Annual Review of Statistics and Its Application

Statistical Data Integration for Health Policy Evidence-Building

Susan M. Paddock, Carolina Franco, F. Jay Breidt, and Brenda Betancourt

NORC at the University of Chicago, Chicago, Illinois, USA; email: paddock-susan@norc.org, franco-carolina@norc.org, breidt-jay@norc.org, betancourt-brenda@norc.org

First published as a Review in Advance on August 19, 2024

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-112723-034507

Keywords

data quality, policy analysis, record linkage, small area estimation, statistical matching, survey

Abstract

Health policy evidence-building requires data sources such as health care claims, electronic health records, probability and nonprobability survey data, epidemiological surveillance databases, administrative data, and more, all of which have strengths and limitations for a given policy analysis. Data integration techniques leverage the relative strengths of input sources to obtain a blended source that is richer, more informative, and more fit for use than any single input component. This review notes the expansion of opportunities to use data integration for health policy analyses, reviews key methodological approaches to expand the number of variables in a data set or to increase the precision of estimates, and provides directions for future research. As data quality improvement motivates data integration, key data quality frameworks are provided to structure assessments of candidate input data sources.

1. INTRODUCTION

Data are essential for obtaining evidence of the magnitude of effects of health policies (Baicker & Chandra 2017). The importance of data to evidence-building more generally is underscored by the Foundations for Evidence-Based Policymaking Act of 2018 (also known as the Evidence Act), which requires federal agencies in the United States to facilitate greater use of data to guide their policy- and decision-making and greater data sharing among US federal agencies. For example, the 2023 Evaluation Plan for the US Department of Health and Human Services summarizes priority areas and strategic goals along with the accompanying evaluation questions, study designs, and a broad array of methodologies from descriptive statistics to advanced modeling. Multiple data sources such as health care claims, electronic health records (EHRs), survey data, epidemiological surveillance databases, and administrative data are integral to building the evidence required for policymakers to address the priority areas. These data are captured for different purposes and are typically designed, collected, managed, and stored separately.

Health policy researchers have long relied on data integration, which is the combining of multiple data sources to create a source that is richer and more informative than any single component. Classic examples include linkages such as the Surveillance, Epidemiology and End Results–Medicare database, leveraging external (auxiliary) data sources when imputing missing survey item data using administrative data or other surveys, and combining population surveys of national importance with domain-level estimates from sources such as the American Community Survey (ACS) to facilitate estimation for small demographic groups (Rein et al. 2024). The National Institute on Aging established its Data LINKAGE Program to link data from studies it has funded with claims, mortality data, and other administrative data from the Centers for Medicare & Medicaid Services (CMS). Such linkages allow for examining relationships among measures from registries or surveys with health care cost, utilization, and other claims-based measures. Enhancing data quality, or the degree to which data capture useful and trustworthy information for an intended analytic purpose (Fed. Comm. Stat. Methodol. 2020), thus motivates statistical advances and applications of data integration.

Recent and ongoing developments aim to facilitate data integration that is increasingly done on-demand and customized to one's own analytic needs. The National Secure Data Service demonstration project authorized by the CHIPS and Science Act of 2022 in the United States aims for a model to enhance data sharing and linkage of data from customized linkages of governmental and other data. The private sector also offers new opportunities for customized data integration; companies such as Datavant and HealthVerity provide data ecosystems for securely linking data across multiple data providers.

The purpose of this article is to advance health policy evidence-building by guiding readers to recognize and benefit from expanding opportunities for statistical data integration. In particular, we review key statistical approaches currently in use that underpin ongoing methodological developments in the field. As data quality is a central motivator for data integration, we discuss employing a data quality framework in Section 2 and illustrate its use to evaluate the strengths and limitations of data on public perceptions of actions taken by health authorities and elected officials in response to COVID-19. We anchor our methodological presentation to the survey statistics literature, given its importance for increasing the policy relevance of estimates by improving their representativeness of the target population. The total survey error (TSE) framework (Groves et al. 2004) and a more general data quality framework are reviewed in Section 2. Section 3 turns toward blending probability and nonprobability samples in a principled way to facilitate statistical inference to a policy-relevant target population, noting that concepts of representativeness and coverage of policy-relevant populations are also important to consider when using nonsurvey data.

ILLUSTRATION: THE NEED FOR HEALTH DATA QUALITY ASSESSMENT

Consider a policymaker who would like information about social determinants of health (SDOH) for Medicare beneficiaries. SDOH are nonmedical conditions in one's environment that affect health and include economic conditions, reliable access to food, and familial support. As of the mid-2010s, providers were allowed to enter a new code in traditional Medicare claims to indicate whether they discussed SDOH with their patients; however, only 1.6% of claims in 2019 included such codes (Maksut et al. 2021). In contrast, the Medicare Current Beneficiary Survey (MCBS) includes SDOH measures, for which the percentage of beneficiaries reporting SDOH is much higher. MCBS is a longitudinal survey of a nationally representative sample of the Medicare population in the United States. To understand the strengths and limitations of each data source for examining SDOH, data users might consider questions such as why and when codes would not be recorded in claims and what that means with respect to the quality of the available data, or whether the MCBS includes SDOH of interest or whether its sample size would support analyses of SDOH for small and important subgroups.

We also briefly compare and contrast the blending of probability and nonprobability samples with the combining of randomized controlled trial data and observational data to examine the generalizability and transportability of health policy intervention effects in Section 4. Blending probability and nonprobability data can improve the precision of estimates, as does small-area estimation for domains such as geographic areas, demographic groups, or patient subgroups, as reviewed in Section 5. Methods to increase the number of variables available for analysis are discussed in Section 6, specifically record linkage of units appearing in multiple data sources and statistical matching of data sets with partially overlapping variable lists and containing distinct units. We conclude and provide future directions in Section 7.

2. CHARACTERIZING DATA QUALITY

The first step in determining the need for and feasibility of data integration is to assess the quality of each available data source in the context of a focal health policy research question (see the sidebar titled Illustration: The Need for Health Data Quality Assessment for an example of data quality considerations when using administrative versus survey data sources to examine social determinants of health). This section includes some data quality frameworks useful for more formally structuring such data quality assessments.

2.1. Total Survey Error Framework

Many general questions of statistical uncertainty, pertaining to data obtained from surveys or other sources (blended or otherwise), can be placed within the TSE framework (Groves et al. 2004). This framework divides the creation of a statistic into two parallel paths: measurement and representation. Some measurement of a theoretical construct of interest is proposed, and the implemented measurement protocol may result in missing values and measurement errors. Edited and imputed responses are then used in computing a statistic. What does this statistic represent? A target population is envisioned, and a sampling frame gives access to some or all of that population. Undercoverage error occurs when the entire target population is not accessible through the frame: Population entities not in the frame have no representation in subsequent samples. Samples are representative of the frame from which they are drawn if they are selected with known, positive probabilities. The representation of a sample that does not meet these conditions is unknown. Not all sampled entities respond, increasing the variance of computed statistics and possibly leading to

nonresponse bias if the propensity to respond is related to the characteristic of interest. Weighting adjustments attempt to account for errors of coverage, sampling, and nonresponse, reducing bias without overly increasing variance. The TSE framework forms a conceptual basis not only for designed observational studies like survey samples but also for undesigned big data alternatives such as those described by Amaya et al. (2020).

2.2. Federal Committee on Statistical Methodology Framework for Data Quality

Motivated by the increasing use of existing data and data integration by federal statistical agencies in the United States, the Federal Committee on Statistical Methodology (FCSM) developed an 11-dimensional framework for data quality applicable to all data (Fed. Comm. Stat. Methodol. 2020). The 11 dimensions are organized into three domains:

- Utility: relevance, accessibility, timeliness, punctuality, granularity
- Objectivity: accuracy and reliability, coherence
- Integrity: scientific integrity, credibility, computer and physical security, confidentiality

The FCSM framework provides a common vocabulary for describing the quality of data sources and their fitness for use for statistical purposes (e.g., Mirel et al. 2023). It encourages systematic and transparent reporting of data quality threats, mitigation strategies, and remaining limitations. Most of the statistical considerations encapsulated by the TSE framework are placed together under the objectivity domain. Several of the FCSM dimensions are directly relevant to blended data and the unique challenges that might arise in combining data sources: for example, the logistical challenges of punctuality and physical security, and the scientific challenges of granularity and coherence.

2.2.1. Example: trust in public health decision-makers. Consider two data sources collected by Statistics Canada that were used by Bahamyirou & Schnitzer (2021) in an analysis of Canadians' trust in elected officials, health authorities, and others shortly after the start of the COVID-19 pandemic: Impacts of COVID-19 on Canadians, a crowdsourced web-based data collection, and the Labour Force Survey (LFS), a probability sample whose target population is composed of noninstitutionalized persons aged 15 years or older. Considerations of the data quality dimensions by domain include the following:

- Utility: Unlike the LFS, the crowdsourced data had high utility for analyses conducted in 2020—the inclusion of trust measures of interest made it relevant. It was also timely and punctual, with real-time measurement of COVID-19 impacts and data that were rapidly collected and disseminated to health officials. Both sources include individual-level data, providing a high degree of granularity.
- Objectivity: This domain is unclear for the crowdsourced data; the lack of a sampling frame makes it impossible to assess whether estimates based on the data reflect true values (accuracy) and whether the data are consistent or comparable with other relevant data (coherence). In contrast, the LFS is based on a known sampling frame.
- Integrity: The crowdsourced nature of the Impacts of COVID-19 on Canadians data puts its scientific integrity into question for obtaining population estimates, but its administration by a Canadian federal statistical agency lends credibility to it for many data users. In contrast, LFS is designed to meet scientific integrity standards for official statistical probability surveys. Users have confidence that LFS-based estimates reflect the target population. Confidentiality of respondents is safeguarded in both data sources.

Bahamyirou & Schnitzer (2021) implemented data integration using methods along the lines of those described in Section 3 that anchored the crowdsourced data to the target population represented in the LFS, thus enabling statistical inference about COVID-19 impacts that could not have been obtained from either data set alone.

