Identifying and addressing constraints to fair de-identification and data sharing

By

J. Thomas Brown

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Biomedical Informatics

May 10, 2024

Nashville, Tennessee

Approved:

Bradley A. Malin, Ph.D.

Ellen W. Clayton, M.D., J.D.

Michael Matheny, M.D., M.S., M.P.H.

Murat Kantarcioglu, Ph.D.

Zhijun Yin, Ph.D., M.S.

DEDICATION

To my wife.

ACKNOWLEDGMENTS

I would like to express my gratitude to my mentor Brad Malin for his guidance and mentorship throughout my PhD training. Brad has a unique ability to meet students where they are and propel them further than they imagined, and I was no exception. Every time I brought a new goal to Brad, he immediately and unreservedly committed to supporting me. He is also very precise when giving feedback. Constructive criticism tends to be difficult to receive or not so "constructive." Brad's approach is different. He continually challenged me in ways that both improved my thinking and increased my confidence – a gift I will always be grateful for. Finally, Brad is an infectiously enthusiastic teacher. I hope to emulate similar enthusiasm throughout my career.

I would also like to thank the members of my dissertation committee – Ellen Clayton, Michael Matheny, Murat Kantarcioglu, and Zhijun Yin – for their support throughout this process. I would not have finished in the time that I did without their encouragement and guidance. Their diverse expertise also broadened my perspective and greatly improved this work.

I am also thankful to the adminstrators of the Department of Biomedical Informatics at Vanderbilt University. I would like to thank Rischelle Jenkins, Cindy Gadd, Kim Unertl, and Jessica Ancker in particular. In addition to organizing a wonderful training program, they supported me as mentors and friends in every phase of my training.

I am also grateful to National Library of Medicine for providing me with funding throughout my tenure as a graduate student. The structure of the T15 grant provided me the flexibility to pursue the questions I found most interesting, which made graduate school more fun and more rewarding.

Finally, I am grateful to my wife and my two daughters. They are my motivation to excel, my reason to persist, and my loudest and proudest cheerleaders. They have also listened to me talk about privacy and AI more than anyone else – hopefully they like it because this is just the beginning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Overview

1.1 Introduction

As biomedical data sources continue to grow in size, complexity, and diversity, so does their potential to support research. Frequently collected for health care operations and delivery, such data can be reused to accelerate advancements in medicine, genomics, artificial intelligence, and public health[1]. However, realizing data's full potential depends on its availability. Even the best dataset produces no value if it cannot be analyzed. As such, calls for and efforts to share data continue to increase.

Yet sharing more data cannot come at the expense of an individual's right to privacy[2]. Biomedical data may contain patients' sensitive health information that could be misused. Therefore, data sharing must include appropriate privacy safeguards that are legally compliant, ethically justifiable, and technically feasible.

In the United States, federal and state regulations – such as the Common Rule[3], the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[4], and a rapidly growing collection of consumer data protection laws – such as the California Consumer Privacy Act (CCPA)[5] – aim to support data sharing while preventing privacy intrusions. These laws permit several avenues to share data, depending on whether the data subjects can reasonably be identified and the associated risks. First, to share individually identifiable information, researchers ideally would obtain data subjects' informed consent. While respecting subjects' autonomy, obtaining informed consent may be impractical, particularly for large datasets. Restricting the shared dataset to consented records may also bias representation, as consenters and non-consenters frequently differ on important demographic features[6,7]. In cases in which the impracticality of obtaining consent hinders beneficial research, the data is already available, and the risks to patients is sufficiently low, the Common Rule and HIPAA allow identified data to be used for research without consent when the governing institution review board (IRB) waives such requirement[3,8]. HIPAA also contains many exceptions permitting the disclosure of unconsented identified data – such as for public health activities, judicial proceedings, and law enforcement[8]. Nevertheless, such exceptions cannot support large-scale data sharing while minimizing the privacy risks. Second, HIPAA permits sharing a Limited data set without informed consent, in which particular attributes are removed and data users are required to sign a data use agreement[8]. However, the prescribed transformations do not rely on tailored privacy risk assessments, such that they

could inadvertently expose patients to privacy risks[9], and restricting access to those who sign a data use agreement limits the speed and breadth at which knowledge can be gained from the data. Finally, regulations permit sharing person-level data without obtaining informed consent and without data use agreements if there is little basis to believe the information can be used to identify an individual; that is, if the data is de-identified.[10] Even though de-identification requires obscuring information and cannot ensure individual anonymity[11], in that de-identification cannot guarantee an individual cannot be re-identified, data stewards increasingly turn to de-identification for sharing large datasets. De-identification provides legal flexibility – as de-identified data is not considered personally identifiable information or protected health information[8] – to facilitate broad data dissemination and access.

Nevertheless, de-identification's promise depends on its ability to balance privacy protections while supporting the data's intended use case. Complicating the process is the fact that data transformations that reduce subjects' identifiability also typically degrade data utility (or usefulness). And despite this privacy-utility tradeoff receiving considerable attention in the past decades – spurring the development of diverse privacy models, definitions, implementations, and optimizations[12] – new solutions may be required to meet emerging needs and support data-driven technological innovation.

De-identification's propensity for creating data utility disparities presents another challenge to supporting meaningful research. De-identification algorithms disproportionately distort more distinguishable records compared to less distinguishable records[13], where the difference in data utility can be so great that significant health disparities among minority populations may be masked[14,15] and algorithmic discrimination may be exacerbated[16,17]. Disparities in privacy risks may also exist, where the minority populations retain greater privacy risk than the majority population in the de-identified dataset[13,15]. While such inequities carry substantial ethical implications, the privacy research community has not prioritized addressing them. The constraints to achieving equitable de-identification are not well understood and de-identification methods that explicitly consider the distribution of both privacy risk and data utility between subgroups of patients have not been developed.

To better support equitable research in diverse applications, this dissertation identifies and addresses several limitations of traditional de-identification methods, as illustrated in Figure 1.1. The first research aim (shown in blue) focuses on developing and validating a de-identification method that can flex with a dynamic dataset, or a dataset that regularly accumulates records at a varying rate, while enabling timely updates. This aim was motivated by the need to make COVID-19 pandemic data publicly available to support a data-driven pandemic response. However, the experimental results highlight de-identification's

tendency to unequally preserve data utility between subgroups of patients. To better understand the unequal benefits afforded by current de-identification practices, the second research aim (shown in red) formalizes the data-based constraints to equalizing the privacy protections and data utility retention across patient subgroups. I discuss the implications of the fairness constraints and propose a data sharing model that relies on external deterrents to privacy intrusions to relax them. To complement the data sharing model, the third research aim (shown in green) develops a de-identification method that breaks transformation conventions and relaxes privacy guarantees to preserve minority subgroups' representation better than standard approaches.

