Implications of Data Anonymization on the Statistical Evidence of Disparity

Heng Xu and Nan Zhang

Kogod School of Business, American University, Washington DC

Research and practical development of data anonymization techniques has proliferated in recent years. Although the privacy literature has questioned the efficacy of data anonymization at protecting individuals against harms associated with re-identification, this paper raises another new question: whether anonymization techniques themselves can mask statistical disparities and thus conceal evidence of disparate impact that is potentially discriminatory. If so, the choice of data anonymization technique to protect privacy, and the specific technique employed, may pick winners and losers. Examining the implications of these choices on the potentially disparate impact of privacy protection on underprivileged sub-populations is thus a critically important policy question.

The paper begins with an interdisciplinary overview of two common mechanisms of data anonymization and two prevalent types of statistical evidence for disparity. In terms of data-anonymization mechanisms, the two common ones are **data removal** (e.g., k-anonymity), which aims to remove the part of a dataset that could potentially identify an individual; and **noise insertion** (e.g., differential privacy), which inserts into a dataset carefully designed noises that block the identification of individuals yet allow the accurate recovery of certain summary statistics. In terms of the statistical evidence for disparity, the two commonly accepted types are **disparity through separation** (e.g., the "two or three standard deviations" rule for a prima facie case of discrimination), which is grounded in the idea of detecting the separation between the outcome distributions for different sub-populations; and **disparity through variation** (e.g., the "more likely than not" rule in toxic tort cases), which concentrates on the magnitude of difference between the mean outcomes of different sub-populations.

Our work demonstrates that the two data anonymization mechanisms have distinctive impacts on the identifiability of disparity, which also varies based on its statistical operationalization. Specifically, under the regime of disparity through separation, data removal tends to produce more false positives (i.e., detecting false disparity when none exists) than false negatives (i.e., failing to detect an existing disparity); while noise insertion rarely produces any false positives at all. Meanwhile, noise insertion does produce false positives (equally likely as false negatives) under the regime of disparity through variation; while the likelihood for data removal to produce false positives and false negatives depends on the underlying data distribution.

Keywords: Privacy, Data Anonymization, Discrimination, Statistical Disparity

To cite this article:

Xu, H., Zhang, N. (2022). Implications of Data Anonymization on the Statistical Evidence of Disparity. *Management Science* (68:4), 2600-2618.

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 4, April 2022, pp. 2600–2618
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Implications of Data Anonymization on the Statistical Evidence of Disparity

Heng Xu,[a] **Nan Zhang**[a]

[a] Kogod School of Business, American University, Washington, District of Columbia 20016

Contact: xu@american.edu, https://orcid.org/0000-0001-5642-6543 (HX); nzhang@american.edu, https://orcid.org/0000-0002-0454-7885 (NZ)

Received: May 21, 2020
Revised: October 13, 2020
Accepted: January 18, 2021
Published Online in Articles in Advance: June 4, 2021

https://doi.org/10.1287/mnsc.2021.4028

Copyright: © 2021 INFORMS

Abstract. Research and practical development of data-anonymization techniques have proliferated in recent years. Yet, limited attention has been paid to examine the potentially disparate impact of privacy protection on underprivileged subpopulations. This study is one of the first attempts to examine the extent to which data anonymization could mask the gross statistical disparities between subpopulations in the data. We first describe two common mechanisms of data anonymization and two prevalent types of statistical evidence for disparity. Then, we develop conceptual foundation and mathematical formalism demonstrating that the two data-anonymization mechanisms have distinctive impacts on the identifiability of disparity, which also varies based on its statistical operationalization. After validating our findings with empirical evidence, we discuss the business and policy implications, highlighting the need for firms and policy makers to balance between the protection of privacy and the recognition/rectification of disparate impact.

History: Accepted by Chris Forman, information systems.
Funding: Financial support from the Defense Advanced Research Projects Agency [Grant HR00111920023] and National Science Foundation [Grants 1850605 and 1851637] is gratefully acknowledged.
Supplemental Material: The data files and e-companion are available at https://doi.org/10.1287/mnsc.2021.4028.

Keywords: privacy • data anonymization • discrimination • statistical disparity

> The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.
>
> —General Data Protection Regulation,
> GDPR Recital 26

1. Introduction

The emergence of privacy laws and regulations, like the European Union's General Data Protection Regulation (GDPR), spurred the proliferation of data-anonymization techniques in recent years. In general, such techniques are conceptualized and developed as an organizational-level (or population-level) solution balancing two countervailing interests. One is to disassociate the data from the data subjects (i.e., individuals) for the purpose of privacy protection. The other is to maintain the utility of the data. For example, Apple deployed differential privacy techniques in its collection of user keystrokes in iOS devices (Apple Inc. 2020). On one hand, the collected keystrokes can no longer be easily linked back to an individual, whereas on the other hand, Apple can still use the collected data to improve its autocorrection and predictive text-entry features. Because disassociating data from subjects necessitates a reduction of data utility (Kifer and Machanavajjhala 2011, Dwork et al. 2017), many data-anonymization techniques operationalize the privacy–utility trade-off as a tunable parameter that is set at the organizational level.

Although the implementation of data anonymization as an organizational-level solution meets the regulatory requirements (e.g., the above-quoted recital in GDPR), it may not align with the idiosyncratic nature of individuals' privacy preferences and demands (Acquisti et al. 2015). Similarly, the benefits (or harms) derived from improved (or diminished) data utility also vary drastically from one individual to another (Pitoura et al. 2018). Considering the organization-level solution and the individual-level impacts in tandem, a natural question emerges: Could an organization-level, one-size-fits-all solution to privacy protection stimulate disparate impacts on different individuals? The emergence of privacy laws and regulations gives primacy to answering this question, because if such disparate impacts do exist, legislators and regulators would be essentially picking winners and losers by requiring or incentivizing the use of data-anonymization techniques.

Unfortunately, the existing studies on the implications of data anonymization were largely focused on the reduction of overall data utility, again defined at the organizational level over all data records in the data set, leaving the individual-level impact an unanswered question that represents a considerable gap in the literature. Answering this question is obviously challenging. The design of data anonymization and the utility of anonymized data are usually governed by proprietary technologies and processes that are opaque to researchers and the public (Tang et al. 2017). The impact of data anonymization on individuals is also difficult to grapple with, given that even privacy experts are often confused by what a data-anonymization technique offers in terms of privacy protection (Bambauer et al. 2013). To take the first step toward answering this question, we focus on a specific type of disparate impacts that is analytically accessible, yet practically salient: whether data anonymization could mask *gross statistical disparities* between subpopulations in the data. The presence of such a disparity-masking effect may have far-reaching business, social, and policy implications. For example, it could prevent Apple from detecting and rectifying the subpar accuracy its autocorrection feature offers to minority populations with distinct language patterns (Chang et al. 2019). In a healthcare context, it could preclude the identification of health disparities pertaining to gender, race, ethnicity, income, sexual orientation, etc., which represent one of the most pressing social justice issues facing the United States (Kelley et al. 2005). For the U.S. Census, with which a public discourse of how to apply data anonymization is currently under way (Macagnone 2019), the masking of disparities could have detrimental impacts on public policy for the next decade.

To examine the potential impact of data-anonymization techniques on the detection of statistical disparity, one must first identify the mechanisms through which the current techniques anonymize private data and define the statistical evidence required to identify disparities between subpopulations. Unfortunately, researchers and practitioners today frequently and casually use the term *data anonymization* to refer to a wide variety of techniques, from those that directly manipulate data with predefined rules (Texas Department of State Health Services 2019) to those that dynamically determine whether and how to answer queries posed to the data (Dwork et al. 2019). The term *statistical disparity* also requires closer examination and refinement, as its meaning in one context could be grounded in statistical significance (e.g., unlawful discrimination; Garaud 1990), yet in another could focus on frequency comparison (e.g., public health[1]).

Seeking clarity to the notions of data anonymization and statistical disparity, we first propose a typology categorizing the mechanisms of data anonymization as *data removal* or *noise insertion*; and a typology operationalizing statistical disparity as *through separation*[2] or *through variation*. After offering a detailed presentation of the two typologies, we explore the interplay between the two—that is, the implications of applying each anonymization mechanism on the identifiability of each disparity operationalization. We emphasize the distinct outcomes of these four (2 × 2) combinations: When disparity is operationalized through separation, although the noise-insertion mechanism only tends to mask disparities, the data-removal mechanism could more likely produce false positives than masking disparities. In contrast, when disparity is operationalized through variation, both anonymization mechanisms could mask disparities or produce false positives, even flipping the direction of observed disparities on occasion. We develop these distinct conclusions through conceptual development and mathematical formalism, before validating them with empirical evidence. In the final section of the paper, we discuss the practical implications of our findings, the limitations of our work, and the potential directions for future research.

2. Related Work

Closely related to our work is a recent stream of research on how anonymization affects the fairness of predictions made from the anonymized data. For example, Pujol et al. (2020) studied the question that if we were to allocate vital resources based on an anonymized data set (e.g., epidemiological data), whether the error of allocated resources would be unfairly large for some individuals (or subpopulations) than others. Similarly, Bagdasaryan et al. (2019) examined the fairness of predictions from a recurrent neural network if it were trained on a differentially private data set. Dwork and Mulligan (2013) and Agarwal (2020) further challenged the fundamental fairness of any *downstream products* built from an anonymized data set. Agarwal (2020), for example, proved the infeasibility of building a "fair" machine-learning model (under a broad notion of fairness) from an anonymized data set that satisfies ϵ-differential privacy.

Developed in parallel to the above research stream is a recent body of research in machine learning that focuses on examining whether removing certain social determinants from the input data could improve (or impede) the fairness of predictions generated by a particular type of downstream product, a (supervised) machine-learning model (Ekstrand et al. 2018, Lipton et al. 2018, Kleinberg and Mullainathan 2019, Cowgill and Tucker 2020, Rambachan et al. 2020). Although the machine-learning algorithms and the notions of fairness varied considerably in this body of research, its recent development unequivocally established that the removal of social determinants is always detrimental to the fairness of downstream machine learning (Lipton et al. 2018, Kleinberg and Mullainathan 2019, Rambachan et al. 2020), regardless of the input data set, the machine-learning algorithm being used, or the notion of fairness.

Although these two streams of research and our work are all related to the unintended consequences of privacy protection on issues related to disparity/fairness, their *referents* for disparity are fundamentally different. Our work focuses on the detection (or masking) of disparity in the original data from their anonymized version. The two existing research streams, on the other hand, examined the disparity of *downstream* data products built from the anonymized data set. Interestingly, despite this fundamental difference, the policy implications of our work and the existing research are remarkably relevant, as elaborated on later in the paper. For example, whereas our findings show how different types of anonymization mechanisms could mask disparity detection (to different extents) over the (upstream) original data set, the two existing research streams showed that anonymization could also blunt the learning of fair prediction models from the anonymized data, making the downstream data products susceptible to incurring disparate impacts in practice.

Finally, sharing the same theoretical roots with our work was the research on how anonymization reduces the utility of the data being released (e.g., Dwork et al. 2016, Abowd and Schmutte 2019) and the implications of such reduced utility on empirical research (e.g., Santos-Lozada et al. 2020). Santos-Lozada et al. (2020), for example, demonstrated that the noises inserted to population counts could reduce the accuracy of mortality rates calculated based on such counts, which, in turn, could affect our understanding of health disparities across racial/ethnic groups. Although the study of anonymization-induced error in inferential statistics is an important theoretical underpinning of our work, the focus of our work is *not* to demonstrate that anonymization-induced errors could alter empirical findings, such as the outcomes of disparity detection, but to examine how such outcomes specifically relate to the complex *interplay* between different mechanisms of data anonymization and different types of (statistical evidence used in) disparity detection.

3. Conceptual Development

In this section, we develop conceptual foundation demonstrating the potential implications of data anonymization on disparity detection. We first introduce two typologies for the design of data anonymization and the statistical operationalization of disparity, respectively, before explicating the mechanisms through which an anonymization mechanism affects the detection of disparity according to its operationalization.

3.1. A Typology of Anonymization Mechanisms: Data Removal and Noise Insertion

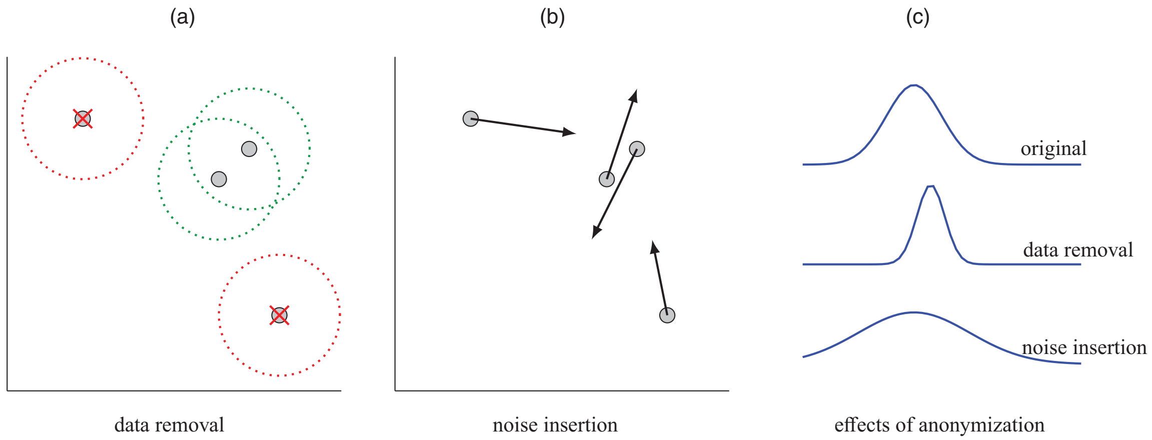
In what follows, we first describe the two types of anonymization mechanisms, data removal and noise insertion, before explicating the key differences between the two.