2.2.2. Example: impact of a health care payment system change. Paddock et al. (2007) examined changes in patient severity before and after the 2002 implementation of a new Medicare payment system for inpatient rehabilitation facilities in the United States. Unlike the prior payment system, payments under the new system were functions of patient conditions and functional and cognitive scores. The new system was designed to better align payment with costliness of care, thereby improving access to care for relatively costly patients. Paddock et al.'s (2007) chief data quality concern was coherence: Key data elements took on different levels of importance (and potentially coding practices) pre- versus post-implementation. Since it was problematic to make direct comparisons of these data elements under the prior versus new system, Paddock et al. (2007) used a predictive approach similar to statistical matching (Section 6.2) for their analysis.

3. BLENDING SAMPLES TO IMPROVE REPRESENTATIVENESS

3.1. The Role of Surveys in Health Policy

Nationally representative surveys provide policy-relevant descriptive statistics for the general population, patients or other subgroups, health care providers, hospitals, or other entities. For example, the ACS provides estimates for detailed population demographics, economic characteristics, and housing information on the general population, along with health-related data on fertility, disability, and health insurance coverage. Those who are invited to complete the ACS are legally obligated to do so and are the focus of extensive outreach efforts.

The ACS is an example of a data source that provides trusted population information on the distribution of demographic and other key characteristics, against which respondents to other surveys (discussed in this section) or participants in randomized trials and/or evaluations (Section 4) can be benchmarked. The Canadian LFS in the COVID-19 example in Section 2.2.1 provides benchmarks that facilitate data integration with a more timely but nontraditional nonprobability survey. Many surveys likewise rely on integrating information from one or more samples to address the key challenges facing surveys of mitigating nonresponse bias and the increasing demand for subgroup estimates. The need to meet such goals in a time- and cost-conscious manner leads to consideration of statistical methods to blend probability and nonprobability samples.

3.2. Example: National Hospital Care Survey

The 2020 National Hospital Care Survey (NHCS) provides an example of blending probability and nonprobability samples to improve the representativeness of estimates. The survey was conducted by the National Center for Health Statistics (NCHS) as a stratified simple random sample of hospitals (nonfederal, noninstitutional, six or more staffed inpatient beds). The survey is subject to hospital-level nonresponse. Responding hospitals provided (essentially) complete encounter data for all patients for 2020. The NHCS is thus an invaluable research resource for understanding patterns of health care delivery and utilization in the United States during the first year of the COVID-19 pandemic.

The NHCS is a probability sample from a US federal statistical agency with established scientific integrity, but relatively high nonresponse gave the agency concerns about reliability of estimates published from the NHCS alone. NCHS thus sought to combine the 2020 NHCS

with other data sources to address nonresponse, produce nationally representative estimates, and improve granularity (Breidt et al. 2023). A proprietary commercial data source provided 2020 patient encounter data for another set of participating hospitals, but with an unknown participation mechanism. Both sources of encounter data were comparable in their utility, yielding highly granular data with relevance to the first year of COVID-19 hospitalizations. The encounter data have excellent coherence between the two sources, so that reliability of the NHCS national estimates could be improved by addressing possible selection biases in the commercial data. Data protection concerns necessitated novel weighting methods to create a restricted-use blended data file. Methodological approaches for accomplishing these goals by blending data sources in this complex setting are reviewed in Sections 3.3–3.5 and summarized for the NHCS application in Section 3.6.

3.3. Probability Samples

Many questions of health policy can be answered with descriptive inference, in which characteristics of a real, observable, finite population are estimated from a sample (see Section 3.1). Other questions may be addressed with analytic inference, in which characteristics of a hypothetical population are modeled and predicted. In the NHCS application, both inference types are of interest: A descriptive example is estimating the total number of patient encounters with a given diagnosis code during 2020, while an analytic example is fitting a model to predict length of stay as a function of diagnosis code and patient demographics.

Let $U = \{1, 2, \ldots, k, \ldots, N\}$ denote a finite population relevant to a policy question of interest. We consider estimation of finite population totals T_y using weighted estimates, $\widehat{T}_y = \sum_{k \in s} \omega_k y_k$, where $s \subset U$ is a subset of the entities of interest and ω_k are constructed so that $\widehat{T}_y$ is (at least approximately) unbiased for T_y for any characteristic y_k. In design-based descriptive inference, y_k is treated as nonrandom, and statistical inference is based on stochastic properties of the sample s and possibly the weights ω_k.

By contrast, analytic inference relies on statistical models (often parametric) that are assumed to have generated the finite population of values $\{y_k\}_{k \in U}$. These models are referred to in the survey literature as superpopulation models. Under mild conditions, the weighted survey estimators remain valid as estimators of superpopulation model parameters. A chaining argument explains the validity behind this reasoning. First, the weighted survey estimator consistently targets a finite population parameter, with asymptotic rate given as the inverse square root of sample size, via survey sampling theory. Second, the finite population parameter is the hypothetical estimator of the superpopulation parameter that would be computed if the entire finite population were observed. It is consistent, and its asymptotic rate is the inverse square root of population size, by classical statistical theory. The weighted survey estimator is thus a valid point estimator of the superpopulation parameter, with uncertainty that is dominated by the sampling error because the sample size is usually much smaller than the population size. The same argument works when superpopulation parameters are defined as solutions to finite-population estimating equations, which are themselves estimated with survey-weighted versions. Survey-weighted estimators and design-based estimation thus cover a wide range of inferential tasks, both descriptive and analytic, and are the basis for general-purpose statistical software that handles complex survey data (e.g., Stata `svy` methods, SAS survey procedures, and the R `survey` package).

Some further notation is needed to describe a variety of scenarios of combining information across samples. First, define two sampling frames, A and B, which list entities in the population of interest and might be incomplete. Let $F_k^A = 1$ if entity $k \in U$ is in frame A and $F_k^A = 0$ otherwise, and $F_k^B = 1$ if element $k \in U$ is in frame B and $F_k^B = 0$ otherwise.

Independent samples s_A and s_B are selected from the two frames, with sample membership indicators $I_k^A = 1$ if $F_k^A = 1$ and $k \in s_A$ and $I_k^A = 0$ otherwise, and $I_k^B = 1$ if $F_k^B = 1$ and $k \in s_B$ and $I_k^B = 0$ otherwise. Following the design-based approach, the indicators I_k^A and I_k^B are treated as stochastic. Define the inclusion probabilities:

$$\pi_k^A = \mathrm{P}\left[k \in s_A\right] = \mathrm{E}\left[I_k^A\right] \text{ when } F_k^A = 1; \ \pi_k^B = \mathrm{P}\left[k \in s_B\right] = \mathrm{E}\left[I_k^B\right] \text{ when } F_k^B = 1.$$

Several scenarios relevant in combining information can then be defined. First, s_A is a probability sample if frame A has complete coverage ($F_k^A \equiv 1$), $\pi_k^A > 0$ for all $k \in U$, and π_k^A is known for all $k \in s_A$. The Horvitz & Thompson (1952) (HT) estimator,

$$\widehat{T}_{y,\mathrm{HT}} = \sum_{k \in s_A} \frac{y_k}{\pi_k^A} = \sum_{k \in U} y_k \frac{I_k^A F_k^A}{\pi_k^A}, \qquad 1.$$

is then unbiased for T_y under repeated sampling, because $\mathrm{E}\left[I_k^A\right] = \pi_k^A$. A census is a special case of a probability sample, in which $\pi_k^A \equiv 1$ for all $k \in U$.

A frame may suffer from undercoverage, in which $F_k^A = 0$ for some $k \in U$. Entities not in the frame have no chance of being selected for the sample, and hence they are not represented in the study. In this case, the HT estimator has bias

$$\mathrm{E}\left[\widehat{T}_{y,\mathrm{HT}}\right] - T_y = \mathrm{E}\left[\sum_{k \in U} y_k \frac{I_k^A F_k^A}{\pi_k^A}\right] - \sum_{k \in U} y_k = -\sum_{k \in U} y_k (1 - F_k^A).$$

Even a well-designed and well-conducted survey will typically fail to meet the conditions of a probability sample exactly, either due to undercoverage or differential nonresponse, so that π_k^A are not known for all responding elements. But surveys intending to be representative of populations are careful to minimize coverage error, to use formal probability sampling designs, and to mitigate the effects of nonresponse. This mitigation includes reducing differential nonresponse through various follow-up and incentive strategies, followed by weighting adjustments to downweight those more likely to respond and upweight those less likely. The NHCS sample has no undercoverage because its frame of US hospitals is complete, but not all sampled hospitals respond, so that weighting adjustments are necessary.

3.4. Nonprobability Samples

Nonprobability samples, in contrast, are not representative of the population of interest, nor are they designed to be. They may arise from administrative processes; systems with voluntary participation, such as the crowdsourced COVID-19 survey in the Section 2.2.1 example; systems with unknown participation, such as the proprietary commercial data in the NHCS application; or other mechanisms not under the control of the researcher. A standard set of approaches for dealing with nonprobability samples is to treat them as if they arose from a random mechanism, with I_k^B independent, Bernoulli random variables with inclusion probabilities $\pi_k^B = \mathrm{P}\left[I_k^B = 1\right]$ that might be zero even when $F_k^B = 1$, or with π_k^B unknown for $k \in s_B$. The minimal conditions for unbiased estimation are therefore not met, and using these nonprobability data for population inference requires additional information to avoid bias.

3.5. Weighting Methods for Data Integration

Three key weighting methods for data integration are then calibration, inverse propensity weighting, and dual-frame estimation. These methods can be combined in various ways depending on the integration problem, including via doubly-robust estimation methods. We assume throughout that a common set of covariates, $\mathbf{x}_k$, is available on both samples. Unless otherwise specified, we assume complete coverage by both frames ($F_k^A \equiv 1, F_k^B \equiv 1$).

