PNAS PERSPECTIVE

# Balancing data privacy and usability in the federal statistical system

V. Joseph Hotz[a,1], Christopher R. Bollinger[b], Tatiana Komarova[c], Charles F. Manski[d,1], Robert A. Moffitt[e], Denis Nekipelov[f], Aaron Sojourner[g], and Bruce D. Spencer[h]

Edited by James Poterba, Massachusetts Institute of Technology, Cambridge, MA; received November 13, 2021; accepted June 15, 2022

**The federal statistical system is experiencing competing pressures for change. On the one hand, for confidentiality reasons, much socially valuable data currently held by federal agencies is either not made available to researchers at all or only made available under onerous conditions. On the other hand, agencies which release public databases face new challenges in protecting the privacy of the subjects in those databases, which leads them to consider releasing fewer data or masking the data in ways that will reduce their accuracy. In this essay, we argue that the discussion has not given proper consideration to the reduced social benefits of data availability and their usability relative to the value of increased levels of privacy protection. A more balanced benefit–cost framework should be used to assess these trade-offs. We express concerns both with synthetic data methods for disclosure limitation, which will reduce the types of research that can be reliably conducted in unknown ways, and with differential privacy criteria that use what we argue is an inappropriate measure of disclosure risk. We recommend that the measure of disclosure risk used to assess all disclosure protection methods focus on what we believe is the risk that individuals should care about, that more study of the impact of differential privacy criteria and synthetic data methods on data usability for research be conducted before either is put into widespread use, and that more research be conducted on alternative methods of disclosure risk reduction that better balance benefits and costs.**

federal statistical system | data disclosure risk | data access

The United States prides itself on having a broad, extensive federal statistical system. The federal government has 13 official statistical agencies collecting an impressive range of data including classic population counts, monthly unemployment rates, the number of children receiving subsidized school lunches, and myriad other topics. In addition to official statistical agencies, dozens of other federal agencies gather data and furnish statistics on government programs and operations. This system provides descriptive evidence on the population of the nation, its states and local communities, data used to make important decisions such as congressional apportionment, data used for research on determinants of social and economic outcomes, and information used in the important, practical evaluation of public programs' effectiveness (1).

The federal statistical system is currently experiencing major pressures for change from two opposing directions. On the one hand, the US Commission for Evidence-Based Policymaking pointed out, in a 2017 report, that massive amounts of data currently held by federal agencies are not made available to researchers (or even to other government agencies) or are made available only under onerous conditions, but could usefully drive improvements in policy making and research (2). The commission convincingly argued that major increases in federal data availability would have great benefits to society. For example, data from confidential tax records have enormous value in illuminating levels of income and wealth inequality, their trends, and how barriers to upward social mobility vary across the country (3, 4), but those data have been unavailable to all but a few researchers. Recognizing this, Congress passed legislation in 2018 (5) creating an Advisory Committee on Data for Evidence Building to recommend how federal agencies should increase the research and scientific communities' access to their data (6). Some agencies, such as the IRS's Statistics of Income program, have begun to devise ways to release more data from individual tax returns to credentialed researchers than ever before (7).

On the other hand, many federal agencies, policy makers, and researchers see new challenges to safeguarding the confidentiality of information the data agencies currently release, and in protecting the privacy of subjects in those released databases. Significant advances in computational infrastructure and methods combined with the growing availability of commercial databases containing detailed information on individuals add risks. For example, the US Census Bureau conducted a simulated "reconstruction attack" on the 2010 data it had publicly released, data that had already been subjected to its conventional "disclosure avoidance system" (DAS). A DAS is a set of methods to

Author affiliations: [a]Department of Economics, Duke University, Durham, NC 27708; [b]Department of Economics, University of Kentucky, Lexington, KY 40503; [c]The London School of Economics and Political Science, London WC2A 3PH, United Kingdom; [d]Department of Economics, Northwestern University, Evanston, IL 60208; [e]Department of Economics, Johns Hopkins University, Baltimore, MD 21211; [f]Department of Economics, University of Virginia, Charlottesville, VA 22904; [g]W. E. Upjohn Institute for Employment Policy, Kalamazoo, MI 49007; and [h]Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208

Author contributions: V.J.H., C.R.B., T.K., C.F.M., R.A.M., D.N., A.S., and B.D.S. designed research; V.J.H., C.R.B., T.K., C.F.M., R.A.M., D.N., A.S., and B.D.S. performed research; and V.J.H., C.R.B., T.K., C.F.M., R.A.M., D.N., A.S., and B.D.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]To whom correspondence may be addressed. Email: v.joseph.hotz@duke.edu or cfmanski@northwestern.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2104906119/-/DCSupplemental.

mask certain elements of the data to prevent attribution of survey response values to particular survey respondents. The bureau found that published tabulations from these data could be used to produce a "highly accurate" microdata file that violated the bureau's previous disclosure avoidance rules. In combination with commercial databases, these could be used to correctly reidentify between 52 million and 179 million of the almost 309 million 2010 Census respondents (8). The bureau concluded, based on these findings among others, that it needed a new DAS for the 2020 Census data. The bureau's subsequent evaluations of its new DAS, which is based on the concept of differential privacy (DP) (9), produced evidence suggesting that it has greater privacy protections (10). However, the bureau's findings and adoption of a new DAS sparked a vigorous debate, contesting the appropriate interpretation of the bureau's findings and the impacts of this new DAS on the data released from the 2020 Census (11–18).

Discussions of how to balance the objectives of data usability and privacy have a long history, beginning in US law with the Privacy Act of 1920 (19, 20). Newly emerging threats to privacy have led to much discussion in the research community and the federal government about how best to strike a balance. Some argue that the increased privacy protections that some federal agencies envisage will harm the US federal statistical system and result in major losses to society's research and knowledge base.

This essay provides a critical overview of the issues facing the federal statistical system in balancing release of accurate, usable data with privacy protections. While there are previous discussions of how to achieve the right balance, and of the advantages and disadvantages of particular methods of privacy protection, our discussion differs in a number of respects. First, we aim to make progress in defining the value of accurate, "usable" data, going beyond the measures used in the past, to provide a different and broader conceptual framework for understanding the problem. Second, rather than adopt the perspective of computer science, which mostly focuses on algorithms for privacy protection, we take a perspective rooted in economics and statistics. This perspective provides a comprehensive framework to understand the trade-offs at hand. We advocate weighing the benefits of privacy protection methods in a meaningful way against their costs, in terms of how much they would degrade the quality of public decisions given the consequences of privacy protection for the properties of statistical estimation and inference. From this perspective, we analyze two topics related to disclosure avoidance currently under discussion or in use by statistical agencies: the method of synthetic data as an approach to limiting disclosure risk and the use of DP as a metric to measure the level of disclosure risk.

