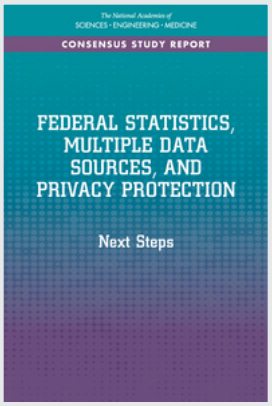


# Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps



## DETAILS

194 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-46537-3 | DOI 10.17226/24893

## CONTRIBUTORS

Robert M. Groves and Brian A. Harris-Kojetin, Editors; Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

# 2

# Statistical Methods for Combining Multiple Data Sources

In the panel's first report, we described the multiple types of additional data sources—federal and state administrative data, electronic health records, web scrapings, credit card transactions, satellite images, and sensor data, among others—that might be used to improve the level of detail, timeliness, and cost of federal statistics. Federal statistical agencies have long used administrative data to improve the efficiency of the design of probability surveys and to adjust for nonresponse, and we noted how a number of agencies are currently investigating nonsurvey data sources to supplement or replace data from probability surveys. These investigations share common features, in which the information from the different sources needs to be evaluated and combined.

In this chapter, we review statistical methods for combining information, identify research needs, and propose steps that can be taken to facilitate a new paradigm for producing federal statistics. As noted in Chapter 1, that paradigm would shift from sole reliance on probability surveys to a system that relies on probability surveys along with administrative and private-sector data, making use of the strengths of each data source. We begin by describing statistics that are currently produced or might be desired and summarizing some of the features of data sources that might be combined to produce those statistics. We next summarize the statistical methods that have been proposed for combining information, where the choice of method depends on the statistical purpose, the nature of the available data, and privacy and other considerations. When individual records from multiple datasets for each person or entity are available, they can sometimes be linked through statistical models. When aggregate

statistics are available or linkage cannot be done, multiple frame methods or modeling can be used. We also outline research that is needed in the area of statistical methodology and describe a framework for promoting the development of methods for combining data sources. Citro (2014) and Lohr and Raghunathan (2017) provide more detailed discussions of statistical methods and possible research directions.

## DEMANDS FOR MORE GRANULAR STATISTICS

The usefulness of a data source for federal statistics depends on the type of information that is desired, which includes

- Information for the United States as a whole: What is the national unemployment rate? How many people were victimized by violent crime in 2016? How many people have diabetes in the United States, and what are the associated health care costs?
- Information for regions or states: How many children are eligible to receive assistance from the Supplemental Nutrition Assistance Program (SNAP) in Arkansas? What is the forecast yield of the winter wheat crop in Kansas?
- Information for local jurisdictions: What is the violent crime victimization rate in Chicago? What is it in Fresno? What percentage of 4th-grade students in the Chicago Public Schools is at or above the "proficient" level in mathematics? What effect did Hurricane Katrina have on poverty in New Orleans?
- Information for demographic groups or other subpopulations: What is the 2016 unemployment rate among adults without a high school degree? What percentage of people ages 65 and older worked full time in January 2017? What is the job creation rate among businesses that are less than 5 years old? How many business establishments in Hawaii with four or fewer employees closed in 2016? Information may be desired for race or ethnicity groups, men or women, and specific age or education groups. Information may be desired for cross-classifications of demographic and geographic subpopulations.

Large-scale probability sample surveys have long been the foundation for producing many national statistics for the United States. Probability surveys can be designed to measure the specific concepts of interest, but they are expensive, particularly those conducted through face-to-face interviews. As discussed in the panel's first report, both costs and nonresponse rates for probability surveys have increased in recent years.

Policy makers and data users are demanding ever-increasing granular-

ity for statistics, wanting more geographic detail, more frequent releases of statistics, and more information about subpopulations. Some probability surveys have been designed and others modified to allow for the release of more detailed or frequent statistics. Before the American Community Survey (ACS) was launched in 2005, detailed geographic-level information on poverty, disability, employment, family relationships, and other characteristics of the U.S. population was available only from the "long form" of the decennial census at 10-year intervals. The ACS produces direct annual estimates[1] for areas with populations of at least 65,000, and estimates based on the past 5 years of data collection for areas with populations of at least 7,000. These estimates produce 11 billion statistics each year (see Hedrick and Weister, 2016).

Similarly, the Current Population Survey (CPS) publishes monthly estimates of national unemployment and labor participation rates, with separate statistics given for subpopulations that include cross-classifications by race and ethnicity, sex, and age (for an example, see Bureau of Labor Statistics, 2017b). However, labor force estimates for subpopulations of smaller geographic regions—census regions and divisions, states, metropolitan areas, and principal cities—are produced annually by aggregating the monthly surveys (see Bureau of Labor Statistics, 2015).

For the ACS and CPS, as for other probability samples, there is a tradeoff between geographic or subpopulation detail and timeliness.[2] A direct estimate of a subpopulation characteristic from the survey requires a sample size for the subpopulation that is large enough for the statistic to be reliable. In order to do so, data must be accumulated either over time

[1]A direct estimate is one that is produced using the data from the survey. Other data sources may be used to calibrate the weights of the survey for undercoverage and nonresponse, but the data about the characteristic being estimated come from the survey.

[2]Tradeoffs are also made in the surveys' design to enable production of state-level estimates. To produce state-level statistics, the CPS takes a sample of households from every state, with sample sizes ranging from 500 to 4,600 households (Bureau of Labor Statistics, 2012). Some states have a higher share of the sample than their share of the adult population (people ages 16 and older) and this oversampling of smaller states allows the CPS to produce reliable state-level estimates, but it makes the design less efficient for producing national estimates because adults in large states are less likely to be included in the sample than adults in small states.

The design of the National Crime Victimization Survey (NCVS) has also been modified to enable production of selected state-level estimates. The original survey design, tailored to produce national estimates of victimization, gave every household in the United States roughly an equal chance of being selected for the survey, but this design could result in some states having no one in the sample. In response to an increasing need for crime statistics at the state and local level, the Bureau of Justice Statistics redesigned the survey to produce direct estimates of victimization for 22 states (Planty and Langton, 2014; Langton and Fay, 2016; Bureau of Justice Statistics, 2016). This was done by augmenting the sample size in those states as needed to produce 3-year rolling-average estimates of victimization with the desired precision.

(through repeated data collections) or over space (collapsing across different geographic areas).

The designs of the ACS and CPS have been formed or modified so that they can produce direct estimates at more frequent intervals or at finer levels of geography, but the high costs of data collection limit their sample sizes in small geographic areas. By combining these surveys with other data sources, it may be possible to produce reliable estimates for even smaller subpopulations or with greater frequency. In addition, other data sources may measure variables not found in a survey, which may give a richer picture of the relationship between, say, poverty and health outcomes. Administrative and private-sector data sources already exist and the cost to use them for statistical purposes may be lower than the cost to collect additional data from probability surveys.

Nonsurvey data sources can also provide a fresh perspective on the redesign of federal surveys. In some cases, questions can be eliminated from a survey if equivalent and reliable measurements are available from another data source. Nonsurvey data sources can also be used to construct or refine the sampling frame used to draw the samples to improve efficiency and reduce respondent burden. The measurements available in nonsurvey data sources may also be useful in determining the most efficient mode for data collection. Thus, nonsurvey data sources are not just useful for estimation purposes but also for the possible redesign of many federal surveys.