3.5.1. Calibration. Consider two probability samples, with y_k observed on the smaller sample, s_A, but not observed on the larger sample, s_B (which may be a census). Information on $\mathbf{x}_k$ may be predictive of y_k and can improve the precision of weighted estimates. This information can be incorporated into weighted estimation for A-only by modifying the HT weights to encode the B information. Deville & Särndal (1992) proposed a general class of calibration estimators with weights that are as close as possible, for a given distance measure, to the original HT weights $\{1/\pi_k^A\}_{k\in s_A}$ while satisfying the calibration constraints,

$$\sum_{k\in s_A} \omega_k \mathbf{x}_k^\top = \sum_{k\in s_B} \frac{\mathbf{x}_k^\top}{\pi_k^B} = \widehat{T}_{\mathbf{x},B}^\top.$$

For linear calibration (assuming without loss of generality that the specified matrix inverse exists), the weighted estimator is the generalized regression estimator,

$$\widehat{T}_{y,\text{cal}} = \sum_{k\in s_A} \left\{ \frac{1}{\pi_k^A} + \left(\widehat{T}_{\mathbf{x},B} - \widehat{T}_{\mathbf{x},A}\right)^\top \left(\sum_{k\in s_A} \frac{1}{\pi_k^A} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \frac{\mathbf{x}_k}{\pi_k^A} \right\} y_k = \sum_{k\in s_A} \omega_k y_k. \quad 2.$$

By changing the distance measure, other forms of the calibrated weights are possible (raking weights are one standard variant), though all yield asymptotic variance equivalent to the generalized regression estimator (Deville & Särndal 1992).

If frame A is incomplete, $\widehat{T}_{y,\text{HT}}$ is biased. If frame B has complete coverage, the calibration estimator not only improves precision by making predictions within the A-frame but also reduces undercoverage bias by making predictions outside the A-frame. The potential for both reducing undercoverage bias and increasing precision makes calibration standard practice in complex surveys. The methodology relies critically on high-quality data from censuses or large surveys (like the ACS). The calibration methodology can also be extended to novel applications like the proprietary microdata protection described in Section 3.6.

3.5.2. Inverse propensity weighting. If a study variable of interest is available on the nonprobability sample s_B, as with the proprietary patient encounter data in the NHCS application, then any population-level inference needs to address potential selection bias. One defensible approach is to borrow representation from a probability sample, s_A, by constructing inverse propensity weights. Since I_k^B are modeled as independent Bernoulli(π_k^B), the log-likelihood of self-selection into s_B is

$$\sum_{k\in U} \ln\left(\frac{\pi_k^B}{1-\pi_k^B}\right) I_k^B + \sum_{k\in U} \ln\left(1-\pi_k^B\right), \quad 3.$$

where π_k^B is specified to follow a parametric model, like logistic regression: $\text{logit}(\pi_k^B) = \mathbf{x}_k^\top \boldsymbol{\theta}$.

The first term in Equation 3 is observed for $I_k^B = 1$. The log-likelihood in Equation 3 would be feasible if the second term could be computed, which requires observed covariates $\{\mathbf{x}_k\}_{k\in U}$. If the covariates are not available for all $k \in U$, the second term can be estimated unbiasedly with the probability sample, yielding the pseudo-log-likelihood,

$$\sum_{k\in U} \ln\left(\frac{\pi_k^B}{1-\pi_k^B}\right) I_k^B + \sum_{k\in U} \ln\left(1-\pi_k^B\right) \frac{I_k^A}{\pi_k^A},$$

which can be maximized to obtain estimated π_k^B (Chen et al. 2020). These estimates can then be used to form inverse propensity weights, either directly as $\omega_k = (1/\widehat{\pi}_k^B)$, or as averages within groups with comparable participation propensities, formed by sorting and dividing the $\widehat{\pi}_k^B$ at quantile levels. (Similar methods are used for probability samples with nonresponse like the NHCS application, though it suffices if $\mathbf{x}_k$ are observed for the entire sample, not the entire population.)

The resulting weighted estimators are then sums over the s_B sample, $\widehat{T}_{y,\text{ipw}} = \sum_{k \in s_B} \omega_k y_k$. These estimators are approximately unbiased under correct specification of the propensities. Variations of the pseudo-log-likelihood approach are possible (Elliott & Valliant 2017, Valliant 2020), depending on the degree of known or expected overlap between s_A and s_B. This is an ongoing area of research.

3.5.3. Doubly-robust estimators. The possibility of incorrect specification of the propensity model has led to consideration of doubly-robust estimators (Chen et al. 2020), which guard against model misspecification. An example combining elements of $\widehat{T}_{y,\text{cal}}$ and $\widehat{T}_{y,\text{ipw}}$ is

$$\widehat{T}_{y,\text{doubly}} = \sum_{k \in s_B} \frac{y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_N}{\widehat{\pi}_k^B} + \sum_{k \in s_A} \frac{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_N}{\pi_k^A}.$$

If the propensity model is correctly specified, then $\widehat{T}_{y,\text{doubly}}$ is approximately unbiased even if the regression model does not correctly describe the mean structure of y_k, while if the regression model is correctly specified, then $\widehat{T}_{y,\text{doubly}}$ is approximately unbiased even if the propensity model is incorrect.

3.5.4. Dual-frame estimators. Methods for reducing bias due to undercoverage include using auxiliary information to predict the missing part of the population, as described in Section 3.5.1, or supplementing with one or more additional frames. Assume that

$$F_k^A + F_k^B - F_k^A F_k^B = 1 \text{ for all } k \in U,$$

so that neither frame necessarily covers U, but together the frames have complete coverage. This is the setting of dual-frame sampling, for which there are multiple valid estimators, including combined, separate, and multiplicity (for a review, see Mecatti & Singh 2014).

Because entities can enter the combined sample, $s_A \cup s_B$, via two independent paths, the combined probability of selection is

$$\mathrm{P}\left[k \in s_A \cup s_B\right] = \pi_k^A F_k^A + \pi_k^B F_k^B - \pi_k^A \pi_k^B F_k^A F_k^B.$$

If this combined probability is known for all $k \in s_A \cup s_B$, we can plug it into Equation 1 to construct the unbiased combined dual-frame estimator. While this estimator can have excellent statistical properties, including stable weights and low variance, its challenges include the need to identify which entities in s_A, if any, are duplicated with those in s_B, the need to know F_k^A and π_k^A for $k \in s_B$, and the need to know F_k^B and π_k^B for $k \in s_A$.

The separate dual-frame estimator avoids these challenges. It is obtained by forming three disjoint domains and summing the corresponding estimators: the HT estimator of the domain covered by frame A and not frame B, the HT estimator of the domain covered by frame B and not frame A, and the composite estimator for the overlap domain covered by both (λ times the A-frame estimator for the overlap plus $(1 - \lambda)$ times the B-frame estimator, for some compositing parameter $0 < \lambda < 1$). The separate estimator requires F_k^A (but not π_k^A) for entities in the s_B sample and F_k^B (but not π_k^B) for entities in the s_A sample. The parameter λ might be chosen in an ad hoc way (e.g., $\lambda = 0.5$), or to reflect relative sample sizes ($\lambda = n_A/(n_A + n_B)$), or to minimize the expected variance for one or more key y variables.

The multiplicity estimator does not require choice of λ but otherwise uses the same frame membership information as the separate estimator. Each design weight is divided by its multiplicity, meaning the number of frames on which the entity is present: $\{\pi_k^A(F_k^A + F_k^B)\}^{-1}$ if sampled from the A-frame or $\{\pi_k^B(F_k^A + F_k^B)\}^{-1}$ if sampled from the B-frame. An entity present on only one frame receives its full weight while an entity present on both frames receives half its weight. With complete coverage for both frames, this is the separate estimator with $\lambda = 0.5$, the estimator used

for the NHCS application. The multiplicity estimator is easily extended to any number of frames, as long as we can compute the multiplicity of each sampled element.

If s_A is a probability sample and s_B is a nonprobability sample, then any of the dual-frame estimators might be employed, once the π_k^B are estimated as described in Section 3.5.2.

3.6. Example: National Hospital Care Survey (Continued)

We now summarize our approach for the 2020 NHCS example introduced in Section 3.2. The analytic goal was to combine responding 2020 NHCS hospitals, s_A, with other data sources in order to produce nationally representative estimates (Breidt et al. 2023). Census-level information $\{\mathbf{x}_k\}_{k\in U}$ about hospitals was obtained from the Healthcare Cost and Utilization Project. A proprietary commercial data source, s_B, provided 2020 patient encounter data for participating hospitals. The participation mechanism for s_B is unknown.

Data integration then proceeded by (*a*) using s_A and $\{\mathbf{x}_k\}_{k\in U}$ to construct nonresponse-adjusted inverse propensity weights (Section 3.5.2), (*b*) modeling s_B as nonprobability and using $\{\mathbf{x}_k\}_{k\in U}$ to construct participation-adjusted propensity weights (Section 3.5.2), and (*c*) using a separate dual-frame estimator (Section 3.5.4) to combine the reweighted s_A and s_B and produce national estimates. An additional analytic step was to create a restricted-use data file that included none of the proprietary microdata from s_B. Instead, a reweighted version of s_A was constructed, using calibration methods (Section 3.5.1) to ensure that weighted estimates from s_A only agree with a vector of key national estimates from the combined data. Key national estimates include numbers of encounters by coarsened diagnosis codes, discharge status, length of stay, age group, sex, and newborn status.

4. BLENDING SAMPLES TO EXAMINE GENERALIZABILITY AND TRANSPORTABILITY OF HEALTH POLICY INTERVENTIONS

Dahabreh et al. (2020) and Shi et al. (2023) review methods to blend randomized controlled clinical trial (RCT) data with external sources to examine how generalizable RCT findings are to a target population from which the RCT participants are drawn. The target population is characterized by using a data source such as a survey, a patient population registry, or a real-world population such as all patients with a certain health condition in a health system. Carefully designed RCTs yield estimands with high internal validity but lacking external validity if the RCT sample and target population differ in important ways. Transportability—whether findings from one study population are relevant for another that is fully or partially outside of the target population—is also of interest. For example, Subbiah (2023) explores whether treatment effects estimated and used to guide health policy in one country are transportable to another. Degtiar & Rose (2023) comprehensively review generalizability and transportability.

Similar to blending probability and nonprobability survey samples, methods such as modeling, weighting, and doubly-robust estimation are used when assessing RCT generalizability. The latter literature also focuses on causal inference for RCTs, including explication of causal estimands and their associated identifiability assumptions. Despite the similarities, these literatures are somewhat parallel [see reviews by Shi et al. (2023) and Yang & Kim (2020)]. However, Keiding & Louis (2016, 2018) connect these two literatures when examining nonprobability web surveys.