3.1.1. Data Removal. Because the goal of data anonymization is to prevent any individual from being identified from the anonymized data set, a natural idea for anonymizing a data set is to remove the part of data that could be used to identify an individual. The data-removal mechanism is rooted in this idea. Its initial implementations focused on removing variables that are obvious identifiers, such as name, address, Social Security number, etc. These implementations were challenged by the discovery (Sweeney 2000) that 87% of Americans can be uniquely identified by a combination of ZIP code, gender, and date of birth—none of which is traditionally deemed an "identifier." Following this dramatic discovery, a plethora of data-removal techniques were developed in the computer science literature to detect and rectify the issues caused by such "quasi-identifiers" (e.g., Sweeney 2002b, Machanavajjhala et al. 2007), forming the bulk of the existing literature for the data-removal mechanism.

Although these techniques differ considerably in their design (see Fung et al. 2010 for a comprehensive review), a common procedure they follow is to first determine which individuals are at risk for being identified, before removing the minimum amount of information necessary to block such identifications. Figure 1(a) demonstrates a simple example premised on the notion that an individual is at risk for being identified if his or her record in the data set is like no other (i.e., if the record has an empty neighborhood like the non-overlapping dotted circles in the figure). The idea for anonymizing a data set is then to remove all records with an empty neighborhood (i.e., the "suppression" technique; Sweeney 2002a), as depicted in the figure. Beyond this simple example, many other forms of data removal have been developed, such as removing selected variables of an individual or obscuring the variables with coarser values—for example, by changing ZIP code to city or state (Fung et al. 2010).

Despite its intuitive appeal and the innumerable efforts spent on its development, technical research on

Figure 1. (Color online) Illustration of Data-Anonymization Mechanisms



Notes. (a) Data removal. (b) Noise insertion. (c) Effects of anonymization.

data removal all but ceased in the late 2010s. Part of this is because of the (re)emergence of noise-insertion mechanisms, which we will discuss next. The more important reason, however, is that researchers realized that it is untenable to block all possible identifications without making substantial assumptions about what other data sources might be linked with the anonymized data set to identify an individual (Ganta et al. 2008). Interestingly, this concern did not deter data removal from gaining firm footing in practice. As of today, it is not only widely adopted by firms and government agencies (e.g., Ali 2018, Rocher et al. 2019), but frequently included as recommended practices for complying with privacy laws and regulations[3] (U.S. Department of Health and Human Services 2012, Article 29 Data Protection Working Party 2014, Finnish Social Science Data Archive 2020). Table 1, for example, depicts the data-removal mechanism adopted by the Texas Department of State Health Services (2019) to anonymize their state-wide inpatient discharge data set for regulatory compliance. Similar rules were adopted by many other states, such as New York (U.S. Agency for Healthcare Research and Quality 2018). The practical pertinence of such techniques makes data removal an important anonymization mechanism to examine in this paper.

3.1.2. Noise Insertion.

Like data removal, the idea of noise insertion has been extensively studied in several disciplines, such as computer science, information systems, and statistics (Traub et al. 1984; Muralidhar et al. 1995, 1999; Agrawal and Srikant 2000; John et al. 2018). Researchers have long recognized the feasibility of accurately recovering certain summary statistics from a noise-ridden data set using methods akin to statistical calibration (Osborne 1991). Early efforts on noise insertion, though, were blunted by the finding that simply adding independent Gaussian noise to all variables in a data set could allow the inserted noise to be disentangled from the anonymized data using spectral methods (Huang et al. 2005), effectively re-enabling the identification of an individual. The development of differential privacy (Dwork et al. 2016) addressed this issue and provided a rigorous anonymity guarantee in the form of *statistical indistinguishability* (Goldwasser et al. 1989) between a data set that includes an individual's information and a data set that does not. More importantly, this guarantee holds, no matter what other data sources might be available to be linked with the anonymized data. This avoids the aforementioned pitfall of the data-removal techniques, makes differential privacy the de facto standard for modern noise-insertion techniques, and helps noise insertion gain

Table 1. Texas Healthcare Information Collection Anonymization Rules

Rule	If a patient meets the following criterion	Then perform the following action on his/her record
1	<30 patients with the same ZIP code	Remove last two digits of ZIP code
2	State ≠ Texas or an adjacent state	Remove ZIP code
3	ICD-10 codes indicate alcohol/drug use or HIV	Remove ZIP code and gender, change age to age group
4	<5 patients with the same gender and hospital ID	Remove ZIP code and hospital ID
5	<50 patients with the same hospital ID	Remove ZIP code and hospital ID
6	<5 patients with the same country	Remove country
7	<5 patients with the same county	Remove county
8	<10 patients with the same hospital ID and race	Remove race and ethnicity

Note. In Rule 3, the age is generalized to one of five age segments (<20, [20,40), [40, 60), [60,80), >80).

wide acceptance in research (Hay et al. 2016) and a solid footing in practice, especially among high-tech companies (Apple Inc. 2020) and in statistical agencies such as the U.S. Census Bureau (Abowd and Schmutte 2019).

Techniques for noise insertion in general, and differential privacy in particular, take many ways, shapes, and forms. Random noises could be directly added to the original data (Agrawal and Srikant 2000, John et al. 2018), like in Figure 1(b), or be added to when answering a query over the data set (Dwork et al. 2016). The statistical estimates after noise insertion could remain unbiased (e.g., with the standard Laplace mechanism for differential privacy; Dwork et al. 2016) or include a small bias determined by the input data (e.g., the data- and workload-aware algorithm for differential privacy; Li et al. 2014). Similarly, the inserted noise could be independent of the original data set (e.g., Dwork et al. 2016) or be generated according to the data (e.g., Li et al. 2014). Although the implementations differ, their conceptual underpinning is remarkably consistent: The confidence interval for any summary statistics inferred from the anonymized data must be *wider* than the inference over the original data set, so as to make the inferred statistics indistinguishable, whether an individual is in the original data set or not.

3.1.3. Comparison.

Figure 1(c) illustrates the comparison between the two anonymization mechanisms: Data removal often introduces bias to the estimated statistics, and likely reduces the observed standard deviation due to its tendency to remove "outlier" records, such as those with empty neighborhoods in Figure 1(a). In contrast, noise-insertion techniques are usually unbiased or introduce minimal bias to the estimated statistics. Nonetheless, the inserted noise tends to increase the observed standard deviation considerably. In the e-companion (section EC.1), we also provide a more detailed summary of the typical anonymization algorithms in each category.

Before concluding our conceptual discussions of the anonymization mechanisms, we offer the caveat that neither mechanism is a silver bullet for privacy protection. As mentioned before, the literature has unequivocally confirmed the feasibility of reidentifying an individual after data removal by linking the anonymized data set with another external data source (Fung et al. 2010). Noise insertion cannot block all possible identifications either. For example, it has long been known that a differentially private data set may still disclose a considerable amount of information about an individual (Kifer and Machanavajjhala 2011). Further, it is impossible to directly compare the degree of anonymity offered by different anonymization mechanisms because such a comparison depends on myriad factors, including what external data sources might be linked with the anonymized data (Du et al. 2008). To this end, it is important to note that we do not attempt to compare the two anonymization mechanisms vis-à-vis in the paper. Instead, our imperative is to illuminate and contrast their *qualitatively* distinct impacts on the detection of disparity from the anonymized data.

3.2. A Disparity Typology: Disparity Through Separation and Disparity Through Variation

There is no dispute that disparities between subpopulations exist in a wide range of social and economic outcomes, from rates of poverty and unemployment to quality of education and healthcare. It is not surprising, therefore, that the detection of disparity is a long-standing research problem in a variety of disciplines, from sociology and criminology to epidemiology and medicine (Pager and Shepherd 2008). Equally unsurprising is the fact that courts are frequently called upon to decide the existence and magnitude of disparities, such as in cases pertaining to labor and toxic tort laws (King 2006). Given the range of domains that have examined the issue, we do not attempt to be exhaustive in our presentation of the typology. Instead, our intent in developing the typology is to highlight two conceptually distinct, yet equally prevalent, types of disparity operationalizations that could give rise to distinctive findings once data anonymization is applied. Table 2 summarizes the key differences between the two types of operationalizations, disparity through separation and disparity through variation. In the passages that follow, we first describe the two types, respectively, before explicating the differences between the two.

3.2.1. Disparity Through Separation.

One stream of disparity operationalizations originated in the examination of racial discrimination in sociology (Pager and Shepherd 2008) and naturally extended to domains related to employment discrimination (e.g., Barnum et al. 1995, Gupta et al. 2020) and court cases pertaining to labor markets, such as pattern-and-practices cases alleging systemic discrimination in a workplace. In these domains, the main driver behind the detection of disparity is to affirm or reject the existence of underlying discrimination based on a focal social determinant such as race or gender (Garaud 1990, National Research Council 2004). Indicatively, disparity was operationalized with the goal of discerning its presence from chance. For example, courts have long applied a threshold of 5% significance level when establishing a prima facie case of discrimination (Barnett 1982), meaning that there must be less than 5% probability for the observed disparity to result from chance. Translating the 5% threshold to observed disparity, the Supreme Court opined in *Castaneda v. Partida*[4] that the disparity must exceed "two or three standard

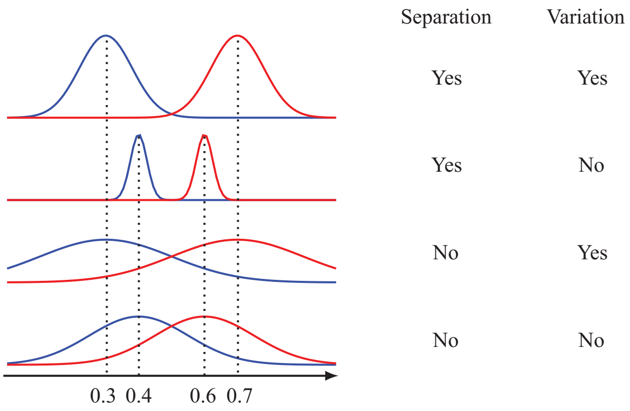
Table 2. Meanings and Properties of Disparity Operationalizations

Type	Meaning	Properties at maximum disparity	Statistical evidence	Primary disciplines	Court cases
Disparity through separation	Separation between outcome distributions for subpopulations	Maximum distance between means; minimum standard deviation	Regression analysis	Sociology, management, criminology, education	Pattern and practice of discrimination
Disparity through variation	Differences between mean outcomes of subpopulations	Maximum distance between means	Odds ratio	Epidemiology, medicine	Toxic tort

deviations." In essence, this is identical to the operationalization used by researchers to detect disparity through sample–mean comparisons (e.g., two-sample *t*-test; Castilla 2008, Acquisti and Fong 2020), even though the operationalizations in research tend to be more complex, bringing to bear not only the focal social determinant, but also other relevant variables (e.g., job performance), as well as the interaction effects between the social determinant and the other variables (National Research Council 2004).

We refer to this stream of disparity operationalizations as "disparity through separation" because, in both research and legal domains, these operationalizations were grounded in the idea of detecting the *separation* between outcome distributions for different subpopulations. Consider a simple example depicted in Figure 2, where there are two subpopulations and the outcome variable is binary—for example, representing whether an employee was promoted to a managerial position. Assuming independence between the promotion decisions for different employees, the percentage of employees who were promoted forms a binomial distribution for each subpopulation, like the curves in the figure. As can be seen from the figure, when disparity is operationalized through separation, its detection hinges on the degree of separation between the promotion-rate distributions of different subpopulations, rather than the raw difference between the observed promotion rates. For example, when the data set has 10 samples for each subpopulation, the imputed disparity cannot meet the 5% threshold when the observed promotion rates are 30% and 70%, respectively ($t = 1.95$, $p = 0.067$, with two-tailed *t*-test). Yet the disparity could meet the threshold for a much closer pair of observed promotion rates, like 40% and 60%, when the sample sizes are larger (e.g., when $n = 50$, there is $t = 2.04$, $p < 0.05$). The implication of this property, as summarized in Table 2, is that maximum disparity occurs when the distance between mean outcomes is maximized across subpopulations, *and* the standard deviation of outcomes within each subpopulation is minimized. We will discuss later in the paper how data-anonymization mechanisms, which could modify both the mean and standard deviation of the outcome variable, impact the detection of disparity when the latter is operationalized through separation.

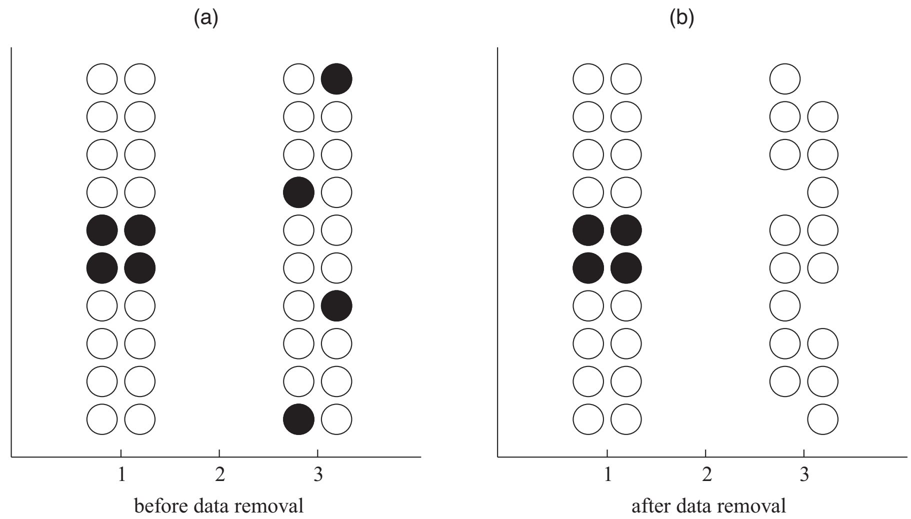
Figure 2. (Color online) Illustration of Disparity Operationalizations



Notes. Each curve depicts the probability density function for the sample mean of the outcome variable for a subpopulation. When the outcome variable is binary, such a sample mean follows a binomial distribution. In the top two charts, the two distributions have little overlap, enabling the detection of disparity through separation. The first and third chart from the top have the expected sample means differ considerably from each other, enabling the detection of disparity through variation.