Yet nonsurvey data sources have their own problems (see Chapter 6 for a fuller disussion of quality issues). Administrative data, such as tax records, are collected for a specific purpose that may not match the needs for the statistics being produced. For example, the tax entity represented in the records could be a large business enterprise with multiple locations, which could be different from the statistical entity of interest: a single business establishment location. The administrative data often have limited variables, and these do not necessarily measure the characteristics of interest. Data sources may be missing important parts of the population: for example, electronic medical records may be less likely to contain information about people who do not have health insurance or people who have not recently used any medical services. Data sources such as social media may be vulnerable to external manipulation through "bots" or organized campaigns. The quality of the responses given in data sources may be unknown, and protocols for data collection may change without notice or documentation. Finally, to be useful, an alternative data source must have continued accessibility and availability for federal statistical purposes. Despite these shortcomings, it would be valuable to investigate and implement strategies to combine information from survey and nonsurvey data sources to improve efficiency and meet the ever-growing need for more information.

Ultimately, a framework is needed for combining different data sources that draws on the strengths and counterbalances the weaknesses of each source, resulting in more useful information, or lower costs, than what would be achievable from a single source. For example, Horrigan (2013a, 2013b) describes data sources that the Bureau of Labor Statistics (BLS) uses when producing the Consumer Price Index (CPI) and the Producer Price Index (PPI), which include the following:

- data from the Billion Prices Project (Cavallo and Rigobon, 2016),
- retail scanner data,
- information on used cars from J.D. Power and Associates,
- stock exchange bid and ask prices and trading volume data,
- data on hospitals from the American Hospital Association,
- diagnosis codes from the Agency for Healthcare Research and Quality,
- administrative data on crude petroleum from the Energy Information Administration,
- administrative data on baggage fees from the U.S. Department of Transportation,
- SABRE data on airline pricing, and
- Medicare Part B reimbursement information.

The economic concepts for the CPI and PPI provide the framework for integrating different data sources, and BLS can create more accurate and cost-effective indexes by relying on multiple sources rather than on a single source. In a similar vein, the Medical Expenditure Panel Survey incorporates data from multiple survey and administrative records sources as part of its design (see Box 2-1).

Combining survey data with other data sources, or combining multiple administrative data sources, has many potential advantages over the survey paradigm. A number of recent studies have identified information domains that would benefit from drawing on alternative data sources to provide key statistics beyond what is possible or practical through a federal survey. For example, one study (National Research Council, 2014a) recommended that the National Center for Science and Engineering Statistics (NCSES) engage in a program of research to explore and experiment with a variety of existing alternative datasets quickly and inexpensively to understand aspects of innovation in science and engineering. Similarly, another study that considered measuring social and civic engagement and social cohesion (National Research Council, 2014b) concluded that only a limited number of variables can be included on national surveys and that combining survey data with other sources can provide useful explanatory variables and

**BOX 2-1**
**Use of Multiple Data Sources in the Medical Expenditure Panel Survey**

The Medical Expenditure Panel Survey (MEPS) is sponsored by the Agency for Healthcare Research and Quality (AHRQ). It is designed to give accurate and reliable information about the U.S. population's health care coverage, utilization, expenditures, and access to care. Created in 1996, MEPS is designed by a combination of three different interrelated surveys: the household component (MEPS-HC), the medical provider component (MEPS-MPC), and the insurance component (MEPS-IC). MEPS is cosponsored by the National Center for Health Statistics (NCHS) and uses Westat, Research Triangle Institute (RTI) International, and the U.S. Census Bureau as main data collection organizations. MEPS provides a model for combining data sources, combining information across person, household, and provider level, and using information from parts of one component as a source of information for other components.

The MEPS-HC is designed by selecting a subsample of households from the National Health Interview Survey (NHIS) conducted by NCHS. During the survey, information is collected about health conditions, health status, demographic characteristics, employment, and income, in addition to information regarding health insurance coverage, access to care, and changes and source of payment.

Following the MEPS-HC, the MEPS-MPC is used to collect information from providers for the individuals who provided responses in the MEPS-HC (see Cohen and Cohen, 2013). The information is collected from health care providers, including physicians, hospitals, health agencies, and pharmacies, and includes dates of visits, charges, and medical care services.

These two components are then linked using statistical probabilistic matching procedures. Ideally, these two components should match as they contain information about the same individual, but sometimes there are inconsistencies in the reporting of the same medical events between MEPS-HC and MEPS-MPC. When there is an inconsistency, the MEPS-MPC information is preferred because providers' data are generally considered superior in accuracy to household responses (Cohen et al., 2009). This linked dataset is then used as the primary source of information regarding expenditure estimation.

Finally, the MEPS-IC obtains information from a sample of private- and public-sector employers on the health insurance plans they offer their employees, including health insurance plans offered, premiums, contributions by employer and employees, and employer characteristics. The purpose of this component is to better understand what health insurance is available on both a national and state level; these data are not linked with data from the MEPS-HC.

greater geographic detail needed for research on social capital, social cohesion, and civic engagement.

Citro (2014, p. 152) summarized the advantages of using multiple data sources to produce official statistics, listing eight ways in which administra-

tive data sources could be used to improve the quality of household survey data:

1. Assist in the evaluation of survey data quality by using comparisons with aggregate estimates, appropriately adjusted for differences in population universes and concepts, and by exact matches of survey and administrative records.
2. Provide control totals for adjusting survey weights for coverage errors.
3. Provide supplemental sampling frames for use in a multiple frame design.
4. Provide additional information to append to matched survey records to enhance the relevance and usefulness of the data.
5. Provide covariates for model-based estimates for smaller geographic areas than what the survey can support directly.
6. Improve models for imputations for missing data in survey records.
7. Replace "no" for survey respondents who should have reported an item, replace "yes" for survey respondents who should not have reported an item, and replace reported values for survey respondents who misreport an item.
8. Replace survey questions and use administrative records values directly.

Her arguments can be extended to nonfederal and non-administrative data sources as well. Most household surveys currently use methods 1 and 2, and some surveys use or are exploring methods 3 through 8 to make more efficient use of data from other sources.

**CONCLUSION 2-1 New data sources have emerged during the last few years, providing opportunities to develop a new paradigm for statistical design and analysis systems that can improve timeliness, geographic or subpopulation detail, statistical efficiency, and reduce costs of producing federal statistics.**

**RECOMMENDATION 2-1 Multiple data sources should be used to redesign current data collection efforts and estimation tasks to improve the utility, timeliness, and cost-efficiency of federal statistics.**

In the panel's first report we noted several examples of statistical agencies that are currently making efforts along these lines (see National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 2), but more research is needed to understand these new approaches and to evaluate specific sources for use in particular applications. We recognize that alter-

ing major federal surveys by combining data sources (such as administrative data and survey data) requires substantial work both in planning and research and in the design phase. Agencies should be careful and deliberative in implementing changes based on this research, to understand the implications of substituting an administrative data source for particular survey data.

In some cases, items currently collected in a survey could be available from an administrative source. The Census Bureau has been exploring the usefulness of tax information from the Internal Revenue Service to replace the income questions in the American Community Survey (see O'Hara, 2016). In other cases, it may be possible to considerably redesign or even discontinue a survey based on the possibilities of obtaining and using administrative data and data from other sources. The National Center for Health Statistics was able to replace the National Nursing Home Survey and the National Home and Hospice Survey beginning in 2012 with administrative data from the Centers for Medicare & Medicaid Services. In yet other situations, it may be possible to combine information from administrative data sources with information from surveys. The remainder of this chapter summarizes statistical methods that can be used for combining data from different sources.

## STATISTICAL METHODS FOR COMBINING DATA

### Record Linkage

Record linkage refers to any method by which records from different data sources that are thought to belong to the same entity are associated, and records that are thought to belong to different entities are distinguished. Linking variables are variables that are used to match or distinguish records from different sources, most commonly: Social Security number (SSN), name, address, date of birth, age, race, sex, family relationships, and use of social services. However, almost any variable that is present on two (or more) data sources can be used as a linkage variable.