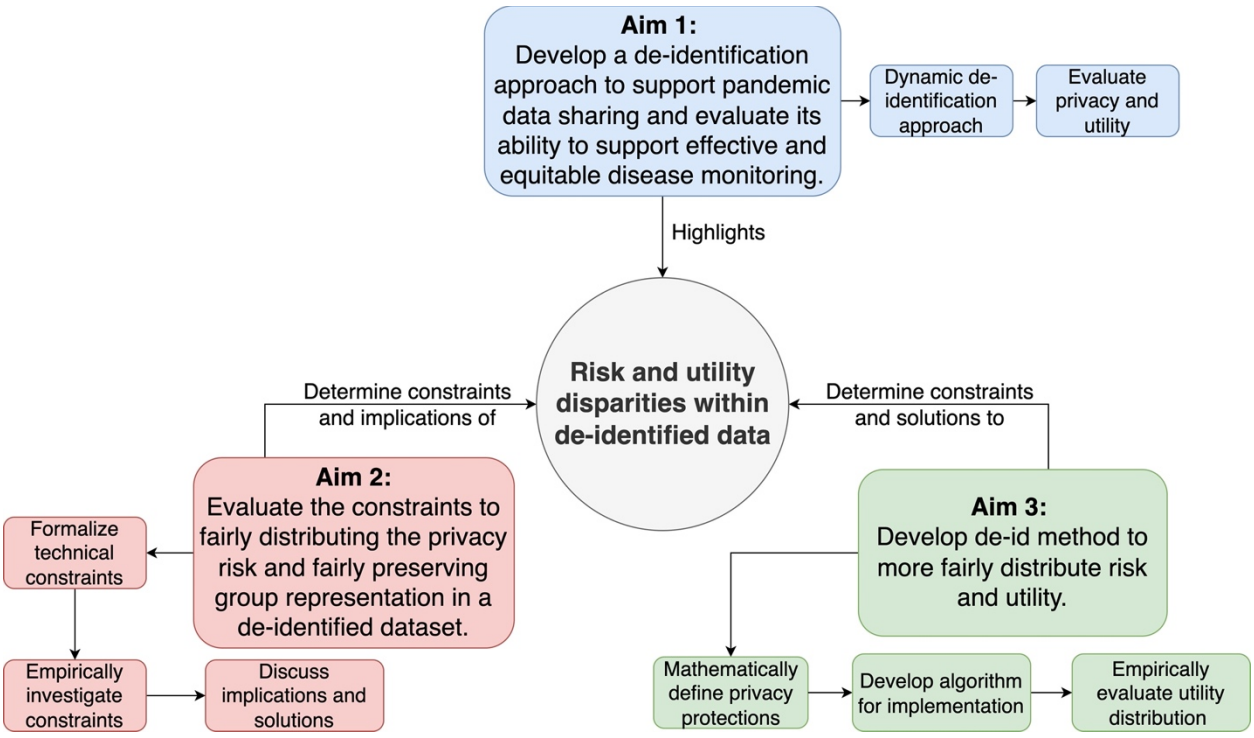

Figure 1.1. Overview of research aims.

I would like to note that alternative privacy enhancing technologies can also support data analytics while preserving patient privacy. Examples include secure multi-party computation and homomorphic encryption, which mitigate privacy intrusions by letting users compute over the data without giving users access to the data itself. While these technologies are gaining in popularity, they suffer from increased computational overhead and limit user's ability to conduct exploratory analyses, preventing the technologies from

replacing data sharing altogether. As such, in this dissertation, I focus on how to use de-identification and data sharing frameworks to support privacy-preserving data sharing.

1.2 Summary of contributions

To support data-driven pandemic responses as well as biomedical research's need for growing datasets, the first aim of this dissertation is to develop an approach to prospectively de-identify dynamic person-level datasets. Driven by a privacy risk estimation framework, the approach enables near-real time data sharing, while adjusting to the particular privacy risks of a given data sharing scenario and incorporating information priorities (i.e., prioritizing which variables should be preserved). Using a combination of real-world COVID-19 infection counts[18], United States (U.S.) Census statistics[19], and simulated data, I empirically validate the approach's ability to both decrease the re-identification risk[20] and preserve the evidence of underlying infection disparities (i.e., preserve data utility)[21] when publicly sharing a COVID-19 disease case registry. I also evaluate the fairness of privacy protections and disparity detection utility, where fair is considered equality between subgroups of data subjects. I show that standard de-identification methods as well as the dynamic de-identification method can unequally expose subgroups to re-identification risk and/or mask evidence of their infection disparities. Particularly, I look at racial inequalities as actual disparities in infection[22], hospitalization[23], and mortality[24] rates existed among racial minorities in the COVID-19 pandemic.

The second aim of this dissertation is to evaluate the constraints to simultaneously achieving the fair distribution of risk and utility across groups in a de-identified dataset. Expanding upon preliminary fairness investigations for alternative transformation strategies[16,25], I formally show that when records start with different re-identification risks, it becomes impossible for standard de-identification transformations to simultaneously equalize privacy risk and data utility. In fact, the unequal starting points with respect to risk imposes a tradeoff between achieving fair privacy and fair utility. Hence, I call the formalization the "fairness tradeoff theorem". I then illustrate how the constraints necessarily induce privacy risk and data utility inequalities between racial subgroups in the United States. The mathematical impossibility of achieving fairness across both risk and utility forces data stewards to choose between prioritizing equal privacy protections and equal representation when sharing de-identified data. While how to solve this dilemma merits broader discussion from researchers, community representatives, and policy makers, I propose an initial solution here. Specifically, I outline the conceptual design of a data sharing framework, called the passport-visa model, in which de-identification data transformations paired with sociotechnical

safeguards allow more equal representation in a de-identified dataset while facilitating access to trustworthy users.