We do not address legal restrictions embodied in current statutes such as Title 13, Title 26, the Confidential Information and Protection Statistical Efficiency Act, and other regulations (21). The Commission on Evidence-Based Policy-Making referenced above (5) suggested that some of these statutes and regulations should be revisited by Congress to allow government agencies to better manage the data usability–privacy protection trade-off. We leave such legal issues to others, although we note substantial disagreement in current discussions of whether the specific disclosure limitation methods noted above align with the privacy protections guaranteed under these statutes (22).

## 1. Privacy and Usability for Federal Statistical Data

**1.1. Broad Concepts.** Researchers commonly consider privacy and data confidentiality as distinct concepts. Privacy refers to an individual's right to control what information is collected about her or him (23, 24). Confidentiality refers to the responsibility of a data steward (i.e., the agent in charge of data development, such as a statistical agency) not to divulge subjects' personal information and/or identity to unauthorized parties (24, 25). The distinction between privacy and data confidentiality gets blurred in practice. For example, individuals may not always provide informed consent for use of their data, such as when statistical agencies use administrative records to measure earnings or program benefits. Nonetheless, agencies may still have legal obligations to maintain the confidentiality of this information and protect the identities of individuals. In this essay, we use the terms privacy and data confidentiality interchangeably for simplicity.

The term "disclosure risk" broadly means the probability that specified information about a particular data subject in a particular database and presumed private will be obtained by an unauthorized party and associated with the data subject. In some cases, the data subject will have provided the data under an explicit pledge of confidentiality from the data collector; in other cases, a government agency will have gathered the data subject's information without such a pledge but, nevertheless, with legal or presumed confidentiality. Disclosure could refer to association of a known individual with a particular microdata record ("identification") or to association of information in a microdataset (e.g., the value of a particular variable) with a particular individual ("attribution") (26). Government agencies collect data on individuals, households, businesses, governmental units, and nonprofit entities. In this essay, we focus on situations where individual or household privacy is at issue and exclude consideration of the protection of privacy for data on business establishments. (See ref. 27 for discussion of privacy protection for business data.)

Data usability refers to the potential for data to effectively help answer specific questions of interest. In research, the question of interest commonly is to learn the value of an empirical quantity (or estimand) deemed important from the perspective of science or society. The quantity, for example, might be the poverty rate, life expectancy in a specified population, or a correlation coefficient or a regression coefficient in a linear regression model. When data are used to inform policy making, one objective is to evaluate the effects of different government policies and programs and to use this information to make a decision that yields the best possible social welfare (21). Data systems that reduce data usability are likely to worsen policy-making decisions and reduce social welfare. See section 1.2 for further discussion.

In recognition of disclosure risks, data may not be released at all or, if released, may be so extensively altered to limit disclosure risks that they lose value in addressing

some questions effectively. And the effectiveness of their use also depends on the quality of the original data, including sampling error, measurement error, nonresponse error, and the level of geographic detail.

Aiming to measure how data quality affects research, statisticians and econometricians study the accuracy with which a quantity of interest can be estimated. Statisticians often use mean square error (MSE), the expected squared difference between a quantity of interest and its estimate, as an omnibus measure of accuracy (28). This measure may be reasonable in some applications, such as shares of a sample in various race/ethnic groups, but not in others. For example, MSE may be the suitable metric to measure accuracy when one aims to monitor a phenomenon (e.g., extent of poverty or racial and ethnic segregation), but not be adequate when evaluating the social welfare gain of policy impacts (e.g., effects of government training programs on later life outcomes or of pollutants on child death rates).

For any quantity of interest, achievable accuracy may be limited, for two different reasons: statistical imprecision and identification problems.* In *SI Appendix, Appendix A*, we characterize both of these problems and how they differ from one another. In *SI Appendix, Appendixes B* and *C*, we discuss how these problems manifest when using data subject to alternative forms of privacy protection.

**1.2. Evaluating the Social Welfare of Disclosure Limitation Policies.** Broad conceptualization of privacy and data usability does no more than introduce the difficult problem that federal statistical agencies face in choosing disclosure limitation policies. Reasonable policy formation requires much more. It needs sensible measurement of concepts and coherent evaluation of the benefits and costs of alternative feasible policies. Furthermore, it should recognize that the magnitudes of the relevant benefits and costs may be highly uncertain. While we do not attempt to solve the difficult problem of operationalization, we propose a framework taken from economics that gives a focused way to understand the problem.

The subfield of public economics has long strived to inform policy formation through specification of what is known as a "social welfare function" that determines how the level of social well-being, or welfare, is affected by public policies. By using a univariate function, it evaluates the social benefits and costs of alternative policies on a common scale. Then, society may aim to choose a policy that maximizes social welfare defined as social benefits minus social costs. Practical implementation of this general idea is commonly called benefit–cost analysis.

While we will not conduct a benefit–cost analysis for any particular disclosure limitation scheme, it is useful to characterize the elements of such analyses to help frame our discussion. Benefit–cost analysis requires assessments of both facts and values. First, the factual changes in both data usability and privacy protection need investigation, listing the research and decisions that the data should inform, quantifying the extent to which its usability is reduced, and quantifying the reduction in the risk of disclosure. Broader impacts, such as the willingness of individuals to participate in surveys and the information loss if they do not, also should be included. Second, an assessment of the value that society puts on both the reduction in data usability and reduction in disclosure risk needs to be conducted, which then informs an assessment of whether social benefits of the disclosure limitation scheme exceed social costs.[†]

In addition, any evaluation of a disclosure reduction strategy needs to determine how it will address uncertainty in the severity of disclosure risk, in the loss to welfare with disclosure, and in the social value of the data made public. Whether it should (unrealistically) ignore such uncertainty, as has been common in past applications of benefit–cost analysis to transportation and environmental policy, as well as the data dissemination programs of federal statistical agencies, must be determined, as well as whether it should adopt something like a Bayesian perspective, in which the evaluator places a subjective probability distribution on unknown quantities and seeks to maximize subjective expected social welfare. A determination also must be made on whether the evaluator, recognizing the possibility of deep uncertainty, assesses policy using a criterion for decision-making under ambiguity, such as the maximin or minimax regret criterion, which, in different senses, aim to achieve satisfactory social welfare performance whatever values unknown quantities may take. Explicit recognition of uncertainty in benefit–cost analysis has been rare, but precedents exist (29).

To illustrate the issues that a proper benefit–cost analysis should address, consider the US Census Bureau's decision to subject all 2020 Census data products released to the public to a new DAS intended to satisfy a differentially private criterion. As we see it, the questions that should be addressed in a full assessment of this change include the following: What types of research and governmental decisions (e.g., legislative districting, federal allocation of funds to states and localities) that rely on Census data will be materially affected under this DAS as implemented? How does this approach change the probabilities of identification of specific individuals or households, and what are the social costs if disclosures of identities or attributes occur? What changes to social welfare will result? Finally, how should this new DAS incorporate the many uncertainties involved in its implementation?