3.2.2. Disparity Through Variation. The other stream of disparity operationalizations has its root in epidemiology, but has been applied to a wide variety of domains, including the detection of income disparity (Éltetö and Frigyes 1968, Siegel and Hambrick 2005), the argument of tort cases in courts (King 2006), etc. A classic example is *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[5] where statistical evidence was used to determine whether the ingestion of a drug during pregnancy has a disparate impact on birth defects—that is, whether there was a disparity in birth-defect rate between those who took the drug and those who did not. In these domains, the purpose of operationalizing disparity is often not only to establish its existence, but to quantify its magnitude (e.g., in order to assess monetary relief; King 2006). Indicatively, the operationalizations in this stream were explicitly designed to measure the

Figure 3. Illustration of a False Positive Created by Data Removal



Notes. Each circle represents an individual, with color marking its subpopulation membership. The x-axis represents the outcome variable. The y-axis represents another variable in the data set. The data-removal anonymization is conducted based on all three variables (i.e., x, y, and color). The four dark circles with $x \approx 3$ are removed after anonymization because, for each of them, there is no other circle that is close on all three variables. (a) Before data removal. (b) After data removal.

magnitude of disparity. For example, courts have long applied a "more likely than not" requirement on establishing disparate impact in civil cases, meaning that the odds of the undesired outcome in one subpopulation must be at least twice the odds in another[6] (Gastwirth 2019). This relative ratio resembles the *odds ratio* metric used in epidemiology research to detect disparity (e.g., Hebert et al. 2008).

We refer to this stream of disparity operationalizations as "disparity through variation" because, in both research and legal domains, these operationalizations were grounded in the idea of contrasting the mean outcomes across subpopulations. Consider again the example in Figure 2. When disparity is operationalized through variation, its detection depends only on observed promotion rates, but not the standard deviation of their distributions. For example, an observed pair of promotion rates of 30% and 70% always meets the "more likely than not" criterion (because $0.7/0.3 > 2$), no matter if it satisfies the aforementioned 5% threshold. In contrast, if the observed rates are 40% and 60%, their ratio is below the cutoff ($0.6/0.4 = 1.5 < 2$), no matter how large the samples are and whether the two distributions overlap. As summarized in Table 2, when disparity is operationalized through variation, the maximum disparity occurs when the distance between mean outcomes is maximized and is irrelevant to the standard deviation of outcomes within each subpopulation, forming a sharp contrast to the operationalization of disparity through separation.

It is important to note that, when the data are *not* anonymized, this obliviousness to standard deviation is unlikely to make the findings less robust, so long as the sample size is sufficiently large (Agresti 2002). Further, because the distribution of disparity-through-variation measures, like odds ratios, tend to have a positive skew[7] (Agresti 2002), the likelihood of them masking an existing disparity by chance is fairly small. Nonetheless, as we will elaborate in the mathematical formalism section, the situation could change drastically once the data set is anonymized. For example, many noise-insertion algorithms introduce the same degree of uncertainty to a statistical estimate, regardless of the sample size (e.g., Dwork et al. 2016). In this case, even a large sample is no longer a safeguard allowing one to gloss over the standard errors of statistical estimates in disparity detection. As we will discuss next, this leads to qualitatively distinct impacts of data anonymization on the two disparity operationalizations.

Table 3. Implications of Anonymization Mechanisms on Disparity Operationalizations

Mechanism	Disparity through separation	Disparity through variation
Data removal	FP more likely; FN less likely	Likelihood of FP and FN depend on data distribution
Noise insertion	FP highly unlikely; FN likely	FP and FN are equally likely

Notes. FN, false negatives (i.e., type II errors); FP, false positives (i.e., type I errors). The table assumes the direct use of the anonymized data set or the differential-privacy algorithm in disparity detection, which is the state of the practice today (Rocher et al. 2019). It does not preclude the future design of dedicated algorithms to compensate for the effect of data anonymization.

3.3. Implications of Anonymization on Disparity Detection

In presenting the typology of the data-anonymization mechanisms, we outlined two important distinctions between data removal and noise insertion. First, whereas data removal tends to reduce the standard deviation (of the outcome distribution) for a subpopulation, noise insertion almost always increases it. Second, whereas data-removal techniques rarely make any guarantee on the bias of statistics estimated from the anonymized data, many noise-insertion techniques guarantee certain estimated statistics (e.g., mean) to be unbiased. In what follows, we discuss how these two distinctions interact with the two disparity operationalizations to stimulate different outcomes of disparity detection over anonymized data. Table 3 summarizes the key differences.

Consider disparity through separation first. Given that data removal and noise insertion tend to shift the standard deviation in opposite directions, we can expect their ramifications on disparity detection to differ correspondingly. For example, it is highly likely for noise insertion to mask disparity because the increased standard deviation reduces the significance level of the difference between subpopulations. For the same reason, it is highly unlikely for noise insertion to trigger a false positive[8] when no disparity exists in the original data set. In contrast, a data-removal technique could more likely occasion false positives, as it reduces the observed standard deviation. Meanwhile, it may not be as likely to mask disparities unless the technique biases the outcomes in a way that reduces the observed differences between subpopulations.

Now, consider the operationalization of disparity through variation. In this case, the change of standard deviation does not affect the detection of disparity, but the bias of the observed outcomes does. This brings to the fore the potential bias introduced by data-removal techniques. Figure 3 depicts such an example. As can be seen from the figure, the two subpopulations have the same mean outcome in the original data set. Yet, after data removal, the mean outcome for one subpopulation becomes twice as much as the other, giving rise to a false positive. The opposite scenario can be constructed, where data removal masks an existing disparity.[9] Thus, with the data-removal mechanism, whether the bias would manifest as false positives or false negatives heavily depends on the underlying data distribution. This stands in contrast with the case of noise insertion, where most existing techniques guarantee the absolute or asymptotic unbiasedness of mean estimates (e.g., Dwork et al. 2016). Although the increased standard deviation could still shift the observed ratio in unpredictable directions, we are equally likely to observe false positives and false negatives, regardless of the underlying data distribution.

4. Mathematical Formalism

4.1. Preliminaries

In what follows, we present the preliminaries needed to launch our mathematical inquiry into the implication of anonymization on disparity detection. Specifically, we first introduce the formal data model before using it to present the two disparity operationalizations and the two anonymization mechanisms, respectively.

4.1.1. Model of Data. A common method for exploring the issue of disparity in an observed outcome (e.g., promotions in a workplace) is to develop a regression model that includes the focal social determinant (e.g., race) as one variable and the other relevant observed characteristics (e.g., job performance) as the other variables (National Research Council 2004). That is,

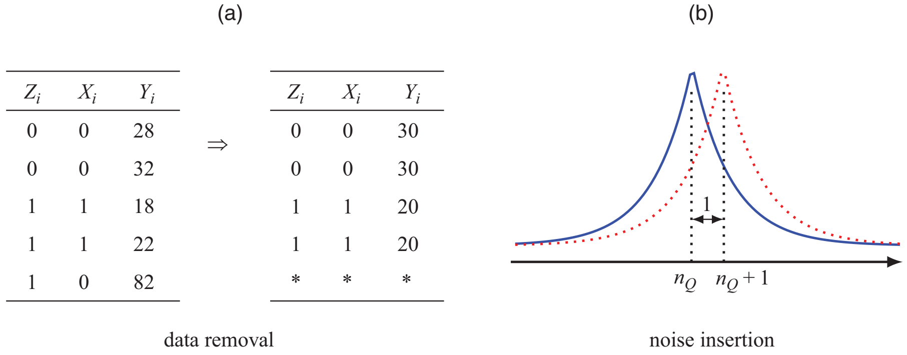
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i Z_i + \beta_3 Z_i + \xi_i, \qquad (1)$$

where Y_i is the outcome of interest, Z_i is the social determinant variable, X_i is a set of variables[10] deemed relevant to the outcome Y_i, and ξ_i is the stochastic error term (with a zero mean). A simple variant of this model (e.g., Everett and Wojtkiewicz 2002) is given by $\beta_2 = 0$, so that it captures only the direct effect of Z_i, but not the interaction between Z_i and X_i. The model can also allow for nonlinear relationships—for example, by applying the log-odds transformation when the outcome is binary (Rosenbaum 2010).

4.1.2. Disparity Operationalizations. Regardless of the operationalization, disparity is clearly captured in Equation (1) by $\beta_2 X + \beta_3$, the multiplicative factor for Z. This factor can be interpreted as the magnitude of disparity on the outcome variable Y given X. For example, when Z is binary (i.e., $Z \in \{0,1\}$), $\beta_2 X + \beta_3$ is the difference we would observe in Y if an individual were moved from one subpopulation ($Z = 0$) to another ($Z = 1$). Consistent with this interpretation, the average magnitude of disparity for an underprivileged subpopulation can be expressed as $\beta_2 \bar{X}_{\mathrm{u}} + \beta_3$, where $\bar{X}_{\mathrm{u}}$ is the average value of X for the subpopulation (National Research Council 2004). For the simpler variant of the model with $\beta_2 = 0$, the magnitude of disparity on Y becomes simply β_3.

When estimating the regression coefficients (e.g., β_3), we obtain both a point estimate and an estimate of its standard error. The two operationalizations of disparity differ in terms of which estimate(s) they take into account when detecting disparity. When disparity is operationalized through separation, the focus is on determining whether $\beta_2 \bar{X}_{\mathrm{u}} + \beta_3$ can be separated from zero,[11] according to a given level of statistical significance. Such a determination obviously depends on both the point estimate and the standard error. When

Figure 4. (Color online) Illustration of Anonymization Mechanisms


(a)

Z_i	X_i	Y_i
0	0	28
0	0	32
1	1	18
1	1	22
1	0	82

⇒

Z_i	X_i	Y_i
0	0	30
0	0	30
1	1	20
1	1	20
*	*	*

data removal

(b)

noise insertion

Notes. (a) Data removal. Depicted is a data set before and after data removal. Each row is a record, while each column is a variable. Following the principle of grouping similar records together, data removal assembles the two records with the same value combination of Z_i and X_i and replaces their values of Y_i with a common value. The last record (1,0,82) is removed because no other record shares the same (Z_i, X_i) combination of (1, 0). (b) Noise insertion. Depicted is the probability density function of two Laplace distributions with scale 1 and mean n_Q and n_Q+1, respectively. These two distributions represent the distributions of the (noise-inserted) query answers provided by the Laplace mechanism to any COUNT query with true answer n_Q and n_Q+1, respectively, in order to guarantee $(1,0)$-differential privacy. Note that the two distributions largely overlap with each other, indicating that the change of one record (i.e., shifting the COUNT query answer from n_Q to n_Q+1 or vice versa) is difficult to detect from the noise-inserted query answer.

disparity is operationalized through variation, only the magnitude is at play. For example, the "more likely than not" criterion discussed before is essentially testing whether $\beta_3 > \beta_0$—that is, whether the focal determinant Z bears a higher weight than the other factors captured in the intercept β_0—when X is empty and $Z \in \{0,1\}$. Clearly, such an operationalization depends on only the point estimates, but not the standard errors.

4.1.3. Anonymization Mechanisms.

Figure 4 illustrates the design of the two anonymization mechanisms. For both mechanisms, the input to anonymization is the set of all variables in the data set—that is, (X_i, Y_i, Z_i) in the model. With the data-removal mechanism, the goal is to prevent an individual from being identified from the output data. There are two common methods for existing techniques to achieve this goal. One is to *generalize* the values of certain variables. For example, in Figure 4(a), we generalized the values of Y_i for the first two records from 28 and 32 to both being 30. By doing so, we made the two records identical to each other, so that neither could be uniquely identified from the output data. The second method is called *suppression*, which is to remove certain records that cannot be easily made similar to other records. The last record in Figure 4(a) is an example. Given how far its value of Y_i is from the other records, in order to use generalization to make the last record identical with any other record, we have to make significant changes to Y_i for both records, limiting the usefulness of both in the anonymized data. Instead of doing so, we could simply remove the last record from the anonymized data and save other records from being modified, like is shown in Figure 4(a). Note that the anonymized data in the figure meet a popular data-removal guarantee called k-anonymity ($k = 2$), which requires that, for each record in the anonymized data set, there must be at least $k-1$ other records with the exact same value combination (Sweeney 2002b).

Compared with data removal, the existing noise-insertion techniques have provided a wider variety of outputs, from a noise-inserted data set to a way of generating randomly perturbed answers to queries over the data. Correspondingly, anonymity guarantees for the noise-insertion mechanism, like the aforementioned differential privacy guarantee, were broadly conceived to support any noise-insertion algorithm M that maps the input data set to (a value in) an arbitrary range Θ. For example, the popular (ϵ, δ)-differential privacy guarantee (Dwork et al. 2016) requires that, for any two data sets D and D' differing by one record and for any $S \subseteq \Theta$, the probability[12] for $M(D) \in S$ and $M(D') \in S$ must not differ significantly, with the difference bounded by a function of the two parameters ϵ and δ:

$$\mathcal{P}(M(D) \in S) \le e^{\epsilon}\mathcal{P}(M(D') \in S) + \delta. \tag{2}$$

Note from the equation that the smaller ϵ and δ are, the harder it is to distinguish D from D' after applying M, meaning that M provides a higher degree of anonymity.

Researchers have developed many techniques that can guarantee (ϵ, δ)-differential privacy (see review in Hay et al. 2016). A simple, yet popular, one is the *Laplace mechanism* (Dwork et al. 2016) depicted in Figure 4(b), which inserts noise when answering queries posed over the data. For example, when answering a count query Q that asks for the number of records n_Q satisfying the conditions specified in Q, the Laplace mechanism adds to n_Q a random variable drawn from the Laplace distribution with mean zero and scale $1/\epsilon$, so that the probability density of the perturbed query answer at point $n_Q + r$ is

$$f(n_Q + r) = \frac{\epsilon}{2} e^{-r\epsilon}, \tag{3}$$

which is depicted by the solid curve in Figure 4(b). Note from Equation (3) that the probability density varies by a multiplicative factor of e^{ϵ} when n_Q varies by one, which is the maximum possible difference between two data sets D and D' that differ by one record (as specified in the definition of differential privacy). According to Equation (2), this means the Laplace mechanism always achieves $(\epsilon, 0)$-differential privacy, no matter what the data set D or the query Q is.