While the passport-visa model relaxes the fairness constraints defined by the fairness tradeoff theorem, it relies on sociotechnical safeguards that can reduce data accessibility. To complement this data sharing model and relax data accessibility constraints, for the third aim of this dissertation, I develop a de-identification method, called Altruistic Masking, that focuses on improving minorities' representation in a de-identified dataset. This method is informed by the fairness tradeoff theorem, which makes a particular assumption derived from standard de-identification models and algorithms: that data transformations are deterministic. For example, all records corresponding to 20-year-old females would be transformed in the same manner by an algorithm. The deterministic constraint prevents potential cooperation between subgroups of records in a manner that more equally distributes the privacy and utility benefits of de-identification. As such, Altruistic Masking leverages non-deterministic transformations to allow the majority subgroups' records to contribute to the minority subgroups' privacy protections such that the minority subgroups retain greater data utility. Notably, however, this comes at the cost of certain privacy guarantees provided by standard de-identification models. I develop an algorithm to implement such an approach and show how the resulting data more equally preserves group representation and subsequently better supports outcome disparity detection compared to state-of-the-art de-identification methods.

1.3 Dissertation structure

The remainder of this dissertation is as follows. Chapter 2 reviews the related work. In Chapter 3, I develop and validate the dynamic de-identification approach. The contents expand upon the publications dedicated to the development of the approach[20] and to the evaluation of its ability to support infection disparity detection[21]. In Chapter 4, I investigate and define constraints to achieving fairness with respect to privacy risk and data utility in de-identified data and propose alternative data sharing strategies. Chapter 5 develops the Altruistic Masking de-identification method and evaluates the utility of the resulting data. Finally, in Chapter 6, I summarize the contributions of the dissertation, discuss their implications, and highlight future directions.

Chapter 2

Literature Review

2.1 Privacy legislation

HIPAA was initially designed to ensure the continuation of individuals' health insurance coverage between jobs. The U.S. law additionally sets standards for the electronic transfer of health information. Among those standards is the Privacy Rule, which was finalized in 2002 and implemented in April 2003. Intended to strike a balance between preserving patient privacy and supporting meaningful research, the Privacy Rule outlines standards to regulate "the use and disclosure of individuals' health information—called 'protected health information' by organizations subject to the Privacy Rule — called 'covered entities'"[8]. The Rule also specifies "standards for individuals' privacy rights to understand and control how their health information is used."[8] Covered entities include health plans, health care clearing houses, and health care providers. Since the passage of the HIPAA Omnibus Rule in 2013, the regulations additionally extend to covered entities' "business associates," or those who enter a contractual relationship as a business associate with a covered entity[26].

HIPAA provides for several approaches to share personal health information for research purposes. Identifiable health information can be shared if 1) the patients provide authorization to use and disclose their information or 2) an institutional review board (IRB) approves a waiver of individuals' authorization.

Alternatively, HIPAA permits sharing a Limited data set of person-level information without individual authorization. Instead, to share a Limited data set, the attributes outlined in Table 2.1 must be removed and the data recipient must sign a data use agreement prior to gaining access.

Finally, HIPAA permits sharing individual health information without individual authorization or a data use agreement when the data is de-identified. HIPAA does not consider de-identified data protected health information[8]. The de-identification standard may be achieved by one of two alternative implementations: Safe Harbor and Expert Determination. Safe Harbor requires the removal of an expanded set of identifiers, relative to the Limited data set standard (Table 2.1). In addition to removing these fields, for a dataset to meet the Safe Harbor standard, "the covered entity [must] not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of

the information."[27] Expert Determination affords a more flexible approach to de-identification, where instead of a list of fields to be removed, the data steward may determine health information is not individually identifiable if " [a] person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

- Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
- Documents the methods and results of the analysis that justify such determination"[27]

HIPAA's federal regulation preempts state privacy laws, unless state laws require more stringent privacy protection[28]. In other words, HIPAA provides a minimum standard for protection. In the U.S., a growing number of states have passed comprehensive privacy laws, including California, Virginia, Utah, Colorado, Connecticut, and Tennessee[5,29–32]. These laws grant consumers the opportunity to control certain aspects of their personal information collected by businesses, such as the right to access and delete such data maintained by certain businesses[33]. Notably, all state laws provide exemptions for data covered by HIPAA and permit the dissemination of de-identified data.

It could be argued that de-identifying data according to HIPAA's standard also satisfies the Common Rule's regulations, the primary regulations governing human subjects research in the U.S. Data subjects that are not identifiable in the data may be recategorized as "non-human" subjects under the Common Rule, and therefore the requirement to obtain human research subjects' informed consent may not apply[34].

Outside the U.S., the European Union's General Data Protection Regulation (GDPR) is arguably the most influential privacy legislation. Inspiring the legislation passed by several states, GDPR comprehensively regulates the storage and use of any data related to people in the European Union[35]. Similar to the U.S. laws permitting the dissemination of de-identified data, GDPR regulations do not apply to anonymous data. While de-identification as defined by HIPAA and anonymization as defined by GDPR both involve transforming data in a way that reduces the risk a data subject can be re-identified, GDPR defines anonymized data as that which has irreversibly rendered data subjects unidentifiable[36,37]. De-identification, on the other, only requires the re-identification risk to be "very small."[27] In this dissertation I focus on de-identification as defined by HIPAA for sharing person-level information.

Table 2.1. Suppressed attributes for Limited data set and Safe Harbor standards[27,38]

Suppressed Attribute	Limited data set	Safe Harbor
Names	X	X
Telephone number	X	X
Fax numbers	X	X
E-mail addresses	X	X
Social Security numbers	X	X
Medical record numbers	X	X
Health-plan beneficiary numbers	X	X
Account numbers	X	X
Certificate and license numbers	X	X
Vehicle identifiers and serial numbers, including license plate numbers	X	X
Device identifiers and serial numbers	X	X
Web Universal Resource Locators (URLs)	X	X
Internet Protocol (IP) address numbers	X	X
Biometric identifiers including fingerprints and voice prints	X	X
Full-face photographic images and any comparable image	X	X
Postal address information, other than town or city, State, and Zip code	X	
All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: A. The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and B. The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000		X
All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older		X
Any other unique identifying number, characteristic, or code, unless: A. The code or other means of record identification is not derived from or related to information about the individual and is not otherwise capable of being translated so as to identify the individual; and B. The covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.		X

2.2 Privacy risk

Individual privacy can be infringed upon through several types of disclosures[11]. First, identity disclosure occurs when a data recipient can re-identify an individual within a dataset. Second, attribute disclosure occurs when the recipient learns confidential information, such as HIV or cancer status, pertaining to a data subject. Learning such information may or may not require an accompanying identity disclosure. Third, membership disclosure occurs when the data recipient can determine an individual is a subject in the dataset, even if they do not know which record corresponds to the individual. This can also reveal potentially confidential information regarding the individual. For instance, if the dataset contains records of COVID-19 disease cases, then learning an individual resides in the dataset reveals they have had COVID-19.