The Census Bureau has stated its intention "to preserve the utility of our legally mandated data products while also ensuring that every respondents' personal information is fully protected" (30). However, the new DAS has primarily concentrated on the privacy loss side, and the above issues have not yet been adequately addressed. As we discuss in the next section, many disclosure limitation methods focus primarily on the privacy loss side and too little on the benefit side of data release. Hence, their adoption decisions are not adequately grounded in balanced benefit–cost analyses.

## 2. Disclosure Avoidance Strategies

Statistical agencies have long struggled with how to make their data available to various user communities while

*This use of "identification" is rooted in econometrics and closely related to consistency in statistics and is different from reidentification risks in the context of privacy protection.

[†]The literature on the privacy–utility trade-off commonly uses "R–U" curves to illustrate the trade-off between risk and utility. Abowd and Schmutte (28) pose this trade-off using the language and framework of economics, as we do. However, while the papers in this literature often discuss R–U curves, they rarely make them concrete, and almost never operationalize the framework in the way we are suggesting in this paragraph.

protecting the privacy of individual subjects. Some agencies have long histories of using different strategies, or DASs, to limit the risk of disclosing individual identities or attributes. Some agencies have decided not to release their data at all.

Statisticians were the leading group developing the initial approaches, under the heading of statistical disclosure limitation (SDL) or statistical disclosure control. Systematic reviews of SDL are Duncan et al. (24) and Matthews and Harel (31). With the development of databases that contain massive amounts of information on individuals ("big data") and growing computational power, computer scientists have taken an increasing role in developing strategies for privacy-preserving analysis of data. Much of this work focuses on databases in the commercial and health sectors, but governmental agencies, especially statistical ones, increasingly turn to computer science to devise ways to meet their legal requirements to protect privacy in the data they disseminate. Below, we discuss three topics: traditional methods, synthetic data methods, and the use of DP as a measure of disclosure risk. Before doing so, we summarize the modes of access that users can have to confidential, privatized data and outline the key concept of disclosure risk and how it is measured, which cuts across all of the SDLs.

**2.1. Modes of Access.** Discussion of disclosure limitation strategies is usefully organized in terms of the ways in which statistical agencies limit the access of users to the "original" confidential microdata that is potentially disclosive if released in its unaltered form.[‡] Karr (32, 33) distinguishes three modes of access that characterize alternative disclosure limitation strategies agencies use.

One mode is dissemination-based access. In this case, statistical agencies provide microdata files to the general public for whatever analyses they wish to conduct with those data. This is the most common form of access to the research community, and government agencies release hundreds of public use microdata files of this type. However, because the risk of disclosure is particularly great given the wide variety of users and types of analyses that can occur, agencies always apply some type of SDL strategies to the data before release. We will discuss the types used below. Applying such strategies also will inevitably reduce the usability of the data to some extent.

Karr also includes in this category what are called synthetic data files, which are pseudoversions of microdata files equivalent to a surrogate version of the original data, as we will discuss below. While this form of data access has existed for decades, federal agencies have adopted it only rarely, to date. (See section 2.4 for exceptions.)

A second mode is what Karr terms direct access to most or all of the original data. Because of the obvious risk of disclosure in providing the original data in almost unaltered form, direct access is limited in one of two ways: through licensing or so-called research data centers (RDCs). In both cases, users enter into legally binding agreements with agencies for use of specified data elements (samples and variables) and face penalties for misuse of data. In the licensing case, users are required to maintain security arrangements at their home institutions and usually have to submit output to review prior to publication. In RDC access, approved users are allowed to analyze approved data elements of confidential microdata files in secure enclaves. All output removed from an enclave is subjected to disclosure review by the statistical agency. RDCs have historically been physical facilities, but enclaves also can be virtual, online entities.

The penalties for misuse of this mode of access are an important feature. Penalties range from significant monetary fines to loss of access to the data, both currently and for some future period. Penalties have not been used in the other forms of access that we discuss, but they could be considered for more widespread use. Penalties are a method of increasing the cost to intruders who wish to conduct a reidentification attack, and could be an additional tool in a government agency's toolkit to reduce the risk of disclosure.

Several statistical agencies provide either licensing or RDC access. In contrast to open public provision of microdata files, which necessarily require strong SDL limitations, licensing and RDC access can allow users access to virtually the entire original dataset. Typically, users of data under restricted access are credentialed researchers employed by research institutions, and not the general public. The disadvantage of the licensing and RDC mode is that barriers to their use are currently very high, often requiring physical travel to a site (including the requirement of working only at the agency itself) and/or with onerous disclosure limitation procedures (such as manual checking of output) that can delay release of results for weeks or months. For this reason, licensing and RDC access is currently quite limited relative to the use of publicly available open access.

A third mode is query-based access. Here, users are not given access to either the original data—either in public use form or restricted form—or to microdata that have been significantly altered. Rather, users are allowed to pose queries, such as requests for summary statistics or estimates from statistical analyses (e.g., regression coefficients). While this mode greatly limits the ability of users to conduct analyses, it does make strong privacy protections possible, because access to a microdataset itself is not permitted. It makes exploratory data analysis more complicated and potentially greatly limits users' ability to assess properties of estimates they seek. At the same time, it greatly enhances the ability of data providers to control disclosive information contained in a confidential database. Some statistical agencies provide query-based access through their online query systems—such as the Census Bureau's American Factfinder system and its successor https://data.census.gov—that allow users to obtain reports on summary statistics, such as state or county employment rates or mean income, from underlying microdata. These summary statistics are subjected to some form of disclosure avoidance to protect privacy. In addition, computer scientists have developed query-based strategies that aim to achieve DP limits on disclosure risks, as we will discuss below. However, to date, query-based access is almost never used by researchers because it does not

[‡]Statistical agencies do release original data to users when disclosure risks are minimal, such as with release of census records 72 years after their collection.

provide sufficient information to conduct a typical research exercise.

**2.2. Statistical Measures of Disclosure Risk.** Statisticians have developed ways to assess the disclosure risks associated with released tabular and microlevel data (34–36). A measure of disclosure risk, first developed by Duncan and Lambert (34, 35) over 30 years ago, characterizes the probability that an "intruder" can identify individuals in a dataset released by a statistical agency and can determine the confidential values of some of their data by combining information the intruder has with the released data (20, 37). Using a notation that combines elements of Reiter (38) and McClure and Reiter (39), let $J$ denote a target individual in the intruder's dataset, $A$, and let $j$ be an individual in a released dataset, $D^*$. Let $Y_j$ be data on individual $j$ that are in the confidential dataset, $D$ (which has no perturbed variables), but not in the intruder's dataset ($A$), which the intruder would like to know. The absolute disclosure risk, $\Pr(J=j, Y_j|D^*, A)$, is defined as the probability that the intruder can link an individual in their dataset to an individual in the released data and determine the value of $Y_j$. This is a direct measure of the disclosure risks facing individuals and feeds directly into the benefit–cost evaluations discussed in section 1.2, because it is needed to calculate individuals' expected utility losses.