4.2. Implications of Anonymization on Disparity Testing

We start by considering how the data-removal mechanism affects the outcomes of disparity testing. When disparity is operationalized through separation, the standard errors of the estimated regression coefficients are salient to whether the test identifies a statistically significant separation between different populations. Consequently, the following theorem examines how a generalization method for data removal—designed specifically to achieve the aforementioned k-anonymity guarantee—affects the standard errors of the regression outputs. Note that the theorem assumes the direct use of the anonymized data set in regression analysis, which is the way these data sets are used in practice today (Rocher et al. 2019). Although it is possible to modify the regression analysis to compensate for the effect of data removal, the design of such dedicated algorithms is beyond the scope of the paper.

Theorem 1. *When the data set has k records for each value combination of X_i and Z_i, and anonymity is achieved by replacing the values of Y_i in each anonymous group with their average, the standard error for the estimation of each regression coefficient (i.e., $\beta_0, \beta_1, \beta_2, \beta_3$) is reduced by a multiplicative factor of $1/\sqrt{k}$ after anonymization.*

Because of the space limit, please refer to the e-companion (section EC.2) for the proof of the theorem. Consistent with our earlier conceptual findings, Theorem 1 shows that the use of the data-removal mechanism, specifically the popular generalization technique, reduces the standard errors of regression coefficients and may produce false positives in identifying disparities. Although the mathematical proof is subtle, the finding of the theorem has a simple intuitive explanation. Note that data removal in general, and generalization in particular, tends to group similar records together to elide their differences, so as to prevent any single record from being uniquely identified. A direct consequence of this design is that records belonging to the same subpopulation are more likely to be grouped together. Consider a case where all records for the same subpopulation are placed into one group, and their outcome variable values are all replaced by the group mean. It clearly makes any testing of disparity-through-separation more likely to declare a positive result, because the within-subpopulation variance is artificially reduced to zero.

This issue no longer applies when disparity is operationalized through variation, because the identification now depends only on the point estimates, and not the standard errors. Nonetheless, it gives primacy to the potential bias introduced by data removal to the point estimates. We demonstrated an example in the conceptual development section (Figure 3), where data removal substantially alters the observed outcome distribution. The following theorem extends the example to highlight the severity of the problem when the outcome distribution is skewed, like the heavy-tailed distributions commonly present in practice (Nolan 2003). With a heavy-tailed distribution, the largest values are associated with the lowest probability density, meaning that removing records with the sparsest neighborhoods tends to reduce the mean estimate considerably. The theorem considers the exponential distribution as a conservative example, because its skewness serves as a lower bound for the skewness of heavy-tailed distributions.[13]

Theorem 2. *When Y_i follows the exponential distribution $Y_i \sim \mathrm{Exp}(\lambda)$, suppressing m out of n records according to the density of Y_i shifts the sample mean of Y_i by an expected value of*

$$E(\bar{Y} - \bar{Y}_0) = \frac{m \log n - \log m!}{\lambda n}, \tag{4}$$

where $\bar{Y}$ and $\bar{Y}_0$ are the sample mean of Y_i before and after suppression, respectively.

As discussed earlier, please refer to the e-companion (section EC.3) for the proof of the theorem and a discussion of its extensions beyond the exponential distribution. The theorem confirms our conceptual discussions by demonstrating how a few suppressed records could have considerable influence on point estimates, such as the sample mean. For example, removing 10 records from a 100-record data set changes the sample mean by an expected amount of $0.45/\lambda$. Because the mean of an exponential distribution $\mathrm{Exp}(\lambda)$ is $1/\lambda$, this expected change represents 45% of the real value, clearly large enough to flip the outcome of disparity-through-variation measures, such as the aforementioned "more likely than not" criterion.

Finally, we turn our attention to the noise-insertion mechanism for anonymization. Interestingly, unlike data removal, the designers of noise-insertion techniques often provide statistical guarantees on how the inserted noise affects the output of a regression analysis. For example, when a noise-insertion technique directly modifies query answers (e.g., the aforementioned Laplace mechanism), we can construe a regression coefficient as a complex query posed to the data. The random noise added to the query answer then directly reveals the statistical properties of our estimate of the regression coefficient. This makes our analysis here considerably easier. Specifically, many existing techniques for noise insertion, including the Laplace mechanism, produce estimates that are guaranteed to be unbiased (Dwork et al. 2016). Although other techniques might introduce a small degree of statistical bias in exchange for a substantially reduced standard error, such biases tend to be small and asymptotically close to zero as the data set size grows (Li et al. 2014). As such, noise insertion could obviously produce both false positives and false negatives when disparity is operationalized through variation. For disparity through separation, the following theorem establishes an upper bound on the statistical power of *any* disparity test for *any* noise-insertion algorithm that is (ϵ, δ)-differentially private.

Theorem 3. *With any* (ϵ, δ)*-differentially private algorithm, when* $Z_i \in \{0,1\}$, X_i *follows an independent and identically distributed (univariate or multivariate) Gaussian distribution, and* $Y_i = \beta_0 + \beta_1 X_i + \beta_3 Z_i + \xi_i$ $(\beta_3 > 0)$, *the statistical power of any disparity-through-separation test must satisfy*

$$\text{Power} \leq 1 - e^{\frac{-6\epsilon n|\beta_3|}{\sigma}}(1-\alpha) + \frac{4n\delta|\beta_3|}{\sigma}, \tag{5}$$

where α *is the significance level for the disparity test,* n *is the number of records in the data set, and* σ *is the standard deviations for* $\beta_1 X_i + \xi_i$.

The proof of the theorem is available in the e-companion (section EC.4). The bound in Inequality (5) offers important insights into the trade-off between privacy protection and disparity identification. For example, the smaller ϵ and δ are (i.e., the more stringent the privacy guarantee), the lower the statistical power will be. As a result, stringent privacy protection comes at the cost of disparity detection. For example, requiring a privacy budget of $\epsilon = 0.001$ means that the statistical power of disparity detection could drop from 0.80 for an original data set with 100 records to, at most, 0.32 after anonymization ($\alpha = 0.05$, $|\beta_3|/\sigma = 0.63$). This confirms our conceptual finding that noise insertion could mask a considerable number of disparities in the data set when disparity is conceptualized through separation.

5. Empirical Examination

5.1. Data Set

We obtained an inpatient data set from one of the five most populated states in the United States. The data set contains 486,924 records of patients who were admitted and discharged by one of the healthcare facilities in the state during a calendar quarter. It covers 244 healthcare facilities, which represent all privately owned facilities in the state admitting inpatients, except three types of exempted ones: long-term acute-care facilities; psychiatric and rehabilitation facilities; and facilities in noncompliance (e.g., due to excessive error rate). The number of patients discharged from a hospital ranges from 5 to 9,104. The variables included for each patient cover information about demographics, diagnosis, treatment, and financial arrangements. Table EC.2 in the e-companion (section EC.5) lists the summary statistics for the key variables of the data set used in the study.

An important reason why we used this data set is its close resemblance to the inpatient data set processed by the Texas Department of State Health Services (2019) under the procedure in Table 1. Because Texas does not permit the release of its data set before anonymization, our data set becomes an ideal device to examine the implications of applying the Texas procedure.[14] Although we also studied a number of other, more technically sophisticated, anonymization techniques (as elaborated later in the section), we considered the examination of the Texas procedure important because it represents a rare case where a government agency explicitly specifies the step-by-step procedure for anonymizing a data set with highly sensitive private information.

5.2. Design of Empirical Study

5.2.1. Disparity Measurement. In the passages that follow, we describe the dependent variables, independent variables, and disparity-detection methods used in the study, respectively.

5.2.1.1. Dependent Variables. We used two dependent variables that have been frequently examined in the context of health disparity (e.g., Xu and Zhang 2019, Danziger et al. 2020): admission severity (*SERV*) and nonresponder indicator (*NONRES*). Admission severity (Steen and Cherney 1996) was measured on a five-point scale (from zero, no clinical instability, to four, maximal instability). The nonresponder indicator captures whether a patient is responding to treatment during the hospital stay. It is determined by comparing the clinical variables collected at midstay with those collected at admission. A patient is deemed a nonresponder if the probability of in-hospital mortality (as predicted based on the clinical variables) is higher at midstay than at admission.

An important reason why we chose the two dependent variables is their distinctive characteristics. All hospitals in the state of our data set are required to collect the admission severity information, making its coverage near 100% in the data set. The nonresponder indicator, in contrast, is optional to report. Further, patients deemed to have moderate or lower clinical instability at admission are ineligible for the calculation. As a result, only 4.3% records in the data set contain the binary (Yes/No) determination of whether the patient is a nonresponder. This sharp contrast between the two dependent variables allows us to examine two distinct scenarios: (1) where the outcome variable applies to all individuals in the data set (e.g., income disparity), and (2) where the outcome applies only to a small fraction of the individuals (e.g., disparity in promotion to executive positions; rare-disease disparity studies, Holtzclaw Williams 2011; or in the Apple iOS case, where only a small percentage of keystrokes need correction).

5.2.1.2. Social Determinants. For the purpose of disparity detection, we examined gender, race, ethnicity, and age as the focal social determinants, respectively. There are two main reasons for selecting these four variables. First, these variables were frequently the focal social determinants in disparity studies (Zhang et al. 2017). Second, they were also often treated as "quasi-identifiers" in the privacy context (Sweeney 2000), and therefore (selectively) removed or obscured for the purpose of anonymization, like in the Texas procedure in Table 1. Given their prominence in both data anonymization and disparity detection, focusing on these variables allows us to better explicate the implications of the former on the latter.

5.2.1.3. Control Variables. To further emulate the analyses commonly carried out in the disparity research literature, we also included three individual-level variables as controls in the empirical study: (1) insurance status (*INS*; a binary variable indicating whether an individual is covered by a health insurance); (2) cancer history (*CANCER*; a binary variable indicating whether an individual has a history of cancer diagnosis); and (3) length of the hospital stay in days (*LOS*; an integer variable and a widely used proxy for the complexity of the individual's medical condition; May et al. 2016). These control variables were selected because of their relevance to either the financial situation or the medical condition of a patient, which were frequently used as controls in the health-disparity literature (see references in Zhang et al. 2017). We stress at the outset that we do not attempt to establish any causal relationships between a social determinant and an outcome variable in this paper. As we elaborate on in the discussion, doing so would require theoretical development that is beyond the scope of this paper and a careful scrutiny of which clinical and socioeconomic variables to use as controls, an issue that is still under intensive debates in the disparity-research literature (National Research Council 2004, Pager and Shepherd 2008).

5.2.1.4. Disparity Detection. We considered two disparity-analysis methods corresponding to the two operationalizations, respectively. For disparity through separation, we turned to regression analysis for estimating the model in Equation (1). Specifically, the disparity in an outcome variable (e.g., admission severity) over a focal social determinant (e.g., race) was estimated with the outcome being Y, the focal social determinant being Z, and the other three social determinants together with the three control variables[15] forming X. We created dummy variables for *Sex* and *Race* and estimated the model using ordinary least squares when the dependent variable is admission severity. Because the other dependent variable (i.e., nonresponder indicator) is binary, we used logistic regression with the maximum-likelihood estimator.[16] For disparity through variation, we considered the frequently used measure of odds ratio (Bland and Altman 2000):

$$\mathrm{OR} = \frac{P(Y=1 \mid Z \in V_1, X)/(1-P(Y=1 \mid Z \in V_1, X))}{P(Y=1 \mid Z \in V_0, X)/(1-P(Y=1 \mid Z \in V_0, X))}, \tag{6}$$

where X, Y, and Z are as defined in Equation (1), and V_0 and V_1 are two subsets of the domain of Z. Intuitively, the odds ratio captures the impact of shifting Z from V_0 to V_1[17] on the odds of $Y = 1$ when holding X constant. The estimation of odds ratio can be done through logistic regression, specifically, as e^{β}, where β is the regression coefficient for Z. To make the dependent variable Y binary, when calculating the odds ratio for admission severity, we grouped its five values into two groups divided at the median: {0, 1} as one level and {2, 3, 4} as the other.

5.2.2. Data-Anonymization Techniques. To examine the distinct implications of different data-anonymization mechanisms on disparity detection, we implemented a total of four data-anonymization algorithms, two in data removal and the other two in noise insertion. The first algorithm we implemented was the aforementioned rules used by the Texas Department of State Health Services (2019) to anonymize their state-wide inpatient discharge data set (Table 1). Although all rules are applicable to our data set, two minor adjustments are in order. First, we changed the state of Texas in rule 2 to the state in our data set. Second, because our data set contains ICD-9 instead of ICD-10 codes, we identified and used the ICD-9 codes indicating alcohol/drug use or HIV[18] when applying rule 3. We found in our study that the only rules (in Table 1) with material impacts on disparity detection are rules 3 and 8, because they removed social determinants included in our disparity analysis. Because rule 8 has a tunable parameter (i.e., a threshold of 10 patients), we also tested a variant of the rules when the threshold is 20 instead of 10.

Next, we considered *k*-anonymity, a data-removal mechanism recommended by an European Union advisory body for removing the risk of individual identifications (Article 29 Data Protection Working Party 2014). Notwithstanding recent debates on the correctness of such a recommendation (Cohen and Nissim 2020), *k*-anonymity is clearly popular among practitioners (e.g., Ali 2018). For our implementation, we used the local suppression algorithm included in the sdcMicro R package (Templ et al. 2015). The algorithm was designed to remove as few variable values as possible to achieve *k*-anonymity. To examine how minimal anonymization (i.e., $k = 2$) affects disparity detection, we tested the cases of $k = 2$ and 5 in the study.