Of the three main types of disclosures, the HIPAA Privacy Rule primarily regulates against identity disclosures, which can reveal the greatest amount of personal information. As such, de-identification via Expert Determination requires a re-identification risk assessment. The assessment measures the likelihood a data recipient can successfully re-identify data subjects, with respect to the recipient's assumed background knowledge and how that knowledge can exploit the distinguishability of individual records. The assessment informs how to develop and tune de-identification models to share useful data while minimizing the re-identification risk. The assessment, and subsequently the protections, can be extended to consider attribute and membership disclosures as well[39]. Though it is often assumed that an adversary has perfect background knowledge when designing data sharing policies, several studies have demonstrated the inherent difficulty to obtain such information[9,40]. In fact, Xia et al.[41] showed how such worst-case assumptions effectively overestimate the privacy risk. Thus, de-identification models need not provide perfect protection to reasonably mitigate the privacy risk.

The attributes present in the dataset vary in terms of the re-identification risk they pose. Directly identifying attributes, such as name or address, must be removed from the dataset entirely. Quasi-identifying attributes are those that may combine to distinguishably represent individual data subjects and can be found in external, identified datasets[42]. For example, it has been estimated that about 63% of the United States population can be uniquely represented by their combination of sex (Male/Female), 5-digit ZIP code, and full date of birth[43]. Were a bad actor to obtain a dataset containing people's names, sex, ZIP code, and date of birth (such as in a voter registration list[9]), or were they to know that information for a target individual, the bad actor could attempt re-identification on those quasi-identifying attributes. Figure 2 provides an example of such a re-identification attack. Finally, non-identifying attributes are those that can be shared without increasing an individual's re-identification risk.

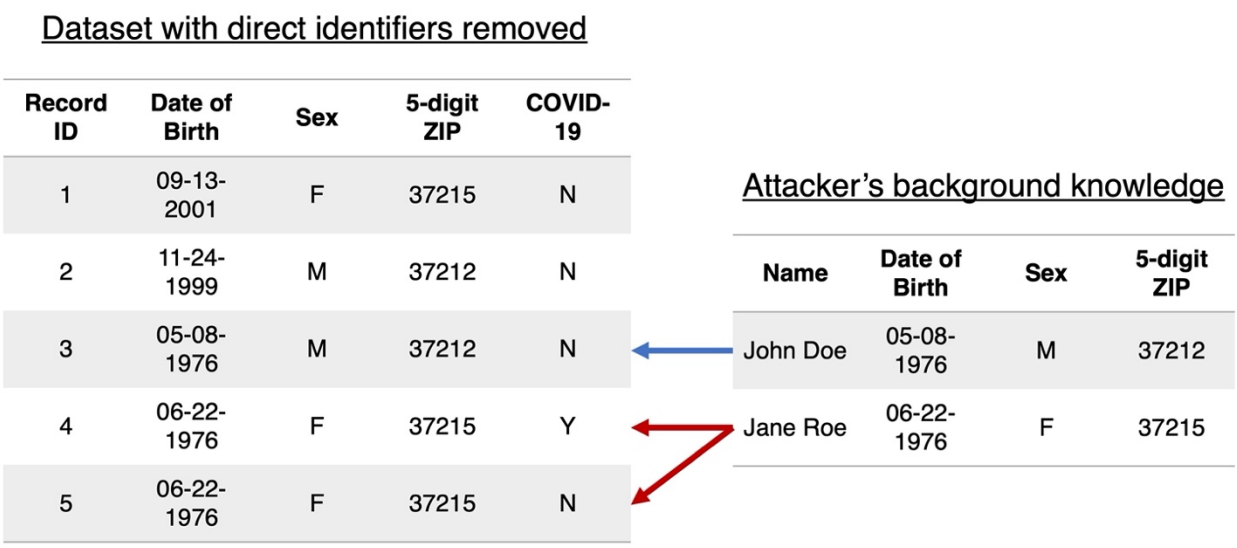

Figure 2.1. Example of a re-identification attack. (Left) Dataset wih direct identifiers, such as name, removed. (Right) Attacker's background knowledge, which includes the name, date of birth, sex, and 5-digit ZIP code for two individuals in the dataset. The unique combination of the quasi-identifying attributes {date of birth, sex, 5-digit ZIP} allows the attacker to re-identify record 3 as John Doe. The fact that records 4 and 5 share the same set of quasi-identifying attributes makes it more difficult for the attacker to correctly re-identify Jane Roe.

2.3 De-identification models

One of the more well-studied and applied de-identification models is *k*-anonymity[44]. The *k*-anonymity model is designed to mitigate re-identification of individual records by ensuring that each record is indistinguishable, in terms of their combination of quasi-identifying values, from at least $k-1$ other records. In other words, if we define the quasi-identifier as an individual's set of quasi-identifying feature values (e.g., age, race, county of residence) and define each group of records with the same quasi-identifier as an equivalence class, a *k*-anonymous dataset is one in which each equivalence class contains *k* or more records. As it has been shown that the combination of only a few quasi-identifying features can uniquely represent the majority of large population datasets[43,45,46], *k*-anonymity is often achieved by generalizing quasi-identifiers to coarser representations and/or suppressing quasi-identifiers corresponding to small equivalence classes[44,47]. The generalization options follow a hierarchical structure (see Figure 3.4), where moving up the hierarchy generalizes the information to increase privacy at the cost of utility[47].

A *k*-anonymous dataset guarantees that every record in a dataset falls into an equivalence class size of *k* or greater. This, in turn, guarantees the probability an adversary can re-identify any individual in the dataset

on the first attempt is less than or equal to $1/k$; and this upper bound holds against adversaries' of varying background knowledge. For instance, the worst-case scenario against a strong adversary assumes 1) the adversary knows a target individual's record is in the dataset, 2) the adversary knows all of the target individuals' quasi-identifying features, and 3) the target individual resides in an equivalence class of size k. Since the target individual's quasi-identifier looks like that of $k - 1$ other records, the probability the adversary re-identifies the individual on the first attempt is $1/k$. Alternatively, the adversary could link the quasi-identifiers between the shared dataset to an identified population register, such as a voter registration list[44,48]. This re-identification method is often referred to as a marketer attack in the privacy literature[48]. Here, each record's probability of being re-identified on the first attempt is one over the size of the equivalence class in the population. As the size of each equivalence class inside the dataset must be equal to or larger in size in the population, the probability an individual record is correctly re-identified on the first attempt is bounded to $1/k$. Regardless of the attack method, were an adversary to be able to repeat a re-identification attack, the probability of success may be greater than $1/k$ in subsequent attempts. Still, as shown by Xia et al.[49], the expected number of attempts required to correctly re-identify a patient is directly proportional to k.