As noted by Duncan and Lambert (34), Reiter (38), and McClure and Reiter (39), it is instructive to use Bayes theorem to express the absolute disclosure probability in the form

$$\Pr(J=j, Y_j|D^*, A) = [\Pr(D^*|J=j, Y_j, A)/\Pr(D^*|A)] \cdot \Pr(J=j, Y_j|A). \quad [1]$$

On the right-hand side of [**1**], $\Pr(J=j, Y_j|A)$ is the intruder's prior probability that individual $J$ corresponds to record $j$ in $D^*$ and that the value of $j$'s confidential data is $Y_j$.[§] Absolute disclosure risk is the product of this prior probability and the ratio in brackets, which indicates how the probability of $D^*$ would change if $J=j$ and $j$ has data value $Y_j$. Absolute disclosure risk depends on both terms, not on either alone.

Dividing both sides of [**1**] by the prior probability yields

$$\Pr(J=j, Y_j|D^*, A)/\Pr(J=j, Y_j|A) = \Pr(D^*|J=j, Y_j, A)/\Pr(D^*|A). \quad [2]$$

The left-hand side, the ratio of the posterior probability of disclosure to the prior probability, is what McClure and Reiter (39) refer to as the relative risk of disclosure associated with release of $D^*$. As discussed in section 2.5 below, the right-hand side turns out to be what the DP approach to disclosure avoidance seeks to limit. This demonstrates why it is insufficient, by itself, to bound absolute disclosure risk.

We argue that individuals should care about absolute disclosure risk and not relative risk alone. We illustrate this point with an example from public health discussed in Manski (40). Consider the risk of an individual dying from a particular illness or health condition. Some epidemiological studies calculate only the relative risk of dying from one cause versus another. But what the individual cares about is the probability of dying. The individual will feel very differently if the magnitude of the probability of dying from either cause is small, in which case the individual may not care much about even a large relative increase in the probability. In contrast, if the magnitude of the probability of dying is high, the individual is likely to care a great deal about even a small increase in the relative risk. Likewise, in data release, the individual will care a great deal about small increases in the disclosure risk if the probability of disclosure is already high, but may not be bothered by even a large relative increase in risk from data release if the probability of disclosure remains low in absolute terms after release.

The Bayes theorem expression in [**1**] highlights the important fact that an individual faces some disclosure risk, quantified by the prior probability, even if the dataset is not released at all. Release of dataset $D^*$ modifies the prior disclosure risk, multiplying it by the right-hand side of [**2**]. As we will discuss further in our essay, the Bayes theorem expression also makes plain that measuring absolute disclosure risk is difficult, as it depends on the intruder's information, $A$. Learning this information, or making assumptions about it, can be challenging.

**2.3. Traditional SDLs.** There is a wide variety of traditional methods for SDL. These include methods like top coding and bottom coding, suppressing cells with small numbers of observations, setting geographic population thresholds for disclosure of microdata, rounding of values, noise infusion, data "swapping," and data masking (24, 31, 41–43). All these methods aim to reduce the probability of disclosure risk as defined in the last section, while preserving data usability. But, with the increased availability of external, often commercial, databases noted in the Introduction, researchers within and outside of the statistical agencies (44–46) have documented that these methods are vulnerable to disclosure and reidentification risks. Furthermore, some of these methods rely on agencies not disclosing key features of their implementation, for example, not disclosing the rates at which individuals' race, ethnicity, or age are swapped to public use files, which compromises the transparency of these SDL methods. As noted in the Introduction, these concerns and findings have led agencies like the US Census Bureau to turn to new SDLs that are the subject of the remainder of this essay.

While not typically classified as traditional SDL methods, the disclosure limitation rules imposed on output used by agencies who provide data under licensing agreements or in RDCs are similar in type. Those rules consist of a long list of how tables are to be restricted, minimum cell sizes, restrictions on the number of observations used in regressions and what type of variables can be used in them, and so on. These rules do not account for disclosure risk as just defined, especially those arising from an intruder's use of external data sources.

**2.4. Synthetic Data.** Construction of synthetic data to limit disclosure risk was first proposed by Rubin (47). It has been extensively discussed, but has seen limited use by government agencies. It is used in the first mode of access

[§]To clarify, the intruder may have data on a large set of individuals, including some who will be in the released dataset. The prior probability denotes the probability that the intruder can identify a person who will be in the released data and that individual's value of $Y$.

referred to above ("dissemination-based access"). Rubin proposed that the data steward use the original microdata to estimate a model approximating the probability distribution of sensitive data *y* conditional on nonsensitive data *x*. Random draws from this modeled distribution provide imputed values of sensitive data. In the synthetic dataset, they replace the sensitive data; hence, the synthetic data attempt to have no confidential values that appear in the original data. This can be repeated, yielding multiple imputations and multiple synthetic datasets (48) or, alternatively, just a single synthetic dataset can be constructed from a single draw from the distribution. It is important to understand that the core concept is the modeled distribution. Multiple imputation is simply an operational procedure to approximate the modeled distribution by repeated simulation and to facilitate use of readily available statistical software. We discuss this issue in *SI Appendix*, section B.2.

There are two primary issues with synthetic data and disclosure risk. The first concerns the extent to which traditional synthetic data procedures, in fact, do reduce the risks of disclosing the data for, or identity of, an individual in a dataset. A key principle of the synthetic data approach is that, assuming an appropriate model is used to create synthetic data, it produces a dataset that is randomly different from the confidential dataset. As such, proponents argue that the identity of individuals and their characteristics are protected, since "no unit in the released data has sensitive data from an actual unit in the population" (49).

However, this logic does not guarantee that synthetic datasets, either by themselves or combined with external information and information about the synthesizer, cannot be used to determine individuals' characteristics or identities. To make that determination, a formal analysis of the disclosure risk probability as defined above needs to be undertaken. For example, the more accurate the synthesis in representing the actual data, which may contain outliers or extreme values, the greater the potential for high risks of disclosure, so that even synthetic datasets could be used by an intruder to make a high-probability guess of the identity or attributes of a real person.

The literature on the development of the synthetic data approach has recognized this issue and has suggested that specific privacy protection measures be built into the synthetic dataset creation from the beginning, rather than simply drawing random numbers from a distribution (38, 49). These include synthetic data methods that satisfy the DP criterion, referred to as differentially private synthetic methods (50). We discuss the implications of this development for statistical inference in *SI Appendix*, section B.2.