For linear regression, we implemented a recently developed variant (Wang 2018) of the differentially private *sufficient statistics perturbation* (SSP) algorithm (Foulds et al. 2016), which guarantees (ϵ, δ)-differential privacy in solving a linear model $y = X\beta + \xi$ by first computing the differentially private versions of $X^\top X$ and Xy, respectively, before generating the estimated coefficients as $(\widehat{X^\top X})^{-1}\widehat{Xy}$. Compared with the original SSP algorithm, the variant developed by Wang (2018) further exploits data-dependent quantities to achieve near-optimal data utility and was shown to substantially outperform the other existing solutions for differentially private linear regression. Because δ is usually set as a negligible value, we followed Wang (2018) in setting $\delta = \min(10^{-6}, 1/n^2)$, where n is the data set size, and varied ϵ between 0.1 and 1.

For logistic regression (and the related odds-ratio estimation), we implemented the differentially private algorithm for regularized empirical risk estimation (Chaudhuri et al. 2011), which produces more accurate coefficient estimates than traditional output perturbation algorithms (like the aforementioned Laplace mechanism) because it achieves differential privacy by perturbing the objective function of the optimization process instead of the final output of coefficient estimates. The algorithm was designed to achieve $(\epsilon, 0)$-differential privacy and features only two parameters, ϵ and λ, which is the regularization parameter controlling the ℓ_2-regularization term. Following the recommendations by Chaudhuri et al. (2011), we set in our implementation $\lambda = 1/(4n(e^{\epsilon/20} - 1))$, where n is the input size, and varied ϵ between 0.1 and 1.

5.3. Empirical Results

Table 4 reports how applying the data-removal mechanism affected the detection of disparity operationalized through separation. As can be seen from the table, the anonymization methods used in practice, like the Texas procedure, could substantially interfere with identifying disparity through separation. In the case of *k*-anonymity, even the weakest form of anonymity (i.e., $k = 2$) produced a false-positive disparity in admission severity for Asians. The interference could be toward either direction. For example, for people of Hispanic origins, the Texas procedure identified a significantly lower admission severity, whereas *k*-anonymity ($k = 5$) identified a significantly higher severity, yet the original data cannot support either. Also note from the table that, consistent with Theorem 1, the *k*-anonymity algorithm tends to produce more false positives than false negatives.

Table 5 reports how applying the data-removal mechanism affected the detection of disparity operationalized through variation. Note that, although we reported the regression coefficients for the control variables in Table 4, we do not include these variables in Table 5 onward, given that the odds-ratio metric is only applicable to the social determinants. As can be seen from Table 5, both data-removal techniques affected disparity detection substantially, even reversing its direction in several cases. This is consistent with our earlier conceptual discussions and Theorem 2. Also note a sharp contrast with Table 4: When disparity was operationalized through variation, *k*-anonymity masked the severity of disparity *in addition to* producing false positives. For example, achieving 2-anonymity entailed a reduction of the odds ratio for Asian to be a nonresponder from 2.79 to 1.52, incurring a false negative if the "more likely than not" criterion is used.

Table 6 reports how noise insertion affected the detection of disparity. The left part of the table confirms the results in Theorem 3—that is, differential privacy is likely to produce false negatives, but not false positives, when disparity is operationalized through separation. Remarkably, even when $\epsilon = 1$, a level widely perceived as weak in practice (Tang et al. 2017, Dwork et al. 2019), the differential privacy algorithm still masked the (only) statistically significant disparity for the nonresponder indicator, with a false negative rate of 99%. The right half of the table shows a result similar to Table 5. Like the data-removal mechanism, the noise-insertion algorithms shifted the estimated magnitude of disparity in unpredictable ways, amplifying the odds ratio for some, weakening it for others, and even reversing the direction in several cases. This is, again, consistent with our conceptual discussions for Table 3.

We also examined the robustness of the empirical findings when varying the size of the input data set. Because of the space limit, please refer to the e-companion (section EC.6) for the results.

Table 4. Effects of Data Removal on Detecting Disparity Through Separation

	SERV					NONRES				
	Original	Texas Procedure		*k*-Anonymity		Original	Texas Procedure		*k*-Anonymity	
	Data	$c = 10$	$c = 20$	$k = 2$	$k = 5$	Data	$c = 10$	$c = 20$	$k = 2$	$k = 5$
AGE	0.01***	0.01***	0.01***	0.01***	0.01***	0.00	0.01	0.01	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
HISPAN	-0.01	-0.15***	-0.17***	0.00	0.05***	0.93	13.43	0.27	0.73	0.72
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.61)	(1175.14)	(4026.27)	(0.61)	(0.74)
RACEA	-0.03	-0.00	-0.03	-0.06***	-0.09***	1.03	-13.47	-13.81	0.42	0.70
	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)	(0.56)	(703.57)	(1405.41)	(0.76)	(1.06)
RACEB	0.05***	0.09***	0.09***	0.05***	0.05***	0.60**	1.12	-13.50	0.61**	0.62**
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.21)	(0.76)	(747.98)	(0.21)	(0.21)
RACEI	-0.65***	0.16	0.16	-0.71***	-0.79***	-7.52			-7.86	-7.87
	(0.02)	(0.15)	(0.15)	(0.02)	(0.02)	(137.98)			(196.97)	(196.97)
RACEN	-0.03**	-0.27***	-0.29***	-0.03**	-0.07***	0.19	1.14	-12.80	0.40	0.56
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.43)	(1664.78)	(4439.25)	(0.42)	(0.45)
SEXF	-0.06***	-0.06***	-0.06***	-0.06***	-0.06***	-0.09	-0.44*	-0.02	-0.08	-0.08
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.07)	(0.20)	(0.27)	(0.07)	(0.07)
INS	0.07***	0.09***	0.10***	0.07***	0.07***	-0.23	0.02	13.60	-0.23	-0.28
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.33)	(1.05)	(876.60)	(0.33)	(0.33)
CANCER	0.10***	0.07***	0.08***	0.10***	0.10***	0.05	0.19	0.32	0.05	0.05
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.08)	(0.24)	(0.31)	(0.08)	(0.08)
LOS	0.03***	0.04***	0.04***	0.03***	0.03***	0.05***	0.05***	0.05***	0.05***	0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)	(0.00)

Notes. Dependent variable: *SERV*, admission severity; *NONRES*, nonresponder indicator. Columns under *SERV* are estimated by ordinary least squares. Columns under *NONRES* are estimated by logistic regression with the maximum-likelihood estimator. Dark-gray background marks false negatives (under significance level of 0.05), while light gray marks false positives. *RACEx* and *SEXF* are dummy variables, with their names concatenating the original variable name with the value explained in Table EC.2 in the e-companion. *RACEI* is empty under *NONRES*/"Texas Procedure" because the procedure removes the race information for all American Indians with *NONRES* = 1 (i.e., "Y") in the data set.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

6. Discussion

6.1. Policy and Managerial Implications

The protection of consumer privacy has emerged as a task front and center for firms and policy makers in today's digitally connected world. Similarly, the recognition and rectification of disparate impact is increasingly regarded as a societal imperative in employment, housing, healthcare, etc. The primacy afforded to the two issues will only be reinforced in the future by the rapidly growing collection of consumer data and the diversifying landscape of technological issues, from which both privacy and disparity concerns frequently arise. This makes it all the more important for researchers, practitioners, and policy makers to be mindful of the potentially complex interplay between the two, which is the focus of this paper. With this

Table 5. Effects of Data Removal on Detecting Disparity Through Separation

	SERV					NONRES				
	Original	Texas Procedure		*k*-Anonymity		Original	Texas Procedure		*k*-Anonymity	
	Data	$c = 10$	$c = 20$	$k = 2$	$k = 5$	Data	$c = 10$	$c = 20$	$k = 2$	$k = 5$
AGE	1.03	1.03	1.03	1.03	1.03	1.00	1.01	1.01	1.00	1.00
HISPAN	0.92	0.60	0.59	0.92	0.96	2.54	$> 10^5$	1.31	2.08	2.06
RACEA	0.87	0.90	0.89	0.85	0.82	2.79	0.00	0.00	1.52	2.01
RACEB	1.18	1.31	1.33	1.18	1.18	1.83	3.07	0.00	1.83	1.85
RACEI	0.07	1.58	1.56	0.04	0.00	0.00			0.00	0.00
RACEN	0.90	0.40	0.37	0.90	0.85	1.21	3.11	0.00	1.49	1.75
SEXF	0.94	0.95	0.97	0.94	0.93	0.92	0.65	0.98	0.92	0.92

Notes. Odds ratio estimated by logistic regression as $\widehat{\mathrm{OR}} = e^{\widehat{\beta}}$, where $\widehat{\beta}$ is the estimated regression coefficient. Dark-gray background marks cases where the magnitude of disparity decreases by more than 25% (i.e., $\max(\widehat{\mathrm{OR}}, 1/\widehat{\mathrm{OR}}) \le 0.75\max(\mathrm{OR}, 1/\mathrm{OR})$). Light-gray background marks cases where the magnitude increases by more than 25% (i.e., $\max(\widehat{\mathrm{OR}}, 1/\widehat{\mathrm{OR}}) \ge 1.25\max(\mathrm{OR}, 1/\mathrm{OR})$). Inverted color marks cases where the estimated effect has a reversed direction (i.e., $\widehat{\mathrm{OR}} > 1$ and $\mathrm{OR} < 1$ or vice versa).

Table 6. Effects of Noise Insertion on Detecting Disparity

	Separation						Variation					
	SERV			NONRES			SERV			NONRES		
	Original	$\epsilon = 0.1$	$\epsilon = 1$	Original	$\epsilon = 0.1$	$\epsilon = 1$	Original	$\epsilon = 0.1$	$\epsilon = 1$	Original	$\epsilon = 0.1$	$\epsilon = 1$
AGE	$p < .001$	1.00	1.00	$p > .05$	0.00	0.00	1.03	1.02	1.02	1.00	1.00	1.00
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
HISPAN	$p > .05$	0.00	0.00	$p > .05$	0.00	0.00	0.92	1.04	0.98	2.54	0.34	0.45
		(0.00)	(0.00)		(0.00)	(0.00)		(1.22)	(0.42)		(0.61)	(0.27)
RACEA	$p > .05$	0.00	0.00	$p > .05$	0.01	0.00	0.87	1.29	0.63	2.79	1.00	1.43
		(0.00)	(0.00)		(0.01)	(0.00)		(5.20)	(0.51)		(12.29)	(2.01)
RACEB	$p < .001$	1.00	1.00	$p < .01$	0.01	0.01	1.18	1.15	1.01	1.83	1.11	1.57
		(0.00)	(0.00)		(0.01)	(0.01)		(2.20)	(0.17)		(1.99)	(0.48)
RACEI	$p < .001$	0.94	1.00	$p > .05$	0.00	0.00	0.07	1.02	1.02	0.00	0.93	1.08
		(0.02)	(0.00)		(0.00)	(0.00)		(29.16)	(2.51)		(13.28)	(3.69)
RACEN	$p < .01$	0.00	0.00	$p > .05$	0.00	0.01	0.90	1.25	1.14	1.21	0.89	0.65
		(0.00)	(0.00)		(0.00)	(0.01)		(2.28)	(0.42)		(3.26)	(0.65)
SEXF	$p < .001$	1.00	1.00	$p > .05$	0.01	0.00	0.94	0.80	0.80	0.92	0.89	0.91
		(0.00)	(0.00)		(0.01)	(0.00)		(0.12)	(0.02)		(0.31)	(0.03)

Notes. Because both algorithms are randomized ones, we ran each algorithm/parameter combination 100 times and reported the statistics of the results in the table. Columns under *Separation*/$\epsilon = 0.1$ or 1 depict the frequency for the algorithm to return $p < 0.05$ in the 100 runs, with the standard error in parentheses. Columns under *Variation* depict the median odds ratio estimation from all runs, with the standard deviation in parentheses. Note that median is reported to highlight the fact that the change of odds ratio must be at least as severe in 50% of all runs. Color coding follows Table 5.

backdrop, our results highlighted the importance of examining the disparate impact of privacy protection on different individuals and illuminated the intricacies of identifying disparity in privacy-preserved data. In the following, we lean on our findings to provide actionable suggestions for ensuring a proper alignment between the design of anonymization and the operationalization of disparity.

First, when a data set used for identifying disparity has already been anonymized, it is of paramount importance to explicate the anonymization mechanism that has been applied *before* examining the statistical evidence of disparate impacts. For example, if noise insertion has been applied, operationalizing disparity through separation tends to produce conservative findings, in which false positives are highly unlikely. Thus, such results are at least as valid (as the results over the original data set) in establishing a prima facie case of discrimination. In contrast, if a data-removal mechanism like *k*-anonymity has been applied, then false positives become much more likely than false negatives. As a result, these may better serve as exploratory steps for determining whether further research is warranted for a particular type of disparate impacts. Understanding this subtle interaction between anonymization and disparity detection is increasingly important, given the growing popularity among firms to either anonymize data at collection time (e.g., Google's RAPPOR system; Erlingsson et al. 2014) or base analytical decisions on privacy-preserved data (e.g., Uber's Flex system; Johnson et al. 2018).

Second, if a data set has the potential to be used for an examination of disparate impacts, then the design of anonymization should consider, in tandem, the protection of privacy and the utility of the anonymized data for disparity detection. The literature has repeatedly noted the necessary trade-off between the two goals (e.g., Kifer and Machanavajjhala 2011). More importantly, there are existing techniques for noise insertion (e.g., an algorithm used in this paper for differential privacy; Chaudhuri et al. 2011) that were proven to achieve the Pareto optimality on this trade-off (under certain assumptions). For data removal, although achieving optimality was proven difficult (e.g., see Meyerson and Williams 2004 for the NP-hardness proof for the case of *k*-anonymity), researchers have developed approximation algorithms that reach within a constant factor of the optimal trade-off (Aggarwal et al. 2005). These results not only provide anonymization mechanisms that suit the purpose of disparity detection, but help to illustrate what *can* be achieved in terms of disparity detection when the data set must be anonymized to satisfy certain privacy guarantees. Knowledge of this achievable trade-off will, in turn, enable regulators and policy makers to properly appraise the varying impacts of privacy protection (and disclosure) on different subpopulations before mandating or incentivizing either privacy protection (e.g., through privacy legislation such as GDPR) or the collection of social-determinant information for disparity detection (Adler and Stead 2015).