Notwithstanding its intuitive approach to preserving patient privacy while sharing accurate information, k-anonymity has some notable drawbacks. Namely, k-anonymity is susceptible to homogeneity attacks and background knowledge attacks[39]. In such attacks, an adversary can still learn potentially sensitive information (e.g., cancer or HIV status) about a target individual without correctly re-identifying them. For example, if 1) the adversary knows the target individual's quasi-identifying features and 2) each record in the target individual's equivalence class is reported to have cancer, the adversary can infer the target individual must have cancer from the homogeneous distribution. The adversary may also possess sufficient background knowledge to correctly infer the target individual's sensitive attribute, even when the distribution of sensitive values within the equivalence class is non-homogenous. To alleviate such privacy disclosures, models such as l-diversity are applied in conjunction with k-anonymity to reduce the distinguishability and homogeneity of sensitive values[39].

An alternative to k-anonymity is statistical confidentiality methods, one of the most popular being the differential privacy model[50]. Instead of generalizing quasi-identifiers to make individual records less distinguishable, the model protects patient privacy by injecting a parameterized amount of noise into the data. Initially designed for sharing statistical aggregates, differential privacy provides formal privacy guarantees to every individual in a dataset. Namely, when an adversary queries a database, it is guaranteed the adversary cannot learn much more about any individual when the individual's data is included in the

query calculation than when the individual's data is not included. The difference in knowledge gained is controlled by a tunable parameter, ε.

A third alternative for privacy-preserving data sharing is synthetic data generation. Recently gaining in popularity, synthetic data generation involves synthesizing records that mimic the statistical properties of a dataset of interest that can be shared in place of the actual records. One of the growing collection of data synthesis methodologies is generative adversarial networks (GAN), in which one neural network model synthesizes fake records while the other discriminates between the real and fake records[51]. As the models compete, the synthesizing model improves performance until the fake records are relatively indistinguishable from the original records.

Each of the aforementioned privacy models has inherent weaknesses. At a cursory level, *k*-anonymity may excessively reduce the granularity of the data while potentially exposing individuals to sensitive attribute disclosure[39,52]. Differential privacy was initially designed for sharing statistical aggregates instead of person-level data, and injecting noise may not be appropriate for every data sharing scenario[53,54]. And while synthetic data is attractive for its aim to simulate the statistics of a dataset without sharing real patient records, its privacy risks are still not fully understood. Currently, membership and attribute disclosures are the most obvious concern, but there is still potential for the synthesizer to overfit to the real data and subsequently enable re-identification[55–57]. Nevertheless, in this work, I will focus on the *k*-anonymity model and its relaxed counterparts (i.e., marketer risk-guided generalization and suppression[48]) for several reasons. First, *k*-anonymity is commonly applied in practice to person-level datasets to the extent that federal and state legislation have established standard values of *k*[58–60]. Second, *k*-anonymity protects against identity disclosures more consistently than differential privacy and synthetic data generation[57,61], satisfying the HIPAA Privacy Rule's regulations.

2.4 Privacy-preserving data sharing architecture

Since de-identified data is not considered protected health information under HIPAA[8], it can legally be shared without constraint. De-identified data can even be published online such that it is openly accessible and downloadable. While proponents of the sharing data in the public domain emphasize its ability to facilitate research, increasing accessibility in this manner also potentially increases the risk the data will be misused. Bad actors have equal access to such data, and any successful re-identifications would likely go undetected (unless they publish the results in an attempt to discredit the organization that published the

data) and unpunished (as the privacy violation is generally attributed to the publishing organization for not properly de-identifying the data). As such, data sharing initiatives often supplement de-identification with sociotechnical mechanisms to deter and further mitigate the risk of re-identification[11,62]. Examples of such mechanisms include requiring users to sign a data use agreement, requiring users to obtain institutional representation, imposing a financial cost for access, and limiting users' access to the data to be within a monitored analytics environment with constraints on which operations a user can perform.

Modeled in several investigations by Wan et al.[63–65], the addition of sociotechnical mechanisms can effectively increase the penalty for data misuse in a manner that rational actors are disincentivized to attempt re-identification. The sociotechnical safeguards also allow for more granular and sensitive data to be shared, as exemplified by the difference between HIPAA's Limited data set and Safe Harbor requirements[27], while maintaining the re-identification risk below a tolerable threshold. Thus far, the combination of de-identified data with additional sociotechnical safeguards have effectively prevented re-identification[66] and have been adopted by several data sharing initiatives. One example is MIMIC, a freely-available database of de-identified electronic health record, medical imaging, and clinical notes[67–69]. Another is the National Institute of Health's (NIH) All of Us Research Program, which is currently developing a research platform built on de-identified electronic health record, survey, and genomic data from diverse patient populations[70]. However, controlling access reduces the overall accessibility of the resource and may produce access inequities[71]. As such, many data sharing initiatives create tiers of access to provide varied offerings that each maintain patient privacy while varying the tradeoff between data utility and data accessibility. For example, All of Us has three tiers: a public tier, at which any user can access aggregate counts of the database; a Registered Tier, at which approved researchers gain access to a curated dataset of person-level data; and a Controlled Tier, at which researchers who obtain additional approvals gain access to more detailed information than provided in the Registered Tier as well as to genomic data[72].

2.5 De-identification algorithms

There is an inherent tradeoff between patient privacy and data utility. Decreasing patient distinguishability requires distorting the raw data, but distortion degrades the retained information. As such, there has been a substantial amount of research in developing algorithms to minimize the distortion necessary to achieve *k*-anonymity. However, it has been shown that the problem of finding the minimal amount of generalization is an NP-hard problem. As such, heuristic methods are often used to approximate the global optimum.

k-anonymity algorithms generally make several assumptions. First, they assume that generalization options for each quasi-identifying feature follow a hierarchical pattern, where moving up the hierarchy increases privacy at the cost of utility. Second, they often assume that increased generalization degrades utility. Optimization typically involves an information-theoretic cost function, where generalization increases the information lost. The cost function may consider the number of levels up each generalization hierarchy are taken[47], or the divergence between the original data and the generalized data[6,73,74]. Third, the algorithms frequently assume all data records of a static dataset have been accumulated and are ready for dissemination. This assumption is particularly problematic for sharing infectious disease surveillance data or any other dynamic dataset with frequent updates, motivating the work in Chapter 3. Minimizing the generalization of the current version of a dataset may limit the data sharer's ability to share updated information in the future. Moreover, waiting to accumulate records before retrospectively designing the data-sharing policy delays publishing the updated dataset, limiting the public's situational awareness[75–77]. Finally, algorithms assume generalization and suppression transformations are deterministic, in that every record in an equivalence class prior to de-identification is transformed in the same manner by the de-identification algorithm. As I show in Chapter 4, this assumption constrains an algorithm's ability to distribute risk and utility more equally between subgroups of records. In Chapter 5, I show how rescinding this assumption provides the flexibility to improve fairness.