The second issue concerns whether the model for the distribution of the data used for imputation faithfully represents the data, which is called the "accuracy" of the synthetic data in the literature. Matthews and Harel (ref. 31, p. 10) comment on this, observing that inferences made with synthetic data are generically incorrect if the imputation model is incorrect. Reiter (ref. 49, p. 532) notes that "the validity of inferences depends critically on the accuracy of the imputation model." He also explains that "[t]he extent of this dependence is driven by the amount of synthesis. If entire variables are simulated, analyses involving those variables reflect only those relationships included in the data generation models. When these models fail to reflect accurately certain relationships, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses."

Synthetic data must be incorrect in a basic sense because, by design, the approach attempts to contain no confidential values, or at least to do so with only low probability. If confidential values were included in the data, confidentiality could be violated.

What synthetic data producers mean by accuracy is that certain relationships between variables in the data are approximately correct. In the most general case, the imputation model is perfectly correct if it correctly captures the underlying distribution of the variables in the original dataset, with the actual values in the original data simply representing a random finite-sample draw from that distribution, and the synthetic data representing a different finite-sample draw. However, in cases where datasets have a large number of variables, and the number of relationships that researchers might want to investigate is very large, it is essentially impossible for all relationships to be accurately captured by the type of models that are feasible to use for synthetic dataset creation.

Often, data stewards creating synthetic datasets must choose which relationships will be faithfully replicated in the synthetic dataset and which will not. This is a critical issue in evaluating the loss in social welfare from employing synthetic data, because some research will no longer be able to be reliably conducted, and the social value of that research will be lost.[¶]

The synthetic data community has suggested that this problem might be addressed by providing some information to the researcher about the accuracy of their estimates. One method of doing so is to allow the user to submit a query (see definition above) to the original dataset asking for the exact estimate of a relationship on those original data. This approach uses what is called a "validation" server with the original data (51). A second method is for the user not to be given information on the exact estimate from the original data but instead some numerical measure of how accurate the user's synthetic data estimates are (e.g., whether the user's estimates of confidential intervals overlap with those yielded by the original data). This approach uses what is called a "verification" server (52).[#] Both types typically utilize a query access approach where, after initial exploration of the synthetic data, the user submits specific query requests for a small number of outputs. However, in both cases, giving the user information from the original dataset reintroduces privacy protection issues. The information given to the user will necessarily have to be assessed using the disclosure risk probability as defined above, thereby reducing the value of the synthetic data method in the first place. We discuss the implications of this approach for statistical inference in *SI Appendix*, section B2.

[¶]Reiter (ref. 43, p. 175) notes that it can be difficult for researchers to know whether particular relationships they are investigating are adequately captured by the simulation model used.

[#]The Office of Personnel Management Synthetic Data Project uses this approach (52).

**2.5. DP.** DP is a criterion, not a method, for privacy protection that arose from the computer science literature on disclosure risk that began in the late 1990s and early 2000s. Building on earlier work (53), the Dwork et al. (54) proposal of DP has been viewed as revolutionary because it provides a formal way to measure disclosure risk and to decide whether a suggested disclosure approach gives sufficiently small disclosure risk to be acceptable. Furthermore, this approach is not premised on particular assumptions about what intruders know and do not know about individuals and their private data, including what they might learn about them in the future.

The DP approach uses a disclosure risk criterion to apply to data released from original confidential data, or released in response to queries about those data, that bound the sensitivity of the released information to the presence of any individual in the dataset (55, 56). These bounds are set, *ex ante*, by the data steward (54). A commonly used version of the DP criterion, referred to as $\varepsilon$-differential privacy, sets bounds on the ratio of the probability distributions of any released function (e.g., a perturbation function) of any two confidential datasets that differ by, at most, one record. Applied to the generation of releasable microdatasets [see McClure and Reiter (39) for details], the DP criterion is given by

$$\Pr\Big(f(D_1) \in S\Big)/\Pr\Big(f(D_2) \in S\Big) \leq \exp(\varepsilon), \qquad \textbf{[3]}$$

where $f$ is the data generator function, $f(D_i)$ denotes the possible resulting datasets, and $S$ is any set of values for this function. An analogous version of the DP criterion in [**3**] can be expressed for statistics, or queries, produced from $D_1$ and $D_2$ (54). Algorithms that achieve the DP criterion infuse random noise into responses to queries from the confidential data or into cell counts (for categorical data) and histograms (for numerical data) from these data that can be used to produce privacy-protected microdata, where the variance of the infused noise is an inverse function of the product of $\varepsilon$ and the largest change that an inclusion or exclusion of an individual record can have on the data.[‖]

While DP has the advantage of providing an implementable quantitative metric of disclosure risk, it raises multiple important issues. First, although DP users often recognize the importance of data usability, they rarely quantify, in detail, the implications of their DP choices for the subsequent utility of the data released after noise infusion. This would require going far beyond assessing the impact of disclosure methods based on mean-squared error or variances of particular statistics—which is what most of the literature examines—to determining the types of analyses that could no longer be reliably conducted with the altered data. This essential matter, which is central to assessing the trade-off between the value of data and privacy protection as we described in section 1.2, is left for others to consider (57).**

Second, an important issue with repeated releases of data is that information that can be acquired by an intruder cumulates with each release. This is particularly clear with DP, which was initially developed and tailored to applications with query-based access to data. In the query-based approach, the researcher asks a single question, or a set of questions, at a time, and the answer given to the researcher is altered to reduce disclosure risk. Each additional query increases the total risk of disclosure, because even the altered response to a query results in some information leakage. Keeping track of the total information released makes use of what is called "privacy loss accounting," which is required regardless of what definition of disclosure risk is used. The DP criterion has the advantage of a simple additive property for privacy loss, where $k$ successive queries, each protected by an $\varepsilon$ criterion, has a level of privacy bounded by $k\varepsilon$ (59). But the DP approach incorrectly addresses this problem by imposing a total "privacy loss budget" across the queries to preserve privacy, that is, limiting the number of queries allowed. Once the total privacy budget is expended on a set of queries, the data steward is not able to respond to any further queries of the data that have not been released. The problem with this mechanical method of privacy loss accounting is that it does not evaluate the costs and benefits of each subsequent release of data which, as we have emphasized, should be the foundation for decisions of any data release. Imposing a "budget" for privacy loss ignores the benefits of additional releases when their benefits exceed their (privacy loss) costs, and is likely to result in socially incorrect decisions regarding sequential data releases.

Extending the DP approach from query-based data access to the release of noise-infused microdatasets ("dissemination-based access") is problematic (60). Microdata can contain many variables that can be used to calculate an essentially unconstrained number of statistics. To ensure that a DP criterion is met for all possible applications of the noise-infused data may require adding so much noise that the resulting dataset would be unusable (12, 57, 61).

