New Data Sources for Demographic Research

Casey F. Breen and Dennis M. Feehan

We are in the early stages of a new era of demographic research that offers exciting opportunities to quantify demographic phenomena at a scale and resolution once unimaginable. These scientific possibilities are opened up by new sources of data, such as the digital traces that arise from ubiquitous social computing, massive longitudinal datasets produced by the digitization of historical records, and information about previously inaccessible populations reached through innovations in classic modes of data collection. In this commentary, we describe five promising new sources of demographic data and their potential appeal. We identify cross-cutting challenges shared by these new data sources and argue that realizing their full potential will demand both innovative methodological developments and continued investment in high-quality, traditional surveys and censuses. Despite these considerable challenges, the future is bright: these new sources of data will lead demographers to develop new theories and revisit and sharpen old ones.

Introduction

Social and technological change have produced new sources of data that offer exciting opportunities to quantify demographic phenomena at scales and resolutions once unimaginable. These new data sources include digital traces that arise from ubiquitous social computing, massive longitudinal datasets made possible by the digitization of historical and administrative records, large-scale geospatial and genetic datasets, and data about previously inaccessible populations reached through innovations in traditional data collection. In the near term, these new data sources may advance measurement by allowing researchers to observe intricate patterns of human mobility and migration, population-scale measurements of social connectedness, day-by-day records of cultural change, and

Casey F. Breen, Department of Sociology, Leverhulme Centre for Demographic Science, and Nuffield College. University of Oxford, Oxford OX1 2JD, UK. E-mail: casey.breen@demography.ox.ac.uk. Dennis M. Feehan, Department of Demography. University of California, Berkeley 94720, USA. E-mail: feehan@berkeley.edu.

This article is part of PDR's 50th anniversary special issue, **Looking Backward, Looking Forward: Celebrating 50 Years of *Population and Development Review***.

generation-by-generation trajectories of socioeconomic well-being. In the long term, we expect the measurements of these new sources of data enable to lead demographers to revisit and sharpen old theories and to develop and test new ones.

The future looks bright: There have been some immediate gains from these new data sources, and there are more on the horizon; it seems possible that some areas of demographic research will experience a revolution. But this will not happen automatically: realizing the potential of these data sources depends on our ability to develop methods and disciplinary norms to address new, cross-cutting challenges that include understanding data provenance, data access, representativeness, and inference. It will also be critical to continue to put resources into traditional sources of demographic data, as these new sources of demographic data all depend on traditional data sources to be useful.

The first half of this forward-facing commentary begins by describing five exciting "new" data sources. These five new data sources are far from exhaustive, but they were chosen for two reasons: first, they represent, in our subjective assessment, the largest developments in the data infrastructure space; second, they can be used to illustrate the methodological advances needed to harness the full potential of these new data sources. We provide a high-level overview of each of these five new data sources, describe how they may be used to advance demographic measurement and theory, and highlight their reliance on traditional data sources.

The second half of this commentary focuses on identifying and discussing methodological challenges that cut across these new data sources: data provenance, usability, representativeness, and inference. We explain how the challenges arise, what efforts have been made to address them so far, and how we anticipate the field will make progress on these challenges in the coming years. We conclude with a broader discussion of the implications of these new data sources for future prospects and disciplinary norms.

Five exciting new sources of data

To illustrate how new data sources are stimulating exciting research in demography and related fields, we chose five examples: big population microdata, digital trace data, advances in traditional modes of data collection, geospatial data, and genetic data.[1] Our goal is to highlight distinctive features of each type of data[2] and to highlight the promise that each seems to offer (Table 1).

Big population microdata

Big population microdata have arisen from the digitization and linkage of large-scale administrative and census records (Ruggles 2014). For instance,

TABLE 1 Overview of new sources for demographic research and their relationship with traditional data sources

New data sources for demographic research			
Data source	**Areas of impact**	**Typical relationship with traditional data sources**	**Selected citations**
Big population microdata	Life-course dynamics, intergenerational mobility, causal determinants of longevity	Links traditional survey, administrative, and census data together using record linkage techniques	Halpern-Manners et al. (2020); Abramitzky, Boustan, and Eriksson (2012); Graetz et al. (2023)
Digital trace data	Measurement of key demographic rates, immigrant social integration, effects of digitization	Uses traditional surveys and censuses as ground truth to validate samples, calibrate models	Nobles, Cannon, and Wilcox (2022); Bruch and Newman (2018); Rampazzo et al. (2021)
Geospatial data	Small-area population estimates, nowcasting, climate change and population dynamics, demography of crises	Uses traditional data sources as key inputs in models	Leasure et al. (2020); Blumenstock, Cadamuro, and On (2015); Chi et al. (2022)
Genetic data	Fertility, determinants of longevity, genetic underpinnings of social networks and behaviors, social stratification	Relies on traditional data for outcomes of interest, re-weighting	Fletcher et al. (2023); Robinson et al. (2017); Barban et al. (2016)
Advances in Classical Data Collection	Hard-to-reach populations, unpack heterogeneous demographic subgroups (e.g., "Asian Americans")	Uses traditional data sources to construct sampling frames, weighting targets, and benchmarks	Breza et al. (2023); Schneider, and Harknett (2022); Feehan and Cobb (2019); Ruiz, Noe-Bustamante, and Shah (2023)

the U.S. Census Bureau is now digitizing the decennial complete-count censuses from 1950 to 1990 using modern cloud-based computing infrastructure and optical character recognition to process over 4 petabytes (4000 terabytes) of digitized imagery (Genadek and Alexander 2022). In parallel, new machine-learning based methods for record linkage have greatly increased the accuracy and precision of linkages (Helgertz et al. 2022; Abramitzky et al. 2021; Martha Michael Bailey et al. 2020). Together, these advances are facilitating the creation of massive longitudinal datasets combining census records, birth certificates, military enlistment records, education records, voting files, and IRS tax records. In the United States, several large-scale projects have created publicly available datasets tracking millions of individuals over the life course (Abramitzky et al. 2020; J. R. Goldstein et al. 2021; Ruggles et al. 2020). In Europe, where population register data have long been used for demographic research (S. Goldstein 1964; Baccaini and Courgeau 1996), linking together population register data will allow demographers to better study migration decisions, fertility, and marital history (Poulain, Herm and Depledge 2013; Thorvaldsen, Andersen and Sommerseth 2015). Digitization efforts of historical Chinese government records allow for the study of elites and government bureaucracy (Campbell and Lee, 2006; Song and Campbell 2017; Chen et al. 2020). Similar efforts in the private sector are producing massive databases with information on credit history (Hurley and Adebayo 2016), voting and political behavior (Nickerson and Rogers 2014), and more.

Linked big population microdata depend heavily on investment in high-quality traditional survey, census, and administrative data since those data form the fundamental elements that are combined into powerful longitudinal datasets through record linkage. Maintaining robust investments in these foundational data sources is essential to ensure a rich data ecosystem in the future.

The size and scope of big population microdata have already allowed researchers to investigate topics as varied as social mobility across generations (Mare 2011; Song 2021; Ward 2023; Feigenbaum 2018), changes in ethnoracial identification and classification (Saperstein and Gullickson 2013; Liebler et al. 2017; Saperstein and Penner 2012; Loveman and Muniz 2007; B. Duncan and Trejo 2011), changing patterns of intergenerational coresidence (Ruggles 2007), and the social determinants of mortality (Fletcher and Noghanibehambari 2021; Halpern-Manners et al. 2020; Noghanibehambari and Engelman 2022).

A key benefit of linked longitudinal microdata is that it allows social scientists to study the experiences of actual cohorts over the life course instead of relying on synthetic cohorts derived from period data. Historical longitudinal microdata is also not typically subject to the kind of disclosure concerns that prevent individually identifiable information from being used by researchers, making analysis at very fine levels of granularity

feasible. This expanded microdata landscape will likely enable researchers to deepen and improve theories of life-course population processes such as social determinants of health and mortality, environmental change and demographic behavior, migration patterns, fertility intentions and outcomes, heterogeneous demographic outcomes for smaller population subgroups, and more.

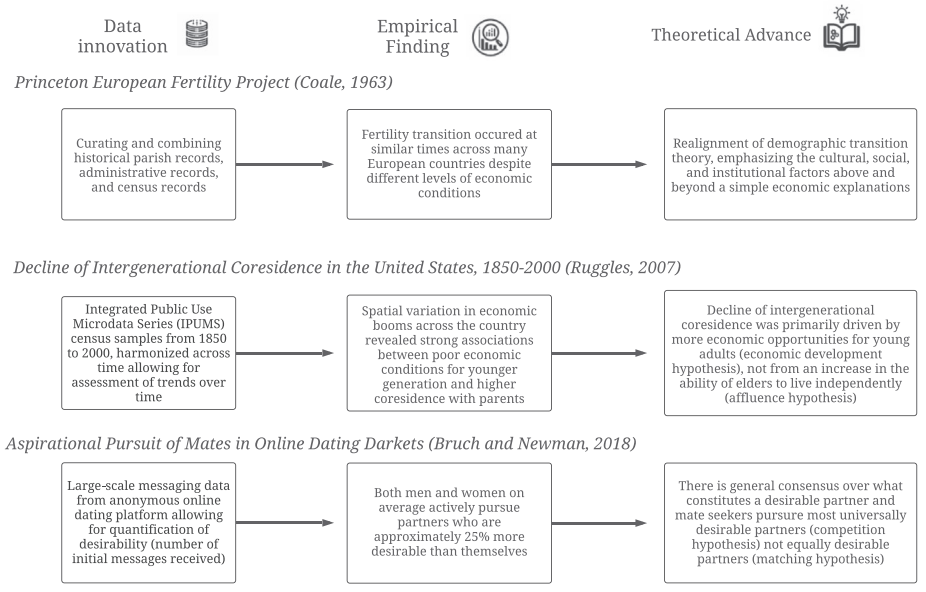
Digital trace data

Digital trace data arise as a by-product of people's interactions with social computing platforms, such as blogs, social media, dating apps, and cell phone networks. These digital traces can contain rich information about movement, social interactions, social networks, culture, and many other phenomena. Digital trace data capture a wide range of information, including social media posts, connections, messaging patterns, smartphone location data, group memberships, and user statistics on various platforms. In some cases, digital trace data can be obtained for free or for a modest cost. However, the richest digital trace data are often only available through some kind of partnership or collaboration with a private company.

Digital trace data have a distinctive feature: the data themselves result from tremendous social change that, in many cases, can directly affect processes of interest to social scientists. For example, digital trace data may prove to be helpful in estimating rates of migration between countries—but the social platforms that produce digital trace data may also begin to play an important role in the process of migration itself (Pesando et al. 2021). Another example: disparities in Internet access and mobile phone ownership have become a salient dimension of population inequality, potentially influencing factors like health and mortality (Flückiger and Ludwig 2023; Byaro, Rwezaula and Ngowi 2023), fertility (Byaro, Rwezaula and Ngowi 2023), migration (Grubanov-Boskovic et al. 2021), wealth (Edquist et al. 2018), well-being (Rotondi et al. 2020), and more.

To date, the primary application of digital trace data has been to improve the measurement of key demographic quantities. Researchers have used both mobile phone call logs and social media data to "nowcast" migration rates and stocks (Blumenstock 2012; Zagheni, Weber, and Gummadi 2017; Alexander, Polimis, and Zagheni 2020; Palotti et al. 2020; Leasure et al. 2023), produce high-resolution estimates of poverty (Blumenstock, Cadamuro and On 2015; Chi et al. 2022), and better measure behavior and friendship within networks (Eagle, Pentland, and Lazer 2009; Michael Bailey et al. 2018; Lewis et al. 2008; Eagle, Macy, and Claxton 2010). These approaches often combine unrepresentative but up-to-date digital trace data with representative—yet outdated—conventional surveys (Rampazzo et al. 2021; Alexander, Polimis, and Zagheni 2020). Without this traditional data for benchmarking, calibrating, and weighting, the practical utility of digital trace data can be limited.

FIGURE 1 Overview of three studies where innovations in data lead to advances in demographic theory



The majority of research to date using digital trace data takes advantage of the timeliness of these data, but less has been done studying trends over time. This represents a promising avenue for future research: exploiting the high-level of temporal detail offered by digital trace data could offer a richer, more detailed picture of demographic changes and trends as they unfold.