Some of the most influential *k*-anonymization algorithms include Sweeney's original Datafly algorithm[78], Sweeney's theoretical MinGen algorithm[47], Bayardo and Aggarwal's heuristic-based search algorithm[79], LeFevre's Mondrian algorithm[80], and El Emam and Dankar's optimal lattice anonymization (OLA) algorithm[6]. The Mondrian algorithm, frequently used as a standard by which new algorithms are compared, uses a greedy search to partition quasi-identifiers into groups with *k* or more records and approximate the optimal local recoding for the dataset[80]. Local recoding variably generalizes individual records' quasi-identifier within the dataset, whereas global recoding applies the same generalization to all records in the dataset. Instead of a greedy search, the OLA algorithm leverages the monotonicity of generalization hierarchies to search a lattice of potential generalizations and identify the globally optimal global recoding[6]. OLA can apply any monotonic information loss measure, can use the loss measure to preferentially preserve certain features according to user-defined information priorities, and supports suppression of complete records to maximize the granularity of the remaining records. Notwithstanding their ability to guarantee *k*-anonymity while minimally distorting the data compared to other algorithms, Mondrian and OLA were not designed for dynamic datasets. They do not account for how additional records may change the optimal generalization.

k-anonymization algorithms designed for dynamic datasets can be categorized into continuous data publishing or sequential data publishing applications. Continuous data publishing considers incrementally updating the shared dataset with the addition and deletion of records. Sequential data publishing involves sequentially sharing different subsets of attributes from the same underlying table. For continuous data publishing, Byun et al.[81] initially demonstrated the disclosures that can occur when repeatedly applying standard *k*-anonymity algorithms to a monotonically increasing dataset. Pei et al.[82] later introduced a method to *k*-anonymize such datasets, called "monotonic incremental anonymization." Similar to the Mondrian algorithm, the accumulating records increase the sizes of the equivalence classes until they can be split into more specific quasi-identifiers. For datasets where records are removed and deleted, Xiao and Tao developed the *m*-invariance model, which adds counterfeit records to achieve *k*-anonymity[83]. Several extensions of *m*-invariance have been proposed to consider more complex dynamic datasets[84,85]. For sequential data publishing, Wang and Fung developed an algorithm to maintain *k*-anonymity against potential join attacks via global recoding[86]. Shmueli et al.[87] extended this work with local recoding, while also allowing for the addition of records to the dataset. These continuous and sequential data publishing algorithms contribute to the theoretical foundations for achieving *k*-anonymity in dynamic datasets; however, they still require obtaining records prior to advising how to generalize the dataset. This can delay dataset updates and prevent long-term policy planning for maximum data utility. Moreover, these algorithms may be difficult to apply in practice.

Notably, no *k*-anonymity algorithm, for static or dynamic datasets, explicitly considers how the de-identification transformations may disproportionately mask or expose particular subgroups. Without considering fairness, the algorithms remain susceptible to inadvertently distributing privacy protections and/or utility in an unequal manner.

2.6 The distribution of risk and benefits

To incorporate the principle of fairness into data sharing, de-identification should explicitly consider both the distribution of the privacy risk across records as well as the extent to which de-identification algorithms mask individuals and groups. For example, Xu and Zhang demonstrated how, in a Pennsylvania inpatient dataset, non-White patients were more uniquely represented than White patients[15]. If the dataset was not de-identified, non-White patients would be exposed to greater re-identification risk. At the same time, they showed how applying the Texas de-identification procedure (a rule-based policy similar to Safe Harbor) induced substantially greater information loss among non-White patients than White patients. Furthermore,

the authors showed in a follow-up study how *k*-anonymity and differential privacy can mask disparities and/or create false disparities in dependent variables[14]. Their work implies an inherent tradeoff between the equitable distribution of the privacy risk and the equitable distribution of de-identification transformations.

The differential accuracy between group representation imposed by de-identification algorithms not only affects the evidence of health disparities; it can also influence the development of fair machine learning (ML) models. For example, consider the scenario in which the de-identification algorithm removes the protected attribute from the dataset. This would impose one of the initial approaches to mitigating algorithmic bias via data pre-processing, referred to as fairness through unawareness[88]. Without seeing it, the model should hypothetically be able to assign a prediction independent of the protected attribute. However, protected attributes may be highly correlated with other features and the outcome (e.g., ZIP code correlating with race and ethnicity), making it difficult to completely blind the algorithm to the protected attribute[89]. Kleinberg et al.[90] proved the mathematical limitations of fairness through unawareness, while showing how defining race-specific case-definition thresholds led to a more racially fair model than a racially unaware model. As such, algorithmic bias mitigation strategies that require access to the protected attribute hold greater promise to support algorithmic fairness. De-identification needs to make protected attributes available for health disparity research and ML model development.

Research investigating the tradeoffs between privacy, utility, and fairness in the context of *k*-anonymity have been limited. Outside this dissertation, Xu and Zhang's studies are the only to consider how generalization and suppression mask health disparities[14,15]. Recently, Wan et al.[63] investigated how to incorporate a fairness constraint with a game theory-driven de-identification algorithm. The results identified a tradeoff between achieving fair privacy protections and fair utility retention between groups. Chester et al.[91,92] also published two studies evaluating the effect of *k*-anonymity via generalization on developing fair ML models. They first investigated the interaction between *k*-anonymization, accuracy bias (unfairness), and inherent class imbalances in the dataset.[91] They then investigated how *k*-anonymization impacted the effect of resampling to mitigate ML model bias.[92] These studies provided evidence that *k*-anonymization can weaken the effect of resampling on improving model fairness, but the interaction is complex. There was no clear correlation, for instance, between the value of *k* and fairness in prediction performance. This is likely due to the variability in correlation between the outcome and the quasi-identifying attributes – it is possible they are not correlated at all. Moreover, if a quasi-identifying attribute is correlated with the outcome, it is still not guaranteed that a more granular representation of the attribute would improve model performance. Similar to other feature engineering, it could be that such an attribute needs to be generalized to a coarser representation for the model to capture its predictive signal. As such,

the impact of *k*-anonymity via generalization and suppression on ML prediction performance is likely to be nuanced and vary between datasets and applications. Nevertheless, the more de-identification can preserve data granularity and group representation, the more flexibility ML practitioners will have to develop accurate and fair predictive models.