4.1. Example: Potential Expansion of Health Care Payment Models

The CMS Innovation Center in the United States is congressionally mandated to develop new Medicare health care payment and service delivery models and demonstrate whether they improve patient care and/or lower cost. Some demonstration projects involve mandatory

provider participation, though many are voluntary. Regardless of whether provider assignment to the demonstration model is randomized or requires the use of causal inference methods for observational data, CMS's Office of the Actuary must determine whether to expand the implementation of successful models from the demonstration model participants to the population. This is fundamentally a generalizability question. Tipton & Olsen (2018) review propensity score subclassification, inverse probability weighting, and model-based approaches for addressing generalizability that are applicable.

5. COMBINING SOURCES OF INFORMATION TO PRODUCE ACCURATE GRANULAR ESTIMATES VIA SMALL AREA ESTIMATION

5.1. Motivation

Understanding where individuals with hearing loss live within the United States, and how their geographic distribution varies across demographic groups, can have important policy implications. It can help prioritize areas to target for raising awareness, allocate resources for prevention and treatment, and shed light on health disparities. However, when estimating hearing loss for counties by gender, age, and race and ethnicity, even the ACS, which samples approximately 3.5 million addresses per year, does not provide a large enough sample to yield accurate estimates. Furthermore, data on hearing loss in ACS is self-reported, and subject to error. The National Health and Nutrition Examination Survey (NHANES) collects hearing loss data measured by trained technicians, but the sample size is much smaller, at about 5,000 persons per year. To overcome these challenges, Rein et al. (2024) combine estimates from NHANES and ACS, as well as various other data sources including administrative records such as Medicare, to obtain more accurate estimates for these detailed levels. The modeling techniques they use to combine these data sources fall under the realm of small area estimation (SAE).

Often, there is an interest in obtaining granular estimates from survey data, but due to budget constraints, the available sample size is not enough to support accurate estimation for all of the desired domains using traditional, direct methods. It is usually not feasible with respect to time or budget to collect additional data at the desired level of granularity, as the cost would be prohibitive. SAE involves borrowing strength from auxiliary information to improve upon direct survey estimates, typically by combining the survey data with other data sources via modeling. For instance, there may be a need to produce estimates of the prevalence of a disease measured by a survey for small geographical entities such as counties, possibly cross-classified with demographic characteristics such as race, sex, and age group, as in the hearing loss example. Typically the survey's sample size would not support this, but if good predictors were available from auxiliary sources, SAE modeling could incorporate them to obtain more reliable estimates. Reviews on SAE include those by Pfeffermann (2013) and Rao & Molina (2015). While the granularity challenge is more pronounced for survey data, it can also surface for other data sources whenever domains are small.

The success of an SAE program depends on the availability of strong auxiliary information. Classic sources are administrative records and censuses. One can also combine information from various vintages of a given survey for surveys where data are collected periodically, or from different surveys that measure related characteristics. Another common strategy is to exploit spatial information among the domains. Other sources of covariates that are increasingly being used are those derived from commercial sources such as cell phone usage, satellite data, etc.

A prominent SAE application in health policy is the Small Area Health Insurance Estimates program administered by the US Census Bureau. Estimates produced for states and counties by demographic and economic characteristics can be used to understand the geographic distribution of health insurance coverage and to analyze disparities. Three additional examples are available on

the National Cancer Institute's dedicated SAE website (**https://sae.cancer.gov/**) on the following important topics: cancer risk factors and screening behaviors, tobacco use and policies, and cancer-related knowledge. More examples can be found in articles by Wakefield et al. (2020) and Zhang et al. (2021) and in the sections that follow.

In terms of the FCSM data quality framework, SAE enhances the utility and objectivity of data by enabling the production of estimates with granularity and accuracy. It can also help with timeliness in the sense that it may be necessary to pool many years of data to obtain reliable direct estimates for small areas, whereas SAE modeling can leverage other sources of information while calibrating the results to the time point of interest. Furthermore, SAE enhances accessibility, because it often enables the publication of estimates where the direct survey estimates would need to be suppressed due to quality filters and confidentiality concerns.

5.2. Types of Small Area Estimation Models

SAE models can typically be classified as area-level or unit-level. Area-level models assume relationships between the survey-weighted direct estimates (e.g., Equation 1 or 2) and auxiliary information. Unit-level models involve the survey microdata, which for health applications often is at the person level. For a discussion of the advantages of each, readers are directed to, for instance, Franco & Maitra (2023). Both types of models can be handled via the frequentist or Bayesian paradigms. The former one typically employs empirical best or empirical best linear unbiased prediction. Alternatively, one can use a hierarchical Bayes approach, where the posterior distributions of the parameters of interest are computed or approximated using techniques such as Markov chain Monte Carlo. Various software packages are available for fitting SAE models, and a listing of many of them is available in the United Nation's SAE toolkit (**https://unstats.un.org/wiki/display/SAE4SDG/**).

5.2.1. Fay–Herriot model. To give a taste of SAE modeling, we introduce a classic and commonly used area-level model, the Fay–Herriot (FH) model (Fay & Herriot 1979). Suppose y_i is a direct estimator of a quantity of interest θ_i, with $i = 1, \ldots, m$. For instance, θ_i could represent the prevalence of a disease in a county i. Denote the covariate vector as $\mathbf{x}_i$. The FH model is

$$y_i = \theta_i + e_i, \qquad i = 1, \ldots, m,$$
$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i,$$

where e_i is the sampling error in y_i, generally assumed to be $\mathrm{N}(0, v_i)$, and u_i is the area i random effect or model error, usually assumed to be independent and identically distributed $\mathrm{N}(0, \sigma_u^2)$ and independent of the e_i. The sampling variance v_i is typically assumed known for identifiability, but in practice it must be estimated from the microdata. In many applications, it is important to smooth the direct estimates of sampling variance as they can yield absurd estimates for small samples; for an example of this, readers are directed to Rein et al. (2024). You & Hidiroglou (2023) and Franco & Bell (2022) also provide examples of ways to smooth.

When the parameters are known, the best predictor (BP) of θ_i is

$$\hat{Y}_i^{\mathrm{BP}} = (1 - \gamma_i)y_i + \gamma_i \mathbf{x}_i'\boldsymbol{\beta},$$

where

$$\gamma_i = \frac{v_i}{v_i + \sigma_u^2}.$$

The smaller the sampling variance, the more weight is placed on the direct estimator. This is a desirable property because for large sample sizes, the direct estimator is very accurate. Note also that shrinkage to the synthetic estimator $\mathbf{x}_i'\boldsymbol{\beta}$ occurs.

In practice, the parameters are not known and must be estimated in an empirical best approach or assigned priors in a hierarchical Bayes approach. For the former, the parameters can be estimated via maximum likelihood, restricted maximum likelihood, etc. For the latter, often noninformative priors are selected. The FH model is just one example of an SAE model that is simple and appropriate in some situations, but a wealth of other choices of models are available in the literature (see Rao & Molina 2015).

5.2.2. Multivariate and measurement error models. The type of auxiliary information available typically dictates the types of models that are suitable. For instance, when combining information from different surveys, ignoring the sampling error of one or more of the surveys can result in suboptimal predictions and incorrect estimates of uncertainty measures (Bell et al. 2019). Appropriate models to capture this uncertainty include measurement error models and multivariate models, among others. One of the pioneering works about measurement error in SAE involved modeling the body mass index for 50 demographic subgroups defined by race and ethnicity using data from NHANES, with data from the National Health Interview Survey as a covariate with measurement error (Ybarra & Lohr 2008). A recent example of a bivariate model applied to address a health policy question uses estimates from the ACS to improve National Health Interview Survey health insurance coverage estimates at the state level (Franco & Bell 2022). A review on combining surveys in SAE is provided by Franco & Maitra (2023).

Bivariate models have also been used to combine estimates based on probability and nonprobability samples, offering an alternative to methods described in Section 3. For example, Ganesh et al. (2017) and Gupta et al. (2019) take this approach to increase the precision of estimates of the prevalence of self-reported and doctor-diagnosed food allergies in children and adults. The study used data from a food allergy survey conducted by NORC at the University of Chicago on behalf of Northwestern University, using NORC's AmeriSpeak® sample. The survey estimates were combined with estimates based on a nonprobability web panel from Survey Sampling International (now Dynata). The domains here were also demographic groups, defined by age, education, gender, and race and ethnicity. **Table 1** shows the subset of results reported by Ganesh et al. (2017) of allergies deemed either convincing or nonconvincing by expert assessment as described by Gupta et al. (2019). The second and third columns are based on descriptive survey statistics for the probability and nonprobability samples, respectively. The third column shows the standard FH model applied to the probability sample only, while the fourth column shows the results from bivariate modeling of the probability and nonprobability domain estimates. For all variables, the nonprobability sample mean estimates are quite far from estimates under the other approaches, though the large sample results in narrower confidence intervals than shown in the first column. As expected, the first and third columns are very similar, though the

Table 1 Convincing and nonconvincing self-reported allergies: comparison of adult prevalence estimates and 95% confidence intervals

Variable	Probability sample	Nonprobability sample	Fay-Herriot: univariate	Fay-Herriot: bivariate
Ever had food allergy	21.6 ± 1.5	28.1 ± 0.6	21.1 ± 1.0	21.3 ± 0.6
Peanut allergy	1.7 ± 0.4	5.1 ± 0.3	1.4 ± 0.2	1.4 ± 0.1
Milk allergy	4.4 ± 0.5	6.5 ± 0.3	4.0 ± 0.4	4.1 ± 0.1
Biological parent has a food allergy	11.9 ± 1.1	14.7 ± 0.5	11.5 ± 0.8	11.5 ± 0.5
Biological parent has environmental allergy	28.6 ± 1.9	29.6 ± 0.7	28.2 ± 1.2	27.9 ± 0.6

Results are from Ganesh et al. (2017).

small-area modeling results in narrow confidence intervals as expected. The bivariate FH model results once again in mean estimates that are very similar to the probability sample but with much narrower confidence intervals, demonstrating the benefit of small area modeling for reducing variance estimates.

5.2.3. Spatial, temporal, and spatiotemporal models. When combining several vintages of the same survey, temporal SAE models may be appropriate (see Rao & Yu 1994, Franco & Bell 2015). Spatial and spatiotemporal SAE models are suitable for exploiting geographic information such as adjacencies; for a review, readers are directed to Pratesi et al. (2023). A closely related topic is Bayesian disease mapping, which also makes use of spatial and spatiotemporal techniques to study diseases, though it is typically not based on survey data. Reviews with further discussion about the difference between spatial SAE models and Bayesian disease mapping include those of Wakefield et al. (2020) and Zhang et al. (2021).