Recently, some applications of digital trace data have begun to expand beyond measurement into building and testing theory. For instance, as shown in Figure 1, Bruch and Newman (2018) used data from an online dating platform to test theories of partner pursuit, finding both men and women pursue partners who are more desirable than themselves ("competition hypothesis"). Another recent study used a collection of geolocated tweets to identify migrants and assess theories of immigrant integration using the time taken for a user to tweet in the host country's native language as an integration metric (Gil-Clavel, Grow, and Bijlsma 2023). And, Nobles, Cannon, and Wilcox (2022) used data on 1.6 million menstrual cycles from a menstrual tracking app to assess physiological limitations in detecting pregnancy before fetal cardiac activity. We hope these contributions represent early-stage applications that demonstrate how digital trace data can be used to build and test demographic theories in the future.

Geospatial data

Advances in geospatial data have significantly enriched our ability to add an important geographic component to analyses of population dynamics. Remotely sensed imagery from satellites, aerial photography, and images captured by drones can offer near real-time data on human settlements, urbanization trends, and environmental contexts that affect populations (Grace et al. 2019). Recent applications of these data include estimates of population displacement in humanitarian emergencies (Checchi et al. 2013); microestimates of wealth (Blumenstock, Cadamuro, and On 2015; Chi et al. 2022); targeted humanitarian aid allocation during the COVID-19 pandemic (Aiken et al. 2022); and high-resolution population estimates (Leasure et al. 2020). One advantage of remotely sensed data is that, by design, they generally have good coverage in remote, data-sparse settings.

The spatial dimension in many central questions in demography is often overlooked due to data limitations (Matthews et al. 2021). Researchers can now better address complex questions about human–environment interactions, such as the impacts of climate change on migration (Hauer, Jacobs, and Kulp 2024). The scale and fine geographic resolution of many new data sources can provide valuable insights into small-area phenomena. For instance, Levy et al. (2022) uses neighborhood-level demographics and digital trace data from 45 million mobile devices to predict COVID-19 infections, finding neighborhood characteristics before the pandemic strongly predicted COVID-19 incidence rates.

This wealth of geospatial data allows population scientists to develop new and refine old theoretical models that consider the spatial variability in demographic processes. However, to date, the application of large-scale geospatial data in demographic theory has been limited; this is an important area for focus in the coming years.

Genetic data

Early studies in behavioral genetics used twins and other kin relations to identify associations between genetic relatedness and behavioral traits; however, these approaches require strong assumptions for their estimates to be interpreted causally and generally do not provide a way to identify the specific genes that are associated with a behavioral trait (Friedman, Banich, and Keller 2021). Over the past few decades, technological advances in molecular genetics have made it increasingly possible to measure detailed characteristics of individual genomes in a cost-effective way. It is becoming common for major data collection projects, like longitudinal social science surveys, to include genetic measurements along with other biomarkers; further, commercial services like 23andMe and ancestry.com have made detailed genetic testing available to the general public. We can therefore

expect to see genetic data and social science genomics playing an increasing role in demography and other social sciences in the coming years.

So far, the expansion of genetic data has been accompanied by major challenges, helping to highlight how the development of methods can be critical to progress. Around the mid-2000s, measurements of some genetic characteristics started to be included in large social science data collection projects such as Add Health (Harris et al. 2013), the Health and Retirement Study (Sonnega et al. 2014), and the Wisconsin Longitudinal Study (Herd, Carr and Roan 2014). In this era, social scientists focused on candidate gene studies, whose goal was to find one or a small number of genes that could explain variation in behavioral traits; typically, these candidate genes were identified on the basis of theory and lab studies in nonhuman animals, such as mice (Gizer, Ficks and Waldman 2009). Unfortunately, the reliability of results from candidate gene studies has been called into question because most candidate gene study designs appear not to have had sufficient statistical power to reliably identify the effects of interest; as a result, associations reported in the literature are likely to be false positives (Chabris et al. 2012; Benjamin et al. 2012; Conley and Fletcher 2017). Despite some successes, this era thus serves as a cautionary tale, illustrating the crucial importance of understanding study design and inference, including issues like statistical power, multiple testing, and meta-analysis (Mills and Tropf 2020; L. E. Duncan, Pollastri, and Smoller 2014).

Contemporary social science genomics has turned away from candidate gene studies and reformed around the idea of polygeneity: most behavioral traits seem to be the product of very small effects from a large number of genes (Conley and Fletcher 2017; Chabris et al. 2015). At the same time, genetic measurement technology has continued to improve and costs have continued to drop, making it possible for social science studies to routinely measure a much larger share of the study participants' genomes. Analyses now focus on genome-wide association studies (GWAS), whose goal is to atheoretically test for an association between a very large number of genes and the trait of interest; GWAS results can be further extended to construct polygenic scores, which are summary indices that estimate an individual's genetic propensity for some behavioral trait (Choi, Mak, and O'Reilly 2020). Methodologically, designing GWAS studies requires careful consideration of sample size, statistical power, and adjustments for multiple tests. This analysis reveals that sample sizes required to conduct a GWAS for a behavioral trait can be very large—so large that, practically speaking, no one study would provide enough statistical power to detect much; therefore, in practice, researchers often construct consortia involving many large cohort studies (Zenebe-Gete, Salowe, and O'Brien 2021; Manolio, Goodhand, and Ginsburg 2020; Okbay et al. 2022). As a result, researchers have had to develop methods that each team in a consortium can apply

to their own data, and that enable the results to be aggregated up using techniques from meta-analysis (e.g., Sullivan et al. 2018).

Other methodological challenges remain (Akimova et al. 2021; Mills and Tropf 2020; Freese, Rauf, and Voelkel 2022; Conley and Fletcher 2017; Becker et al. 2021). For example, a common concern is population stratification, or the so-called "chopstick problem": genetic variation can be correlated with the environment by historical accident, potentially leading to spurious associations between genes and behavioral traits in studies that do not fully adjust for environmental confounders (Conley and Fletcher 2017; Hamer and Sirota 2000). Methods have been developed to try to adjust for population stratification, but this remains an active area of work (Bulik-Sullivan et al. 2015; Hellwege et al. 2017). It also highlights concerns about the lack of diversity in genetic data available to social science researchers; increasingly, samples from outside Western Europe and North America are becoming available, which is a welcome sign. However, the high levels of genetic diversity in African and Latin American populations means that sample-size demands are likely to be even larger for these groups (Campbell and Tishkoff 2008; Ruiz-Linares et al. 2014). The lack of large-scale, representative genetic samples often limits the ability of researchers to make population-generalizable claims.

Despite the challenges of genetic data, there remains a large potential for population researchers to use these data to understand key demographic phenomena. One exciting data infrastructure project in this space is the "All of Us" project, an observational cohort study aiming to gather data on one million diverse U.S. participants (The All of Us Research Program Investigators 2019). Another is the UK Biobank study, which is a prospective cohort of around 500,000 people that collects detailed genetic and behavioral data (Bycroft et al. 2018). Other studies have provided interesting findings with implications for demography. For instance, Mills et al. (2021) used a large genome-wide study to identify 371 genetic variants that influence age at first birth and the number of children ever born. Additional methodological advances, such as for better understanding of the gene interplay (Johnson, Sotoudeh, and Conley 2022) or identifying causal effects and heritability estimates using genetic instrumental variables (DiPrete, Burik, and Koellinger 2018), point towards a promising future for genetic research. Specifically, the interaction of genetics and exposomes, the environmental exposure and individual encounters throughout the life course, is an important direction for future research.

New advances in traditional sources of data

Exciting advances are also being made in traditional approaches to data collection. These advances fall into two general categories: advances in probability-based methods and advances in non-probability-based

methods. Probability-based methods depend on collecting data using a known, stochastic mechanism to decide which members of a population to include in a sample (Särndal, Swensson, and Wretman 2003). For probability-based methods, advances in constructing specialized, probabilistic online panels will improve our ability to study smaller population subgroups. For example, the Pew Research Center recently developed a custom online panel to target Asian Americans, who represent only 7% of the general U.S. population, but who are the fastest growing ethnoracial group in the United States (Ruiz, Noe-Bustamante, and Shah 2023). A collaboration with the National Opinion Research Center is underway to make this online panel of Asian Americans available to the broader research community. Such panels will allow researchers to better target and sample subpopulations to unpack broad and heterogeneous categories such as "Asian American."

Nationally representative, probabilistic cohort studies originally designed to study earlier life conditions have been repurposed for studying health and mortality later in life as the cohorts age. For example, EdSHARe (Educational Studies for Healthy Aging Research) consists of two educational cohort studies that are now reoriented for studying life course influences on health, cognition, and mortality in mid/late adulthood (Grodsky et al. 2022). Similarly, the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative cohort of 20,475 U.S. adolescents first observed in 1994/1995, has begun releasing mortality outcomes, including the cause of death (Trani et al. 2022). These cohort studies will become valuable resources for studying the causal pathways through which early life conditions and social networks influence health and mortality, intracohort mortality selection and frailty dynamics, and more.

Another set of advances in probability-based methods focuses on networks. For example, advances in network reporting—having a probability-based sample of respondents report on others in their broader social network—have shown promise. Researchers have used the *network scale-up method* to estimate the size of hidden populations (Bernard et al. 1991; Killworth et al. 1998; Feehan and Salganik 2016; Maltiel et al. 2015) and to estimate mortality rates (Feehan and Salganik 2023) using moderately sized probability samples. Researchers are also increasingly collecting *aggregate relational data* to learn more about the social network that connects members of a population (McCormick and Zheng 2015; Breza et al. 2020; McCormick, Salganik, and Zheng 2010; Breza et al. 2023), including patterns of segregation in social networks (DiPrete et al. 2011).

Non-probability sampling is playing an increasingly large role in social science research, in part because it is becoming more difficult and expensive to obtain a true probability sample, especially in high-income countries (Leeper 2019). As a result, many researchers are exploring non-probability designs as a basis for recruiting survey respondents or study

participants. These approaches include quota samples from opt-in online panels (Hays, Liu, and Kapteyn 2015) or recruitment via targeted online ads (Schneider and Harknett 2022; Olivier 2011; Yun and Trumbo 2000). As non-probability data become more common, methods for analyzing these data will become crucial. A parallel can be drawn with causal inference: much intellectual effort has been devoted to understanding how to use observational data—in which the researcher does not randomly assign some kind of treatment—to make causal inferences; a similar amount of effort may be needed to understand how it may be possible to use non-probability samples—in which the researcher does not control selection into the sample—to produce reliable population-level inferences. There has been active work in this area; for example, multilevel regression with poststratification (Gao et al. 2021; Gelman and Little 1997; Park, Gelman, and Bafumi 2004; Gelman et al. 2018) seems to increase the feasibility of using non-probability samples for demographic estimation (Pejcinovska et al. 2023; Breen, Mahmud, and Feehan 2022). The framework of Meng (2018), discussed in a later section, is also promising. But there is clearly much to be done: there is tremendous variation in non-probability designs, and relatively little can conclusively be said about best practices for analyzing data collected in this way.

Other advances in non-probability methods are not focused on conventional surveys. For example, respondent-driven sampling (RDS) uses a chain-referral sampling method that produces a non-probability, network-based sample. RDS can be a powerful tool for sampling within networks of hard-to-reach populations, but quantifying uncertainty and bias in RDS estimates has proven to be a major challenge (Goel and Salganik 2009). Fortunately, new methods for uncertainty quantification in RDS estimates have been shown to improve accuracy (Baraff, McCormick, and Raftery 2016; Rohe 2019), and RDS is now increasingly being applied to study stigmatized demographic events, such as abortion (Rossier et al. 2022; Sully, Giorgio, and Anjur-Dietrich 2020).

These advances in traditional methods of data collection will enable researchers to study previously inaccessible populations, especially in parts of the world lacking robust data infrastructure; for example, improvements in traditional probability-based methods may enable researchers to directly estimate death rates in countries that lack complete vital registration systems. And advances in non-probability methods may make it possible to study important groups for whom data are currently lacking, such as people experiencing homelessness, LGBTQ+ individuals, sex workers, remote rural populations, and other hard-to-reach populations. This kind of progress will produce a richer empirical understanding of important but previously understudied groups, and it should also produce theoretical frameworks that are more inclusive and therefore more powerful.