In contrast to *k*-anonymity, there have been many studies investigating the impact of differential privacy on fairness, with a particular focus on algorithmic fairness[16]. Zhu et al.[93] and McGlinchey et al[94]. demonstrated how post-processing functions, such as non-negativity post-processing as applied to the U.S. Census data or histograms of populations counts within a dataset more generally, can induce additional bias into the data. The noise introduces skewed residual errors into the population counts, which Steed et al.[17] showed could lead to inequities in Census-guided funding initiatives. Pujol et al.[95] demonstrated that differentially private versions of the data can disproportionately impact smaller groups. One solution they proposed was to allocate additional privacy budget to groups at risk of disparate utility loss; however, without guaranteeing the initial privacy budget will be met. Otherwise, methods to mitigate the bias induced by differential privacy have focused on modifying the Differential Privacy Stochastic Gradient Descent (DP-SGD) framework[16]. For example, Xu et al.[96] proposed varying levels of privacy transformations across different protected groups to reduce accuracy disparities. However, improving fairness with respect to utility disproportionately distributed the privacy risk. Therefore, Tran et al. [97] proposed the addition of a fairness constraint to the DP-SGD framework to reduce excessive privacy risk differences across groups.

There have also been several studies investigating the fairness of synthetic data generation. Bhanot et al.[98] showed how a data synthesizer may produce data in which historically marginalized groups are under-represented. There have been several efforts to develop GAN-based synthesizers that mitigate bias in representation and ML model performance.[99,100] However, without standard benchmarks for measuring privacy of utility of synthetic data[57], it is unclear how improving fairness in representation may affect the fairness with respect to privacy protections. For example, Cheng et al.[25] explored how incorporating differential privacy into synthetic data generation, with the goal of improving privacy protections, led to less fair ML model performance. If increasing privacy protections reduced model fairness, it is likely the converse is true.

The insights provided from differential privacy and synthetic data generation are instructive, highlighting the tradeoffs between privacy risk, overall utility, fair privacy risk, and fair utility. However, the investigations have not sufficiently formalized the relationship between fairness with respect to privacy and fairness with respect to utility. Also, as mentioned above, the privacy risk as defined by differential privacy

may not directly correlate to re-identification risk as defined by HIPAA, and the privacy risks of synthetic data are still being worked out. The difference between how differential privacy, synthetic data generation, and generalization and suppression may induce bias into the data is also likely to be nuanced. And since generalization and suppression are so frequently applied to real-world biomedical data sharing initiatives, such de-identification methods merit further investigation.

2.7 Health disparities and data representation

The novel coronavirus 2019 (COVID-19) pandemic disproportionately impacted certain groups of people. McLaren[24] found racial and ethnic minorities to have disproportionately high COVID-19 mortality rates in Spring 2020. Rossen et al.[101] observed similar results comparing weekly, all-cause mortality rates in 2020 to those in 2015-2019. They calculated that the number of deaths of Hispanic-Latino individuals increased by 53.6% on average in 2020. American Indian/Alaskan Native (AI/AN) persons, Black persons, and Asian persons experienced 28.9%, 32.9%, and 36.6% average increases, respectively. The authors also found notable increases in deaths for age groups 25+, with a contrasting decrease (over 2%) in deaths for individuals less than 25 years of age. Levin et al.[102] discovered an exponential relationship between age and the infection fatality rate (IFR), where IFR for children under 10 was 0.002%, and 15% for individuals age 85 and older. Other studies found disparities in infection[22] and hospitalization rates[23,103] as well.

In several instances, disparities have been identified early enough to enable targeted interventions. The most common example is the state of Michigan, which found imbalanced infection and mortality rates between racial and ethnic groups early in the pandemic. The state responded by increasing testing resources and access to primary care physicians to minority subpopupulations[104,105]. Thanks in part to these measures, from April to November 2020, the percentage of COVID-19 cases in Michigan corresponding to African Americans dropped from 40.7% to 8%[106].

Yet, investigations into the differential impact of COVID-19 have been stifled limitations in the data. For instance, McLaren's study revealed that the disproportionate mortality rates in minority groups peaked by summer 2020 before dissipating by the end of fall. The study also found that adjusting for occupation, education, income, and poverty rates reduced the effect for Asian Americans, but not for other minorities. The disparities were evolving over the course of the pandemic; however, McLaren could not identify the source of the transient effects. Due to the unavailability of person-level demographic information, he had to rely on cumulative death counts by county and county-level demographic information[24]. The dearth of

publicly available COVID-19 data with racial and ethnic information is widespread. Gross et al.[107] found that only 28 states, and New York City, broke down COVID-19 mortality data by race and ethnicity. Only 8 states provided datasets with <5% missingness. The early detection of disparities by public health researchers, as well as retrospective investigations of their sources, requires the publication of more informative person-level COVID-19 data.

Though the work in this dissertation was initially motivated by the COVID-19 pandemic (see Chapter 3), disparities in health outcomes and data representation extend far beyond it. Minority groups – whether defined by race, ethnicity, gender identity, geographic location, educational attainment, or other demographic features – are often underrepresented in research initiatives. This has been due to differences in participation, recruitment, and consent; motivated, in part, by mistrust in medical research following unethical research practices[6,108–110]. The lack of diversity has subsequently limited the external validity of research findings and the benefit received among such populations[111,112]. Furthermore, underrepresented minorities have relatively poor health care access and suffer worse health outcomes relative to the majority groups[113,114]. Therefore, as representative data is critical to the pursuit of understanding health disparities and achieving health equity[115–117], initiatives such as the NIH's All of Us Research Program have sought to collect and share data on more diverse patient populations[70]. However, the question that remains and the question this dissertation investigates is: Can de-identified data sufficiently preserve minorities' representation to support such pursuits?