6.2. Limitations and Future Research

Our work was limited by its focus on the detection of observable disparity, rather than the identification of any underlying causal discrimination. It is important to note that even large and persistent disparities in a data set do not prove discrimination, as the latter requires substantial prior knowledge about the mechanisms through which the data were generated. For example, one must tease out endogeneity threats from omitted variables and common biases, such as sample-selection bias, before supporting a causal inference of discrimination (Pager and Shepherd 2008). To this end, our work is only the first step toward understanding the implications of data anonymization on identifying discrimination. Future studies could examine how data anonymization affects the subsequent steps of causal inference beyond the establishment of observed disparity.

Another limitation of our work relates to other potential impacts of anonymizing data. Although there are obviously disparate impacts on underprivileged subpopulations when identifiable disparities are masked, which is the focus of this paper, such disparate impacts could also stem from other uses of the anonymized data—for example, when the data are used to allocate resources like education and healthcare (Ekstrand et al. 2018, Pujol et al. 2020). Interestingly, if we switch the unit of analysis from subpopulations to individuals, then anonymization has been shown to *prevent* certain discrimination, simply because no individual is uniquely identifiable from the anonymized data set (Ruggieri et al. 2014, Hajian et al. 2015, Kashid et al. 2015). Future research could further examine these countervailing impacts of data anonymization, so firms could properly balance them when choosing the data-anonymization mechanism to apply.

Finally, we offer the caveat that the typologies presented for the anonymization mechanisms and the disparity operationalizations were meant to highlight their nuanced interplay rather than serving as a strict binary classification. As such, although the characteristics for each type are expected to hold in general, there are bound to be exceptions. For example, we listed employment discrimination and epidemiological disparity as the sample domains operationalizing disparity through separation and variation, respectively. In practice, although most studies and legal cases involving employment discrimination operationalized disparity through separation, the U.S. Equal Employment Opportunity Commission (EEOC) famously suggested a rule of thumb that falls under disparity through variation.[19] Similarly, there were technical attempts to develop data-anonymization techniques that feature both data removal and noise insertion (Li et al. 2012, Li and Sarkar 2013). The existence of these exceptions, however, does not affect our findings pertaining to the implications of data anonymization on disparity identification.

Acknowledgments

The authors express sincere gratitude to the editors and reviewers for their helpful comments. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency or agencies that fully or partially supported the work of the author(s).

Endnotes

[1] See *Daubert v. Merrell Dow Pharm., Inc.*, 43 F.3d 1320 (9th Cir. 1995).

[2] As elaborated on later in the paper, disparity through separation means that, for a given outcome variable (e.g., income), there are (at least) two subpopulations for which the distributions of the outcome variable are clearly separable from each other (e.g., their mean estimates have nonoverlapping confidence intervals). Disparity through variation, on the other hand, means that the mean outcome for one subpopulation differs considerably from the other(s).

[3] For example, the U.S. Department of Health and Human Services (2012) recommended the removal of the last two digits of ZIP code as one of the measures for complying with the Health Insurance Portability and Accountability Act.

[4] 430 U.S. 496 (1977).

[5] 509 U.S. 579 (1993).

[6] The rationale for setting this relative-ratio threshold at 2.0 is because, for individuals in the first subpopulation who suffered the undesired outcome, only when the relative ratio is at least 2.0 can we possibly conclude that the undesired outcome is more likely a result of the disparity than other factors shared with the second subpopulation.

[7] More rigorously, when the population odds ratio $OR > 1$, the probability distribution of the observed odds ratio $\widehat{OR}$ tends to have a positive skew.

[8] Obviously, one can always make a trade-off between false-positive and false-negative rates by adjusting the threshold cutoff in disparity detection. Our discussions here assume the direct use of the anonymized data set without such adjustments, which is the way anonymized data sets are used today (Rocher et al. 2019). It is important to note that, even with an adjusted trade-off, the higher standard deviation after data removal entails a lower statistical power of disparity detection for the same significance level.

[9] For example, we can construct such a scenario by removing all light-colored circles with $x \approx 3$ from Figure 3. With this modification, the ratio before anonymization becomes two, and its value afterward becomes one. In other words, the data-removal process masks a previously existing disparity.

[10] When X_i contains more than one variable, β_1 and β_2 become vectors.

[11] This holds even when the detection of disparity does not directly involve a regression model, e.g., when a two-sample t-test is used to determine whether the mean of two subpopulations are different or when one examines whether the confidence intervals for the mean of different subpopulations overlap. In these two examples, the tests are essentially equivalent with testing a null hypothesis of $\beta_3 = 0$ when X is empty (Cumming 2009).

[12] The notion of probability is taken over the randomness of M, e.g., the randomness of the noise it inserts into the data or the query answers.

[13] Note that while the skewness of the exponential distribution is always two, the skewness of a heavy-tailed distribution may be unbounded.

[14] Specifically, it allows us to compare the results of disparity analysis over the original data with those over the anonymized data, so as to explicate the effect of data anonymization on disparity detection.

[15] For the sake of simplicity and clarity, we did not estimate the interaction item in the model.

[16] For the nonresponder indicator, we also tested the probit model and Firth's penalized likelihood estimator (Firth 1993) for correcting the potential rarity-induced bias in logistic regression, but did not find notable differences.

[17] For example, when Z is binary like *Sex*, we set $V_1 = \{\text{Male}\}$ and $V_0 = \{\text{Female}\}$. When Z is *Race*, we could have $V_1 = \{\text{Black}\}$ and V_0 containing all other values of *Race*. When Z is *Age*, the odds ratio is usually calculated for $V_1 = \{v+1\}$ and $V_0 = \{v\}$—that is, differing by one unit, so as to capture the effect of a one-unit increase in Z on the odds of $Y = 1$.

[18] ICD-9 and ICD-10 codes represent different versions of the *International Statistical Classification of Diseases and Related Health Problems*. Either can be used to represent diseases, symptoms, etc. ICD-9 codes for AIDS/HIV are 042, 79571, V08, and V6544; and for alcohol/drug use, 303-30593, 9445-9446, 9453-9454, 9461-9469, 9800, V6542, and V791.

[19] Specifically, in its Uniform Guideline on Employee Selection (29 C.F.R. §1607.4D), the EEOC considers as probative of discrimination when the promotion rate for one subpopulation is less than "eighty percent" of another, which translates to an odds ratio of 1.25.

References

Abowd JM, Schmutte IM (2019) An economic analysis of privacy protection and statistical accuracy as social choices. *Amer. Econom. Rev.* 109(1):171–202.

Acquisti A, Fong C (2020) An experiment in hiring discrimination via online social networks. *Management Sci.* 66(3):1005–1024.

Acquisti A, Brandimarte L, Loewenstein G (2015) Privacy and human behavior in the age of information. *Science* 347(6221):509–514.

Adler NE, Stead WW (2015) Patients in context: EHR capture of social and behavioral determinants of health. *Obstet. Gynecol. Survey* 70(6):388–390.

Agarwal S (2020) Trade-offs between fairness, interpretability, and privacy in machine learning. Master's thesis, University of Waterloo, Waterloo, Canada.

Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A (2005) Approximation algorithms for *k*-anonymity. *J. Privacy Tech.*

Agrawal R, Srikant R (2000) Privacy-preserving data mining. *SIGMOD'00 Proc. 2000 ACM SIGMOD Internat. Conf. Management Data* (Association for Computing Machinery, New York), 439–450.

Agresti A (2002) *Categorical Data Analysis*, 2nd ed. (John Wiley & Sons, New York).

Ali J (2018) Validating leaked passwords with k-anonymity. Accessed April 29, 2020, https://blog.cloudflare.com/validating-leaked-passwords-with-k-anonymity/.

Apple Inc. (2020) Differential privacy. Accessed April 29, 2020, https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

Article 29 Data Protection Working Party (2014) Opinion 05/2014 on anonymisation techniques. Accessed April 29, 2020, https://ec.europa.eu/justice/article-29/documentation/.

Bagdasaryan E, Poursaeed O, Shmatikov V (2019) Differential privacy has disparate impact on model accuracy. *Adv. Neural Inform. Processing Systems* 32:15479–15488.

Bambauer J, Muralidhar K, Sarathy R (2013) Fool's gold: An illustrated critique of differential privacy. *Vanderbilt J. Entertainment Tech. Law* 16:701–755.

Barnett A (1982) An underestimated threat to multiple regression analyses used in job discrimination cases. *Indust. Relations Law J.* 5(1):156–173.

Barnum P, Liden RC, DiTomaso N (1995) Double jeopardy for women and minorities: Pay differences with age. *Acad. Management J.* 38(3):863–880.

Bland JM, Altman DG (2000) Statistics notes. The odds ratio. *BMJ* 320(7247):1468.

Castilla EJ (2008) Gender, race, and meritocracy in organizational careers. *Amer. J. Sociol.* 113(6):1479–1526.

Chang KW, Prabhakaran V, Ordonez V (2019) Bias and fairness in natural language processing. Baldwin T, Carpuat M, eds. *Proc. 2019 Conf. Empirical Methods Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts* (Association for Computational Linguistics, Stroudsburg, PA).

Chaudhuri K, Monteleoni C, Sarwate AD (2011) Differentially private empirical risk minimization. *J. Machine Learn. Res.* 12:1069–1109.

Cohen A, Nissim K (2020) Toward formalizing the GDPR's notion of singling out. *Proc. Natl. Acad. Sci. USA* 117(15):8344–8352.

Cowgill B, Tucker CE (2020) Algorithmic fairness and economics. Preprint, revised September 25, 2020, http://dx.doi.org/10.2139/ssrn.3361280.

Cumming G (2009) Inference by eye: Reading the overlap of independent confidence intervals. *Statist. Med.* 28(2):205–220.

Danziger J, Ángel Armengol de la Hoz M, Li W, Komorowski M, Octávio Deliberato R, Rush BN, Mukamal KJ, Celi L, Badawi O (2020) Temporal trends in critical care outcomes in United States minority serving hospitals. *Amer. J. Respiratory Critical Care Med.* 201(6):681–687.

Du W, Teng Z, Zhu Z (2008) Privacy-MaxEnt: Integrating background knowledge in privacy quantification. *SIGMOD'08 Proc. 2008 ACM SIGMOD Internat. Conf. Management Data* (Association for Computing Machinery, New York), 459–472.

Dwork C, Mulligan DK (2013) It's not privacy, and it's not fair. *Stanford Law Rev. Online* 66:35–40.

Dwork C, Kohli N, Mulligan D (2019) Differential privacy in practice: Expose your epsilons! *J. Privacy Confidentiality* 9(2):1–22.

Dwork C, McSherry F, Nissim K, Smith A (2016) Calibrating noise to sensitivity in private data analysis. *J. Privacy Confidentiality* 7(3):17–51.

Dwork C, Smith A, Steinke T, Ullman J (2017) Exposed! A survey of attacks on private data. *Annual Rev. Statist. Appl.* 4:61–84.

Ekstrand MD, Joshaghani R, Mehrpouyan H (2018) Privacy for all: Ensuring fair and equitable privacy protections. *Proc. First Conf. Fairness, Accountability and Transparency, Proceedings of Machine Learning Research*, vol. 81 (Microtome Publishing, Brookline, MA), 35–47.

Éltetö Ö, Frigyes E (1968) New income inequality measures as efficient tools for causal analysis and planning. *Econometrica* 36(2):383–396.

Erlingsson Ú, Pihur V, Korolova A (2014) RAPPOR: Randomized aggregatable privacy-preserving ordinal response. *CCS'14 Proc. 2014 ACM SIGSAC Conf. Comput. Comm. Security* (Association for Computing Machinery, New York), 1054–1067.

Everett RS, Wojtkiewicz RA (2002) Difference, disparity, and race/ethnic bias in federal sentencing. *J. Quant. Criminol.* 18(2):189–211.

Finnish Social Science Data Archive (2020) Data management guidelines. Accessed April 29, 2020, https://www.fsd.tuni.fi/aineistonhallinta/en/.

Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* 80(1):27–38.

Foulds J, Geumlek J, Welling M, Chaudhuri K (2016) On the theory and practice of privacy-preserving Bayesian data analysis. Ihler A, Janzing D, eds. *UAI'16 Proc. 32nd Conf. Uncertainty Artificial Intelligence* (AUAI Press, Arlington, VA), 192–201.

Fung BC, Wang K, Chen R, Yu PS (2010) Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surveys* 42(4):1–53.

Ganta SR, Kasiviswanathan SP, Smith A (2008) Composition attacks and auxiliary information in data privacy. *KDD'08 Proc. ACM*

SIGKDD Internat. Conf. Knowledge Discovery Data Mining (Association for Computing Machinery, New York), 265–273.

Garaud MC (1990) Legal standards and statistical proof in Title VII litigation: In search of a coherent disparate impact model. *Univ. Pennsylvania Law Rev.* 139(2):455–503.

Gastwirth JL (2019) The role of statistical evidence in civil cases. *Annual Rev. Statist. Appl.* 7:39–60.

Goldwasser S, Micali S, Rackoff C (1989) The knowledge complexity of interactive proof systems. *SIAM J. Comput.* 18(1):186–208.

Gupta VK, Mortal SC, Silveri S, Sun M, Turban DB (2020) You're fired! Gender disparities in CEO dismissal. *J. Management* 46(4):560–582.

Hajian S, Domingo-Ferrer J, Monreale A, Pedreschi D, Giannotti F (2015) Discrimination- and privacy-aware patterns. *Data Mining Knowledge Discovery* 29(6):1733–1782.

Hay M, Machanavajjhala A, Miklau G, Chen Y, Zhang D (2016) Principled evaluation of differentially private algorithms using DPBench. *SIGMOD'16 Proc. 2016 ACM SIGMOD Internat. Conf. Management Data* (Association for Computing Machinery, New York), 139–154.

Hebert PL, Sisk JE, Howell EA (2008) When does a difference become a disparity? Conceptualizing racial and ethnic disparities in health. *Health Affairs* 27(2):374–382.

Holtzclaw Williams P (2011) Policy framework for rare disease health disparities. *Policy Polit. Nursing Practice* 12(2):114–118.