Our third concern about DP is that adherence to a DP criterion for disclosure risk does not ensure low absolute disclosure risk as defined in section 2.2. Gong and Meng (62) show that the left-hand side of the DP criterion in [**3**] equals $\Pr(J=j, Y_j|D^*, A)/\Pr(J=j, Y_j|A)$, the measure of relative disclosure risk in Eq. **2**. Thus, while the DP bound does imply a bound on the absolute disclosure risk, $\Pr(J=j, Y_j|D^*, A)$, it depends on the intruder's prior, $\Pr(J=j, Y_j|A)$. As noted above, the DP criterion presumes no knowledge or maintained assumptions about the prior. But, as we have argued in section 2.2, it is the absolute disclosure risk that individuals should care about, especially those whose information is included in a confidential dataset, and needs to be used in a proper benefit–cost analysis.[††]

Muralidhar et al. (ref. 57, p. 29) assess this issue in the application of DP to a microdataset with multiple variables. Terming the posterior disclosure probability "confidentiality," they observe that there are multiple choices about where to add the DP-style noise. They show that, for a given $\varepsilon$, alternative methods of implementing DP can yield different

[‖] *SI Appendix, Appendix C* contains a more detailed discussion of how algorithms are calibrated to satisfy the DP criterion.

**See Chetty and Friedman (58) for an example which goes in the direction we suggest.

[††] Dwork and Naor (63) prove that there exists a version of the intruder's information set, $A$, that when combined with the released data, $D^*$, will result in an absolute disclosure risk greater than any arbitrary threshold. However, this does not contravene the fact that it is absolute disclosure risk that individuals should care about, as we argued previously, and which is needed for a proper benefit–cost analysis, and that this requires some assumptions on the intruder's outside information. McClure and Reiter (39) provide an empirical illustration of how a data steward could use a range of possible intruder information to estimate a range of absolute disclosure risks.

reidentification probabilities, presenting examples for which the fraction of records in the original database that can be matched correctly to a record in the differentially privatized database (with $\varepsilon = 1$) ranges from 0.0003 to 0.12, a very wide range.

McClure and Reiter (39) also make several important points about the relationship between absolute disclosure risk and the relative risk measure used in DP, what risks they need to assess, and the implications both have for designing DASs (we have alluded to some of these issues previously). First, the authors find that the two measures do not necessarily provide the same answer to the level of risk. Their simulations show that relative risk can be low while absolute risk remains high. The simulations also demonstrate that absolute risk may be quite low, in which case the data steward can allow a higher level of relative risk as long as absolute risk stays below a chosen level. Second, the authors document the need for data stewards to evaluate disclosure risk over a wide range of possible levels of intruder information to determine the worst-case scenarios for absolute disclosure risk. Third, given their findings, the authors argue that reliance solely on disclosure avoidance schemes that meet the DP criterion is flawed, and argue that data stewards need to use schemes that also account for absolute disclosure risk.

A fourth problem, arising in some important applications of DP, concerns the need to impose constraints on the noise infusion mechanisms that may, in fact, violate the bounds sought by the DP criterion. Consider the policy-related constraint that was imposed in the Census Bureau's implementation of DP for dissemination of the data from the 2020 Census. The bureau has to infuse noise into the tabulations of counts for small geographic units within a state, because the original data could reveal values for individuals. However, it also must ensure that the counts add up to the state's total population. Such adding-up constraints are referred to, in the DP literature, as invariants. To satisfy them, the bureau performs a series of postprocessing steps in its DP-based DAS that alter the unconstrained noise-infused data to meet the invariants (64). Dwork and Roth (ref. 65, p. 228) argue that, under certain conditions, postprocessing should not affect the privacy guarantees of DP; that is, "differential privacy is immune to post-processing." However, Gong and Meng (62) observe that constraints that depend on the confidential data, as is the case for invariants, violate the conditions required in the Dwork–Roth postprocessing theorem, and, as a result, the immunity property need not hold.

The potential for DP privacy guarantees to fail under some adaptations of the DP mechanisms used to generate the public data suggests an important admonition to statistical agencies: Disclosure risks of data releases need to be assessed by data stewards, even when starting with DP mechanisms. Given the current lack of knowledge about exactly how different forms of postprocessing or other adaptations of DP mechanisms are likely to be required in actual applications, it is essential that assessments of disclosure risk be conducted. Such assessments need to be undertaken if agencies are to live up to the existing statutes that require their protection of the information they obtain from individuals.

**2.6. The Usability of Privacy-Protected Data: What We Know and Do Not Know about Inference on Parameters of Interest.** While the impact of the SDL methods we have just discussed on privacy protection has been discussed fairly extensively in the existing literature, their impact on the usability of data has been much less discussed. The key question is inherently a statistical one, which is how a user can employ privacy-protected data to conduct analyses that allow accurate and valid inferences from the data about quantities of interest to users of these data, be they specific questions of public policy or impacts of government decisions.

A complete characterization of statistical inference under the different disclosure avoidance methods being used or considered by statistical agencies is beyond the scope of this essay. However, we do discuss some instructive cases in *SI Appendix, Appendixes B* and *C*. There, we consider what is known and not known about the statistical properties of estimates of several parameters of interest for data produced by SDLs that meet a DP criterion. We also consider inference using synthetic data approaches that do not aim to meet a DP criterion.

## 3. Conclusions and Recommendations

Public policy should seek to manage, in an optimal way, the tension between the social value of data and the potential costs to privacy from release of data. Achieving this requires a suitable framework to conceptualize and measure the trade-offs involved in decisions about data release and privacy protection. We have argued that a suitable framework is benefit–cost analysis, long practiced in the field of economics. This framework embraces a broad concept of the benefits of data for social knowledge. It emphasizes the need to assess marginal gains and losses from better social knowledge against the marginal gains and losses of additional or lesser privacy protection. This requires that any data protection policy predict both the types of research and analysis that can no longer be reliably conducted and how that will affect the quality of decision-making in society along with the reduction in disclosure risk associated with the policy.

New assessments of the trade-off between privacy and data usability are now necessary, particularly by the federal statistical agencies on which this essay focuses. Pressures for a reassessment come from two opposing directions. On the one hand, with the advent of large private databases containing personal information coupled with growing sophistication of algorithms to match those databases to data released by the government, the risk of disclosure of personal information to those who wish to identify individuals has grown dramatically. On the other hand, we live in what may fairly be called a golden age of data, with major advances in the availability of data and tools for data analysis. Congress has recognized that many federal agencies have valuable internal datasets that they have not made accessible to researchers at all or only under highly restrictive conditions, yet whose analysis could yield great benefits to society. In the Evidence Act passed by Congress in 2018, agencies are mandated to make their data available to researchers, albeit with appropriate privacy protections. Addressing the first pressure (for privacy) requires

new disclosure limitation policies that still permit the release of accurate and socially valuable information now being released, but with better-understood privacy protection. Addressing the second pressure (for knowledge) requires new policies that allow more release of socially valuable datasets currently not available at all or available only under restricted conditions, but with adequate privacy protection.