Record linkage methods are typically classified as either deterministic or probabilistic, and these methods are described briefly in Box 2-2 and in greater detail in Herzog et al. (2007), Christen (2012), and Harron et al. (2015). In the remainder of this section, we provide examples for which linkage is or could be used and discuss potential problems with using record linkage techniques.

Record linkage can increase the number of variables for records in a survey or administrative data source. For example, the National Center for Health Statistics (NCHS) routinely links the data from the National Health Interview Survey (NHIS) to records from the Social Security Administra-

tion, the Centers for Medicare & Medicaid Services, and the National Death Index (see Box 2-3); this linkage allows researchers to investigate the relationship between health and sociodemographic information reported in the surveys and medical care costs, future use of medical services, mortality, and other variables found in the administrative data sources (National Center for Health Statistics, 2012).

Record linkage can also decrease the time needed to conduct a survey and increase the amount of information obtained for analysis by obtaining information from other sources instead of asking the survey respondent. When faced with a long questionnaire or interview, respondents may stop answering questions before finishing a survey. Cynamon and Blumberg (2016) reported that for every year since 2007, 20 to 30 percent of NHIS interviews have had incomplete data. A shorter survey reduces the burden on the respondents and can also result in fewer uncompleted interviews.

Record linkage can serve to augment the number of records available for study. Ramaprasan (2015) linked records from the tumor registry of Group Health Cooperative, a health insurance company, with records from the Washington State Cancer Registry. The record linkage enabled the researchers to identify and remove duplicated records from the concatenated databases, adding 35,166 new tumor cases from the registry to the Group Health Cooperative database.

Record linkage can validate responses to a survey, fill in values for missing data, or replace survey items. A Housing and Urban Development pilot project (described in the panel's first report, National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 3) linked American Housing Survey records with tax assessment information. Bucholtz (2015) explored whether tax assessment data could substitute for a respondent's missing data about housing characteristics or replace erroneous information. He also suggested that tax data might be considered for replacing some survey items entirely.

The assessment or improvement of sampling frames is also possible through record linkage. The National Teacher and Principal Survey (NTPS) collects data on teacher and principal preparation, the demographic characteristics of teachers and principals, school characteristics, and other information on elementary and secondary education. The sample of public schools is drawn from the Common Core of Data, which is the U.S. Department of Education's annually updated database on public elementary and secondary schools, and each sampled school is asked to provide a listing of teachers. Brummet et al. (2014) explored using commercial school and teacher lists as an alternative sampling frame for teachers in the NTPS, as these lists could avoid the costs of obtaining the teacher listings from each school. Lists from three vendors were linked to the sampling frame for the Schools and Staffing Survey (the predecessor of the NTPS) to evaluate the

**BOX 2-2**
**Record Linkage Methods**

In deterministic record linkage (DRL), a set of linking variables is specified and records must agree on all of the linking variables in that set to be considered a match. The simplest way to use DRL is to have a single linkage variable, such as a Social Security number. Often, however, a single identifying variable is not available, and multiple variables are used for linkage. For example, two records might be linked if they agree exactly on name, ZIP code, and date of birth; otherwise, they are not linked. Some DRL systems have complex sets of rules specifying that records are linked if, say, they agree on at least four of the six linkage variables. If "nearly exact" matches are allowed on linking variables, rules are needed for specifying how close the variables need to be.

If the linkage variables have no errors or missing values and uniquely identify entities in the population, then DRL works well. But few data sources are without errors, even when there are unique identifiers in the population, and there may be missing values or typographical errors. DRL methods that require exact matches often have missed links.

In practice, every method of linking records is subject to missed or false links. This situation led to probabilistic record linkage (PRL), sometimes called fuzzy matching, in which a quantification is sought for the errors in linking records.

In PRL, an algorithm evaluates the similarity in the linkage variables among records from different sources. Many PRL methods are based on the work of Newcombe et al. (1959) and Fellegi and Sunter (1969). In a simple form of PRL, suppose that there are two data sources, *A* and *B*, and that the linking variables are name, marital status, and date of birth. For each pair of records considered as a potential match, one from source *A* and one from source *B*, the *agreement pattern* is determined for the linking variables. If the records have the same name, marital status, and date of birth, the agreement pattern is (Y, Y, Y); if they have the same name and marital status but different dates of birth, the agreement pattern is (Y, Y, N), and so on. Then two probabilities are calculated: the probability that

lists' coverage of the school and teacher population. Brummet et al. (2014) recommended continuing with the current sampling frame because of its greater coverage but also continuing to investigate the vendor lists for possible future use as sampling frames.

At the same time, record linkage methods come with concerns. Linked records have more information about individuals than the original data sources, which raises privacy concerns. Fellegi (1999, p. 5) noted that record linkage is "intrinsically privacy intrusive, in the sense that information is brought together about a person without his or her knowledge and control." Although records do not need to be physically joined at the same location in order to be linked (see Chapter 3), and encryption can be used

two records would have this agreement pattern if they are a true match, and the probability that two records would have this agreement pattern by chance if they are distinct entities. The ratio (*R)* of these two probabilities is

$$R = \frac{\text{probablity that two records have this agreement pattern if they are a true match}}{\text{probablity that two records have this agreement pattern if they are distinct entities}}.$$

If the pair of records is truly a match and the agreement pattern is (Y, Y, Y), then *R* is expected to be large; that is, the probability is higher that the linking variables agree for two records if they are from the same entity than if they are from two distinct entities. Conversely, if the pair of records are from distinct entities and the agreement pattern is (N, N, N), *R* is expected to be small. PRL uses a decision rule with two cutoff values, $C_U$ and $C_L$, where the pair is deemed to be a match if $R \geq C_U$, the pair is deemed to be from distinct entities if $R \leq C_U$, and further review is needed if *R* is between $C_U$ and $C_L$. The probabilities can be estimated from existing datasets in which the matching status of records is known, or they can be estimated from the data sources of interest as processing is done, with early decisions used to improve the accuracy of matching for later record pairs. Many variations are possible, and the probabilities can depend on the values of the linking variables as well as on the simple agreement/disagreement: it may be desirable to have a higher probability of agreement for nonmatching pairs for a common name, such as Jones, than for a less common name, such as Hoogland.

The probabilities evaluated in the Fellegi-Sunter (1969) method are *not* the probabilities that two records are a true match: they are probabilities that records have a specified agreement pattern if they are a true match (or a true nonmatch). In a Bayesian formulation of the problem (see, e.g., Belin and Rubin, 1995; Tancredi and Liseo, 2011; Steorts et al., 2016), a different probability is calculated: the probability that two records are a true match given that they have a specific agreement pattern. This is a more intuitive formulation of the probability of interest, but calculating this probability from the available data can be challenging.

in the linkage process (see, e.g., Schmidlin et al., 2015), record linkage may represent increased privacy risks to entities in the linked data sources. This issue, and the issue of obtaining consent for record linkage, is discussed further in Chapter 4.

It is often difficult to do record linkage well, particularly when good linkage variables are not available. Linkages can have errors, which can affect conclusions of analyses (see Chapter 6). If records that belong to different entities are mistakenly linked, or if records belonging to the same entity are not linked, then relationships among the variables from different datasets can be distorted.

As the panel described in our first report, the Center for Administra-

**BOX 2-3**
**Linking Records from the National Health Interview Survey: Case Study**

The National Health Interview Survey (NHIS) is the principal source of information on the health of the civilian noninstitutionalized population of the United States. It collects data through in-person interviews from a representative sample of households, adults, and children that covers information on topics, such as health status, medical conditions, health insurance coverage, health care access and utilization, and health behaviors. The 2006–2015 sample design is described in Parsons et al. (2015).