Cross-cutting challenges

These new data sources are accompanied by a new set of challenges (Table 2). To address these challenges will require new methods and careful attention to old-fashioned reliability and validity. The sentiment of King (2016)—"Big data is not about the data!"—is largely applicable here. These exciting new data sources alone offer little promise; they must be paired with new methods and frameworks to produce useful insights. Social scientists must develop new methods, standards, and disciplinary norms before the potential of these new resources is fully realized. Here we identify and discuss four common challenges shared by many new sources of data.

Data provenance: Do researchers understand how these data were created?

The first cross-cutting challenge is data provenance: Do researchers understand how these data were created? Is the accompanying data documentation ("metadata") up to scientific standards? This concern is particularly relevant for data that were not originally designed to be used for academic research. In linked big population microdata, the exact data collection process or the postprocessing steps conducted by one or more administrative agencies is often opaque. For example, the U.S. World War II Army Enlistment records released by the National Archives and Records Administration contain information on many ($N = 9$ million), but not all, U.S. WWII Army enlistees. The reason for including some enlistees and excluding others is only partially understood (Wikle and Osborne 2023). Additionally, the definitions of key variables may change or administrative forms may undergo revisions without corresponding updates to the metadata, adding another layer of potentially undocumented complication. Digital trace data are often generated as a by-product of routine business operations and are typically not documented to scientific standards. For example, while many research teams have used Facebook audience count data obtained from Facebook Marketing API for demographic studies, the exact algorithm used by Facebook to estimate user counts is proprietary and may change over time without notice (Zagheni, Weber, and Gummadi 2017). This lack of transparency is compounded by "drift," where the popularity and use of digital platforms may also shift over time (Salganik 2019; Lazer et al. 2014). In traditional data collection approaches, issues of data provenance are often more subtle but just as important. Complex sampling designs or intricate study protocols can introduce errors and variability. For example, survey enumerators might misinterpret or fail to strictly adhere to study protocol, leading to unknown data collection errors. And some modern non-probability techniques, such as RDS, leave the actual sampling mechanism in the hands of

TABLE 2 Overview of key methodological challenges associated with new data sources

Key methodological challenges	Big population microdata	Digital trace data	Advances in classic data collection	Geospatial data	Genetic data
Provenance Do researchers understand how these data were created? Does the data documentation meet scientific standards?	Unclear processes for why some individuals are included, and others excluded, from administrative records, lost or lacking metadata.	Black-box sampling algorithms, lack of high-quality metadata, often characteristics like age and sex come from an opaque predictive model	Complex sampling designs or nonresponse bias can create unclear underlying sampling processes	Often heavily modeled, privately owned with implications for replicability	Challenging to track and document how data was collected, processed, and used
Usability How can researchers gain access to these data? Are there legal, ethical, and/or practical concerns? Can results based on the data be replicated by the broader scientific community?	Some restricted administrative data can only be accessed in secure data enclaves, may require expensive computing resources	Often challenging to access, sometimes only available through partnerships with private companies	Original data collection is often slow and expensive	Processing and analyzing large-scale geospatial datasets require significant computational resources, expensive to purchase	Restricted access + privacy concerns, Computationally demanding, Need for consortia and meta-analysis to reach large sample sizes

/...

TABLE 2 (Continued)

Key methodological challenges					
	Big population microdata	**Digital trace data**	**Advances in classic data collection**	**Geospatial data**	**Genetic data**
Representativeness Who is included, and who is not included, in these data?	Record linkage may introduce selection bias, original admin records may systematically exclude key groups (e.g., undocumented individuals)	Selection in who uses digital technology platforms and how often	Nonprobability methods may yield unrepresentative samples	Model-based estimates may reflect representativity issues with source data	Generally overrepresented wealthy, Western countries
Inference How can researchers formally relate these new sources of data to the questions they want to answer? What exactly should be estimated, and how should uncertainty be assessed?	Large samples may create misleadingly narrow confidence intervals that only capture part of the uncertainty inherent to estimates with linked data	Platforms may change the way key constructs are measured or users may change the way they interact with a platform	Hard to derive estimators based on complex network reports; method-specific challenges with quantifying uncertainty; methods have hard-to-validate assumptions	Heterogeneity in data quality, resolution, missingness, and more across spatial units can be challenging to account for in analyses	Challenging to isolate the causal effect of genes; polygenicity means very careful assessment of sample size, statistical power, and multiple testing are generally needed

study participants; the researcher typically has little insight into how new participants are chosen.

Geospatial data have a unique set of challenges with data provenance. Geospatial data are often modeled using complex algorithms, and researchers must understand the bottom-line implications of some of these modeling choices for their analysis. Further, geospatial data often incorporate many different sources together using complex pipelines, making metadata management more involved. Genetic data face similar challenges: the pipeline of data collection, genetic processing, and the harmonization and comparison of other data sources often require substantial effort and documentation. Consortia formed from many different cohort studies to conduct a GWAS must develop uniform protocols and devote time and careful attention to quality control to ensure that estimates are conducted in the same way across studies, to allow them to be meaningfully pooled together (e.g., Okbay et al. 2016). If not done carefully, this can jeopardize researchers' ability to understand the origins of data.

Some of the challenges of data provenance can be overcome. By recognizing and addressing the constraints around the original process for creating data, researchers can leverage these repurposed datasets most effectively. For example, data from the Facebook API enable researchers to extract demographic insights that can potentially be calibrated using well-documented traditional datasets. However, this requires careful analyses by researchers and special attention to any changes in the underlying algorithm dynamics of the platform generating the digital trace data, which can cause the measurement to no longer be stable over time (Lazer et al. 2014). More generally, scientific reports based on these new sources of data should clearly state which aspects of data provenance are not known and how this uncertainty might affect the analysis. But, for many new data sources, fully accounting for data provenance as part of analyses remains an important and open challenge.

Usability: How can researchers gain access to these data? Are there legal, ethical, and/or practical concerns? Can the results be replicated?

The second cross-cutting challenge focuses on usability and access. Ideally, these new sources of data would be freely available to researchers everywhere—transparent and accessible sharing of research data, methods, and code are the cornerstones of open science (Freese and Peterson 2017; King and Persily 2020). These new data sources are massive, complex, and messy, often requiring substantial amounts of processing and manipulation prior to formal analysis. Small data cleaning or coding decisions can cumulatively have large downstream impacts on research results. In these settings, the many researcher degrees of freedom can introduce a large amount of unrecognized uncertainty (Breznau et al. 2022). Greater clarity

in reporting scientific decisions and publicly releasing code and data is important for all research, but we see these open-science practices playing an especially critical role in analyses involving new data sources because such analyses should be subjected to extra scrutiny. This additional scrutiny will be key for collectively discerning the limitations, best practices, and broader standards for working with these new data.

However, many of the most exciting new data sources are not owned by academic researchers committed to the principles of open science. For example, many linked big population microdata resources are owned by government agencies or private companies whose primary goal is not the production of scientific knowledge. Confidentiality concerns have led some major data producers, such as the U.S. Census Bureau, to increasingly employ differential privacy as a disclosure avoidance system (Abowd 2018). Differential privacy works by injecting a calibrated amount of noise into a dataset to ensure that any one individual cannot be reidentified. The introduced noise, however, can have substantial effects on the accuracy of the data. In the United States, a series of empirical investigations have demonstrated that the differential privacy measures adopted by the U.S. Census Bureau for their flagship public use data products will greatly limit the research community's ability to study small geographic areas or fine-grained population subgroups (Ruggles et al. 2019). Such differential privacy measures, if not subjected to systematic scrutiny and approved by the broader research community, stand to jeopardize a key demographic data resource.

Building, maintaining, and improving the efficiency of secure data enclaves, which enable researchers to use restricted data under stringent confidentiality safeguards, is a promising next step. The U.S. Census Bureau's Federal Statistical Research Data Centers (FSRDCs) serve as a prime example. Within these enclaves, researchers can access microdata from 16 federal statistical agencies, including the American Community Survey, the American Housing Survey, and the Decennial Censuses, among others. However, the current process for getting FSRDC approval and operating such enclaves is often slow and resource intensive (Committee on Policies and Programs to Reduce Intergenerational Poverty et al. 2023). The use of secure data enclaves could be expanded to other settings; for instance, IPUMS-International has established a research data enclave offering access to restricted international census data.

Digital trace data are most often owned by private companies. Collaborations between academia and industry hold immense potential for advancing scientific knowledge, but navigating these partnerships is complex (King and Persily 2020; Lazer et al. 2020). Companies usually have financial incentives, which may conflict with academic objectives of conducting impartial and open research. The struggle to maintain academic independence becomes even more pronounced when research could be influenced by a company's commercial goals. This impact is not just on the types of

projects that get initiated, which may be limited to those that align with a company's business interests but also on what gets published. There may be biases against publishing results that paint a company in a negative light, thereby skewing the academic literature. Furthermore, the issue of data ownership can impose limitations on reproducibility. Many private companies will not allow data to be shared beyond the approved researcher and in-house collaborators. Researchers handling digital trace data must also confront ethical challenges. The individuals whose activities generated these data typically did not consent to its use for research, implying a need for especially rigorous privacy and confidentiality measures when utilizing these datasets (Oberski and Kreuter 2020).

Similar to digital trace data, many geospatial datasets are owned by private companies, which may restrict the use of their data for academic research or impose substantial fees. For instance, acquiring satellite imagery from various private companies typically involves considerable costs. Practically, working with these large geospatial datasets frequently requires large computing servers with ample memory. Genetic data have similar challenges, often requiring secure, resource-intensive computing environments. Genomics projects also often require complex consortia consisting of many different cohort studies—each with its own ethics and privacy concerns—to be assembled and managed in order to achieve required sample sizes; as a result, the genomics community has developed methods that can be separately applied to individual datasets and then combined in a principled way. Genetic data are also accompanied by ethical challenges; we point readers elsewhere for a more comprehensive discussion (Bliss 2018; Duster 2003; Freese 2018).

The research community would benefit from a shared set of professional norms on data transparency, access, and quality. How should reviewers treat papers that cannot be replicated because data are only available in a restricted setting to approved researchers? What equity issues are introduced when data are available to some and not others?

Representativeness: Who is included, and who is excluded, in these data?

A major challenge cutting across all of these new sources of data is representativeness—in other words, who is included and who is excluded? What methods must be used to ensure that inferences from these new data sources refer to the population of scientific interest?

The core of demographic research relies on population-representative data, but the declining quality of some traditional demographic data sources poses a growing challenge for researchers. In high-income countries, survey response rates have dropped dramatically: contemporary telephone surveys often obtain less than a 10% response rate due to a mix of noncontact and

refusals (Leeper 2019). The resulting sample is often highly select, with biased univariate distributions of demographic characteristics (Rindfuss et al. 2015). In high-income countries, this problem is not going away—rather, the survey burden[3] will likely increase, resulting in even lower response rates and quality of responses. In addition, the quality of some online surveys may also erode as large language models (LLMs), such as Generative Pre-trained Transformer 4 (GPT-4) and Large Language Model Meta AI (LLaMA), become increasingly able to mimic human survey responses.

This issue of representativeness is intrinsic to these new data sources. Digital trace data from a social media platform such as Facebook can generally only tell us about Facebook users, who trend younger and more affluent than the general population (Gil-Clavel and Zagheni 2019). Moreover, these data are often not designed with social science research in mind; instead, they are created to help the goals of the company that makes the social computing platform. Key quantities like age and gender, which are central to most demographic analyses, may come from predictive models whose structure is not disclosed, and which may change over time with no notice. Big population microdata are often unrepresentative: administrative data can also systematically exclude entire population subgroups. For instance, data from the U.S. Social Security Administration does not capture undocumented persons (Finlay and Genadek 2021). Data from nonprobability samples, such as a respondent-driven sample, will generally not be representative of the broader population. Finally, genetic data often overrepresents wealthy, western countries (Mills and Rahal 2020).

Demographers have historically excelled at carefully repurposing unrepresentative data for demographic research (Zagheni and Weber 2015; Rogers, Willekens, and Raymer 2003), and their sensitivity to data quality and representativeness will serve the field well in developing methods and norms around research based on these new data resources. At the very least, researchers must carefully consider, address, and communicate the consequences of nonrepresentativeness for their substantive research results. We hope that methods for helping to quantify and perhaps adjust for issues of nonrepresentativeness will continue to be an active and productive area of methodological research.

