S. General Accounting Office, Washington, DC. Available at archive.gao.gov/d31t10/145906.pdf.

VALLIANT, R. and DEVER, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.* **40** 105–137. MR2758301

VALLIANT, R., DORFMAN, A. H. and ROYALL, R. M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. Wiley, New York.

VOS, T., BARBER, R. M., BELL, B., BERTOZZI-VILLA, A., BIRYUKOV, S., BOLLIGER, I., CHARLSON, F., DAVIS, A., DEGENHARDT, L., DICKER, D. et al. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386** 743–800.

WAGNER, J. and RAGHUNATHAN, T. (2007). Bayesian approaches to sequential selection of survey design protocols. In *Proceedings of the Survey Research Methods Section* 3333–3340. Amer. Statist. Assoc., Alexandria, VA.

WAGNER, J., WEST, B. T., KIRGIS, N., LEPKOWSKI, J. M., AXINN, W. G. and NDIAYE, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *J. Off. Stat.* **28** 477.

WAKEFIELD, J. (2004). Ecological inference for 2×2 tables (with discussion). *J. Roy. Statist. Soc. Ser. A* **167** 385–445. MR2082057

WAKEFIELD, J. and SALWAY, R. (2001). A statistical framework for ecological and aggregate studies. *J. Roy. Statist. Soc. Ser. A* **164** 119–137. MR1819026

WANG, J. C., HOLAN, S. H., NANDRAM, B., BARBOZA, W., TOTO, C. and ANDERSON, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *J. Agric. Biol. Environ. Stat.* **17** 84–106.

WANG, H., WOLOCK, T. M., CARTER, A., NGUYEN, G., KYU, H. H., GAKIDOU, E., HAY, S. I., MILLS, E. J., TRICKEY, A., MSEMBURI, W. et al. (2016). Estimates of

global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2015: The Global Burden of Disease Study 2015. *The Lancet*. *HIV* **3** e361–e387.

WHELDON, M. C., RAFTERY, A. E., CLARK, S. J. and GERLAND, P. (2016). Bayesian population reconstruction of female populations for less developed and more developed countries. *Popul*. *Stud*. (*Camb*.) **70** 21–37.

WINKLER, W. E. (2014). Matching and record linkage. *Wiley Interdiscip*. *Rev*.: *Comput*. *Stat*. **6** 313–325.

YBARRA, L. M. and LOHR, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika* **95** 919–931. MR2461220

YEAGER, D. S., KROSNICK, J. A., CHANG, L., JAVITZ, H. S., LEVENDUSKY, M. S., SIMPSER, A. and WANG, R. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opin*. *Q*. **75** 709–747.

YOU, J., DATTA, G. S. and MAPLES, J. J. (2014). Modeling disability in small areas: An area-level approach of combining two surveys. In *Proceedings of the Survey Research Methods Section* 3770–3784. Amer. Statist. Assoc., Alexandria, VA.

ZHOU, H., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2015). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics* **72** 242–252. MR3500593

ZOLAS, N., GOLDSCHLAG, N., JARMIN, R., STEPHAN, P., OWEN-SMITH, J., ROSEN, R. F., ALLEN, B. M., WEINBERG, B. A. and LANE, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science* **350** 1367–1371.