At the end of the interview for the 2010–2013 NHIS, adult respondents were asked a question about the use of their information (Weissman et al., 2016, p. 3):

> To help us link your survey data with vital statistics and health-related records of other government agencies, we would like the last four digits of your Social Security Number. The National Center for Health Statistics uses this information for research purposes only. Providing this information is voluntary. Federal laws authorize us to ask for this information and require us to keep it strictly private. There will be no effect on your benefits if you do not provide this information. What are the last four digits of your Social Security Number?

Respondents eligible for Medicare were also asked for the last four digits of their Medicare Health Insurance Claim number. A survey respondent who did not provide his or her Social Security number (SSN) or Medicare number was then asked if the agency would be allowed to try to link the survey data without the number. A respondent was considered to have consented to the linkage if he or she either provided the SSN or Medicare number or gave permission to link the survey data without it, and 10 to 12 percent of survey respondents refused to allow linkage (Weissman et al., 2016). This result from the 2010–2013 NHIS can be compared with what occurred in the mid-2000s, when respondents were asked to provide all nine digits of their SSN, and approximately 50 percent of respondents did not consent to linkage (Zhang et al., 2016).[a]

The NHIS *Field Representative Manual* provides guidance to interviewers for how to respond to frequently asked questions from survey participants. If a participant asks about why the SSN is needed, the interviewer can respond by outlining some of the uses and benefits of linking records in some detail or providing a less detailed explanation (U.S. Census Bureau, 2017, pp. F-54-F-55):

> NCHS currently links various records from NHIS with death certificate records from the National Death Index (NDI), Medicare enrollment and claims records collected from the Centers for Medicare and Medicaid Services (CMS), and the Old-Age, Survivors, and Disability Insurance (OASDI) and Supplemental Security Income (SSI) benefit records collected from the Social Security Administration (SSA). Files containing the personally identifying information are sent from NHIS to these federal agencies. Personally identifying information used in linkage includes name, date of birth, Social Security Number and/or Medicare number, race, sex, state of birth, and state of residence. If an agency is able to find a survey participant in its own data files, information can be sent back to NHIS and linked with the original survey data. These files contain-

ing detailed health survey data plus information on costs, mortality, or benefits can be used for more complex research, without having to follow up directly with participants.

Alternatively:

We know that this is a long interview and we don't wish to keep you tied up answering more questions. By having your name and Social Security Number or Medicare number, we can combine these health data with other information from Social Security, Medicare and Medicaid, and death records. These records have information about medical conditions and care, and how much they cost. We can join this information to the information that we get during an interview. This allows us to do more complex types of health research without having to come back or ask you more questions. (p. F-56)

The manual says that the interviewer then can give some examples of research that has been done using the linked data: "Predicting the number of disabled persons in the U.S. based on health conditions reported in the NHIS," "Predicting the costs of Medicare based on health conditions reported in the NHIS," or "Studying the health characteristics of people who retire early" (U.S. Census Bureau, 2017, p. F-56).

Golden et al. (2015) described the procedure used to link the 1994–2005 NHIS survey records with administrative records. Because respondents were asked for their SSNs, they were used as the primary variable for matching. When the SSN could not be verified, a probabilistic linkage procedure was used with other information found in both sources.[b]

Because the survey asked for respondents' SSNs, the linkage rates for respondents who consented to linkage were high. However, care must be used when analyzing linked records because people who consent to linkage may differ in some ways that are unknown from those who refuse to consent. Weissman et al. (2016) reported that NHIS respondents with heart disease, stroke, cancer, hypertension, diabetes, chronic obstructive pulmonary disease, or serious psychological distress were significantly more likely to consent to linkage than respondents without those conditions.

A variety of linked data files have been constructed.[c] Many of the linked files are restricted use and may be accessed only at a Research Data Center.[d] Data users are required to abide by the same rules concerning disclosure of confidential information as agency employees. However, to assist researchers in estimating their maximum available sample for analysis, feasibility files containing a limited set of variables are publicly available. The feasibility files contain information about a survey participant's eligibility for linkage and whether a participant was successfully linked to an administrative data source, but do not contain any information about benefits or payments. Public-use linked mortality files containing a limited set of mortality variables for adult survey participants are also available for download from the NCHS Data Linkage website.[e]

Many researchers have used the linked data sources to investigate mortality and health care costs for NHIS respondents. The linked mortality data have been used to investigate the relationship between mortality and strength training, depression, body mass index, smoking, diabetes, alcohol consumption, and height.[f]

Other researchers have used linkages with other datasets. Miller et al. (2016)

*continued*

**BOX 2-3 Continued**

used the linked NHIS/Medicare data to explore differences in health characteristics between people who enrolled in Medicare fee-for-service plans and those who enrolled in Medicare Advantage plans. Gorina et al. (2015) studied hospitalization, readmission, and death rates among Medicare fee-for-service enrollees using linkages among the NHIS, Medicare, and National Death Index files.

Mortality estimates and other research, however, need to account for potential differences in linkage rates: Lariscy (2011) found that the linkage quality for the 1998–2000 linked files was greater for non-Hispanic white adults and adults born in the United States than for Hispanic and foreign-born adults. Failure to account for different linkage error rates might result in too-low estimates of mortality because the matching records in the National Death Index were not found (see Miller et al., 2017).

---

[a]Please note that the methods for obtaining permission to link also changed.

[b]The text describes the linking procedures used with previous NCHS datasets. The linking methodology has been revised, and a publication describing the revised methodology is forthcoming; see https://www.cdc.gov/nchs/data-linkage/medicare-methods.htm [June 2017].

[c]For example, see https://www.cdc.gov/nchs/data/datalinkage/linkagetable.pdf [June 2017].

[d]See https://www.cdc.gov/rdc/ [June 2017].

[e]See https://www.cdc.gov/rdc/data/b4/disclosuremanual.pdf [June 2017].

[f]A list of publications using the linked mortality data is available at https://www.cdc.gov/nchs/data/datalinkage/linked_mortality_files_citation_list_12_2016.pdf [June 2017].

tive Records Research and Applications (CARRA) at the Census Bureau has developed a probabilistic record linkage system in which a protected identification key (PIK) is created for each entity and the PIK is used to link records from different sources behind a secure firewall. Records are matched against a reference file that contains each person's PIK, which is associated with the SSN, name and variants of the name used, date of birth, sex, and current and previous addresses. The linkages provided by CARRA are used in numerous research projects.[3] Jones (2016), for example, used linked data from the CPS and from W-2 records collected by the Internal Revenue Service to study wages of tipped workers in the restaurant industry.

Linkage also allows for the study of entities that are related but not necessarily the same. In a medical study, it may be desired to link electronic medical records of patients with information about their health care providers or with records of other patients of those providers. Baldwin et al. (2015), for example, linked the records of women who had delivered an

[3]See https://census.gov/library/working-papers/series/carra-wp.html [June 2017].

infant to the records of the infant using the surname, address, and dates of birth and delivery for the purpose of evaluating effects of therapeutic interventions during pregnancy.

Hospitals selected to participate in the National Hospital Care Survey are asked to submit electronic health records for all patient discharges and all emergency department and outpatient department visits. NCHS plans to link these records with other data sources, such as the National Death Index and Medicare and Medicaid data, to measure mortality after discharge and other health outcomes (see DeFrances et al., 2012). Such outcomes would be difficult to study without linking records. Levant et al. (2016) illustrated the types of new analyses possible by linking records from a hospital's emergency department to its inpatient treatment records and its outpatient department to show the outcomes of people with traumatic brain injury.