Inference: How can researchers formally relate these new sources of data to the questions they want to answer? What exactly should be estimated, and how should uncertainty be assessed?

Providing a clear description of inferential goals and accompanying estimates of uncertainty is crucial when working with these large, unrepresentative datasets. The "big data paradox"— larger datasets cause researchers to have more confidence in misleading research results—is applicable to both digital trace data and big population microdata (Lazer et al. 2014). In such

settings, incorporating a formal framework for inference provides helpful structure. For instance, Meng (2018) introduces a framework for assessing relative sample size, combining a measure of data quality, data quantity, and a problem difficulty measure. Such a framework allows researchers to decompose different sources of error, answering questions such as "Should we trust a 1% probability sample with a 100% response rate or an administrative dataset covering 80% of the population?" (Meng 2018; Michael A. Bailey 2024; Bradley et al. 2021). Further, it clearly identifies trade-offs, allowing researchers to more clearly understand the potential pitfalls of this big data ecosystem.

Other frameworks could be adapted to help better understand and communicate uncertainty. Insights from the total survey error framework (Groves et al. 2009; Weisberg 2005; Salganik 2019), which decompose both sampling error and non-sampling error (e.g., error from measurement, nonresponse, or coverage issues) could be borrowed for working with linked survey and administrative data. For administrative data and digital trace data, huge samples will often produce very narrow uncertainty intervals. However, these uncertainty intervals can be misleading, as they fail to capture many important sources of uncertainty (e.g., non-sampling error and error in linkage). Moving forward, social scientists would benefit from more clearly communicating which sources of error their uncertainty intervals can and cannot capture.

A formal framework for causal inference may be especially beneficial for social scientists seeking to use these data sources in the future. Explicitly defining the target population and specific quantity of interest—a theoretical estimand—will be especially important (Lundberg, Johnson, and Stewart 2021). For causal questions, social scientists could also (1) connect this theoretical estimand to an empirical estimand that can be linked by a set of identification assumptions and (2) clearly define an estimation strategy to estimate the empirical estimand from data. Demographers will also have to become more accustomed to justifying their sample sizes by thinking carefully about the statistical power of their analyses. These frameworks can also be developed for more specialized research questions. For instance, the generalized network scale-up method, a method for estimating the size of hard-to-count populations, was developed to account for three assumptions of the basic network scale-up method that have been shown to be problematic in practice (Feehan and Salganik 2016). This framework allows researchers to unpack these sources of error and account for them using data from a secondary sample.

New data sources have new sources of error that are crucial to interrogate, describe, and quantify. This is particularly important for researchers to avoid being misled by the large size of these new data resources. We are hopeful these frameworks will be increasingly applied by social scientists using these new datasets.

Discussion

Changing disciplinary norms

These new data sources may also lead to a shift in disciplinary norms. We anticipate innovations in training will be required to ensure best practices for collecting, processing, and analyzing these new data sources. Traditional demographic training equips researchers with the skills to navigate many of the core methodological issues surrounding these new data sources, such as assessing representativeness, quantifying uncertainty, and investigating overall data quality. The emphasis on these foundational areas of training should remain. However, the size of these new data resources (e.g., high-resolution satellite imagery) will necessitate additional training in computing methods. Further, statistical and/or machine learning[4] methods for taking advantage of these new data sources will be beneficial.

A stronger emphasis on reporting transparency and guidelines for enhancing the reliability and reproducibility of research findings would also help. Several frameworks already exist: the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) initiative provides structured recommendations to improve the reporting quality of observational studies (von Elm et al. 2007); the Overview, Design concepts, and Details (ODD) protocol for effectively describing agent-based models and individual-based models (Grimm et al. 2006); and the REFORM checklist (Reporting Standards For Machine Learning Based Science) for clear reporting standards for machine-learning based science (Kapoor et al. 2024). New data sources in demographic research, and the methods that accompany them, may warrant similar reporting and transparency guidelines. For example, such a guideline might have researchers report on what is known and not known about the provenance of the data being analyzed.

As these new data sources become increasingly specialized, we may see a shift towards team-based science. We expect this move towards collaborative efforts across various disciplines will have some advantages. Interdisciplinary cooperation will drive academic innovation through productive collaborations among geneticists with expertise in population genomics, computer scientists skilled in handling digital trace data, and geographers with strong spatial backgrounds. However, demographers should also be mindful to preserve the value of smaller collaborations and solo efforts, which allow for the investigation of highly original ideas without the consensus or compromise often required in team-based projects. Further, we hope the core of demography—formal analysis linking micro- and macropopulation processes (Lee 2001)—does not erode in this shift towards team-based science.

The COVID-19 pandemic illustrated our changing disciplinary norms and conventions surrounding our data ecosystem. The demand for real-

time data necessitated fast, web-based surveys to better understand key parameters related to transmission dynamics, like social contact patterns (Feehan and Mahmud 2021; Wong et al. 2023). Further, cross-disciplinary collaborations, combining expertise from researchers with different areas of expertise produced timely, high-impact papers (e.g., Block et al. (2020); Dowd et al. (2020); Wrigley-Field et al. (2021)). In some countries, death records allowed us to apply demographic science to study racial, sociodemographic, and geographic disparities in COVID-19 mortality, but the pandemic revealed that much of the relevant data were not always easy to find or re-purpose for academic research (Shadbolt et al. 2022). In sum, the pandemic has both highlighted the value of interdisciplinary collaboration and demographic science in studying the COVID-19 pandemic (Dowd et al. 2020) but also the addressable shortcomings associated with the current data ecosystem.

Future prospects

These are exciting times: new data sources have the potential to revolutionize large parts of demographic research. While immediate benefits include enhanced timeliness and accuracy in measurements, the major longer-term impact will emerge from the development and testing of new demographic theories. However, these new data come with important challenges that researchers must address.

We focused on five exciting sources of new data. In our judgment, these five data sources are poised to have a big impact on demographic research. But they are not exhaustive, and it may well be the case that the coming years will produce additional change. For example, as we write, the pace of advances in artificial intelligence, and especially LLMs, has been stunning. These tools may have implications for creating useful synthetic data, and they will likely become a part of collecting other new types of data. For example, they may be used in collecting survey or interview data, they may help improve methods for linking datasets, and they will become increasingly central to digital trace data.

Predicting exactly how new data sources will shape the field is, of course, impossible. Nevertheless, we can suggest several key areas that warrant increased attention over the course of the next decade. First, these new data sources are complements to traditional data sources—not substitutes; each one depends on traditional data in some way. Therefore, investments in traditional demographic data, such as high-quality censuses and probability-based surveys, need to continue or expand. Second, we would like to see population researchers continue to work on addressing the four key methodological challenges that cut across different types of new data. Of course, the need to address these challenges can be seen as an opportunity, and the culture and history of demography make the discipline well-poised

to have an important impact here. In fact, many of these cross-cutting methodological challenges will not appear to be especially new to demographers, who have a long history of developing methods to extract reliable insights from imperfect data. Indeed, this may prove to be a critical comparative advantage that demographers can bring to the interdisciplinary teams that will be required to make the most use of these new sources of data in the future. Third, it is essential that all population researchers have access to rich, high-quality data. Specifically, more attention should be paid to removing barriers to data access and improving data reliability and quality. Finally, the field would benefit from shifts in disciplinary norms, including more emphasis on computational training and data management, practices of promoting open, reliable science, and expansion of team-based science.

Depending on our ability to successfully adapt to these new challenges, we think it is reasonable to also expect exciting advances in demographic theory. We anticipate linked census and administrative data will help improve our understanding of life course theories; that digital trace data will help to fill in big empirical gaps that have slowed down progress in understanding of migration; that data from dating platforms will help enrich our understanding of partnership formation and homogamy; network-based methods will allow researchers to study stigmatized and underresearched groups; and genetic data will allow us to test theories of frailty and mortality selection. This list represents just a handful of the many potentially exciting theoretical advances.

To summarize, the most effective work with these new data sources will require methodological innovation, a reevaluation of disciplinary norms, and sustained investment in high-quality, probability-based surveys and censuses. However, we are cautiously optimistic these new data sources will allow social scientists to revisit old theories and to develop new theories. Along the way, we expect methodological advances that arise as social scientists work to overcome cross-cutting challenges such as understanding data provenance, data usability, representativeness, and inference, and more. Progress will be hard-won, but these new data sources will ultimately spark substantial advances in demographic measurement, methodology, and theory.

Acknowledgments

For helpful discussions and feedback, we thank Constance Citro, Joshua R. Goldstein, Aashish Gupta, Ayesha Mahmud, and Evelina Akimova. The authors acknowledge financial support from the Leverhulme Trust (Grant RC-2018-003) for the Leverhulme Centre for Demographic Science, the Bill and Melinda Gates Foundation (INV-045370), the Berkeley Population Center (P2C HD 073964), and the Berkeley Center for the Economics and Demography of Aging (5P30AG012839).

Notes

1 These data sources are not exhaustive, and we regret that not all new developments in data infrastructure could be included in this commentary. We also acknowledge that the boundaries between these classifications are blurred, although we treat them as distinct here. For example, big population microdata often includes spatial elements, genetic data is often linked into existing cohort studies, and digitized data from Google Scholar may be categorized as both big microdata and digital trace data.

2 Several more comprehensive reviews of relevant literature are available elsewhere. For example, for more on digital trace data, see Kashyap (2021), Kashyap et al. (2022), Cesare et al. (2018), Salganik (2019), and Alburez-Gutierrez et al. (2019). For more on big population microdata, see Ruggles (2014). For more on advances in classical data collection, see Baker et al. (2013) and Salganik (2019). For more on genetic data, see Mills and Tropf (2020) and Conley and Fletcher (2017). For more on geospatial data, see Matthews et al. (2021).

3 The challenge of survey overload can be thought of as a "common pool resource" problem. As separate researchers attempt to gather excessive data from a restricted pool of participants, refusal rates climb, and the quality of answers diminishes.

4 Machine learning is increasingly being used for demographic research and may grow into an important part of the demographic toolkit. Machine learning and AI may potentially become a source of demographic data in and of themselves; for instance, simulating survey response using LLMs may generate a new source of data.

References

Abowd, John M. 2018. The U.S. Census Bureau Adopts Differential Privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (KDD '18), p. 2867. New York, USA: Association for Computing Machinery.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature* 59 (3): 865–918.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez, and Myera Rashid. 2020. "Census Linking Project: Version 1.0." https://censuslinkingproject.org/.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102 (5): 1832–1856.

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. 2022. "Machine Learning and Phone Data Can Improve Targeting of Humanitarian Aid." *Nature* 603 (7903): 864–870.

Akimova, Evelina T., Richard Breen, David M. Brazel, and Melinda C. Mills. 2021. "Gene Environment Dependencies Lead to Collider Bias in Models with Polygenic Scores." *Scientific Reports* 11 (1): 9457.

Alburez-Gutierrez, Diego, Samin Aref, Sofia Gil-Clavel, André Grow, Daniela V. Negraia, and Emilio Zagheni. 2019. "Demography in the Digital Era: New Data Sources for Population Research." In *Smart Statistics for Smart Applications: Book of Short Papers (SIS2019)*, edited by G. Arbia, S. Peluso, A. Pini, and G. Rivellini, 23–30. New York: Pearson.

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2020. "Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States." *Population Research and Policy Review* 41 (1): 1–28.

Baccaini, Brigitte, and Daniel Courgeau. 1996. "The Spatial Mobility of Two Generations of Young Adults in Norway." *International Journal of Population Geography* 2: 333–359.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data." *Journal of Economic Literature* 58 (4): 997–1044.

Bailey, Michael A. 2024. *Polling at a Crossroads: Rethinking Modern Survey Research*. Cambridge: Cambridge University Press.

Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. "Social Connectedness: Measurement, Determinants, and Effects." *Journal of Economic Perspectives* 32 (3): 259–280.

Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1 (2): 90–143.

Baraff, Aaron J., Tyler H. McCormick, and Adrian E. Raftery. 2016. "Estimating Uncertainty in Respondent-Driven Sampling Using a Tree Bootstrap Method." *Proceedings of the National Academy of Sciences* 113 (51): 14668–14673.