Huang Z, Du W, Chen B (2005) Deriving private information from randomized data. *SIGMOD'05 Proc. 2005 ACM SIGMOD Internat. Conf. Management Data* (Association for Computing Machinery), 37–48.

John LK, Loewenstein G, Acquisti A, Vosgerau J (2018) When and why randomized response techniques (fail to) elicit the truth. *Organ. Behav. Human Decision Processes* 148:101–123.

Johnson N, Near JP, Song D (2018) Toward practical differential privacy for sql queries. *Proc. VLDB Endowment* 11(5):526–539.

Kashid A, Kulkarni V, Patankar R (2015) Discrimination prevention using privacy preserving techniques. *Internat. J. Comput. Appl.* 120(1):45–49.

Kelley E, Moy E, Stryer D, Burstin H, Clancy C (2005) The national healthcare quality and disparities reports: An overview. *Med. Care* 43(3 Suppl.):I3–I8.

Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. *SIGMOD'11 Proc. 2011 ACM SIGMOD Internat. Conf. Management of Data* (Association for Computing Machinery, New York), 193–204.

King AG (2006) Gross statistical disparities as evidence of a pattern and practice of discrimination: Statistical vs. legal significance. *Labor Lawyer* 22(3):271–292.

Kleinberg J, Mullainathan S (2019) Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. *EC'19 Proc. 2019 ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 807–808.

Li XB, Sarkar S (2013) Class-restricted clustering and microperturbation for data privacy. *Management Sci.* 59(4):796–812.

Li N, Qardaji W, Su D (2012) On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. *ASIACCS'12 Proc. 7th ACM Sympos. Inform. Comput. Comm. Security* (Association for Computer Machinery, New York), 32–33.

Li C, Hay M, Miklau G, Wang Y (2014) A data-and workload-aware algorithm for range queries under differential privacy. *Proc. VLDB Endowment* 7(5):341–352.

Lipton Z, McAuley J, Chouldechova A (2018) Does mitigating ML's impact disparity require treatment disparity? *Adv. Neural Inform. Processing Systems* 31:8125–8135.

Macagnone M (2019) Efforts to safeguard census data could muddy federal data. *Government Tech.* (December 17), https://www.govtech.com/analytics/Efforts-to-Safeguard-Census-Data-Could-Muddy-Federal-Data.html.

Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007)ℓ-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowledge Discovery Data* 1(1):3.

May P, Garrido MM, Cassel JB, Morrison RS, Normand C (2016) Using length of stay to control for unobserved heterogeneity when estimating treatment effect on hospital costs with observational data: Issues of reliability, robustness, and usefulness. *Health Services Res.* 51(5):2020–2043.

Meyerson A, Williams R (2004) On the complexity of optimal *k*-anonymity. *PODS'04 Proc. Twenty-Third ACM SIGMOD-SIGACT-SIGART Sympos. Principles Database Systems* (Association for Computing Machinery, New York), 223–228.

Muralidhar K, Batra D, Kirs PJ (1995) Accessibility, security, and accuracy in statistical databases: The case for the multiplicative fixed data perturbation approach. *Management Sci.* 41(9):1549–1564.

Muralidhar K, Parsa R, Sarathy R (1999) A general additive data perturbation method for database security. *Management Sci.* 45(10):1399–1415.

National Research Council (2004) *Measuring Racial Discrimination* (National Academies Press, Washington, DC).

Nolan J (2003) *Stable Distributions: Models for Heavy-Tailed Data* (Birkhauser, New York).

Osborne C (1991) Statistical calibration: A review. *Internat. Statist. Rev.* 59(3):309–336.

Pager D, Shepherd H (2008) The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Rev. Sociol.* 34:181–209.

Pitoura E, Tsaparas P, Flouris G, Fundulaki I, Papadakos P, Abiteboul S, Weikum G (2018) On measuring bias in online information. *SIGMOD Rec.* 46(4):16–21.

Pujol D, McKenna R, Kuppam S, Hay M, Machanavajjhala A, Miklau G (2020) Fair decision making using privacy-protected data. *FAT*'20 Proc. 2020 Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 189–199.

Rambachan A, Kleinberg J, Mullainathan S, Ludwig J (2020) An economic approach to regulating algorithms. NBER Working Paper 27111, National Bureau of Economic Research, Cambridge, MA.

Rocher L, Hendrickx JM, De Montjoye YA (2019) Estimating the success of re-identifications in incomplete datasets using generative models. *Nation Commun.* 10(1):1–9.

Rosenbaum PR (2010) *Design of Observational Studies*, 1st ed., *Springer Series in Statistics* (Springer, New York).

Ruggieri S, Hajian S, Kamiran F, Zhang X (2014) Anti-discrimination analysis using privacy attack strategies. Calders T, Esposito F, Hüllermeier E, Meo R, eds. *Proc. Joint Eur. Conf. Machine Learning Knowledge Discovery Databases*, *Lecture Notes in Computer Science*, vol. 8725 (Springer, Berlin), 694–710.

Santos-Lozada AR, Howard JT, Verdery AM (2020) How differential privacy will affect our understanding of health disparities in the United States. *Proc. Natl. Acad. Sci. USA* 117(24):13405–13412.

Siegel PA, Hambrick DC (2005) Pay disparities within top management groups: Evidence of harmful effects on performance of high-technology firms. *Organ. Sci.* 16(3):259–274.

Steen P, Cherney B (1996) Evolution of analytical tools by Mediqual Systems, Inc. *Amer. J. Med. Qual.* 11(1):S15–S17.

Sweeney L (2000) Simple demographics often identify people uniquely. *Health* 671:1–34.

Sweeney L (2002a) Achieving *k*-anonymity privacy protection using generalization and suppression. *Internat. J. Uncertainty Fuzziness Knowledge-Based Systems* 10(05):571–588.

Sweeney L (2002b) *k*-anonymity: A model for protecting privacy. *Internat. J. Uncertainty Fuzziness Knowledge-Based Systems* 10(05):557–570.

Tang J, Korolova A, Bai X, Wang X, Wang X (2017) Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. Preprint, submitted September 8, https://arxiv.org/abs/1709.02753.

Templ M, Kowarik A, Meindl B (2015) Statistical disclosure control for micro-data using the R package sdcMicro. *J. Statist. Software* 67(1):1–36.

Texas Department of State Health Services (2019) Texas hospital inpatient discharge public use data file. Accessed April 29, 2020, https://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm.

Traub JF, Yemini Y, Woźniakowski H (1984) The statistical security of a statistical database. *ACM Trans. Database Systems* 9(4):672–679.

U.S. Agency for Healthcare Research and Quality (2018) Central distributor SID: Description of data elements. Accessed April 29, 2020, https://www.hcup-us.ahrq.gov/db/vars/siddistnote.jsp.

U.S. Department of Health and Human Services (2012) Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. Accessed April 29, 2020, https://www. hhs. gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

Wang YX (2018) Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. *Proc. 34th Conf. Uncertainty Artificial Intelligence* (AUAI Press, Arlington, VA), 93–103.

Xu H, Zhang N (2019) Privacy in health disparity research. *Med. Care* 57(Suppl. 2):S172–S175.

Zhang X, Pérez-Stable EJ, Bourne PE, Peprah E, Duru OK, Breen N, Berrigan D, et al. (2017) Big data science: Opportunities and challenges to address minority health and health disparities in the 21st century. *Ethnicity Disparities* 27(2):95–106.

Supplemental Materials

EC.1. Summary of Typical Data-Removal and Noise-Insertion Algorithms

Table EC.1 Examples of Data-Removal and Noise-Insertion Algorithms

	Examples
Data Removal	Rule-based suppression (e.g., Texas Department of State Health Services 2019)
	Generalization or suppression to prevent the unique identification of a record (e.g., k-anonymity; Sweeney 2002)
	Generalization or suppression to prevent the disclosure of sensitive attributes (e.g., ℓ-anonymity; Machanavajjhala et al. 2007; t-closeness; Li et al. 2007)
Noise Insertion	Output perturbation mechanisms for differential privacy, e.g., Laplace (Dwork et al. 2016), Gaussian (Dwork et al. 2014), and geometric (Ghosh et al. 2012) mechanism
	Randomized response mechanisms (e.g., Warner 1965), which may be designed to achieve local differential privacy (Kasiviswanathan et al. 2011)
	Synthetic data generation mechanisms, such as posterior sampling for differential privacy (e.g., Wang et al. 2015)
	Additive noise mechanisms for input data, with which the noise for different records may be independent (Agrawal and Srikant 2000) or correlated (Zhang et al. 2005)

Table EC.1 summarizes a number of well-known anonymization algorithms in the category of data removal and noise insertion, respectively. For data removal, the selection of information to be removed may be made according to a set of pre-specified rules, like those in the Texas procedure discussed earlier in the paper. Alternatively, the selection may be designed to prevent the unique identification of a record (e.g., to achieve k-anonymity; Sweeney 2002), or to prevent certain predetermined sensitive attribute(s) from being inferred from the released data (e.g., to achieve ℓ-diversity; Machanavajjhala et al. 2007). For noise insertion, the random noise could be applied to the input data - e.g., by randomly altering their values (e.g., Kasiviswanathan et al. 2011) or by inserting additive random noises into the input data (e.g., Agrawal and Srikant 2000). Alternatively, the noise could be applied to the information (e.g., statistics) being released from the dataset. Examples include the direction insertion of additive random noises into the output statistics (e.g., Dwork et al. 2016), and the use of posterior sampling to generate random outputs that satisfy the given privacy guarantee (e.g., Wang et al. 2015).

EC.2. Proof of Theorem 1

Theorem 1 *When the dataset has k records for each value combination of X_i and Z_i, and anonymity is achieved by replacing the values of Y_i in each anonymous group with their average, the standard error for the estimation of each regression coefficient (i.e., $\beta_0, \beta_1, \beta_2, \beta_3$) is reduced by a multiplicative factor of $1/\sqrt{k}$ after anonymization.*

Proof of Theorem 1. To separate the stochastic component in Equation 1 from the non-stochastic ones, we rewrite the linear model as

$$Y_i = \tau^\top W_i + \xi_i \tag{EC.1}$$

where $\tau^\top = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $W_i = (1, X_i, X_i Z_i, Z_i)$. The regression coefficients τ can be estimated using ordinary least square as[20] $\hat{\tau} = (\mathbf{W}^\top\mathbf{W})^{-1}(\mathbf{W}^\top\mathbf{Y})$. Let $\Omega = \boldsymbol{\xi}\boldsymbol{\xi}^\top$. The variance of the estimator $\hat{\tau}$ can be expressed as

$$\mathbb{V}_0(\hat{\tau}) = (\mathbf{W}^\top\mathbf{W})^{-1}\left(\mathbf{W}^\top\Omega\mathbf{W}\right)(\mathbf{W}^\top\mathbf{W})^{-1} = \sigma_\xi^2(\mathbf{W}^\top\mathbf{W})^{-1}, \tag{EC.2}$$

where σ_ξ is the standard deviation of the error term ξ_i, so $\sigma_\xi^2 = E(\xi_i^2)$ is the expected value of each element on the diagonal of Ω.

After applying the anonymization, the distribution of ξ_i changes. Specifically, it is now constant for the k records in each anonymous group. Let C_i be the anonymous-group ID of the i-th record. Unlike in Equation EC.2, the expected values of elements in $\Omega = \boldsymbol{\xi}\boldsymbol{\xi}^\top$ now vary:

$$E(\Omega_{ij}) = \begin{cases} \sigma_\xi^2/k & \text{if } C_i = C_j \\ 0, & \text{otherwise} \end{cases} \tag{EC.3}$$

Thus, the variance of the estimator $\hat{\tau}$ now becomes

$$\mathbb{V}(\hat{\tau}) = (\mathbf{W}^\top\mathbf{W})^{-1}\left(\sum_{c=1}^{n/k}\mathbf{W}_c^\top\Omega_c\mathbf{W}_c\right)(\mathbf{W}^\top\mathbf{W})^{-1} = \frac{\mathbb{V}_0(\hat{\tau})}{k}, \tag{EC.4}$$

where $\mathbf{W}_c$ and Ω_c are the sub-matrices of $\mathbf{W}$ and Ω corresponding to the records in the c-th anonymous group, respectively. The theorem directly follows from Equation EC.4. ■

EC.3. Proof and Extension of Theorem 2

Theorem 2. *When Y_i follows the exponential distribution $Y_i \sim \text{Exp}(\lambda)$, suppressing m out of n records according to the density of Y_i shifts the sample mean of Y_i by an expected value of*

$$E(\bar{Y} - \bar{Y}_0) = \frac{m\log n - \log m!}{\lambda n}, \tag{EC.5}$$

where $\bar{Y}$ and $\bar{Y}_0$ are the sample mean of Y_i before and after suppression, respectively.

Proof of Theorem 2. According to the results in order statistics, the h-th largest value of n samples of Y_i, denoted by $Y_{(n-h+1)}$, is the sum of $n-h+1$ independent exponential random variables with parameter λn, $\lambda(n-1)$, …, λh. Thus, its expected value taken over the randomness of the sample becomes

$$E(Y_{(n-h+1)}) = \frac{1}{\lambda}\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{h}\right) = \frac{\log n - \log h}{\lambda}. \tag{EC.6}$$

The theorem directly follows from summing up Equation EC.6 for $h = 1, \ldots, m$. ■

[20] Throughout this proof, we use bold symbols like **W** and **Y** to represent the matrix/vector representation of variables.

The qualitative finding from Theorem 2 readily extends to many distributions beyond the exponential distribution. For example, consider the Pareto distribution, a heavy-tailed distribution with a monotonic density function. If Y_i follows the Pareto distribution with scale x_m and shape α, then its log-transformation $\log(Y_i/x_\mathrm{m})$ must follow the exponential distribution $\mathrm{Exp}(1/\alpha)$. Given the concaveness of log, the ratio of change incurred on the sample mean is even larger than what was specified in Equation EC.5. More generally, any one-detailed distribution (e.g., the lognormal distribution, Weibull, and Lévy distributions) tends to have the suppressed records concentrating on one side of the distribution, leading to a systemic shift of the sample mean as indicated by Theorem 2. When the one tail is also heavy-tailed (i.e., when the distribution has an infinite moment generating function), then such a concentration is likely more skewed[21] than the case of exponential distribution in Theorem 2, leading to an even larger shift of the sample mean.