Research conducted for the National Household Food Acquisition and Purchase Survey of the U.S. Department of Agricultire (FoodAPS; see Ver Ploeg et al., 2015)[4] links survey responses from a probability sample of approximately 5,000 households with administrative data on SNAP participation and purchases, as well as information about the food items and prices that are accessible to the surveyed households. The linked information from SNAP is used to determine SNAP eligibility in the 30 days prior to the survey, resolve data discrepancies, and provide information on usage of the electronic benefit transfer card (U.S. Department of Agiculture, 2016).

The U.S. Bureau of Justice Statistics (BJS) is linking records of admissions and releases from state correctional facilities with other administrative record data to better understand why prisoners recidivate. CARRA gives BJS access to numerous data sources that can be used to identify activities and changes in status that can affect both criminal activity and return to prison (Carson, 2015). For example, Social Security data will indicate whether the former inmate has a job, while data from the decennial census or the ACS will indicate whether the former prisoner is married. These data indicate events that can be turning points leading to or away from prison.

All of these examples illustrate the potential benefits of record linkage for more efficient use of information. At the same time, it is not a panacea. Linkage rates vary across studies and for subpopulations within studies. Wagner and Layne (2014) found correct matches for more than 90 percent of the records in the 2010 census and more than 70 percent of the records in two commercial files, but match rates for other sources can be much lower. For example, Bucholtz (2015) found links between American Housing Sur-

[4]Also see http://ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/faqs.aspx [June 2017].

vey records and tax assessment information for more than 70 percent of single-family detached homes but for only 13 percent of condominiums in multifamily buildings. Rates of missed links and false links depend in part on the linkage method used, but they depend even more on the quality of the linkage variables. Better statistical methods and algorithms can reduce linkage errors, but their utility is limited if the data sources have little identifying information about the records.

Harron et al. (2014) wrote that linkage errors can lead to biased conclusions, particularly when the linked and unlinked populations differ. Statistical methods have been proposed that account for linkage bias (see, e.g., Lahiri and Larsen, 2005; Hof and Zwinderman, 2012; Judson et al., 2013), but these, like nonresponse adjustments, are not guaranteed to remove the bias in key variables of interest.

### Multiple Frame Methods

Record linkage usually requires that data for individual entities be available from the data sources, along with sufficient identifying information to allow records to be matched. For example, individual property tax records from county assessors are available on the Internet and can be linked with address-based records from survey data. Often, however, even when individual records are available from different data sources, there is not enough identifying information to allow the records to be linked. In other situations, information may be available only at aggregate levels. Although Census Bureau staff and other approved personnel have access to individual data records from decennial censuses and the ACS, the public and agencies without agreements to access the data can see only aggregate statistics that are produced from these surveys.[5] A business collecting data about customers may be willing to distribute summary statistics but not individual records. Thus, statistical methods are needed that can combine aggregate statistics or can combine individual-level information from different sources when records cannot be linked. The multiple frame methods

[5]The Public Use Microdata System makes a sample of individual records from the ACS available to the public; however, all personal information is removed from these records, and other confidentiality protections are used to ensure "that it is impossible to identify individuals who provide any response" (https://www.census.gov/programs-surveys/acs/technical-documentation/pums/confidentiality.html [June 2017]). In addition, the "72-year rule" specifies that the full census records are made available to the public 72 years after the census date (U.S. Census Bureau, 2008).

Selected information from the decennial census is available for census blocks. Statistics from the ACS are available for block groups, which on average contain about 39 census blocks. Other data are available only for census tracts, which generally contain between 1,200 and 8,000 people (U.S. Census Bureau, 2012).

described in this section, as well as some of the statistical modeling techniques described below, can be used to combine statistics from different data sources.

A multiple frame survey draws samples from two or more sampling frames[6] to improve coverage of the population or to decrease costs. In its simplest form, with frames A and B, estimates are calculated for (1) the units in frame A but not in frame B, (2) the units in frame B but not in frame A, and (3) the units in both frames. The units in group 1 could be sampled from frame A or from frame B and thus have a higher chance of being sampled than if they were only in one frame. Lohr (2011) summarized methods that can be used to obtain unbiased estimates from multiple frame surveys, adjusting for the multiple chances of selection. Most of those methods involve reducing the survey weights for observations that are in both frames so that they represent the "overlap" part of the population and are not double-counted in the estimates.

For example, the Behavioral Risk Factor Surveillance System (BRFSS) measures health-related behaviors, health conditions, and use of medical services. It collects more than 400,000 telephone interviews with adults each year, with samples in every state. In the survey's early years, only landlines were called, but pilot studies indicated that, as the number of households with only cell phones increased, limiting the survey to landline households might result in biased estimates of some health characteristics (Hu et al., 2011). In response, in 2011 BRFSS began including cell phone as well as landline data in the public-use datasets. A dual frame design is used, in which one sample is drawn from a sampling frame for landlines and a second sample is drawn independently from a sampling frame for cell phone numbers (Centers for Disease Control and Prevention, 2016). Some households have both a landline and a cell phone, so they could be selected from either or both frames. Adjustments are made to the weights of households with both landline and cell phones so that they represent that part of the population in the combined samples.

Multiple frame surveys are often used in situations in which the frames

[6]A sampling frame is a list of population units from which the sample is drawn or a method for describing the population. The Current Employment Statistics Survey, which provides the establishment survey data in the monthly news releases on the employment situation (see, e.g., Bureau of Labor Statistics, 2017b), surveys about 147,000 businesses and government agencies, representing approximately 634,000 individual worksites. The sample is drawn from a list of Unemployment Insurance accounts (Bureau of Labor Statistics, 2017a, Ch. 2, p. 1). The NCVS samples areas that are formed from individual counties or groups of counties from the list of all U.S. counties and then subsamples households and group quarters within those areas from a sampling frame built from address lists (Bureau of Justice Statistics, 2014, p. 8). In other situations, a sampling frame may be described algorithmically, without assembling a list of the population, such as sampling every 20th visitor to a website.

cannot be consolidated before sampling. The cell and landline frames used for dual frame telephone surveys do not contain enough information to link the records and eliminate duplicates before sampling. Thus, respondents are asked about telephones in their household, and that information is used to determine whether they have a cell phone only (group *a*), a landline phone only (group *b*), or both cell and landline phones (group *ab*). Then, the population total for the characteristic of interest (e.g., the number of smokers in the population) is calculated as the sum of the estimated total number of smokers from groups *a*, *b*, and *ab*. Because group *ab* is sampled from both frames, the total number of smokers from group *ab* may be estimated as (estimated total number of smokers in group *ab* from the cell sample) + $(1 - \lambda)$ (estimated total number of smokers in group *ab* from the landline sample), where $\lambda$ is often chosen to be 0, 1, or 0.5.

The U.S. Department of Agriculture frequently uses multiple frame surveys. The National Agricultural Statistics Service (NASS) maintains a list frame of farm operations, which attempts to list all of the farms in the United States. The list frame is less expensive to sample from and contains most of the large operations, but it is incomplete because farms go in and out of business. To address this situation, NASS surveys often supplement a sample drawn from the list frame with a sample of land segments drawn from an area frame. The area frame for a state contains all of the land in the state and thus is complete, but it is more expensive to sample from (Davies, 2009). Farm operations in the area frame are matched with the list frame, and those found in the list frame are removed before sampling so they have only one chance of being in the sample. The Farm Labor Survey is an example of a NASS survey using this dual frame design.[7]

The 2015 Local Food Marketing Practices Survey was designed to produce statistics on the number of farms that market food directly, for example, through farmers' markets. Two frames were used for the survey. The first frame was the NASS list frame. The second frame, containing potential local food operations, was derived from web-based information and was used to measure coverage of the first frame.[8]

Multiple frame surveys can increase coverage of the population, and they have the potential to reduce costs if one or more of the frames is inexpensive to sample from. In some cases, an incomplete frame may have the information needed so that the entire frame can be used and sampling is not necessary. However, when one or more of the frames is incomplete, it is necessary to determine whether an entity sampled from one frame could also

[7]See https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Labor/05_2017/LABQM_May2017.pdf [September 2017].