Barban, Nicola, Rick Jansen, Ronald de Vlaming, Ahmad Vaez, Jornt J. Mandemakers, Felix C. Tropf, Xia Shen, James F. Wilson, Daniel I. Chasman, Ilja M. Nolte, Vinicius Tragante, Sander W. van der Laan, John R. B. Perry, Augustine Kong, BIOS Consortium, Tarunveer S. Ahluwalia, Eva Albrecht, Laura Yerges-Armstrong, Gil Atzmon, Kirsi Auro, Kristin Ayers, Andrew Bakshi, Danny Ben-Avraham, Klaus Berger, Aviv Bergman, Lars Bertram, Lawrence F. Bielak, Gyda Bjornsdottir, Marc Jan Bonder, Linda Broer, Minh Bui, Caterina Barbieri, Alana Cavadino, Jorge E. Chavarro, Constance Turman, Maria Pina Concas, Heather J. Cordell, Gail Davies, Peter Eibich, Nicholas Eriksson, Tõnu Esko, Joel Eriksson, Fahimeh Falahi, Janine F. Felix, Mark Alan Fontana, Lude Franke, Ilaria Gandin, Audrey J. Gaskins, Christian Gieger, Erica P. Gunderson, Xiuqing Guo, Caroline Hayward, Chunyan He, Edith Hofer, Hongyan Huang, Peter K. Joshi, Stavroula Kanoni, Robert Karlsson, Stefan Kiechl, Annette Kifley, Alexander Kluttig, Peter Kraft, Vasiliki Lagou, Cecile Lecoeur, Jari Lahti, Ruifang Li-Gao, Penelope A. Lind, Tian Liu, Enes Makalic, Crysovalanto Mamasoula, Lindsay Matteson, Hamdi Mbarek, Patrick F. McArdle, George McMahon, S. Fleur W. Meddens, Evelin Mihailov, Mike Miller, Stacey A. Missmer, Claire Monnereau, Peter J. van der Most, Ronny Myhre, Mike A. Nalls, Teresa Nutile, Ioanna Panagiota Kalafati, Eleonora Porcu, Inga Prokopenko, Kumar B. Rajan, Janet Rich-Edwards, Cornelius A. Rietveld, Antonietta Robino, Lynda M. Rose, Rico Rueedi, Kathleen A. Ryan, Yasaman Saba, Daniel Schmidt, Jennifer A. Smith, Lisette Stolk, Elizabeth Streeten, Anke Tönjes, Gudmar Thorleifsson, Sheila Ulivi, Juho Wedenoja, Juergen Wellmann, Peter Willeit, Jie Yao, Loic Yengo, Jing Hua Zhao, Wei Zhao, Daria V. Zhernakova, Najaf Amin, Howard Andrews, Beverley Balkau, Nir Barzilai, Sven Bergmann, Ginevra Biino, Hans Bisgaard, Klaus Bønnelykke, Dorret I. Boomsma, Julie E. Buring, Harry Campbell, Stefania Cappellani, Marina Ciullo, Simon R. Cox, Francesco Cucca, Daniela Toniolo, George Davey-Smith, Ian J. Deary, George Dedoussis, Panos Deloukas, Cornelia M. van Duijn, Eco J. C. de Geus, Johan G. Eriksson, Denis A. Evans, Jessica D. Faul, Cinzia Felicita Sala, Philippe Froguel, Paolo Gasparini, Giorgia Girotto, Hans-Jörgen Grabe, Karin Halina Greiser, Patrick J. F. Groenen, Hugoline G. de Haan, Johannes Haerting, Tamara B. Harris, Andrew C. Heath, Kauko Heikkilä, Albert Hofman, Georg Homuth, Elizabeth G. Holliday, John Hopper, Elina Hyppönen, Bo Jacobsson, Vincent W. V. Jaddoe, Magnus Johannesson, Astanand Jugessur, Mika Kähönen, Eero Kajantie, Sharon L. R. Kardia, Bernard Keavney, Ivana Kolcic, Päivikki Koponen, Peter Kovacs, Florian Kronenberg, Zoltan Kutalik, Martina La Bianca, Genevieve Lachance, William G. Iacono, Sandra Lai, Terho Lehtimäki, David C. Liewald, LifeLines Cohort Study, Cecilia M. Lindgren, Yongmei Liu, Robert Luben, Michael Lucht, Riitta Luoto, Per Magnus, Patrik K. E. Magnusson, Nicholas G. Martin, Matt McGue, Ruth McQuillan, Sarah E. Medland, Christa Meisinger, Dan Mellström, Andres Metspalu, Michela Traglia, Lili Milani, Paul Mitchell, Grant W. Montgomery, Dennis Mook-Kanamori, Renée de Mutsert, Ellen A. Nohr, Claes Ohlsson, Jørn Olsen, Ken K. Ong, Lavinia Paternoster, Alison Pattie, Brenda W. J. H. Penninx, Markus Perola, Patricia A. Peyser, Mario Pirastu, Ozren Polasek, Chris Power, Jaakko Kaprio, Leslie J. Raffel, Katri Räikkönen, Olli Raitakari, Paul M. Ridker, Susan M. Ring, Kathryn Roll, Igor Rudan, Daniela Ruggiero, Dan Rujescu,

Veikko Salomaa, David Schlessinger, Helena Schmidt, Reinhold Schmidt, Nicole Schupf, Johannes Smit, Rossella Sorice, Tim D. Spector, John M. Starr, Doris Stöckl, Konstantin Strauch, Michael Stumvoll, Morris A. Swertz, Unnur Thorsteinsdottir, A. Roy Thurik, Nicholas J. Timpson, Joyce Y. Tung, André G. Uitterlinden, Simona Vaccargiu, Jorma Viikari, Veronique Vitart, Henry Völzke, Peter Vollenweider, Dragana Vuckovic, Johannes Waage, Gert G. Wagner, Jie Jin Wang, Nicholas J. Wareham, David R. Weir, Gonneke Willemsen, Johann Willeit, Alan F. Wright, Krina T. Zondervan, Kari Stefansson, Robert F. Krueger, James J. Lee, Daniel J. Benjamin, David Cesarini, Philipp D. Koellinger, Marcel den Hoed, Harold Snieder, and Melinda C. Mills. 2016. "Genome-Wide Analysis Identifies 12 Loci Influencing Human Reproductive Behavior." *Nature Genetics* 48 (12): 1462–1472.

Becker, Joel, Casper A. P. Burik, Grant Goldman, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Daniel W. Belsky, Richard Karlsson Linnér, Rafael Ahlskog, Aaron Kleinman, David A. Hinds, Avshalom Caspi, David L. Corcoran, Terrie E. Moffitt, Richie Poulton, Karen Sugden, Benjamin S. Williams, Kathleen Mullan Harris, Andrew Steptoe, Olesya Ajnakina, Lili Milani, Tõnu Esko, William G. Iacono, Matt McGue, Patrik K. E. Magnusson, Travis T. Mallard, K. Paige Harden, Elliot M. Tucker-Drob, Pamela Herd, Jeremy Freese, Alexander Young, Jonathan P. Beauchamp, Philipp D. Koellinger, Sven Oskarsson, Magnus Johannesson, Peter M. Visscher, Michelle N. Meyer, David Laibson, David Cesarini, Daniel J. Benjamin, Patrick Turley, and Aysu Okbay. 2021. "Resource Profile and User Guide of the Polygenic Index Repository." *Nature Human Behaviour* 5 (12): 1744–1758.

Benjamin, Daniel J., David Cesarini, Christopher F. Chabris, Edward L. Glaeser, David I. Laibson, Gene/Environment Susceptibility-Reykjavik Study Age, Vilmundur Gunason, Tamara B. Harris, Lenore J. Launer, Shaun Purcell, Albert Vernon Smith, Swedish Twin Registry, Magnus Johannesson, Patrik K. E. Magnusson, Framingham Heart Study, Jonathan P. Beauchamp, Nicholas A. Christakis, Wisconsin Longitudinal Study, Craig S. Atwood, Benjamin Hebert, Jeremy Freese, Robert M. Hauser, Taissa S. Hauser, Swedish Large Schizophrenia Study, Alexander Grankvist, Christina M. Hultman, and Paul Lichtenstein. 2012. "The Promises and Pitfalls of Genoeconomics." *Annual Review of Economics* 4: 627–662.

Bernard, Russell H, Eugene C Johnsen, Peter D Killworth, and Scott Robinson. 1991. "Estimating the Size of an Average Personal Network and of an Event Subpopulation: Some Empirical Results." *Social Science Research* 20 (2): 109–121.

Bliss, Catherine. 2018. *Social by Nature: The Promise and Peril of Sociogenomics*. Stanford, CA: Stanford University Press.

Block, Per, Marion Hoffman, Isabel J. Raabe, Jennifer Beam Dowd, Charles Rahal, Ridhi Kashyap, and Melinda C. Mills. 2020. "Social Network-Based Distancing Strategies to Flatten the COVID-19 Curve in a Post-Lockdown World." *Nature Human Behaviour* 4 (6): 588–596.

Blumenstock, Joshua E. 2012. "Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda." *Information Technology for Development* 18 (2): 107–125.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–1076.

Bradley, Valerie C., Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. 2021. "Unrepresentative Big Surveys Significantly Overestimated US Vaccine Uptake." *Nature* 600 (7890): 695–700.

Breen, Casey F., Ayesha S. Mahmud, and Dennis M. Feehan. 2022. "Novel Estimates Reveal Subnational Heterogeneities in Disease-Relevant Contact Patterns in the United States." *PLOS Computational Biology* 18 (12): e1010742.

Breza, Emily, Arun G. Chandrasekhar, Shane Lubold, Tyler H. McCormick, and Mengjie Pan. 2023. "Consistently Estimating Network Statistics Using Aggregated Relational Data." *Proceedings of the National Academy of Sciences* 120 (21): e2207185120.

Breza, Emily, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan. 2020. "Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data." *American Economic Review* 110 (8): 2454–2484.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke

Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin B. Busch, Juan Carlos-Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsófia S. Ignácz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti Sinatra, Alexander Langenkamp, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan E. Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Kyle McWagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan Mijs, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna O. Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel R. Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Regine Schmidt, Katja M. Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, and Tomasz Żółtak. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." *Proceedings of the National Academy of Sciences* 119 (44): e2203150119.

Bruch, Elizabeth E., and M. E. J. Newman. 2018. "Aspirational Pursuit of Mates in Online Dating Markets." *Science Advances* 4 (8): eaap9815.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." *Nature Genetics* 47 (3): 291–295.

Byaro, Mwoya, Anicet Rwezaula, and Nicholaus Ngowi. 2023. "Does Internet Use and Adoption Matter for Better Health Outcomes in Sub-Saharan African Countries? New Evidence from Panel Quantile Regression." *Technological Forecasting and Social Change* 191: 122445.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–209.

Campbell, Cameron, and James Lee. 2006. "State Views and Local Views of Population: Linking and Comparing Genealogies and Household Registers in Liaoning, 1749–1909." *History and Computing* 14 (1–2): 9–29. https://doi.org/10.3366/hac.2002.14.1-2.9.

Campbell, Michael C., and Sarah A. Tishkoff. 2008. "African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping." *Annual Review of Genomics and Human Genetics* 9: 403–433.

Cesare, Nina, Hedwig Lee, Tyler McCormick, Emma Spiro and Emilio Zagheni. 2018. "Promises and Pitfalls of Using Digital Traces for Demographic Research." *Demography* 55 (5):1979–1999.

Chabris, Christopher F., Benjamin M. Hebert, Daniel J. Benjamin, Jonathan Beauchamp, David Cesarini, Matthijs van der Loos, Magnus Johannesson, Patrik K. E. Magnusson, Paul Lichtenstein, Craig S. Atwood, Jeremy Freese, Taissa S. Hauser, Robert M. Hauser, Nicholas Christakis, and David Laibson. 2012. "Most Reported Genetic Associations With General Intelligence Are Probably False Positives." *Psychological Science* 23 (11): 1314–1323.

Chabris, Christopher F., James J. Lee, David Cesarini, Daniel J. Benjamin, and David I. Laibson. 2015. "The Fourth Law of Behavior Genetics." *Current Directions in Psychological Science* 24 (4): 304–312.

Checchi, Francesco, Barclay T. Stewart, Jennifer J. Palmer, and Chris Grundy. 2013. "Validity and Feasibility of a Satellite Imagery-Based Method for Rapid Estimation of Displaced Populations." *International Journal of Health Geographics* 12 (1): 4.