Considerable attention has been given to development of new measures and methods of privacy protection. However, the discussion has focused almost entirely on privacy protection, with little assessment of the impact on data usability and, hence, social knowledge. Two leading examples are the introduction of DP as a measure of disclosure risk and the proposal of synthetic datasets as a method of privacy protection. Both DP and synthetic data raise serious concerns about data usability and data quality that have been inadequately addressed, to date. Both threaten to impose major limits on the way research and public policy can be conducted. The potential for greatly reduced social knowledge is clear.

We have expressed strong concern that the notion of DP is based on an inappropriate measure of disclosure risk. The appropriate measure, first developed by statisticians in the 1980s, is discussed in section 2.2. It differs from the disclosure risk criterion of DP rooted in computer science as well as from various informal notions of statistical disclosure risk used in the past. The relationship between these other disclosure risk criteria and the appropriate one is variable, depending on factors that must be quantified to conduct a proper assessment of disclosure risk. We recognize that implementation of the appropriate measure of disclosure risk could be challenging to data stewards, but this does not imply that an inappropriate measure of risk should be used instead.

We have three recommendations for managing the tension between the social value of data and the potential cost of privacy loss, with the twin goals of continuing to allow currently available data to be released and of permitting more release of currently unavailable data. First, we recommend that the disclosure risk criterion used by federal statistical agencies to assess the disclosure protection provided by any SDL method be the risk (defined in section 2.2) that we argue individuals should care about. Second, we recommend much more study of the impacts of using a DP criterion or employing synthetic data on data usability and statistical inference before either is put into widespread use. Third, we recommend further research, building on existing work referenced in this essay, on methods of privacy protection and reduction of disclosure risk that use the appropriate criterion for disclosure risk and seek a balance between usability and privacy protection grounded in the value to social welfare.

**Data Availability.** There are no data underlying this work.

**ACKNOWLEDGMENTS.** We thank Katherine Wallman and Constance Citro for providing us with important insights into the regulations governing federal statistical agencies, and we thank Claire Bowen, Constance Citro, George Duncan, Robin Gong, Jessica Hullman, Krish Muralidhar, Jerry Reiter, and Salil Vadhan for extremely useful comments on an earlier draft of this paper.

1. N. Eberstadt, R. Nunn, D. W. Schanzenbach, M. R. Strain, *'In Order That They Might Rest Their Arguments on Facts': The Vital Role of Government-Collected Data* (Brookings Institution, 2017).
2. Commission on Evidence-Based Policymaking, "The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking" (Commission on Evidence-Based Policymaking, 2017).
3. E. Saez, G. Zucman, Wealth inequality in the United States since 1913: Evidence from capitalized income tax data. *Q. J. Econ.* **131**, 519–578 (2016).
4. R. Chetty, J. N. Friedman, E. Saez, D. Yagan, The SOI Databank: A case study in leveraging administrative data in support of evidence-based policymaking. *Stat. J. IAOS* **34**, 99–103 (2018).
5. *115th Congress, Foundations for Evidence-Based Policymaking Act of 2018*: H. R. 4174 (US Congress, 2019).
6. Advisory Committee on Data for Evidence Building, "Advisory Committee on Data for Evidence Building Year 1 Report" (US Department of Commerce, 2021).
7. C. Bowen *et al.*, *A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications* (Urban Institute, Washington, DC, 2020).
8. M. Hawes, Understanding the 2020 Census Disclosure Avoidance System: Simulated reconstruction-abbetted re-identification attack on the 2010 census. https://www2.census.gov/about/training-workshops/2021/2021-05-07-das-presentation.pdf. Accessed 5 July 2022.
9. J. M. Abowd, Protecting the confidentiality of America's statistics: Adopting modern disclosure avoidance methods at the Census Bureau. https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html Accessed 5 July 2022.
10. P. Leclerc, The 2020 decennial census topdown disclosure limitation algorithm: A report on the current state of the privacy loss-accuracy trade-off. https://www.nationalacademies.org/event/12-11-2019/docs/DF5815D39AEDA0511D717967C7C72A681D1BDA2E8437. Accessed 12 July 2022.
11. A. R. Santos-Lozada, J. T. Howard, A. M. Verdery, How differential privacy will affect our understanding of health disparities in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13405–13412 (2020).
12. S. Ruggles, C. Fitch, D. Magnuson, J. Schroeder, Differential privacy and census data: Implications for social and economic research. *AEA Pap. Proc.* **109**, 403–408 (2019).
13. D. Van Riper, T. Kugler, S. Ruggles, "Disclosure avoidance in the Census Bureau's 2010 demonstration data product" in *Privacy in Statistical Databases*, J. Domingo-Ferrer, K. Muralidhar, Eds. (Springer International, Cham, Switzerland, 2020), pp. 353–368.
14. V. J. Hotz, J. J. Salvo, Assessing the use of differential privacy for the 2020 census: Summary of what we learned from the 2019 CNSTAT workshop. https://www.amstat.org/asa/files/pdfs/POL-CNSTAT_CensusDP_WorkshopLessonsLearnedSummary.pdf Accessed 5 July 2022.
15. National Academies of Sciences, Engineering, Medicine, *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*, D. L. Cork, C. F. Citro, N. J. Kirkendall, Eds. (The National Academies Press, Washington, DC, 2020).
16. S. Ruggles, D. Van Riper, The role of chance in the Census Bureau database reconstruction experiment. *Popul. Res. Policy Rev.* **41**, 781–788 (2022).
17. C. T. Kenny *et al.*, The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. *Sci. Adv.* **7**, eabk3283 (2021).
18. J. M. Abowd, Declaration of John Abowd, State of Alabama v. U.S. Department of Commerce (CASE no. 3:21-cv-211-RAH-ECM-KCN, US District Court, 2021).
19. M. J. Anderson, The census and the federal statistical system: Historical perspectives. *Ann. Am. Acad. Pol. Soc. Sci.* **631**, 152–162 (2010).
20. National Research Council, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, G. T. Duncan, T. B. Jabine, V. A. de Wolf, Eds. (The National Academies Press, Washington, DC, 1993).
21. National Academies of Sciences, Engineering, Medicine, *Principles and Practices for a Federal Statistical Agency*, B. A. Harris-Kojetin, C. F. Citro, Eds. (The National Academies Press, Washington, DC, ed. 7, 2021).
22. J. Mervis, Can a set of equations keep U.S. census data private? *ScienceInsider*, 4 January 2019. https://www.science.org/content/article/can-set-equations-keep-us-census-data-private Accessed 5 July 2022.
23. R. F. Boruch, J. S. Cecil, *Assuring the Confidentiality of Social Research Data* (University of Pennsylvania Press, Philadelphia, PA, 2016).
24. G. T. Duncan, J.-J. Salazar-González, M. Elliot, *Statistical Confidentiality: Principles and Practice* (Springer, New York, 2011).
25. S. E. Fienberg, "Confidentiality and disclosure limitation" in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, Ed. (Elsevier, 2005), pp. 463–469.
26. G. T. Duncan, M. Elliot, J.-J. Salazar-González, "Why statistical confidentiality?" in *Statistical Confidentiality: Principles and Practice*, G. T. Duncan, M. J. Elliot, J.-J. Salazar-González, Eds. (Springer, New York, 2011), pp. 1–26.
27. M. J. Anderson, W. Seltzer, Federal statistical confidentiality and business data: Twentieth century challenges and continuing issues. *JPC* **1**, 7–52 (2009).
28. J. M. Abowd, I. M. Schmutte, An economic analysis of privacy protection and statistical accuracy as social choices. *Am. Econ. Rev.* **109**, 171–202 (2019).
29. C. F. Manski, *Public Policy in an Uncertain World: Analysis and Decisions* (Harvard University Press, Cambridge, MA, 2013).
30. R. Jarmin, Census Bureau adopts cutting edge privacy protections for 2020 census. *Director's Blog*, 15 February 2019. https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html. Accessed 5 July 2022.
31. G. J. Matthews, O. Harel, Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Stat. Surv.* **5**, 1–29 (2011).
32. A. F. Karr, Data sharing and access. *Annu. Rev. Stat. Appl.* **3**, 113–132 (2016).