[8]See https://www.agcensus.usda.gov/Publications/2012/Online_Resources/Local_Food/quality_measures/2015_LFMPS_Methodology.pdf [September 2017].

have been sampled from the other frames. Although record linkage may be used to determine frame membership, typically there is less privacy intrusion for multiple frame surveys than for record linkage. It is also important to account for potential differences in the data collection procedures among frames—for example, if one sample is conducted in person and another by e-mail—when analyzing the data.

Multiple frame methods have great potential when used with some of the newer data sources. Some current multiple frame surveys rely on an expensive area frame to ensure complete coverage of the population. It may be possible in some cases to obtain better (although perhaps incomplete) coverage with less expense by constructing supplemental frames from alternative sources such as data provided by commercial vendors, web-scraping, or imaging data.

### Imputation-Based Methods

Another way to conceptualize combining different data sources is using a missing data framework and imputing (filling in) the missing data. Different data sources often measure different sets of variables, and linking or adding two or more data sources results in a merged dataset that has missing values. For example, suppose data source *A* has an identification (ID) variable, age, and sex; data source *B* has ID, age, medical expenditures, and smoking status; and data source *C* has ID, sex, and smoking status. Some of the people in source *A* are also represented in source *B*, while source *C* has different people. If sources *A* and *B* are linked by ID and added to the records in source *C*, the resulting merged dataset has "holes," as shown in Table 2-1. In this situation, the problem of combining information can be viewed as a missing data problem, and imputation methods can be used to fill in or impute the missing values in the combined dataset.

**TABLE 2-1** Information from Three Sources, *A*, *B*, and *C*

| Source | ID | Age | Sex | Medical Expenditures | Smoking Status |
|---|---|---|---|---|---|
| Records Linked from *A* and *B* | X | X | X | X | X |
| Records from *A* with No Linked Record from *B* | X | X | X | | |
| Records from *B* with No Linked Record from *A* | X | X | | X | X |
| Records from *C* | X | | X | | X |

Many approaches can be used to impute the missing values, some of which are reviewed by van Buuren (2012) and Kim and Shao (2013). In some approaches, a missing value on an item is replaced by the value from another data record. In other approaches, a multivariate model is used to predict the missing set of values using the information in the observed values. Alternatively, the imputation may proceed variable-by-variable through a sequence of regression models using all the variables other than the variable being imputed as predictors. The variable-by-variable approach simplifies the modeling task to finding a good fitting regression model for every variable to be imputed. Multiple imputations can be used to include the extra variability from the imputation predictions in standard errors for statistics.

Often, one wishes to combine data sources containing different sets of individuals or sources in which individuals cannot be deterministically linked. Statistical matching, also called data fusion, is sometimes recommended for these situations. Suppose that data source *A* contains demographic variables and information on health care expenditures, and data source *B* contains demographic variables and information on exercise habits for a different set of people. Statistical matching methods (Rodgers, 1984; Moriarity and Scheuren, 2001) use the correlations between the demographic variables and health care expenditures from source *A* and the correlations between the demographic variables and exercise habits from source *B* to make inferences about the relationship between exercise habits and health care expenditures. Statistical matching methods typically rely on strong assumptions for these inferences because there are no records that have both variables of exercise habits and health care expenditures. An alternative approach imputes the missing variable to one or both datasets using estimated relationships between the demographic variables and the responses of interest. Fosdick et al. (2016) reviewed recent literature on statistical matching and proposed a new method in which an inexpensive online survey is used to provide additional information relating the variables of interest.

Schenker and Raghunathan (2007) described four examples in which multiple survey data sources were combined to (1) extend and enhance the coverage, (2) handle transitions from one approach of measurement of a variable to another, (3) correct errors in self-reported data, and (4) improve small-area estimation. Most of these examples used multiple imputation or a Bayesian modeling approach. See Lohr and Raghunathan (2017) for several other examples that use both non-Bayesian and Bayesian perspectives.

Implementation of an imputation-based approach for combining data from multiple sources is now feasible because of the availability of several software packages that can create model-based imputations (see van Buuren, 2012). However, using these packages to combine data sources

requires expertise in imputation methods and in evaluating the comparability of the data sources. Agencies that do not already have this expertise on staff may need to develop it.

Given sufficient computational resources, it is conceivable that data from multiple sources could be used to create a large, representative population, perhaps even with a longitudinal component. This could be constructed from various surveys and administrative data sources. Spatial and temporal components could be added by linking satellite imageries, environmental monitors, and weather and climate data. This dynamic linking of multiple surveys and administrative data sources could create spatiotemporal data representing the U.S. population. Given the large number of variables and subjects, there would likely be a good deal of missing data; however, machine learning techniques informed by substantive modeling could be used to predict the missing values and capture the associated uncertainty with the predictions. Thus, the predictions and associated uncertainty may be used to create several copies of the populations to construct inferences for population quantities of interest.

This approach of creating a synthetic or modeled micro dataset from partially observed data has not been tried in the federal statistical system except in two instances, both undertaken to protect confidentiality. Both the Survey of Income and Program Participation (SIPP) and the Longitudinal Business Database (LBD) use modeling to produce synthetic datasets. The population creation described in this section, however, may be a useful strategy for protecting confidentiality when the actual observed data are embedded in modeled data, thus affording protection from disclosure.

Using multiple imputations to combine information from multiple data sources presents challenges, including taking into account the complex design of the survey data sources and incomparability between sources. Lohr and Raghunathan (2017) note a number of potential incomparabilities that may arise when combining multiple survey data sources, including:

- the types of respondents and the source of information: self-reported medical information from respondents to a health care survey may differ from medical records obtained from health care providers;
- mode of interview: in-person versus telephone versus self-administered questionnaire;
- survey context and sponsorship, such as a federal or a private-sector entity;
- differences in survey design and measurement, such as asking about recall of exercise as opposed to having respondents keep a diary or obtaining data from a fitness tracker; and
- different questions, question wordings, or question orderings.

Additional sources of incomparability can arise when combining surveys with nonsurvey data sources, such as administrative records and private-sector data. We review and discuss a number of these issues in Chapter 6.

Another important component of imputation methods is their reliance on the model assumptions about the mechanism producing the missing data and predictive models for the missing values. These model assumptions have to be thoroughly checked (see, e.g., Abayomi et al., 2008; Bondarenko and Raghunathan, 2016), and the sensitivity of the inferences to the underlying assumptions (for both the missing data mechanism and predictive models) needs to be explored (see, e.g., Raghunathan, 2015; Permuutt, 2016; Smuk et al., 2017).

Despite these challenges, there are a number of advantages to using imputation to fill in missing data: imputation can provide a complete dataset without any "holes"; the imputations can take advantage of the relationships among all the variables that are present on the files; it provides a means for inferring beyond the scope of each individual data source; and the modeling framework provides an explicit and transparent means for incorporating differences and incomparabilities among data sources (see Lohr and Raghunathan, 2017). We elaborate on additional modeling techniques in the next section.