5.3. Opportunities and Challenges in Applying Small Area Estimation in Health Policy Evidence-Building

The demand for disaggregated statistics is increasing worldwide, and SAE has become an extremely important approach to help meet this need. SAE enables the production of detailed estimates and of maps of diseases and other health-related quantities. These can help identify at-risk communities and help plan targeted interventions. Small area estimates can often be made publicly available in cases where the corresponding direct estimates would be suppressed due to poor quality or confidentiality concerns. There are several areas worthy of additional attention and research to expand SAE's value for health policy research.

5.3.1. Parametric model assumptions. For modeling the prevalence of rare illnesses, normality assumptions like those of the FH model may not be suitable due to skewness and the frequent occurrence of observations of zero, so other distributions should be considered. For example, since moderate or severe hearing loss is rare for younger age groups, Rein et al. (2024) use multinomial and binomial SAE models to model hearing loss at the county level for various demographic groups. Rao (2023) includes a review of nonparametric and semiparametric SAE approaches in the context of health disparities research.

5.3.2. Policy and statistical relevance of the shrinkage target. SAE typically involves shrinkage, which tends to decrease uncertainty, but the quantity to which the model shrinks is a crucial policy-relevant assumption. Consider hospital profiling, or public report cards of health care provider performance. A canonical health care quality reporting program is Medicare Hospital Compare (HC), for which quality-of-care measures for hospitals in the United States that provide care to Medicare beneficiaries are made available to the public on a website. The HC estimates for outcome measures, such as heart attack (myocardial infarction) mortality, are based on a random effects logit model that includes hospital random effects and adjusts for patient risk factors. Normand et al. (2016) provide HC model details and a broader review of hospital profiling. Both Normand et al. (2016) and George et al. (2017) address an issue that has been considered by health policy researchers for nearly 15 years: the lack of hospital-level predictors—most notably hospital volume, defined as the number of patients upon which a quality measure is based—in the HC model. George et al. (2017) show these factors are predictive of hospital performance in the HC data, with prior research indicating they are generally associated with patient outcomes. The issue is exacerbated by the fact that not only is volume an important predictor but it is also a driver of greater shrinkage and reliance on model assumptions. For low-volume hospitals, including volume and other hospital characteristics into the model results in more pessimistic performance

estimates than reported on HC. George et al. (2017) further relax model assumptions by using a nonparametric Bayesian approach. George et al. (2017) advocate for adding hospital volume to the HC model while Normand et al. (2016) present additional issues to investigate prior to making such a change, including endogeneity of volume, adequacy of risk adjustment for low-volume hospitals, and the potential for model specifications that would account for high-performing low-volume hospitals.

5.3.3. Predictive versus explanatory goals. SAE is primarily a predictive task, with the objective of estimating unobserved small area population quantities. Shmueli (2010) provides an overview of how predictive and explanatory models differ and warns against confounding the two—models designed to obtain good predictions may not be suitable to infer causal issues or even yield clear interpretation of the parameters. For the hearing loss SAE project mentioned earlier (Rein et al. 2024), the authors emphasize that the models they develop are designed to predict the small area prevalences but warn against inferring relationships based on the model parameter estimates alone.

5.3.4. Model validation. Model validation provides information that allows health policymakers to gain confidence in, and an understanding of, SAE results. The types of tools that are available are design- and model-based simulations, residual analysis, comparison of model aggregates with direct estimators for sample sizes where the latter are reliable, cross-validation exercises, etc. For more discussion on model validation and diagnostics, readers are directed to, for instance, Franco & Maitra (2023) or Brown et al. (2001). For complex problems, validation can be quite challenging and may require a customized approach.

Model selection, diagnostics, and validation in SAE are areas that require further research. For example, though there are various options available for variable selection, including likelihood-based metrics such as the Akaike information criteria and related metrics, cross-validation, and fence methods (Jiang 1996), there is not a consensus within the SAE literature on which methods are most appropriate or effective. Furthermore, comparing different model forms (e.g., binomial versus FH models) can be challenging, as discussed by Franco & Bell (2022) in the context of evaluating health insurance coverage as well as poverty. In the former applications, various metrics fail to distinguish among alternative model forms, despite the models yielding very different posterior variances. Posterior variances based on completely different models are not themselves comparable, but large differences in these across models indicate that selecting one model over others has practical implications. Design-based simulations can be effective tools to compare disparate model forms, but it can be difficult to perform them in a realistic manner, and they require an artificial population with similar characteristics as the phenomena being studied. For large surveys, the survey microdata can be used, but more complex approaches are often needed for small surveys. Another technique that is often used is to compare model aggregates with direct estimates at a level of aggregation where the latter are reliable. However, for smaller surveys, this strategy can be less informative because noise can be substantial. Such challenges can leave analysts and policymakers to judge the appropriateness of model assumptions without a good diagnostic.

5.3.5. Health disparities. More research is also needed to develop best practices for using SAE to study health disparities and, particularly, for disaggregating by quantities like race and ethnicity, gender assigned at birth, or sexual orientation and gender identity. Models that do this require thought and careful assessment of the underlying assumptions. This is another example where shrinkage can be a potential pitfall. When a model is fit primarily with data from a majority group—which is often the reality in surveys even when oversampling minorities—and the results are used to study disparities, careful validation is needed. Furthermore, researchers should

be aware of issues of measurement error. For example, Medicare records can have racial biases due to underdiagnosing of minorities for medical conditions or due to misclassification of race and ethnicity (Gianattasio et al. 2019, Huang & Meyers 2023). Such errors, if not addressed, can affect the validity of results.

6. METHODS TO EXPAND THE NUMBER OF VARIABLES IN A DATA SET

6.1. Record Linkage

The process of identifying and linking records that are believed to represent the same entity is known as record linkage (or entity resolution) (Christen 2012). The ability to link records across multiple sources of information is a key preprocessing step for the enhancement of data utility in the interest of addressing policy-related questions (Shlomo 2019). Record linkage of EHRs has been widely used for estimation of measures of health and health costs, and overall assessment of health programs, among other applications (Schenker & Raghunathan 2007; Gutman et al. 2013, 2016).

When unique identifiers such as Social Security numbers (SSNs) are available, linking records is a relatively straightforward process. For example, sources such as Medicare and Medicaid enrollment records, survey-collected EHRs like the National Ambulatory Medical Care Survey, and administrative records for vital statistics might share a common unique identifier that facilitates linkage. However, privacy regulations often limit access to SSNs, and this information is not obtainable in some cases. For instance, the Montgomery Cares Program is a group of community-based health care providers in Montgomery County, Maryland, that provides medical care to low-income uninsured adults. For Montgomery Cares, it is expected that the eligible population includes undocumented immigrants for whom unique numerical identifiers are not available.

In the absence of unique identifiers, record linkage becomes more challenging, and its accuracy is strongly tied to data quality and the discriminatory power of the available information. In these situations, the record linkage process relies on quasi-identifying linking variables, such as name, address, gender, and date of birth. However, in many cases, different sources of information contain corrupted, inaccurate, outdated, or missing information. For instance, measurement errors, missing names, and misspellings are common in survey data, leading to linkage error and possible biased results (Lariscy 2011, Harron et al. 2014, Gutman et al. 2016, Harron et al. 2017, Gilbert et al. 2018). When the quality of the source data used for integration is high, deterministic matching is a common approach. Otherwise, using probabilistic methods that account for possible errors in the linking variables, which pose a data quality threat to the accuracy and reliability of linked data sets, is customary.

6.1.1. Deterministic linkage. In practice, deterministic matching is based on a series of deterministic rules involving the comparison of record pairs. A very simple example is exact matching, where two records are linked if they agree exactly on all linkage variables (e.g., first name, last name, sex, date of birth, and zip code). Decision rules for a match can be applied iteratively such that a pair of records can be declared a match, even when they only match on a subset of the linking variables. The following is an example of iterative deterministic linkage:

Step 1: Two records must match on SSN and one of the following:

- First name, last name, and sex
- Last name, year of birth, sex, and zip code
- First name, year of birth, sex, and zip code

Step 2: If SSN is missing or does not match, two records must match on last name, first name, year of birth, sex, and one of the following:

- Seven to eight digits of the SSN
- Two or more of the following: day of birth, month of birth, middle initial, or zip code

In deterministic matching, the variables have equal weights, although some variables provide a higher discriminatory power. For instance, last name provides more identifying information compared with gender. In practice, however, it is useful to carry out deterministic linkage before a probabilistic approach for one-to-one matching to reduce computational costs.

6.1.2. Probabilistic linkage. Records across multiple sources are often noisy, and comparison rules should account for small variations (e.g., different name spellings). Fellegi & Sunter (1969) proposed an approach for probabilistic record linkage that is widely used in practice. This method considers two data sets, A and B, which contain an overlapping set of entities, and K common variables used for linkage. The Fellegi–Sunter model assumes a bipartite linkage structure where the files contain no duplicate records within, and there is a one-to-one correspondence between the overlapping entities in A and B. This approach assigns weights (or scores) to each linkage variable and relies on disagreement comparison vectors, $\boldsymbol{\gamma}_{ij} = (\gamma_{ij1}, \ldots, \gamma_{ijK})$, for each record pair (i, j). String similarity measures such as edit distance, Levenshtein, or Jaro–Winkler are used to compare string variables, e.g., names of people, streets, and institutions (Cohen et al. 2003). The model also assumes independence among record pairs and conditional independence of the linking variables given the matching status of a pair of records.