Chen, Bijia, Cameron Campbell, Yuxue Ren, and James Lee. 2020. "Big Data for the Study of Qing Officialdom: The China Government Employee Database-Qing (CGED-Q)." *Journal of Chinese History* 4 (2): 431–460.

Chi, Guanghua, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. 2022. "Microestimates of Wealth for All Low- and Middle-Income Countries." *Proceedings of the National Academy of Sciences* 119 (3): e2113658119.

Choi, Shing Wan, Timothy Shin Heng Mak, and Paul F. O'Reilly. 2020. "A Guide to Performing Polygenic Risk Score Analyses." *Nature Protocols* 15 (9):2759–2572.

Committee on Policies and Programs to Reduce Intergenerational Poverty, Board on Children, Youth, and Families, Division of Behavioral and Social Sciences and Education and National Academies of Sciences, Engineering, and Medicine. 2023. *Reducing Intergenerational Poverty*. Washington, DC: National Academies Press.

Conley, Dalton, and Jason Fletcher. 2017. *The Genome Factor: What the Social Genomics Revolution Reveals about Ourselves, Our History, and the Future*. Princeton, NJ: Princeton University Press.

DiPrete, Thomas A., Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. 2011. "Segregation in Social Networks Based on Acquaintanceship and Trust." *American Journal of Sociology* 116 (4): 1234–1283.

DiPrete, Thomas A., Casper A. P. Burik, and Philipp D. Koellinger. 2018. "Genetic Instrumental Variable Regression: Explaining Socioeconomic and Health Outcomes in Nonexperimental Data." *Proceedings of the National Academy of Sciences* 115 (22): E4970–E4979.

Dowd, Jennifer Beam, Liliana Andriano, David M. Brazel, Valentina Rotondi, Per Block, Xuejie Ding, Yan Liu, and Melinda C. Mills. 2020. "Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-19." *Proceedings of the National Academy of Sciences* 117 (18): 9696-9698.

Duncan, Brian, and Stephen J. Trejo. 2011. "Tracking Intergenerational Progress for Immigrant Groups: The Problem of Ethnic Attrition." *American Economic Review* 101 (3): 603–608.

Duncan, Laramie E., Alisha R. Pollastri, and Jordan W. Smoller. 2014. "Why Many Geneticists and Psychological Scientists Have Discrepant Views About Gene–Environment Interaction (G×E) Research." *The American Psychologist* 69 (3): 249–268.

Duster, Troy. 2003. *Backdoor to Eugenics*. 2 ed. New York: Routledge.

Eagle, Nathan, Alex (Sandy) Pentland, and David Lazer. 2009. "Inferring Friendship Network Structure by Using Mobile Phone Data." *Proceedings of the National Academy of Sciences* 106 (36): 15274–15278.

Eagle, Nathan, Michael Macy, and Rob Claxton. 2010. "Network Diversity and Economic Development." *Science* 328 (5981): 1029–1031.

Edquist, Harald, Peter Goodridge, Jonathan Haskel, Xuan Li, and Edward Lindquist. 2018. "How Important Are Mobile Broadband Networks for the Global Economic Development?" *Information Economics and Policy* 45: 16–29.

Feehan, Dennis M., and Ayesha S. Mahmud. 2021. "Quantifying Population Contact Patterns in the United States during the COVID-19 Pandemic." *Nature Communications* 12 (1): 893.

Feehan, Dennis M., and Curtiss Cobb. 2019. "Using an Online Sample to Estimate the Size of an Offline Population." *Demography* 56 (6): 2377–2392.

Feehan, Dennis M., and Matthew J. Salganik. 2016. "Generalizing the Network Scale-up Method: A New Estimator for the Size of Hidden Populations." *Sociological Methodology* 46 (1): 153–186.

Feehan, Dennis M., and Matthew J. Salganik. 2023. "Comparing Survey-Based Estimates of Adult Mortality to High-Quality Vital Records: Evidence from 27 Brazilian Cities." SocArXiv. https://doi.org/10.31235/osf.io/x5ywv

Feigenbaum, James J. 2018. "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940." *The Economic Journal* 128 (612): F446–F481.

Finlay, Keith, and Katie R. Genadek. 2021. "Measuring All-Cause Mortality With the Census Numident File." *American Journal of Public Health* 111 (Suppl 2): S141–S148.

Fletcher, Jason, and Hamid Noghanibehambari. 2021. "The Effects of Education on Mortality: Evidence Using College Expansions." Technical Report w29423. Cambridge, MA: National Bureau of Economic Research.

Fletcher, Jason M, Yuchang Wu, Zijie Zhao, and Qiongshi Lu. 2023. "The Production of Within-Family Inequality: Insights and Implications of Integrating Genetic Data." *PNAS Nexus* 2 (4): pgad121.

Flückiger, Matthias, and Markus Ludwig. 2023. "Mobile Phone Coverage and Infant Mortality in Sub-Saharan Africa." *Journal of Economic Behavior & Organization* 211: 462–485.

Freese, Jeremy. 2018. "The Arrival of Social Science Genomics." *Contemporary Sociology* 47 (5): 524–536.

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43 (1): 147–165.

Freese, Jeremy, Tamkinat Rauf, and Jan Gerrit Voelkel. 2022. 'Advances in Transparency and Reproducibility in the Social Sciences'. *Social Science Research* 107: 102770.

Friedman, Naomi P., Marie T. Banich, and Matthew C. Keller. 2021. "Twin Studies to GWAS: There and Back Again." *Trends in Cognitive Sciences* 25 (10): 855–869.

Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. 2021. "Improving Multilevel Regression and Poststratification with Structured Priors." *Bayesian Analysis* 16 (3): 719–744.

Gelman, Andrew, Jeffrey Lax, Justin Phillips, Jonah Gabry and Robert Trangucci. 2018. "Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion." http://www.stat.columbia.edu/~gelman/research/unpublished/MRT%281%29.pdf.

Gelman, Andrew, and Thomas C. Little. 1997. "Poststratification Into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23: 127–135.

Genadek, Katie R., and J. Trent Alexander. 2022. "The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Census Infrastructure." *IEEE Annals of the History of Computing* 44 (4), 57–66.

Gil-Clavel, Sofia, André Grow, and Maarten J. Bijlsma. 2023. "Migration Policies and Immigrants' Language Acquisition in EU-15: Evidence from Twitter." *Population and Development Review* 49 (3), 469–497.

Gil-Clavel, Sofia, and Emilio Zagheni. 2019. "Demographic Differentials in Facebook Usage around the World." *Proceedings of the International AAAI Conference on Web and Social Media* 13: 647–650.

Gizer, Ian R., Courtney Ficks, and Irwin D. Waldman. 2009. "Candidate Gene Studies of ADHD: A Meta-Analytic Review." *Human Genetics* 126 (1): 51–90.

Goel, Sharad, and Matthew J. Salganik. 2009. "Respondent-Driven Sampling as Markov Chain Monte Carlo: RDS AS MCMC." *Statistics in Medicine* 28 (17): 2202–2229.

Goldstein, Joshua R., Monica Alexander, Casey F. Breen, Andrea Miranda-Gonz´alez, Felipe Menares, Maria Osborne, and Ugur Yildirim. 2021. CenSoc Mortality File: Version 2.1. https://censoc.berkeley.edu/.

Goldstein, Sidney. 1964. "The Extent of Repeated Migration: An Analysis Based on the Danish Population Register." *Journal of the American Statistical Association* 59 (308): 1121–1132.

Grace, Kathryn, Nicholas N. Nagle, Clara R. Burgert-Brucker, Shelby Rutzick, David C. Van Riper, Trinadh Dontamsetti, and Trevor Croft. 2019. "Integrating Environmental Context into DHS Analysis While Protecting Participant Confidentiality: A New Remote Sensing Method." *Population and Development Review* 45 (1): 197–218.

Graetz, Nick, Carl Gershenson, Peter Hepburn, Sonya R. Porter, Danielle H. Sandler, and Matthew Desmond. 2023. "A Comprehensive Demographic Profile of the US Evicted Population." *Proceedings of the National Academy of Sciences* 120 (41): e2305860120.

Grimm, Volker, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K. Heinz, Geir Huse, Andreas Huth, Jane U. Jepsen, Christian Jørgensen, Wolf M. Mooij, Birgit Müller, Guy Pe'er, Cyril Piou, Steven F. Railsback, Andrew M. Robbins, Martha M. Robbins, Eva Rossmanith, Nadja Rüger, Espen Strand, Sami Souissi, Richard A. Stillman, Rune Vabø, Ute Visser, and Donald L. DeAngelis. 2006. "A Standard Protocol for Describing Individual-Based and Agent-Based Models." *Ecological Modelling* 198 (1): 115–126.

Grodsky, Eric, Jennifer Manly, Chandra Muller, and John Robert Warren. 2022. 'Cohort Profile: High School and Beyond'. *International Journal of Epidemiology* 51 (5): e276–e284. https://doi.org/10.1093/ije/dyac044.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Grubanov-Boskovic, Sara, Sona Kalantaryan, Silvia Migali, and Marco Scipioni. 2021. "The Impact of the Internet on Migration Aspirations and Intentions." *Migration Studies* 9 (4): 1807–1822.

Halpern-Manners, Andrew, Jonas Helgertz, John Robert Warren, and Evan Roberts. 2020. "The Effects of Education on Mortality: Evidence From Linked U.S. Census and Administrative Mortality Data." *Demography* 57 (4): 1513–1541.

Hamer, D., and L. Sirota. 2000. "Beware the Chopsticks Gene." *Molecular Psychiatry* 5 (1): 11–13.

Harris, Kathleen Mullan, Carolyn Tucker Halpern, Brett C. Haberstick, and Andrew Smolen. 2013. "The National Longitudinal Study of Adolescent Health (Add Health) Sibling Pairs Data." *Twin Research and Human Genetics* 16 (1): 391–398.

Hauer, Mathew E., Sunshine A. Jacobs, and Scott A. Kulp. 2024. "Climate Migration Amplifies Demographic Change and Population Aging." *Proceedings of the National Academy of Sciences* 121 (3): e2206192119.

Hays, Ron D., Honghu Liu, and Arie Kapteyn. 2015. "Use of Internet Panels to Conduct Surveys." *Behavior Research Methods* 47 (3): 685–690.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J. Thompson, Steven Ruggles, and Catherine A. Fitch. 2022. "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55 (1): 12–29.

Hellwege, Jacklyn, Jacob Keaton, Ayush Giri, Xiaoyi Gao, Digna R. Velez Edwards, and Todd L. Edwards. 2017. "Population Stratification in Genetic Association Studies." *Current Protocols in Human Genetics* 95: 1.22.1–1.22.23.

Herd, Pamela, Deborah Carr, and Carol Roan. 2014. "Cohort Profile: Wisconsin Longitudinal Study (WLS)." *International Journal of Epidemiology* 43 (1): 34–41.

Hurley, Mikella and Julius Adebayo. 2016. "Credit Scoring in the Era of Big Data." *Big Data* 18: 148–216.

Johnson, Rebecca, Ramina Sotoudeh, and Dalton Conley. 2022. "Polygenic Scores for Plasticity: A New Tool for Studying Gene-Environment Interplay." *Demography* 59 (3): 1045–1070.

Kapoor, Sayash, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. 2024. "REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science." *Science Advances* 10: eadk3452.

Kashyap, Ridhi. 2021. "Has Demography Witnessed a Data Revolution? Promises and Pitfalls of a Changing Data Ecosystem." *Population Studies* 75 (Sup 1): 47–75.

Kashyap, Ridhi, R. Gordon Rinderknecht, Aliakbar Akbaritabar, Diego Alburez-Gutierrez, Sofia Gil-Clavel, André Grow, Jisu Kim, Douglas R. Leasure, Sophie Lohmann, Daniela Veronica Negraia, Daniela Perrotta, Francesco Rampazzo, Chia-Jung Tsai, Mark D. Verhagen, Emilio

Zagheni, and Xinyi Zhao. 2022. "Digital and Computational Demography." SocArXiv. https://doi.org/10.31235/osf.io/7bvp.

Killworth, Peter D., Eugene C. Johnsen, Christopher McCarty, Gene Ann Shelley, and H. Russell Bernard. 1998. "A Social Network Approach to Estimating Seroprevalence in the United States." *Social Networks* 20 (1): 23–50.