## Modeling Techniques

Record linkage and imputation are suitable methods when individual record-level data from multiple sources are available. When data are available only as aggregated statistics at the national, subnational, or subpopulation level, the multiple frame methods described above can be used to combine summary statistics that all measure the same characteristic. In addition, other statistical modeling methods can be used to combine aggregated statistics with each other or with individual record data when the data sources measure different variables.

Small-area estimation methods are examples of statistical models that combine statistics estimated from a probability survey with statistics calculated from administrative data (see National Academies of Sciences, Engineering, and Medicine, 2017b, Box 3-3). In the Small Area Income and Poverty Estimates Program[9] at the U.S. Census Bureau and the Small Area Estimates for Cancer-Related Measures Program at the National Cancer Institute,[10] models are developed that relate direct estimates of the characteristics of interest (poverty rate or cancer rate) to covariates that are available from administrative data. The models are used to predict the

[9]See http://www.census.gov/did/www/saipe/ [June 2017].
[10]See http://www.sae.cancer.gov [June 2017].

poverty or cancer rate in areas where no direct survey estimates are available to improve the precision of estimates in those areas.

For small-area estimates for cancer-related measures, Bayesian hierarchical or multilevel models have been used to model the direct estimates from multiple surveys rather than using a combination of survey- and nonsurvey-based estimates. The models incorporate differing error structures in the estimates (bias and sampling variance) across surveys and also use rich sets of covariates assembled from administrative data. These types of models can also be used to combine survey and nonsurvey estimates.

One example of this method is the small-area estimation of yield or acreage devoted to a particular crop. The estimates from a farm survey, which may be available only for a subset of areas, and the estimates based on area-level satellite imagery, which may be available for all areas, could be combined to improve the accuracy of small-area estimates, especially for the locations that are not sampled in the farm survey (Bellow, 2007; Cruze, 2015; National Academies of Sciences, Engineering, and Medicine, 2017a). The modeling framework provides a means for incorporating differing sources of error structures in the two estimates (in this example, one is subject to mostly sampling and nonresponse errors and the other is subject to mostly measurement errors). There are numerous examples throughout the federal statistical system, as noted above, in such areas as crime and victimization rates, health status, and economic activity: multivariate hierarchical models can be used to combine data from multiple sources to create a systematic program of small-area estimation. Such combining of information can not only benefit estimation, but also provide information useful for redesigning surveys to fully exploit the correlation between various estimates. For example, more survey data could be collected for areas where the measurement error properties of the nonsurvey estimates are high rather than for areas with small measurement errors.

Currently, every federal statistical agency develops its own system for small-area estimation. Even within one agency, small-area estimation may be compartmentalized across divisions. Thus, the current distributed system of developing small-area estimates may not be fully efficient. For example, consider a case in which small-area estimates of smoking status and poverty are needed. To the extent that smoking behavior patterns differ by socioeconomic status, the correlation structure between these two variables can be exploited by jointly modeling the two outcomes using the multivariate Bayesian hierarchical model framework and, hence, deriving the estimates of the prevalence of smoking status and poverty. This modeling technique can be applied, and can be even more useful, when direct estimates for both outcomes are not available in every area. Suppose that for a subsample of areas both outcomes are measured, for some areas only smoking status is available, and for others only poverty is measured. The correlation between

these two outcomes provides information on the missing outcome. Consider another situation in which the precision of the available direct estimates differ by outcomes across areas. Here, too, the correlation between outcomes improves the precision of the model-based estimates. Thus, borrowing strength not just across areas but also across variables may improve the efficiency of small-area estimates, and, hence, a systemic view of the small-area estimation tasks coordinated across federal agencies could leverage aggregated data from multiple sources through joint estimation procedures.

Data from multiple sources may be of a mixed nature, with some having aggregated data and others having individual-level data. Methods for combining such data have been developed using the hierarchical models. For example, Raghunathan et al. (2003) used aggregated data from a large number of small areas or communities and small samples of individual-level data from a few areas to obtain estimates of the parameters in the individual level model (see also, e.g., Haneuse and Wakefield, 2007; Chatterjee et al., 2016). A general hierarchical framework may be used to develop a constrained estimation of individual-level population parameters given the aggregated data from a large number of areas and the individual-level data from a small number of areas.

The methods discussed in this section rely on models to a greater extent than methods currently used for most surveys in the federal statistical system. For most estimates produced from federal probability surveys, it is not necessary to postulate a statistical model relating the quantities being measured, although models are commonly used when adjusting for nonresponse (see Skinner and Wakefield, 2017).[11] However, when combining data from survey and nonsurvey sources, model assumptions may be needed for inference because the nonsurvey data sources lack a probabilistic selection structure for the units in the dataset (Elliott and Valliant, 2017). When statistical models for combining information from survey and nonsurvey data sources are developed, they will need to be empirically tested and substantively justified. These statistical models can then form the basis of constructing estimates and associated measures of uncertainty. In the Bayesian framework, credible intervals from the posterior distribution of the estimand of interest combines information from both survey- and nonsurvey-based estimates. Using statistical models for inference would comport with the practice used in most other areas of statistics.

Modeling plays a central role in developing estimates using the framework described in this section. But what if the model is misspecified? The federal statistical system has traditionally relied on estimates that are based

[11]In practice, models are used in probability sampling inference to adjust for nonresponse and undercoverage, but inference for a survey with a 100 percent response rate could be based solely on the selection probabilities.

on the sampling design rather than specified statistical models to avoid the problem that model-based inferences can be wrong if the model chosen is not appropriate for the data, and the design-based inference approach works well when there are high response rates and low costs. With increasing nonresponse and a need to combine multiple data sources, however, it is necessary to make modeling assumptions. As George Box (1979, p. 2) wrote, "All models are wrong but some are useful." Thus, one may want to change the question, "Is the model reasonable?" and, therefore, useful. The danger lies in using unreasonable models that yield unreasonable estimates. Thus, a transparent description of the underlying assumptions, model checking or diagnostics, and exploration of sensitivity of inferences to the modeling assumptions need to become integral parts of the estimation framework. Such a transparent framework will build trust, open the models and methods for critical review, and minimize the danger of using unreasonable models. The technical documentation, at minimum, needs to include detailed descriptions of the models, the methods used to support the models, and descriptions of the limitations and methods used to explore sensitivity of the derived estimates to the underlying model assumptions. The documentation needs to be accessible at several levels. For the methods described in this report to succeed, staff are needed who are experts in statistical modeling techniques and traditional survey designs.

**CONCLUSION 2-2 Many statistical methods currently available can be adapted for using combined data sources to develop estimates of target population quantities of interest.**

**RECOMMENDATION 2-2 To achieve transparency, federal statistical agencies should document the processes used to collect, combine, and analyze data from multiple sources and make that documentation publicly available.**

## NEXT STEPS FOR COMBINING DATA SOURCES

### Research Needed

This chapter reviews some of the statistical methods that are currently being used or could be adapted to be used to combine information from different data sources to produce official statistics. Many of those methods have been developed to augment data collected from the probability surveys that currently form the backbone of the federal statistical system. Some of the methods—notably, record linkage—can be applied to administrative and commercial data sources as well as to probability surveys. The record

linkage techniques can be applied to join any datasets with common variables that can be used for linkage.

Much of the current federal agency research on using multiple data sources is exploring linking records from different sources. Nearly all of the technical presentations by federal agency personnel at the panel's December 2015 workshop involved data linkage. Much more research is needed on record linkage methods. In particular, more research is needed on estimating the quality of the links and on how to propagate the uncertainty about linkage to analyses of linked datasets.