EC.4. Proof of Theorem 3

Theorem 3. *With any (ϵ, δ)-differentially private algorithm, when $Z_i \in \{0,1\}$, X_i follows an i.i.d. (univariate or multivariate) Gaussian distribution, and $Y_i = \beta_0 + \beta_1 X_i + \beta_3 Z_i + \xi_i$ ($\beta_3 > 0$), the statistical power of any disparity-through-separation test must satisfy*

$$\text{Power} \leq 1 - e^{\frac{-6\epsilon n|\beta_3|}{\sigma}}(1-\alpha) + \frac{4n\delta|\beta_3|}{\sigma} \tag{EC.7}$$

where α is the significance level for the disparity test, n is the number of records in the dataset, and σ is the standard deviations for $\beta_1 X_i + \xi_i$.

Proof of Theorem 3. For a given set of $D_i = (X_i, Y_i, Z_i)$, we consider a mapping of the dataset to a dataset $F(D_i)$ with equal number of records but only one variable for each record $F(D_i) = (\beta_1 X_i + \beta_3 + \xi_i)$. Since $F(D_i)$ can be derived from D_i based on the knowledge of constants β_0 and β_3, specifically as $F(D_i) = Y_i + \beta_3(1 - Z_i) - \beta_0$, if a noise-insertion algorithm is (ϵ, δ)-differentially private over the set of D_i, it must also have the same guarantee over $F(D_i)$. Without loss of generality, we normalize X_i to have mean 0. Since the mean of error term ξ_i is also 0, $F(D)$ follows a Gaussian distribution with $F(D) \sim \mathcal{N}(\beta_3, \sigma)$.

Now consider the sample mean of $F(D_i)$, denoted by Δ. By the design of the disparity test, a positive identification of disparity always entails $\Delta > 0$. Note that Δ follows a Gaussian distribution with $\Delta \sim \mathcal{N}(\beta_3, \sigma/\sqrt{n})$. For any given ($\epsilon$, δ)-differentially private algorithm, let I_α be the $(1-\alpha)$-level confidence interval for Δ when applying the algorithm. One can see that a

[21] Note that while the skewness of the exponential distribution is always 2, the skewness of a heavy-tailed distribution may be unbounded.

false negative occurs if $0 \in I_\alpha$. We use $\mathcal{P}(0 \in I_\alpha)$ to denote the probability of $0 \in I_\alpha$, with the understanding that Power $\leq 1 - \mathcal{P}(0 \in I_\alpha)$.

Now consider an alternative scenario where $F(D_i)$ is generated from a Gaussian distribution $\mathcal{N}(0, \sigma)$. We use $\mathcal{P}_0(0 \in I_\alpha)$ to denote the probability of $0 \in I_\alpha$ in this scenario, where I_α is again the $(1-\alpha)$-level confidence interval for the mean of $F(D)$. Since the population mean of $F(D_i)$ is now 0, by definition of the significance level, we have

$$\mathcal{P}_0(0 \in I_\alpha) \geq 1 - \alpha. \tag{EC.8}$$

Note that the total variation distance (Dunford and Schwartz 1958, Section III.1) between the two distributions of $F(D_i)$ is $d = |\beta_3|/\sigma$. Since the noise-insertion algorithm is (ϵ, δ)-differentially private over $F(D_i)$, a bounded difference on the distribution of $F(D_i)$ yields a bounded distance between $\mathcal{P}(0 \in I_\alpha)$ and $\mathcal{P}_0(0 \in I_\alpha)$. Specifically, using the "group privacy" notion (Hardt and Talwar 2010, Karwa and Vadhan 2018), we have

$$\mathcal{P}(0 \in I_\alpha) \geq e^{-6\epsilon n d}\mathcal{P}_0(0 \in I_\alpha) - 4n\delta d. \tag{EC.9}$$

Taking Inequality EC.8 into EC.9, we have the upper bound on power as stated in the theorem. ■

EC.5. Summary Statistics for the Inpatient Dataset

Table EC.2 depicts the summary statistics for the key variables in the inpatient dataset that were used in our empirical study. While most of the variables are self-explanatory, we would like to further elaborate on the data-generation process for SERV, the admission severeity indicator. When a patient is admitted to a hospital, the hospital's electronic medical record system uses a proprietary algorithm to predict in-hospital mortality based on the patient's clinical variables collected at the time of admission. This prediction is then discretized into the 5-point scale, 0 (No instability, $p < 0.1\%$), 1 (minimal instability, $0.1\% \leq p < 1.2\%$), 2 (moderate instability, $1.2\% \leq p < 5.8\%$), 3 (severe instability, $5.8\% \leq p < 50\%$), and 4 (maximal instability, $p \geq 50\%$), before being reported to the state-level database. Thus, SERV in the dataset is a discrete variable on a 5-point scale.

Table EC.2 Summary Statistics for the Inpatient Dataset

Variable	Levels	n	%	∑%
SERV	0	10475	2.1	2.1
	1	172360	35.4	37.5
	2	202427	41.6	79.1
	3	81980	16.8	96.0
	4	19682	4.0	100.0
	all	486924	100.0	
NONRES	N/A	100528	20.6	20.6
	I	365209	75.0	95.7
	N	20144	4.1	99.8
	Y	1043	0.2	100.0
	all	486924	100.0	
SEX	F	279251	57.4	57.4
	M	207668	42.6	100.0
	all	486919	100.0	
RACE	U	62839	12.9	12.9
	W	359010	73.7	86.6
	B	53115	10.9	97.6
	A	2293	0.5	98.0
	I	1436	0.3	98.3
	N	8178	1.7	100.0
	all	486871	100.0	
HISPAN	1	11491	2.4	2.4
	0	475429	97.6	100.0
	all	486920	100.0	
AGE	< 20	74596	15.3	15.3
	[20,40)	80691	16.6	31.9
	[40,60)	99991	20.5	52.4
	[60,80)	149332	30.7	83.1
	≥ 80	82306	16.9	100.0
	all	486916	100.0	

Variable	Levels	n	%	∑%
INS	0	9771	2.0	2.0
	1	477153	98.0	100.0
	all	486924	100.0	
CANCER	0	429644	88.2	88.2
	1	57280	11.8	100.0
	all	486924	100.0	
LOS	< 2	80660	16.6	16.6
	[2,4)	180808	37.1	53.7
	[4,6)	88976	18.3	72.0
	[6,8)	49018	10.1	82.0
	[8,10)	27924	5.7	87.8
	[10,20)	44174	9.1	96.8
	≥ 20	15364	3.2	100.0
	all	486924	100.0	

Note. SERV = Admission severity. NONRES = Non-responder indicator. HISPAN = Hispanic. INS = Insurance status. CANCER = Cancer history. LOS = Length of Stay (in days). In SERV: 0 = No clinical instability; 1 = minimal instability; 2 = moderate instability; 3 = severe instability; 4 = maximal instability. In NONRES: N/A = missing value; I = Ineligible for calculation; N = not a non-responder (i.e., responding to treatment); Y = non-responder. In SEX: F = Female; M = Male. In RACE: U = Unknown; W = White; B = Black; A = Asian or Pacific Island; I = Native American or Eskimo; N = Other. In INS: 0 = Self paid; 1 = Insurance is the primary payer. In CANCER: 0 = no history of cancer diagnosis; 1 = current or historic cancer diagnosis.

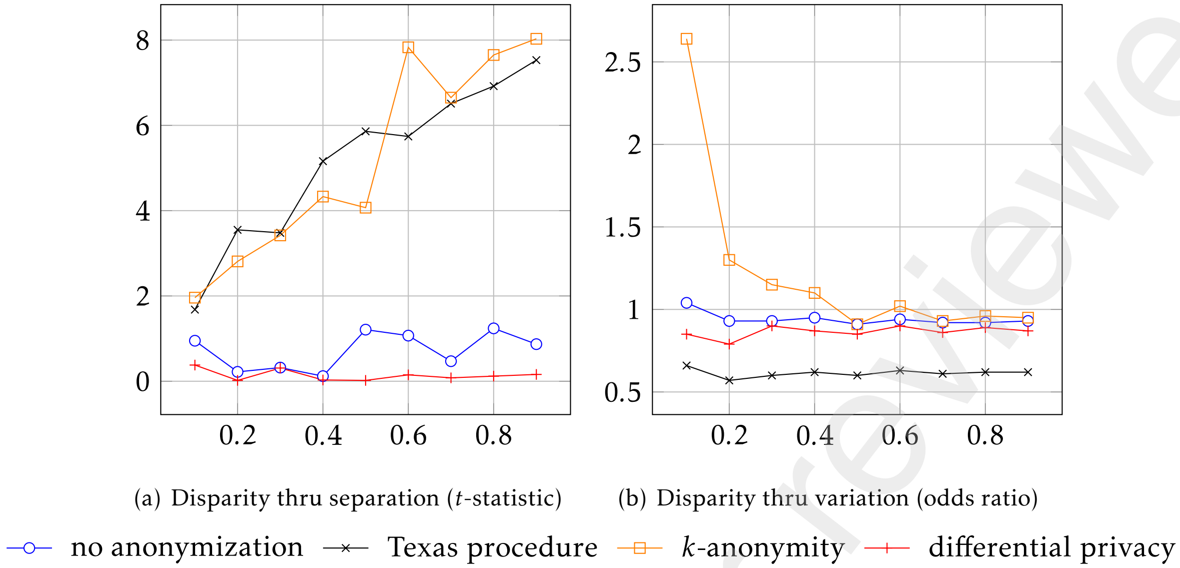
EC.6. Implications of Varying the Dataset Size

To evaluate the robustness of our findings, we examined how the outcome of disparity detection varied with the dataset size, specifically by sub-sampling the inpatient dataset (without replacement) with a sampling rate that ranged from 0.1 to 0.9. For disparity through separation, we tracked how the size affected the t-statistic of the regression coefficient corresponding to a social determinant. For disparity through variation, we tracked the change of the estimated odds ratio. As an example, Figure EC.1 depicts the results when the dependent variable was SERV and the social determinant being examined was HISPAN[22]. Two important observations emerged from the study:

First, when the input dataset grew larger, the data-removal mechanisms (i.e., Texas procedure and k-anonymity) exhibited distinct trends on the t-statistic (i.e., the statistical evidence

[22] Like in the earlier regression analyses, these results were obtained when controlling for AGE, RACE, SEX, INS, CANCER, and LOS. We also studied the same for NONRES and for the other social determinants. While the quantitative results necessarily differed, the qualitative trends and findings discussed later remained unchanged.

Figure EC.1 Change of Disparity Detection with Dataset Size



(a) Disparity thru separation (t-statistic)

(b) Disparity thru variation (odds ratio)

Note. Both subplots reflect the case where the outcome variable is SERV and the social determinant being examined is HISPAN, controlling for the other variables AGE, RACE, SEX, INS, CANCER, and LOS. The x axis is the sampling rate ([0.1, 0.9]) in both subplots. The y axis in (a) is the absolute value of the t-statistic for the regression coefficient of HISPAN. The y axis in (b) is the estimated odds ratio for HISPAN in (b). The parameter settings for the anonymization mechanisms are: $c = 10$ for the Texas procedure; $k = 5$ for k-anonymity; $\epsilon = 1$ for differential privacy.

for detecting disparity through separation) *vs.* the estimated odds ratio (i.e., for disparity through variation). While the t-statistic deviated further from the ground truth, the estimated odds ratio converged closer to it. This is consistent with the contrast between Theorems 1 and 2: While Theorem 2 suggests that the bias introduced by data removal to the estimated odds ratio tends to shrink with a larger dataset size n (specifically, by a factor of approximately $\log n/n$), Theorem 1 suggests that the amplification factor introduced by data removal to the t-statistic does not[23]. In other words, the extent to which data removal affects the outcome of disparity detection could be exacerbated by a larger input dataset if disparity is operationalized through separation (Theorem 1), but ameliorated if disparity is operationalized through variation (as Theorem 2).

Second, when disparity was operationalized through separation, data removal consistently overestimated the t-statistic, while noise insertion consistently underestimated it (under all input sizes). This further confirmed the qualitative findings discussed in Table 3. That is, with disparity through separation, data removal tends to produce more false positives than false negatives, while noise insertion likely produces false negatives only, with false positives being highly unlikely.

[23] More specifically, since Theorem 1 suggests a factor constant to n, an increasing n likely yields a larger t-statistic after data removal because the expected value of t over the original data grows approximately linearly with $\sqrt{n}$.

References

Agrawal R, Srikant R (2000) Privacy-preserving data mining. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 439–450.

Dunford N, Schwartz JT (1958) *Linear operators part I: general theory*, volume 1 (Interscience Publishers New York).

Dwork C, McSherry F, Nissim K, Smith A (2016) Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality* 7(3):17–51.

Dwork C, Roth A, et al. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9(3-4):211–407.

Ghosh A, Roughgarden T, Sundararajan M (2012) Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing* 41(6):1673–1693.

Hardt M, Talwar K (2010) On the geometry of differential privacy. *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, 705–714.

Karwa V, Vadhan S (2018) Finite sample differentially private confidence intervals. *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS)*.

Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A (2011) What can we learn privately? *SIAM Journal on Computing* 40(3):793–826.

Li N, Li T, Venkatasubramanian S (2007) t-closeness: Privacy beyond k-anonymity and ℓ-diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 106–115 (IEEE).

Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1(1):3–es.

Sweeney L (2002) k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.

Texas Department of State Health Services (2019) Texas hospital inpatient discharge public use data file. `https://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm`, accessed: 2020-04-29.

Wang YX, Fienberg S, Smola A (2015) Privacy for free: Posterior sampling and stochastic gradient monte carlo. *International Conference on Machine Learning*, 2493–2502.

Warner SL (1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60(309):63–69.

Zhang N, Wang S, Zhao W (2005) A new scheme on privacy-preserving data classification. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 374–383.