King, Gary. 2016. *Computational Social Science: Discovery and Prediction*. Cambridge: Cambridge University Press.

King, Gary, and Nathaniel Persily. 2020. "A New Model for Industry–Academic Partnerships." *PS: Political Science & Politics* 53 (4): 703–709.

Lazer, David M. J., Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. 2020. "Computational Social Science: Obstacles and Opportunities." *Science* 369 (6507): 1060–1062.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–1205.

Leasure, Douglas R., Ridhi Kashyap, Francesco Rampazzo, Claire A. Dooley, Benjamin Elbers, Maksym Bondarenko, Mark Verhagen, Arun Frey, Jiani Yan, Evelina T. Akimova, Masoomali Fatehkia, Robert Trigwell, Andrew J. Tatem, Ingmar Weber, and Melinda C. Mills. 2023. "Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data." *Population and Development Review* 49 (2): 231–254.

Leasure, Douglas R., Warren C. Jochem, Eric M. Weber, Vincent Seaman, and Andrew J. Tatem. 2020. "National Population Mapping from Sparse Survey Data: A Hierarchical Bayesian Modeling Framework to Account for Uncertainty." *Proceedings of the National Academy of Sciences* 117 (39): 24173–24179.

Lee, Ronald. 2001. "Demography Abandons Its Core."

Leeper, Thomas J. 2019. "Where Have the Respondents Gone? Perhaps We Ate Them All." *Public Opinion Quarterly* 83 (S1): 280–288.

Levy, Brian L., Karl Vachuska, S. V. Subramanian and Robert J. Sampson. 2022. "Neighborhood Socioeconomic Inequality Based on Everyday Mobility Predicts COVID-19 Infection in San Francisco, Seattle, and Wisconsin." *Science Advances* 8 (7): eabl3825.

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. 2008. "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.Com." *Social Networks* 30 (4): 330–342.

Liebler, Carolyn A., Sonya R. Porter, Leticia E. Fernandez, James M. Noon, and Sharon R. Ennis. 2017. "America's Churning Races: Race and Ethnic Response Changes between Census 2000 and the 2010 Census." *Demography* 54 (1): 259–284. https://doi.org/10.1007/s13524-016-0544-0.

Loveman, Mara, and Jeronimo O. Muniz. 2007. "How Puerto Rico Became White: Boundary Dynamics and Intercensus Racial Reclassification." *American Sociological Review* 72 (6): 915–939.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–565.

Maltiel, Rachael, Adrian E. Raftery, Tyler H. McCormick, and Aaron J. Baraff. 2015. "Estimating Population Size Using the Network Scale up Method." *The Annals of Applied Statistics* 9 (3): 1247–1277.

Manolio, Teri A., Peter Goodhand, and Geoffrey Ginsburg. 2020. "The International Hundred Thousand Plus Cohort Consortium: Integrating Large-Scale Cohorts to Address Global Scientific Challenges." *The Lancet Digital Health* 2 (11): e567–e568.

Mare, Robert D. 2011. "A Multigenerational View of Inequality." *Demography* 48 (1): 1–23.

Matthews, Stephen A., Laura Stiberman, James Raymer, Tse-Chuan Yang, Ezra Gayawan, Sayambhu Saita, Sai Thein Than Tun, Daniel M. Parker, Deborah Balk, Stefan Leyk, Mark Montgomery, Katherine J. Curtis, and David W. S. Wong. 2021. "Looking Back, Looking Forward: Progress and Prospect for Spatial Demography." *Spatial Demography* 9 (1): 1–29.

McCormick, Tyler H., Matthew J. Salganik and Tian Zheng. 2010. "How Many People Do You Know?: Efficiently Estimating Personal Network Size." *Journal of the American Statistical Association* 105 (489): 59–70.

McCormick, Tyler H., and Tian Zheng. 2015. "Latent Surface Models for Networks Using Aggregated Relational Data." *Journal of the American Statistical Association* 110 (512): 1684–1695.

Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (i): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *The Annals of Applied Statistics* 12 (2): 685–726.

Mills, Melinda C., and Charles Rahal. 2020. "The GWAS Diversity Monitor Tracks Diversity by Disease in Real Time." *Nature Genetics* 52 (3): 242–243.

Mills, Melinda C., and Felix C. Tropf. 2020. "Sociology, Genetics, and the Coming of Age of Sociogenomics." *Annual Review of Sociology* 46 (1): 553–581.

Mills, Melinda C., Felix C. Tropf, David M. Brazel, Natalie van Zuydam, Ahmad Vaez., BIOS Consortium Management Team Heijmans Bastiaan T. 15 32 't Hoen Peter AC 62 63 van Meurs Joyce 66 Isaacs Aaron 72 Jansen Rick 34 Franke Lude 20 24, and Data Generation van Meurs Joyce 66 Jhamai P. Mila 66 Verbiest Michael 66 Suchiman H. Eka D. 32 Verkerk Marijn 66 van der Breggen Ruud 32 van Rooij Jeroen 66 Lakenberg Nico 32. 2021. "Identification of 371 Genetic Variants for Age at First Sex and Birth Linked to Externalising Behaviour." *Nature Human Behaviour* 5 (12): 1717–1730.

Nickerson, David W., and Todd Rogers. 2014. "Political Campaigns and Big Data." *Journal of Economic Perspectives* 28 (2): 51–74.

Nobles, Jenna, Lindsay Cannon, and Allen J. Wilcox. 2022. "Menstrual Irregularity as a Biological Limit to Early Pregnancy Awareness." *Proceedings of the National Academy of Sciences* 119 (1): e2113762118.

Noghanibehambari, Hamid, and Michal Engelman. 2022. "Social Insurance Programs and Later-Life Mortality: Evidence from New Deal Relief Spending." *Journal of Health Economics* 86: 102690.

Oberski, Daniel L., and Frauke Kreuter. 2020. "Differential Privacy and Social Science: An Urgent Puzzle." *Harvard Data Science Review* 2 (1): 1–21.

Okbay, Aysu, Jonathan P. Beauchamp, Mark Alan Fontana, James J. Lee, Tune H. Pers, Cornelius A. Rietveld, Patrick Turley, Guo-Bo Chen, Valur Emilsson, S. Fleur W. Meddens, Sven Oskarsson, Joseph K. Pickrell, Kevin Thom, Pascal Timshel, Ronald de Vlaming, Abdel Abdellaoui, Tarunveer S. Ahluwalia, Jonas Bacelis, Clemens Baumbach, Gyda Bjornsdottir, Johannes H. Brandsma, Maria Pina Concas, Jaime Derringer, Nicholas A. Furlotte, Tessel E. Galesloot, Giorgia Girotto, Richa Gupta, Leanne M. Hall, Sarah E. Harris, Edith Hofer, Momoko Horikoshi, Jennifer E. Huffman, Kadri Kaasik, Ioanna P. Kalafati, Robert Karlsson, Augustine Kong, Jari Lahti, Sven J. van der Lee, Christiaan deLeeuw, Penelope A. Lind, Karl-Oskar Lindgren, Tian Liu, Massimo Mangino, Jonathan Marten, Evelin Mihailov, Michael B. Miller, Peter J. van der Most, Christopher Oldmeadow, Antony Payton, Natalia Pervjakova, Wouter J. Peyrot, Yong Qian, Olli Raitakari, Rico Rueedi, Erika Salvi, Börge Schmidt, Katharina E. Schraut, Jianxin Shi, Albert V. Smith, Raymond A. Poot, Beate St Pourcain, Alexander Teumer, Gudmar Thorleifsson, Niek Verweij, Dragana Vuckovic, Juergen Wellmann, Harm-Jan Westra, Jingyun Yang, Wei Zhao, Zhihong Zhu, Behrooz Z. Alizadeh, Najaf Amin, Andrew Bakshi, Sebastian E. Baumeister, Ginevra Biino, Klaus Bønnelykke, Patricia A. Boyle, Harry Campbell, Francesco P. Cappuccio, Gail Davies, Jan-Emmanuel De Neve, Panos Deloukas, Ilja Demuth, Jun Ding, Peter Eibich, Lewin Eisele, Niina Eklund, David M. Evans, Jessica D. Faul, Mary F. Feitosa, Andreas J. Forstner, Ilaria Gandin, Bjarni Gunnarsson, Bjarni V. Halldórsson, Tamara B. Harris, Andrew C. Heath, Lynne J. Hocking, Elizabeth G. Holliday, Georg Homuth, Michael A. Horan, Jouke-Jan Hottenga, Philip L. de Jager, Peter K. Joshi, Astanand Jugessur, Marika A. Kaakinen, Mika Kähönen, Stavroula Kanoni, Liisa Keltigangas-Järvinen, Lambertus A. L. M. Kiemeney, Ivana Kolcic, Seppo Koskinen, Aldi T. Kraja, Martin Kroh, Zoltan Kutalik, Antti Latvala, Lenore J. Launer, Maël P. Lebreton, Douglas F. Levinson, Paul Lichtenstein, Peter Lichtner, David C. M. Liewald, LifeLines Cohort Study, Anu Loukola, Pamela A. Madden, Reedik Mägi, Tomi Mäki-Opas, Riccardo E. Marioni, Pedro Marques-Vidal, Gerardus A. Meddens, George McMahon, Christa Meisinger, Thomas Meitinger, Yusplitri Mi-

laneschi, Lili Milani, Grant W. Montgomery, Ronny Myhre, Christopher P. Nelson, Dale R. Nyholt, William E. R. Ollier, Aarno Palotie, Lavinia Paternoster, Nancy L. Pedersen, Katja E. Petrovic, David J. Porteous, Katri Räikkönen, Susan M. Ring, Antonietta Robino, Olga Rostapshova, Igor Rudan, Aldo Rustichini, Veikko Salomaa, Alan R. Sanders, Antti-Pekka Sarin, Helena Schmidt, Rodney J. Scott, Blair H. Smith, Jennifer A. Smith, Jan A. Staessen, Elisabeth Steinhagen-Thiessen, Konstantin Strauch, Antonio Terracciano, Martin D. Tobin, Sheila Ulivi, Simona Vaccargiu, Lydia Quaye, Frank J. A. van Rooij, Cristina Venturini, Anna A. E. Vinkhuyzen, Uwe Völker, Henry Völzke, Judith M. Vonk, Diego Vozzi, Johannes Waage, Erin B. Ware, Gonneke Willemsen, John R. Attia, David A. Bennett, Klaus Berger, Lars Bertram, Hans Bisgaard, Dorret I. Boomsma, Ingrid B. Borecki, Ute Bültmann, Christopher F. Chabris, Francesco Cucca, Daniele Cusi, Ian J. Deary, George V. Dedoussis, Cornelia M. van Duijn, Johan G. Eriksson, Barbara Franke, Lude Franke, Paolo Gasparini, Pablo V. Gejman, Christian Gieger, Hans-Jörgen Grabe, Jacob Gratten, Patrick J. F. Groenen, Vilmundur Gudnason, Pim van der Harst, Caroline Hayward, David A. Hinds, Wolfgang Hoffmann, Elina Hyppönen, William G. Iacono, Bo Jacobsson, Marjo-Riitta Järvelin, Karl-Heinz Jöckel, Jaakko Kaprio, Sharon L. R. Kardia, Terho Lehtimäki, Steven F. Lehrer, Patrik K. E. Magnusson, Nicholas G. Martin, Matt McGue, Andres Metspalu, Neil Pendleton, Brenda W. J. H. Penninx, Markus Perola, Nicola Pirastu, Mario Pirastu, Ozren Polasek, Danielle Posthuma, Christine Power, Michael A. Province, Nilesh J. Samani, David Schlessinger, Reinhold Schmidt, Thorkild I. A. Sørensen, Tim D. Spector, Kari Stefansson, Unnur Thorsteinsdottir, A. Roy Thurik, Nicholas J. Timpson, Henning Tiemeier, Joyce Y. Tung, André G. Uitterlinden, Veronique Vitart, Peter Vollenweider, David R. Weir, James F. Wilson, Alan F. Wright, Dalton C. Conley, Robert F. Krueger, George Davey Smith, Albert Hofman, David I. Laibson, Sarah E. Medland, Michelle N. Meyer, Jian Yang, Magnus Johannesson, Peter M. Visscher, Tõnu Esko, Philipp D. Koellinger, David Cesarini, and Daniel J. Benjamin. 2016. "Genome-Wide Association Study Identifies 74 Loci Associated with Educational Attainment." *Nature* 533 (7604): 539–542.