Based on the previous assumptions, the Fellegi–Sunter model assigns the following estimated weight to each record pair:

$$\hat{w}_{ij} = \ln\left(\frac{\hat{m}_{\boldsymbol{\gamma}_{ij}}}{\hat{u}_{\boldsymbol{\gamma}_{ij}}}\right),$$

where $m_{\boldsymbol{\gamma}_{ij}}$ and $u_{\boldsymbol{\gamma}_{ij}}$ represent the conditional probabilities of observing the comparison vector, $\boldsymbol{\gamma}_{ij}$, given a match and a nonmatch, respectively. These probabilities are unknown and are estimated using the expectation–maximization algorithm (Winkler 2014). In order to classify record pairs as matches, nonmatches, and possible matches, the weights, $\hat{w}_{ij}$, are compared with two fixed thresholds (Binette & Steorts 2022). Manual review is often necessary to further classify the pairs in the set of possible matches, and sometimes it is used to choose the optimal threshold values (Dusetzina et al. 2014). More often, the threshold values are selected by testing different values and choosing the ones that minimize the linkage error. Linkage error can arise from declaring a match between records representing different entities (false positives) or a nonmatch between records representing the same entity (false negatives). There is a trade-off between the two types of errors. One error type can be more acceptable than the other on a specific application, but in general, the goal is to minimize them.

Objectivity standards in the practice of record linkage for health data call for performance evaluation of the linkage method and sensitivity analysis to model assumptions. Many recent approaches for health data linkage applications rely on the Fellegi–Sunter method and its extensions (e.g., Goldstein et al. 2017, Li et al. 2022, Xu et al. 2022, Vo et al. 2023). However, a known limitation of the Fellegi–Sunter framework is the lack of transitive closures. That is, pairwise linkages need to be reconciled to resolve transitive matches. Moreover, the strong independence assumptions and the use of subjective thresholds can be restrictive. These caveats have motivated the development of alternative Bayesian approaches that overcome some of these limitations.

6.1.3. Bayesian record linkage. The Bayesian generalization of the Fellegi–Sunter model of Sadinle & Fienberg (2013) introduces a prior distribution on the linkage structure, which imposes transitive closures and can be applied to more than two files. Other Bayesian approaches include those of Tancredi & Liseo (2011), Steorts et al. (2014a, 2016), Steorts (2015), and Sadinle (2014, 2017). The main advantage of a Bayesian framework is the uncertainty quantification obtained through posterior distributions and subsequent error propagation. For instance, Sadinle (2018) proposes an uncertainty propagation method based on linkage-averaging that can be applied to various multi-file Bayesian partitioning record linkage approaches. Other recent work based on partitions includes that of Zanella et al. (2016), Betancourt et al. (2022a,b), and Aleshin-Guendel & Sadinle (2022).

Statistical analysis with linked data is an active area of research. In Bayesian settings, the propagation of linkage error can be achieved by jointly modeling and sampling the linkage structure and the parameters associated with the downstream task of interest, including, among others, population size estimation (Sadinle 2018, Tancredi et al. 2020), regression analysis (Gutman et al. 2013, Tancredi & Liseo 2015, Dalzell & Reiter 2018, Tang et al. 2020), and causal inference (Wortman & Reiter 2018, Guha et al. 2022, Guha & Reiter 2024).

Despite this progress, scalability remains a challenge for big data tasks. Recent work on scalability and efficient sampling algorithms for Bayesian methods has been done by McVeigh et al. (2020), Marchant et al. (2021), and Taylor et al. (2024). In practice, however, the use of blocking techniques is a standard procedure used to improve linkage accuracy and speed.

6.1.4. Blocking techniques. With the increasing size of health databases in terms of the number of both records and data elements, the number of comparisons to be made rapidly expands. The computational cost of record linkage can be ameliorated through blocking techniques. Blocking is used to prune the set of all possible record pairs to yield a smaller set of candidate matches a priori (Murray 2016). The most basic blocking method involves the selection of reliable attributes and places records in the same block if and only if they agree on a criterion based on those fields of information, such as gender, birth year, first name initial, or combinations thereof, for blocking in multiple passes. After that, record linkage is performed independently within each block.

A caveat of traditional blocking methods is that it is possible to exclude pairs that represent the same entity a priori (Steorts et al. 2014b). However, the basic approach to blocking is widely used in health care applications, and it is known to provide reliable results when proper blocking criteria are used. Recently, Campbell et al. (2021) presented a case study using a supervised machine learning approach for blocking in the health care context. The method is a modified version of the sequential covering algorithm that utilizes labeled training data to find an optimal blocking scheme (Michelson & Knoblock 2006). Campbell et al. (2021) show an application involving the linkage of two data sources: the 2016 NHCS, with 5.6 million records, and the CMS enrollment database, with 84.6 million records (National Center for Health Statistics 2020). The large size of the CMS data motivated the comparison of common blocking techniques and the sequential covering algorithm.

6.1.5. Privacy-preserving record linkage. In recent years, the use of EHRs has increased dramatically. While record linkage is a necessary tool to enhance data utility and inform the policy-making process in health care applications, it is important to point out that record linkage directly counteracts the purpose of data privacy and confidentiality by design. Linked data in combination with other sources can be vulnerable to linkage attacks by adversaries or hackers. As such, privacy-preserving record linkage (PPRL) attempts to identify records that refer to the same entities across multiple data sources while protecting the privacy of the entities represented by these

records (Hall & Fienberg 2010). PPRL techniques are key to achieving data quality standards for health policy under the integrity domain of the FCSM framework.

A variety of PPRL methods have been proposed in the literature, and perturbation-based approaches that change data elements through adding noise or making other modifications have become more commonly used in practice. Some perturbation methods use encryption or encoding to protect sensitive information. For instance, hashing is a form of cryptographic security that is used to transform keys or strings into different values. Kho et al. (2015) use hashing of patient identifiers to implement PPRL of EHR data across multiple health care sites in Chicago. Mirel et al. (2022) evaluate the potential of using PPRL techniques with NCHS survey data and administrative records. In this case, the privacy of identifiable data is protected through encryption by assigning hashes/tokens to groups of identifier fields. For example, a single token is used to represent the concatenation of last name, first name, sex, and date of birth.

6.1.6. Example: National Center for Health Statistics data linkage program. The work of Campbell et al. (2021) for blocking and Mirel et al. (2022) for PPRL are part of the NCHS data linkage program. The NHCS is only one of the NCHS health surveys that has been enhanced through record linkage as part of this program. The main goal of the program is to maximize the scientific value of NCHS population-based surveys to explore the factors that influence disability, chronic disease, health care utilization, morbidity, and mortality (Golden & Mirel 2021). In addition to CMS, NHCS has been linked to National Death Index, Housing and Urban Development, and Veterans Affairs data. Currently, the National Ambulatory Medical Care Survey federally qualified health center data are being linked to National Death Index, Housing and Urban Development, and CMS data.

6.2. Statistical Matching

Statistical matching, or data fusion, involves combining two (or more) data sets, possibly observed on different entities (D'Orazio et al. 2006). Statistical matching has been employed to produce synthetic data sets of individuals endowed with extensive sets of characteristics that subsequently become inputs to health policy microsimulations (Abraham 2013). Gutman et al. (2013) combine record linkage and statistical matching to integrate Medicare claims and cause of death from death certificates to analyze end-of-life medical costs.

A statistically matched data set has an incomplete data structure. The joint distribution can be obtained by invoking identifying assumptions. Specifically, let $\mathbf{Z}$ and $\mathbf{Y}$ be variables missing from data sets A and B, respectively, with $\mathbf{X}$ in both A and B. Standard identifying assumptions address the comparability of key distributions across the samples and conditional independence of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$. Weights can be developed and applied in analyses of the matched file to obtain estimates for a target population. Raghunathan et al. (2021) use multiple imputation to account for the uncertainty of statistically matching the Medicare Current Beneficiary Survey (MCBS) and NHANES data to conduct analyses that draw upon variables available only by blending the two surveys. Shi et al. (2023) discuss assumptions and methods in the context of causal inference.

6.2.1. Example (continued): impact of a health care payment system change. The statistical matching framework applies to the inpatient rehabilitation facility payment system example of Section 2.2.2 (Paddock et al. 2007). Let $\mathbf{X}$ be variables from claims and patient assessments that are believed to be consistently coded over time (e.g., demographics) in data sets A and B before and after implementation of the new system, respectively; let $\mathbf{Y}$ be relative cost and relative severity variables under prior-system coding practices available in A; and let $\mathbf{Z}$ be these variables but under new system coding practices and available in B.

6.2.2. Example: long-term health policy impacts. Evaluations of health policy interventions often terminate before important long-term outcomes can be observed, with practitioners defaulting to analyzing intermediate outcomes. Robbins et al. (2024) statistically matched the Oregon Health Insurance Experiment (OHIE) (Baicker et al. 2013) and the National Longitudinal Mortality Study (NLMS) data to examine the effect of the OHIE's lottery of a health insurance offer on long-term mortality as measured in NLMS 11 years after the first OHIE follow-up. NLMS contains the relevant long-term mortality data element, and integrating both data sets facilitates timeliness of statistical inference.

Use of surrogate markers from the clinical trials literature (Prentice 1989) provides an alternative that has recently received attention from policy researchers (Athey et al. 2019). Robbins et al. (2024) note that the analytic focus in the clinical surrogate markers context is a treatment effect on an intermediate outcome (surrogate) itself versus a long-term outcome. The connection between statistical matching and surrogate markers is illustrated by the similarity of identification assumptions made by Robbins et al. (2024) and Athey et al.'s (2019) surrogate index of multiple intermediate outcomes that proxies for a long-term outcome of interest: The intervention effect is explained entirely through the intermediate variable $\mathbf{X}$, and the conditional distribution of the long-term outcome given the intermediate variable is the same in both data sets A and B.

Statistical matching might enhance the ability of health policy makers to make timely programmatic decisions that incorporate information about critical long-term outcomes. However, the identifying assumptions are not testable, and so their plausibility and sensitivity of results to such assumptions, and also the risks should those assumptions not hold, must be carefully assessed.

7. DISCUSSION AND FUTURE DIRECTIONS

This review is motivated by the expanding opportunities for data integration in health policy and the need for principled approaches to blending multiple data sources. The value of data integration is particularly pronounced when data characterizing the population of interest can be used as an anchor for inferences when blending with observational data. As each data source has its strengths and limitations, considering data quality in the context of the focal health policy application is essential and provides a starting point from which to develop a statistical data integration approach using the methods described above. While a blended data set can provide better quality for a given analysis compared with any single input source, practitioners should assess the policy relevance of potential impacts of model assumptions on analytic conclusions.

Some areas for future research specific to methods are covered in the sections above. Here, the focus is on four directions for future research that are broadly applicable to data integration methods.