Most methods assume that some sources of data (or combination of sources) produce approximately unbiased estimates of some characteristics of the population of interest—that is, the expected value of an estimate of a characteristic is approximately equal to the true population value. In theory, when there is no nonresponse or undercoverage, probability samples produce unbiased estimates and, historically, that unbiasedness has been a major reason for their use. But as discussed in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b), decreasing response rates may be threatening that assumption: although survey analysts attempt to adjust for nonresponse through weighting or imputation (usually based on demographic information), there is no guarantee that these methods remove the bias in the key variables from a survey. Administrative or private-sector records can be similarly weighted or imputed; again, it is anticipated that such adjustments will reduce the bias due to records that are not in the data, but it is always possible that the individuals not present in the dataset have different characteristics than the demographically similar individuals who are in the dataset.

Both probability sample survey records and administrative records have large amounts of missing data by design: sample surveys include only those people or entities selected into the sample and who responded, while administrative records include only those in the program (e.g., SNAP recipients). A key area for which more research is needed is on using the information in all of the data sources to identify potential biases and fill in data that are missing from some of the sources. This can be done through record linkage; through multiple frame methods, in which it may be possible to identify the overlapping parts of the population; or through modeling and imputation methods, in which relationships among variables can be used both to study biases in different sources and to fill in missing values. More research is needed on other methods that can deal with missing data from multiple sources. With multiple sources of data come (possibly) multiple estimates of population characteristics. A framework is needed for evaluating the quality of data sources, and this is discussed in Chapter 6.

Even if alternative data sources are used for some purposes, there will likely be continued reliance on probability samples as a primary source of

federal statistics, at least for some indicators. The National Crime Victimization Survey, for example, measures both crimes reported to the police and those not reported to the police. It is difficult to see how the latter could be measured accurately without using a survey, although police agency reports may be helpful in improving estimates of the former. Even if new data sources are integrated into federal statistics, we anticipate that traditional probability surveys will still be needed to cover parts of the population not in the other sources and to provide a check on their quality. The decreasing response rates of surveys continue to be a concern, and ongoing research is needed on ways to promote response and to deal with nonresponse. How does the public view using alternative sources of data for official statistics, and do those views affect willingness to provide data?

Statistical methods in use for surveys typically produce static estimates: for example, the National Crime Victimization Survey produces estimates of victimization rates in each calendar year, and the CPS produces monthly unemployment statistics. Administrative data records and sensor data, however, may be updated much more frequently. Sensor data, in particular, are collected continuously, as are other automatically collected data, such as data collected from smart phones, fitness monitors, and "smart cars." Challenges arise on how to integrate information that is collected in different time frames. In particular, little research has been done to date on statistical methods for combining some of the "big data" sources with administrative records or probability samples.

As we discussed in our first report (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 4), many of the private-sector data sources that might be used for the statistics of the future are generated as by-products of electronic activity and are massive. Electronic health records, which may contain information on all utilizations of health care; credit card transactions; traffic sensors; cell phone location records; web-scrapings; smart meters; and other data sources produce exabytes of data every day. Machine learning methods—techniques in which algorithms search for patterns in data—are frequently used with these types of large, organically collected datasets to uncover correlations among variables. New statistical methods are needed for interpreting and merging such data. Machine learning techniques have also been used for record linkage and imputation. More research is needed on using and developing machine learning methods to combine data sources. There are additional research needs on the privacy issues associated with these data sources (Froomkin, 2016), which we cover in Chapter 5.

Many of the statistical models for combining data sources discussed in this chapter start with the structure of the existing data sources and then specify how to combine them. More research is needed on designing new systems of data collection that make use of multiple sources to provide

federal statistics. This would represent a shift from the current framework, where a probability survey is designed to serve as the primary source of information and other sources are used as auxiliary information, to a model in which the "best" data source is used for particular aspects of the data. In some cases, this approach may mean systematically redesigning a data collection so that inexpensive data sources are used for the parts of the population they can capture, and more expensive probability samples are used for the parts of the population that cannot be measured any other way.

Finally, research is needed on the robustness of statistical systems. One advantage of the probability sampling framework is that it is difficult for an external actor to manipulate the system. Participants in the survey are selected randomly, and although people or entities that are sampled may decline to participate, no external actor can decide which units are sampled or which choose to respond. With administrative records, commercial data, or convenience data sources, however, it may be possible for external actors to modify the data; for example, social media data could be flooded with responses if those data were to be used to inform policy. In addition, there is no guarantee that the data sources available today will be available next year or any time in the future. A data vendor may stop collecting data or choose to keep the data private.

Many of the methods described above rely on modeling the relationship among variables in different data sources. If the probability sample is replaced by other sources, safeguards need to be put in place to ensure that the models continue to be valid, as relationships among variables may change.

**CONCLUSION 2-3 Research is needed on designing new systems to collect and process multiple data sources to create and enhance federal statistics.**

**RECOMMENDATION 2-3 Current statistical methods should be adapted to the extent possible and new methods should be developed to harness the statistical information from multiple data sources for analysis.**

### Structure Needed for Implementation

Though statistical methods for record linkage, multiple frame surveys, imputation, machine learning techniques, and hierarchical models for combining data are available, many of them need further research and adaptation. Such research is currently being done at many agencies and by academic researchers. This research can be facilitated by better communica-

tion and, perhaps, coordination of the research projects and the knowledge gained from the projects.

As detailed in Chapter 1, the panel's first report recommended the creation of a new entity with the authority to access multiple data sources for blending. Such an entity could achieve this communication through summarizing data linkage projects and other projects involving the combination of data sources; we discuss this topic in Chapter 7. A publicly accessible website could supply basic information about ongoing research projects through the entity, including their purpose, the datasets being combined, an outline of the statistical methodology, results, and lessons learned.

The statistical methods described in this chapter assume that agencies have access to the data sources needed. As described in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b), obtaining this access is a challenging process. In addition, even with access, the data may not be in a form that is amenable for producing statistical estimates. A partnership among agencies is needed to make such data accessible for combining.

Of the research that is needed for establishing a new paradigm for producing federal statistics, one of the primary areas is systemic redesign of data collection methods that rely on multiple sources to produce federal statistics. Such research will require the resources of multiple agencies, as well as cooperation with academia and businesses. The skills needed for the research include the traditional skills in probability survey design and analysis, but they also include knowledge of record linkage, machine learning, new statistical modeling techniques, and privacy expertise. Training is needed both in the statistical agencies and the broader research community to ensure that research staff have the skills and adaptability needed to advance the field of combining data. In addition, development of algorithms and user-friendly software is needed for implementing some of the methods for combining data. A multidisciplinary approach would be ideal, drawing on and developing expertise in statistics, computer science, economics, engineering, and the fields related to the substance of statistics that are produced.

Federal agencies also need to continue to develop partnerships with research organizations and businesses to develop new data sources and new statistical methods. Two important parts of these partnerships are evaluating the quality of different data sources and the quality of the statistics produced by combining data from different sources.

In Chapter 7, we further discuss how the entity could guide and serve as a resource for research and training. The structure for revolutionizing federal statistics has to be dynamic and innovative, willing to explore new frontiers and modern modeling methods. The federal statistical system needs to be empowered to capitalize on modern computational and statisti-

cal developments and the plethora of emerging data sources and to continually improve the methods for producing statistics.

**CONCLUSION 2-4 Federal statistical agencies are currently combining information from multiple data sources for specific projects. Systematic coordination and dissemination of their results will help advance knowledge and promote the use of appropriate statistical methods.**

**RECOMMENDATION 2-4 Federal statistical agencies should ensure their statistical staff receive training for the new skills needed for combining data from different sources.**

**RECOMMENDATION 2-5 Federal statistical agencies should develop partnerships with academia and external research organizations to develop methods needed for design and analysis using multiple data sources.**