Okbay, Aysu, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Seyed Moeen Nehzati, Julia Sidorenko, Hyeokmoon Kweon, Grant Goldman, Tamara Gjorgjieva, Yunxuan Jiang, Barry Hicks, Chao Tian, David A. Hinds, Rafael Ahlskog, Patrik K. E. Magnusson, Sven Oskarsson, Caroline Hayward, Archie Campbell, David J. Porteous, Jeremy Freese, Pamela Herd, Chelsea Watson, Jonathan Jala, Dalton Conley, Philipp D. Koellinger, Magnus Johannesson, David Laibson, Michelle N. Meyer, James J. Lee, Augustine Kong, Loic Yengo, David Cesarini, Patrick Turley, Peter M. Visscher, Jonathan P. Beauchamp, Daniel J. Benjamin, and Alexander I. Young. 2022. "Polygenic Prediction of Educational Attainment within and between Families from Genome-Wide Association Analyses in 3 Million Individuals." *Nature Genetics* 54 (4): 437–449.

Olivier, Lex. 2011. "River Sampling Non Probability Sampling in an Online Environment." *LEX OLIVIER*. https://lexolivier.blogspot.com/2011/11/river-sampling-non-probability-sampling.html.

Palotti, Joao, Natalia Adler, Alfredo Morales-Guzman, Jeffrey Villaveces, Vedran Sekara, Manuel Garcia Herranz, Musa Al-Asad, and Ingmar Weber. 2020. "Monitoring of the Venezuelan Exodus through Facebook's Advertising Platform." *PLoS ONE* 15 (2): e0229175.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–385.

Pejcinovska, Marija, Monica Alexander, Jessie Yeung and Alison Gemmill. 2023. "MRP as a Tool in the Population Sciences: Potential Benefits and Challenges." https://www.monicaalexander.com/pdf/mrp_chapter.pdf.

Pesando, Luca Maria, Valentina Rotondi, Manuela Stranges, Ridhi Kashyap and Francesco C. Billari. 2021. "The Internetization of International Migration." *Population and Development Review* 47 (1): 79–111.

Poulain, Michel, Anne Herm and Roger Depledge. 2013. "Central Population Registers as a Source of Demographic Statistics in Europe." *Population* 68 (2): 183–212.

Rampazzo, Francesco, Jakub Bijak, Agnese Vitali, Ingmar Weber and Emilio Zagheni. 2021. "A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom." *Demography* 58 (6): 2193–2218.

Rindfuss, Ronald R., Minja K. Choe, Noriko O. Tsuya, Larry L. Bumpass, and Emi Tamaki. 2015. "Do Low Survey Response Rates Bias Results? Evidence from Japan." *Demographic Research* 32: 797–828.

Robinson, Matthew R., Aaron Kleinman, Mariaelisa Graff, Anna A. E. Vinkhuyzen, David Couper, Michael B. Miller, Wouter J. Peyrot, Abdel Abdellaoui, Brendan P. Zietsch, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, Harold Snieder, Sarah E. Medland, Nicholas G. Martin, Patrik K. E. Magnusson, William G. Iacono, Matt McGue, Kari E. North, Jian Yang, and Peter M. Visscher. 2017. "Genetic Evidence of Assortative Mating in Humans." *Nature Human Behaviour* 1 (1): 1–13.

Rogers, Andrei, Frans Willekens, and James Raymer. 2003. "Imposing Age and Spatial Structures on Inadequate Migration-Flow Datasets." *The Professional Geographer* 55 (1): 56–68.

Rohe, Karl. 2019. "A Critical Threshold for Design Effects in Network Sampling." *The Annals of Statistics* 47 (1): 556–582.

Rossier, Clémentine, Onikepe Owolabi, Seni Kouanda, Martin Bangha, Caron R. Kim, Bela Ganatra, Dennis Feehan, Casey Breen, Moussa Zan, Rachidatou Compaoré, Adama Baguiya, Ramatou Ouédraogo, Clement Oduor, Vincent Bagnoa, and Sherine Athero. 2022. "Describing the Safety of Abortion at the Population Level Using Network-Based Survey Approaches." *Reproductive Health* 19 (1): 231.

Rotondi, Valentina, Ridhi Kashyap, Luca Maria Pesando, Simone Spinelli, and Francesco C. Billari. 2020. "Leveraging Mobile Phones to Attain Sustainable Development." *Proceedings of the National Academy of Sciences* 117 (24): 13413–13420.

Ruggles, Steven. 2007. "The Decline of Intergenerational Coresidence in the United States, 1850 to 2000." *American Sociological Review* 72 (6): 964–989.

Ruggles, Steven. 2014. "Big Microdata for Population Research." *Demography* 51 (1): 287–297.

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. "Differential Privacy and Census Data: Implications for Social and Economic Research." *AEA Papers and Proceedings* 109: 403–408.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Mathew Sobek. 2020. "IPUMS USA: Version 10.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.

Ruiz-Linares, Andrés, Kaustubh Adhikari, Victor Acuña-Alonzo, Mirsha Quinto-Sanchez, Claudia Jaramillo, William Arias, Macarena Fuentes, María Pizarro, Paola Everardo, Francisco de Avila, Jorge Gómez-Valdés, Paola León-Mimila, Tábita Hunemeier, Virginia Ramallo, Caio C. Silva de Cerqueira, Mari-Wyn Burley, Esra Konca, Marcelo Zagonel de Oliveira, Mauricio Roberto Veronez, Marta Rubio-Codina, Orazio Attanasio, Sahra Gibbon, Nicolas Ray, Carla Gallo, Giovanni Poletti, Javier Rosique, Lavinia SchulerFaccini, Francisco M. Salzano, Maria-Cátira Bortolini, Samuel Canizales-Quinteros, Francisco Rothhammer, Gabriel Bedoya, David Balding, and Rolando Gonzalez-José. 2014. "Admixture in Latin America: Geographic Structure, Phenotypic Diversity and SelfPerception of Ancestry Based on 7,342 Individuals." *PLoS Genetics* 10 (9): e1004572.

Ruiz, Neil G, Luis Noe-Bustamante, and Sono Shah. 2023. "Diverse Cultures and Shared Experiences Shape Asian American Identities." Washington, DC: Pew Research Center.

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Saperstein, Aliya and Aaron Gullickson. 2013. "A "Mulatto Escape Hatch" in the United States? Examining Evidence of Racial and Social Mobility During the Jim Crow Era." *Demography* 50 (5): 1921–1942.

Saperstein, Aliya, and Andrew M. Penner. 2012. "Racial Fluidity and Inequality in the United States." *American Journal of Sociology* 118 (3): 676–727.

Särndal, Carl-Erik, Beng t Swensson and Jan Wretman. 2003. *Model Assisted Survey Sampling*. Berlin: Springer Science & Business Media.

Schneider, Daniel, and Kristen Harknett. 2022. "What's to Like? Facebook as a Tool for Survey Data Collection." *Sociological Methods & Research* 51 (1): 108–140.

Shadbolt, Nigel, Alys Brett, Min Chen, Glenn Marion, Iain J. McKendrick, Jasmina Panovska-Griffiths, Lorenzo Pellis, Richard Reeve, and Ben Swallow. 2022. "The Challenges of Data in Future Pandemics." *Epidemics* 40: 100612.

Song, Xi. 2021. "Multigenerational Social Mobility: A Demographic Approach." *Sociological Methodology* 51 (1): 1–43.

Song, Xi, and Cameron D. Campbell. 2017. "Genealogical Microdata and Their Significance for Social Science." *Annual Review of Sociology* 43 (1): 75–99.

Sonnega, Amanda, Jessica D Faul, Mary Beth Ofstedal, Kenneth M Langa, John WR Phillips, and David R Weir. 2014. "Cohort Profile: The Health and Retirement Study (HRS)." *International Journal of Epidemiology* 43 (2): 576–585.

Sullivan, Patrick F., Arpana Agrawal, Cynthia M. Bulik, Ole A. Andreassen, Anders D. Børglum, Gerome Breen, Sven Cichon, Howard J. Edenberg, Stephen V. Faraone, Joel Gelernter, Carol A. Mathews, Caroline M. Nievergelt, Jordan W. Smoller, and Michael C. O'Donovan. 2018. "Psychiatric Genomics: An Update and an Agenda." *American Journal of Psychiatry* 175 (1): 15–27.

Sully, Elizabeth, Margaret Giorgio and Selena Anjur-Dietrich. 2020. "Estimating Abortion Incidence Using the Network Scale-up Method." *Demographic Research* 43: 1651–1684.

The All of Us Research Program Investigators. 2019. "The "All of Us" Research Program." *The New England Journal of Medicine* 381 (7): 668–676.

Thorvaldsen, Gunnar, Trygve Andersen, and Hilde L. Sommerseth. 2015. "Record Linkage in the Historical Population Register for Norway." In *Population Reconstruction*, edited by Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 155–171. Cham: Springer International Publishing.

Trani, Elyssa A., Robert A. Hummer, Mary Jane Hill, Eric A. Whitsel, and Laura R. Loehr. 2022. "Mortality Outcomes Surveillance, Part I: Ascertaining Decedents." https://doi.org/10.17615/ST8V-TG84.

von Elm, Erik, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche and Jan P Vandenbroucke. 2007. "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *BMJ: British Medical Journal* 335 (7624): 806–808.

Ward, Zachary. 2023. "Intergenerational Mobility in American History: Accounting for Race and Measurement Error." *American Economic Review* 113 (12): 3213–3248. https://doi.org/10.1257/aer.20200292.

Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago, IL: University of Chicago Press.

Wikle, Anna, and Maria Osborne. 2023. "CenSoc WWII Army Enlistment Dataset." https://censoc.berkeley.edu/wp-content/uploads/2024/07/military_enlistment_technical_report-4.pdf

Wong, Kerry L. M., Amy Gimma, Pietro Coletti, Daniela Paolotti, Michele Tizzani, Ciro Cattuto, Andrea Schmidt, Gerald Gredinger, Sophie Stumpfl, Joaquin Baruch, Tanya Melillo, Henrieta Hudeckova, Jana Zibolenova, Zuzana Chladna, Magdalena Rosinska, Marta Niedzwiedzka-Stadnik, Krista Fischer, Sigrid Vorobjov, Hanna Sõnajalg, Christian Althaus, Nicola Low, Martina Reichmuth, Kari Auranen, Markku Nurhonen, Goranka Petrović, Zvjezdana Lovric Makaric, Sónia Namorado, Constantino Caetano, Ana João Santos, Gergely Röst, Beatrix Oroszi, Márton Karsai, Mario Fafangel, Petra Klepac, Natalija Kranjec, Cristina Vilaplana, Jordi Casabona, Christel Faes, Philippe Beutels, Niel Hens, Veronika K. Jaeger, Andre Karch, Helen Johnson, Wjohn Edmunds, Christopher I. Jarvis., and CoMix Europe Working Group. 2023. "Social Contact Patterns during the COVID-19 Pandemic in 21 European Countries – Evidence from a Two-Year Study." *BMC Infectious Diseases* 23 (1): 268.

Wrigley-Field, Elizabeth, Mathew V. Kiang, Alicia R. Riley, Magali Barbieri, Yea-Hung Chen, Kate A. Duchowny, Ellicott C. Matthay, David Van Riper, Kirrthana Jegathesan, Kirsten Bibbins-Domingo, and Jonathon P. Leider. 2021. "Geographically Targeted COVID-19 Vaccination Is

More Equitable and Averts More Deaths than Age-Based Thresholds Alone." *Science Advances* 7 (40): eabj2099.

Yun, Gi Woong, and Craig W. Trumbo. 2000. "Comparative Response to a Survey Executed by Post, E-mail, & Web Form." *Journal of Computer-Mediated Communication* 6 (1): 0–0.

Zagheni, Emilio, and Ingmar Weber. 2015. "Demographic Research with Non-Representative Internet Data." *International Journal of Manpower* 36 (1): 13–25.

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. 2017. "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants." *Population and Development Review* 43 (4): 721–734.

Zenebe-Gete, Selam, Rebecca Salowe, and Joan M. O'Brien. 2021. "Benefits of Cohort Studies in a Consortia-Dominated Landscape." *Frontiers in Genetics* 12: 801653.