33. A. F. Karr, Why data availability is such a hard problem. *Stat. J. IAOS* **30**, 101–107 (2014).
34. G. T. Duncan, D. Lambert, Disclosure-limited data dissemination. *J. Am. Stat. Assoc.* **81**, 10–18 (1986).
35. G. T. Duncan, D. Lambert, The risk of disclosure for microdata. *J. Bus. Econ. Stat.* **7**, 207–217 (1989).
36. D. Lambert, Measures of disclosure risk and harm. *J. Off. Stat.* **9**, 313–331 (1993).
37. Federal Committee on Statistical Methodology, Report on statistical disclosure limitation methodology (2nd version). https://www.hhs.gov/sites/default/files/spwp22.pdf. Accessed 5 July 2022.
38. J. P. Reiter, Estimating risks of identification disclosure in microdata. *J. Am. Stat. Assoc.* **100**, 1103–1112 (2005).
39. D. McClure, J. P. Reiter, Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.* **5**, 535–552 (2012).
40. C. F. Manski, The lure of incredible certitude. *Econ. Philos.* **36**, 216–245 (2020).
41. L. McKenna, *Disclosure avoidance techniques used for the 1960 through 2010 decennial censuses of population and housing public use microdata samples* (US Census Bureau, Washington, DC, 2019).
42. T. Dalenius, S. P. Reiss, Data-swapping: A technique for disclosure control. *J. Stat. Plan. Inference* **6**, 73–85 (1982).
43. J. P. Reiter, Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opin. Q.* **76**, 163–181 (2012).
44. W. E. Winkler, *Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified*. https://www.census.gov/library/working-papers/2007/adrm/rrs2007-21.html. Accessed 5 July 2022.
45. J. Drechsler, J. P. Reiter, Sampling with synthesis: A new approach for releasing public use census microdata. *J. Am. Stat. Assoc.* **105**, 1347–1357 (2010).
46. J. M. Abowd, I. M. Schmutte, Economic analysis and statistical disclosure limitation. *Brookings Pap. Econ. Act.* **2015**, 221–293 (2016).
47. D. B. Rubin, Statistical disclosure limitation. *J. Off. Stat.* **9**, 461–468 (1993).
48. T. E. Raghunathan, J. P. Reiter, D. B. Rubin, Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19**, 1–16 (2003).
49. J. P. Reiter, Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.* **18**, 531–543 (2002).
50. C. M. Bowen, F. Liu, Comparative study of differentially private data synthesis methods. *Stat. Sci.* **35**, 280–307 (2020).
51. L. E. Burman *et al.*, "Safely expanding research access to administrative tax data: Creating a synthetic public use file and a validation server" (Tech. Rep., Statistics of Income Branch, US Internal Revenue Service, 2019).
52. J. P. Reiter, A. Oganian, A. F. Karr, Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Comput. Stat. Data Anal.* **53**, 1475–1482 (2009).
53. A. Blum, C. Dwork, F. McSherry, K. Nissim, "Practical privacy: The Sulq framework" in *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (Association for Computing Machinery, 2005), pp. 128–138.
54. C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis. *JPC* **7**, 17–51 (2016).
55. C. Dwork, "Differential privacy" in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, I. Wegener, Eds. (Springer-Verlag, Berlin, 2006), pp. 1–12.
56. C. Dwork, F. McSherry, K. Nissim, A. Smith, "Calibrating noise to sensitivity in private data analysis" in *Theory of Cryptography*, S. Halevi, T. Rabin, Eds. (Springer-Verlag, Berlin, 2006), pp. 265–284.
57. K. Muralidhar, J. Domingo-Ferrer, S. Martínez, "ε-differential privacy for microdata releases does not guarantee confidentiality (let alone utility)" in *International Conference on Privacy in Statistical Databases*, J. Domingo-Ferrer, K. Muralidhar, Eds. (Springer International, Cham, 2020), pp. 21–31.
58. R. Chetty, J. N. Friedman, A practical method to reduce privacy loss when disclosing statistics based on small samples. NBER [Preprint] (2019). https://www.nber.org/papers/w25626 (Accessed 5 July 2022).
59. C. Dwork, "Differential privacy: A survey of results" in *International Conference on Theory and Applications of Models of Computation,* M. Agrawal, D. Du, Z. Duan, A. Li, Eds. (Springer, 2008), pp. 1–19.
60. S. L. Garfinkel, J. M. Abowd, S. Powazek, "Issues encountered deploying differential privacy" in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society* (Association for Computing Machinery, 2018), pp. 133–137.
61. J. Domingo-Ferrer, D. Sánchez, A. Blanco-Justicia, The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM* **64**, 33–35 (2021).
62. R. Gong, X.-L. Meng, "Congenial differential privacy under mandated disclosure" in *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference* (Association for Computing Machinery, 2020), pp. 59–70.
63. C. Dwork, M. Naor, On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *JPC* **2**, 93–107 (2010).
64. J. M. Abowd *et al.*, "Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge" (Tech. Rep., US Census Bureau, 2019).
65. C. Dwork, A. Roth, The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).