First, data integration increases privacy and confidentiality risks owing to increased granularity of information for those included in the data. Methods to access, integrate, and analyze blended data will continue to evolve accordingly. An example of this is CMS's current proposal to cease its distribution of health care claims data for storage and use on approved users' computing platforms in favor of virtual data access. Readers are directed to National Academies of Sciences, Engineering, and Medicine (2024) for a review of challenges and opportunities for addressing privacy and confidentiality of blended data, including synthetic data generation and PPRL, as mentioned in Section 6.1. One illustrative technical challenge is the proper application of methods that satisfy formal privacy criteria (e.g., differential privacy, or DP) to survey data. This is an open research area (Drechsler 2023) and one that cascades into further questions about DP when blending a survey with another data source. Another challenge is that DP and synthetic data generation require adding (modeled) noise to the data, which can reduce data utility, particularly

for small demographic subgroups (Brummet et al. 2022) including for health disparities analysis (Santos-Lozada et al. 2020). Slavković & Seeman (2023) provide a review of DP and traditional statistical disclosure limitation.

A second area in need of further research is methodology for obtaining informed consent from research participants for data integration. Struminskaya & Sakshaug (2023) summarize the few consistent findings between data privacy concerns and participant propensity to consent to blending of their data sources—specifically, willingness to provide consent is a function of respondents' own privacy and confidentiality concerns and their understanding of the linkage request. Respondents' propensity to consent to record linkage is also associated with their opinions about the data collection sponsor and the subject matter focus of the data to be linked. Thus, health policy researchers might want to evaluate whether available guidance on how to ask for consent adequately addresses these items.

A third future research area encompasses facilitating data integration at scale. Examples include enabling practical workflows to support integrated Bayesian record linkage and analysis (Binette & Steorts 2022) and data integration more generally (Steorts 2023), and emerging approaches such as nonparametric mass imputation to relax parametric assumptions in statistical matching (Chen et al. 2022).

Finally, effective communication of statistical concepts and methods (Schneider et al. 2024) will be essential for expanding the acceptance of data integration among health policymakers, most critically about the implications of modeling assumptions, as in the example in Section 6.2. Robust discussion of assumptions and methodological developments aimed toward exploring and relaxing critical assumptions underlying differences-in-differences modeling in recent years (Roth et al. 2023) provides a potential model for data integration.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abraham JM. 2013. Using microsimulation models to inform U.S. health policy making. *Health Serv. Res.* 48(2):686–95

Aleshin-Guendel S, Sadinle M. 2022. Multifile partitioning for record linkage and duplicate detection. *J. Am. Stat. Assoc.* 118(543):1786–95

Amaya A, Biemer PP, Kinyon D. 2020. Total error in a big data world: adapting the TSE framework to big data. *J. Surv. Stat. Methodol.* 8(1):89–119

Athey S, Chetty R, Imbens GW, Kang H. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely*. NBER Work. Pap. 26463

Bahamyirou A, Schnitzer ME. 2021. Data integration through outcome adaptive LASSO and a collaborative propensity score approach. arXiv:2103.15218 [stat.ME]

Baicker K, Chandra A. 2017. Evidence-based health policy. *New Engl. J. Med.* 377(25):2413–15

Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, et al. 2013. The Oregon experiment—effects of Medicaid on clinical outcomes. *N. Engl. J. Med.* 368(18):1713–22

Bell WR, Chung HC, Datta GS, Franco C. 2019. Measurement error in small area estimation: functional versus structural versus naïve models. *Surv. Methodol.* 45(1):61–80

Betancourt B, Sosa J, Rodríguez A. 2022a. A prior for record linkage based on allelic partitions. *Comput. Stat. Data Anal.* 172:107474

Betancourt B, Zanella G, Steorts RC. 2022b. Random partition models for microclustering tasks. *J. Am. Stat. Assoc.* 117(539):1215–27

Binette O, Steorts RC. 2022. (Almost) all of entity resolution. *Sci. Adv.* 8(12):eabi8021

Breidt FJ, Resnick D, Jackson G, White D. 2023. *Combining data sources to produce nationally representative estimates of hospital encounter characteristics*. Presented at the Federal Committee on Statistical Methodology 2023 Research and Policy Conference, Oct. 24–26, Hyattsville, MD

Brown G, Chambers R, Heady P, Heasman D. 2001. Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium 2001*. Ottawa, ON, Can.: Stat. Can.

Brummet Q, Mulrow E, Wolter K. 2022. The effect of differentially private noise injection on sampling efficiency and funding allocations: evidence from the 1940 census. *Harv. Data Sci. Rev.* Spec. Issue 2. **https://hdsr.mitpress.mit.edu/pub/ft9fhmku**

Campbell SR, Resnick DM, Cox CS, Mirelb LB. 2021. Using supervised machine learning to identify efficient blocking schemes for record linkage. *Stat. J. IAOS* 37(2):673–80

Chen S, Yang S, Kim JK. 2022. Nonparametric mass imputation for data integration. *J. Surv. Stat. Methodol.* 10(1):1–24

Chen Y, Li P, Wu C. 2020. Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.* 115(532):2011–21

Christen P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer Sci. Bus.

Cohen W, Ravikumar P, Fienberg S. 2003. A comparison of string distance metrics for name-matching tasks. In *IIWeb'03: Proceedings of the 2003 International Conference on Information Integration on the Web*, ed. S Kambhampati, CA Knoblock, pp. 73–78. Cambridge, MA: AAAI

Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernan MA. 2020. Extending inferences from a randomized trial to a new target population. *Stat. Med.* 39(14):1999–2014

Dalzell NM, Reiter JP. 2018. Regression modeling and file matching using possibly erroneous matching variables. *J. Comput. Graph. Stat.* 27(4):728–38

Degtiar I, Rose S. 2023. A review of generalizability and transportability. *Annu. Rev. Stat. Appl.* 10:501–24

Deville JC, Särndal CE. 1992. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* 87:376–82

D'Orazio M, Di Zio M, Scanu M. 2006. *Statistical Matching: Theory and Practice*. New York: Wiley

Drechsler J. 2023. Differential privacy for government agencies—are we there yet? *J. Am. Stat. Assoc.* 118(541):761–73

Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. 2014. *Linking data for health services research: a framework and instructional guide*. Tech. Rep., Agency Healthc. Res. Qual., Rockville, MD

Elliott MR, Valliant R. 2017. Inference for nonprobability samples. *Stat. Sci.* 32(2):249–64

Fay RE, Herriot RA. 1979. Estimation of income from small places: an application of James-Stein procedures to census data. *J. Am. Stat. Assoc.* 74:269–77

Fed. Comm. Stat. Methodol. 2020. *A framework for data quality*. Tech. Rep. FCSM 2020–04, Fed. Comm. Stat. Methodol., Washington, DC

Fellegi IP, Sunter AB. 1969. A theory for record linkage. *J. Am. Stat. Assoc.* 64(328):1183–210

Franco C, Bell WR. 2015. Borrowing information over time in binomial/logit normal models for small area estimation. *Stat. Transit. New Ser.* 16(4):563–84

Franco C, Bell WR. 2022. Using American Community Survey data to improve estimates from smaller US surveys through bivariate small area estimation models. *J. Surv. Stat. Methodol.* 10(1):225–47

Franco C, Maitra P. 2023. Combining surveys in small area estimation using area-level models. *Wiley Interdiscip. Rev. Comput. Stat.* 15(6):e1613

Ganesh N, Pineau V, Chakraborty A, Dennis JM. 2017. Combining probability and non-probability samples using small area estimation. In *JSM Proceedings, Survey Research Methods Section*, pp. 1657–67. Alexandria, VA: Am. Stat. Assoc.

George EI, Ročková V, Rosenbaum PR, Satopää VA, Silber JH. 2017. Mortality rate estimation and standardization for public reporting: Medicare's Hospital Compare. *J. Am. Stat. Assoc.* 112(519):933–47

Gianattasio KZ, Prather C, Glymour MM, Ciarleglio A, Power MC. 2019. Racial disparities and temporal trends in dementia misdiagnosis risk in the United States. *Alzheimer Dementia Transl. Res. Clin. Interv.* 5:891–98

Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, et al. 2018. GUILD: GUidance for Information about Linking Data sets. *J. Public Health* 40(1):191–98

Golden C, Mirel L. 2021. Enhancement of health surveys with data linkage. In *Administrative Records for Survey Methodology*, ed. AY Chun, MD Larsen, G Durrant, JP Reiter, pp. 271–96. New York: Wiley
Goldstein H, Harron K, Cortina-Borja M. 2017. A scaling approach to record linkage. *Stat. Med.* 36(16):2514–21
Groves R, Fowler F, Couper M, Singer E, Tourangeau R. 2004. *Survey Methodology*. New York: Wiley
Guha S, Reiter JP. 2024. Regression-assisted Bayesian record linkage for causal inference in observational studies with covariates spread over two files. *J. Stat. Plan. Inference* 229:106090
Guha S, Reiter JP, Mercatanti A. 2022. Bayesian causal inference with bipartite record linkage. *Bayesian Anal.* 17(4):1275–99
Gupta RS, Warren CM, Smith BM, Jiang J, Blumenstock JA, et al. 2019. Prevalence and severity of food allergies among US adults. *JAMA Netw. Open* 2(1):e185630
Gutman R, Afendulis CC, Zaslavsky AM. 2013. A Bayesian procedure for file linking to analyze end-of-life medical costs. *J. Am. Stat. Assoc.* 108(501):34–47
Gutman R, Sammartino C, Green T, Montague B. 2016. Error adjustments for file linking methods using encrypted unique client identifier (eUCI) with application to recently released prisoners who are HIV+. *Stat. Med.* 35(1):115–29
Hall R, Fienberg SE. 2010. Privacy-preserving record linkage. In *Privacy in Statistical Databases (PSD 2010)*, ed. J Domingo-Ferrer, E Magkos, pp. 269–83. New York: Springer
Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, et al. 2017. Challenges in administrative data linkage for research. *Big Data Soc.* 4(2). **https://doi.org/10.1177/2053951717745678**
Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. 2014. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med. Res. Methodol.* 14:36
Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663–85
Huang AW, Meyers DJ. 2023. Assessing the validity of race and ethnicity coding in administrative Medicare data for reporting outcomes among Medicare advantage beneficiaries from 2015 to 2017. *Health Serv. Res.* 58(5):1045–55
Jiang J. 1996. REML estimation: asymptotic behavior and related topics. *Ann. Stat.* 24(1):255–86
Keiding N, Louis TA. 2016. Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. R. Stat. Soc. Ser. A* 179(2):319–76
Keiding N, Louis TA. 2018. Web-based enrollment and other types of self-selection in surveys and studies: consequences for generalizability. *Annu. Rev. Stat. Appl.* 5:25–47
Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, et al. 2015. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *J. Am. Med. Inf. Assoc.* 22(5):1072–80
Lariscy JT. 2011. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *J. Aging Health* 23(8):1263–84
Li X, Xu H, Grannis S. 2022. The data-adaptive Fellegi-Sunter model for probabilistic record linkage: algorithm development and validation for incorporating missing data and field selection. *J. Med. Internet Res.* 24(9):e33775
Maksut J, Hodge C, Van C, Razmi A, Khau M. 2021. *Utilization of Z codes for social determinants of health among Medicare fee-for-service beneficiaries, 2019*. Data Highlight 24, Cent. Medicare Medicaid Serv. Off. Minor. Health, Baltimore, MD
Marchant NG, Kaplan A, Elazar DN, Rubinstein BIP, Steorts RC. 2021. d-blink: Distributed end-to-end Bayesian entity resolution. *J. Comput. Graph. Stat.* 30(2):406–21
McVeigh BS, Spahn BT, Murray JS. 2020. Scaling Bayesian probabilistic record linkage with post-hoc blocking: an application to the California Great Registers. arXiv:1905.05337 [stat.ME]
Mecatti F, Singh AC. 2014. Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. *J. Soc. Fr. Stat.* 155(4):51–69
Michelson M, Knoblock CA. 2006. Learning blocking schemes for record linkage. In *AAAI'06: Proceedings of the 21st National Conference on Artificial Intelligence*, Vol. 6, pp. 440–45. Cambridge, MA: AAAI
Mirel LB, Banks J, Breidt J, Finamore J, Mannshardt E, et al. 2023. A data quality scorecard to assess a data source's fitness for use. In *2023 Big Data Meets Survey Science (BigSurv)*. New York: IEEE

Mirel LB, Resnick DM, Aram J, Cox CS. 2022. A methodological assessment of privacy preserving record linkage using survey and administrative data. *Stat. J. IAOS* 38(2):413–21

Murray J. 2016. Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *J. Priv. Confidentiality* 7(1):3–24

National Academies of Sciences, Engineering, and Medicine. 2024. *Toward a 21st century national data infrastructure: managing privacy and confidentiality risks with blended data*. Tech. Rep., Natl. Acad., Washington, DC

National Center for Health Statistics. 2020. *The linkage of the 2016 National Hospital Care Survey to the 2016/2017 Centers for Medicare & Medicaid Services Medicare enrollment, claims/encounters and assessment data: matching methodology and analytic considerations*. Tech. Rep., Div. Anal. Epidemiol., Hyattsville, MD

Normand SLT, Ash AS, Fienberg SE, Stukel TA, Utts J, Louis TA. 2016. League tables for hospital comparisons. *Annu. Rev. Stat. Appl.* 3:21–50

Paddock SM, Escarce JJ, Hayden O, Buntin MB. 2007. Did the Medicare inpatient rehabilitation facility prospective payment system result in changes in relative patient severity and relative resource use? *Med. Care* 45(2):123–30

Pfeffermann D. 2013. New important developments in small area estimation. *Stat. Sci.* 28(1):40–68

Pratesi M, Marchetti S, Giusti C, Salvati N. 2023. The use of spatial information in area-level models: an evaluation based on auxiliary data availability. *Calcutta Stat. Assoc. Bull.* 75(2):155–72

Prentice R. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* 8(4):431–40

Raghunathan T, Ghosh K, Rosen A, Imbriano P, Stewart S, et al. 2021. Combining information from multiple data sources to assess population health. *J. Surv. Stat. Methodol.* 9(3):598–625

Rao JN, Molina I. 2015. *Small Area Estimation*. New York: Wiley

Rao JN, Yu M. 1994. Small-area estimation by combining time-series and cross-sectional data. *Can. J. Stat.* 22(4):511–28

Rao JS. 2023. *Statistical Methods in Health Disparity Research*. Boca Raton, FL: CRC

Rein DB, Franco C, Reed NS, Herring-Nathan ER, Lamuda PA, et al. 2024. The prevalence of bilateral hearing loss in the United States in 2019: a small area estimation modelling approach for obtaining national, state, and county level estimates by demographic subgroup. *Lancet Reg. Health Am.* 30:100670

Robbins MW, Bauhoff S, Burgette L. 2024. Data fusion for predicting long-term program impacts. *Stat. Med.* 43:3702–22

Roth J, Sant'Anna PH, Bilinski A, Poe J. 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J. Econom.* 235(2):2218–44

Sadinle M. 2014. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* 8(4):2404–34

Sadinle M. 2017. Bayesian estimation of bipartite matchings for record linkage. *J. Am. Stat. Assoc.* 112(518):600–12

Sadinle M. 2018. Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Ann. Appl. Stat.* 12(2):1013–38

Sadinle M, Fienberg SE. 2013. A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Am. Stat. Assoc.* 108(502):385–97

Santos-Lozada AR, Howard JT, Verdery AM. 2020. How differential privacy will affect our understanding of health disparities in the United States. *PNAS* 117(24):13405–12

Schenker N, Raghunathan T. 2007. Combining information from multiple surveys to enhance estimation of measures of health. *Stat. Med.* 26:1802–11

Schneider CR, Kerr JR, Dryhurst S, Aston JA. 2024. Communication of statistics and evidence in times of crisis. *Annu. Rev. Stat. Appl.* 11:1–26

Shi X, Pan Z, Miao W. 2023. Data integration in causal inference. *Wiley Interdiscip. Rev. Comput. Stat.* 15(1):e1581

Shlomo N. 2019. Overview of data linkage methods for policy design and evaluation. In *Data-Driven Policy Impact Evaluation*, ed. N Crato, P Paruolo, pp. 47–65. New York: Springer

Shmueli G. 2010. To explain or to predict? *Stat. Sci.* 25(3):289–310

Slavković A, Seeman J. 2023. Statistical data privacy: a song of privacy and utility. *Annu. Rev. Stat. Appl.* 10:189–218
Steorts RC. 2015. Entity resolution with empirically motivated priors. *Bayesian Anal.* 10(4):849–75
Steorts RC. 2023. A primer on the data cleaning pipeline. *J. Surv. Stat. Methodol.* 11(3):553–68
Steorts RC, Hall R, Fienberg SE. 2014a. SMERED: a Bayesian approach to graphical record linkage and de-duplication. *Proc. Mach. Learn. Res.* 33:922–30
Steorts RC, Hall R, Fienberg SE. 2016. A Bayesian approach to graphical record linkage and de-duplication. *J. Am. Stat. Assoc.* 111(516):1660–72
Steorts RC, Ventura SL, Sadinle M, Fienberg SE. 2014b. A comparison of blocking methods for record linkage. In *Privacy in Statistical Databases (PSD 2014)*, pp. 253–68. New York: Springer
Struminskaya B, Sakshaug JW. 2023. Ethical considerations for augmenting surveys with auxiliary data sources. *Public Opin. Q.* 87(S1):619–33
Subbiah V. 2023. The next generation of evidence-based medicine. *Nat. Med.* 29(1):49–58
Tancredi A, Liseo B. 2011. A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* 5(2B):1553–85
Tancredi A, Liseo B. 2015. Regression analysis with linked data: problems and possible solutions. *Statistica* 75:19–35
Tancredi A, Steorts R, Liseo B. 2020. A unified framework for de-duplication and population size estimation. *Bayesian Anal.* 15:633–82
Tang J, Reiter J, Steorts R. 2020. Bayesian modeling for simultaneous regression and record linkage. In *Privacy in Statistical Databases (PSD 2020)*, ed. J Domingo-Ferrer, K Muralidhar, pp. 209–23. New York: Springer
Taylor I, Kaplan A, Betancourt B. 2024. Fast Bayesian record linkage for streaming data contexts. *J. Comput. Graph. Stat.* 33:833–44
Tipton E, Olsen RB. 2018. A review of statistical methods for generalizing from evaluations of educational interventions. *Educ. Res.* 47(8):516–24
Valliant R. 2020. Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* 8(2):231–63
Vo TH, Chauvet G, Happe A, Oger E, Paquelet S, Garès V. 2023. Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system. *Comput. Stat. Data Anal.* 179:107656
Wakefield J, Okonek T, Pedersen J. 2020. Small area estimation for disease prevalence mapping. *Int. Stat. Rev.* 88(2):398–418
Winkler WE. 2014. Matching and record linkage. *Wiley Interdiscip. Rev. Comput. Stat.* 6(5):313–25
Wortman JH, Reiter JP. 2018. Simultaneous record linkage and causal inference with propensity score subclassification. *Stat. Med.* 37(24):3533–46
Xu H, Li X, Grannis S. 2022. A simple two-step procedure using the Fellegi-Sunter model for frequency-based record linkage. *J. Appl. Stat.* 49(11):2789–804
Yang S, Kim JK. 2020. Statistical data integration in survey sampling: a review. *Jpn. J. Stat. Data Sci.* 3(2):625–50
Ybarra LM, Lohr SL. 2008. Small area estimation when auxiliary information is measured with error. *Biometrika* 95(4):919–31
You Y, Hidiroglou M. 2023. Application of sampling variance smoothing methods for small area proportion estimation. *J. Off. Stat.* 39(4):571–90
Zanella G, Betancourt B, Wallach H, Miller J, Zaidi A, Steorts RC. 2016. Flexible models for microclustering with application to entity resolution. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1425–33. Red Hook, NY: Curran
Zhang X, Liu B, Yun S, Phillips RL. 2021. Small area estimation and Bayesian disease mapping for minority health and health disparities. In *The Science of Health Disparities Research*, ed. I Dankwa-Mullan, EJ Pérez-Stable, KL Gardner, X Zhang, AM Rosario, pp. 203–20. New York: Wiley