Chapter 3

Dynamically adjusting case-reporting policies

3.1 Introduction

The COVID-19 pandemic put a spotlight on infectious disease surveillance systems[118] and the importance of making such information widely accessible[119]. The data produced by these systems contains important information regarding who is infected and when they were diagnosed and may additionally include information regarding potential risk factors and outcomes. Such data can fuel a wide variety of public health research endeavors. For instance, the data can be used to model disease transmissibility and simulate potential interventions[76,120–122]. It can be used to identify the pandemic's disproportional impact on certain subpopulations and the sources of such disparities[24,123]. Furthermore, it can provide the public with situational awareness of outbreaks[75–77]. As such, an effective data-driven pandemic response depends on the accessibility of up-to-date infectious disease surveillance data.

Despite the rapid growth in the volume and diversity of epidemiological resources and the significant efforts to advance surveillance infrastructure during the pandemic[124,125], public data sharing on a wide scale remained limited[126]. As described Section 2.7, much of the publicly available data in the United States (U.S.) lacked important demographic information (e.g., race or ethnicity)[75]. The data that included such information were typically limited to aggregate counts at the state level[24,75,123]. Moreover, most of the initiatives that formed patient-level COVID-19 data repositories – such as the NIH's National COVID Cohort Collaborative (N3C)[127], the Datavant COVID-19 Research Database[128], the Centers for Disease Control and Prevention's (CDC) COVID-19 Case Surveillance datasets[129–131], and the Global.health data science initiative[132] – were not readily open to the public or did not include data shared in real time[124].

One of the primary factors that limited the public availability of person-level surveillance data with demographic information was concerns about an individual's right to privacy. Public health authorities often lacked the resources to de-identify the data in-house and thus, citing privacy concerns, refused to share data directly with researchers[133,134]. Hence, the reliance on sharing aggregated counts through dashboards.

Data sharing initiatives that had access to the person-level data and the resources to de-identify the data were also stifled by the rigidity of standard de-identification methods to publish useful data. One of the data

access tiers for N3C, for example, de-identified the data following Safe Harbor requirements[135]. However, Safe Harbor requires hiding epidemiologically critical factors, such as reducing the granularity of dates of events to their year[27], which renders such a policy useless for characterizing infectious disease transmission. As such, N3C created another access tier that shared a Limited data set, thus requiring, under HIPAA, that users first sign a data use agreement to gain access.[127] The alternative to Safe Harbor or sharing a Limited data set is to de-identify the data according to HIPAA's Expert Determination implementation. For example, the CDC followed Expert Determination by *k*-anonymizing its COVID-19 surveillance datasets[131]. However, for reasons described in Section 2.5, standard de-identification methods that follow Expert Determination do not consider the nuances of pandemic data sharing and thus hinder the ability to share de-identified data with maximal public health utility. The CDC's datasets, for example, were consistently generalized in the same manner, regardless of the opportunity to share more granular information, and were updated on a monthly or bi-monthly basis[131]. To meet the needs of pandemic sharing, a more tailored de-identification method must be developed.

In this chapter, I introduce an approach to adaptively generate policies to publicly share de-identified patient-level epidemiological data in near real-time. Here, a policy defines the level at which each quasi-identifying attribute is generalized. The approach is driven by a privacy risk estimation framework, which simulates disease cases to estimate the longitudinal privacy risk of sharing infected individuals' data under a given generalization policy, in the absence of actual patient data. The approach periodically adjusts the policy applied, according to the forecasted re-identification risk, to allow the data sharer to adapt data granularity according to the influx of new patient records while simultaneously allowing periods of consistent quasi-identifier representation. I specifically apply the framework to illustrate how policies could be developed to share COVID-19 patient health information against adversaries who attempt patient re-identification with varying levels of background knowledge.

The chapter is structured as follows. First, I describe the privacy risk estimation framework and the dynamic policy approach. I then evaluate the privacy protections afforded by the dynamic policy using real-world COVID-19 disease case counts. I then evaluate the fairness of the privacy protections the dynamic policy provides, in terms of the distribution of privacy risk across racial subpopulations. Next, I evaluate the utility of the data shared via the dynamic policy approach. Specifically, I determine how well the data enables the detection of disproportionately elevated infection rates within a specific subpopulation. Such COVID-19 disparities fluctuated longitudinally, emerging and dissipating as subpopulation outbreaks[23,24]. As such, the utility evaluation applies an outbreak detection algorithm to measure the timeliness and accuracy at which disparities can be detected. I then evaluate the fairness of detection performance, in terms of enabling

similar disparity detection times and accuracy between regions and subpopulations. For the evaluations, I compare several versions of the dynamic policy to policies resembling those applied to two publicly available COVID-19 datasets: 1) the CDC's COVID-19 Case Surveillance Public Use Data with Geography[129] and 2) the aggregated case counts that have been used in several disparity investigations[107]. Finally, I summarize the findings and describe future work.

While the work presented in this chapter was originally motivated by the need to publish COVID-19 data, it should be recognized the framework applies to any type of dynamic dataset. For example, it could apply to any type of epidemiological disease spread and be used to address emerging data sharing needs, such as for vaccine registries[136,137]. The framework could also help large data sharing initiatives develop a de-identification plan for how to adjust the generalization with the influx of new records as well as inform participant recruitment in order to optimize the resulting data's utility (i.e., "How many more data subjects from group *x* do I need to share data under policy *y* while still meeting the risk threshold of *z*?"*)*.

3.2 Dynamic policy approach

Due to the challenge of predicting exactly who will be infected, prospectively fixing a data sharing policy requires probabilistic risk assessment. The privacy risk estimation framework provides longitudinal privacy risk estimates for a data generalization policy within a specified geographic region. Given the appropriate population statistics, the framework can utilize any geographic level of detail (e.g., state, county, or ZIP code). In this chapter, I apply the framework to simulate disease spread on a county level to match the format of the COVID-19 surveillance data made accessible by the CDC[129,130].

3.2.1 Privacy risk estimation framework

Figure 3.1 summarizes the framework. In the first step, a data generalization policy is selected, which defines the generalization of each quasi-identifying feature considered. Here, I consider basic demographic features and the date of diagnosis as quasi-identifying features, shown in Table 3.1, as they are typical features organizations have been requested to share. The actual quasi-identifier depends on the adversary's assumed background knowledge and is a subset of the features presented in Table 3.1.

The second step generates the county-level population across the quasi-identifying features per the selected policy. I use population count data from the U.S. Census Bureau to calculate the number of people in the county that fall into each demographic group[19], where each group is defined by a unique combination of quasi-identifier values, excluding date of diagnosis.

The third step applies a Monte Carlo simulation (represented by the black box in Figure 3.1) to generate synthetic patient datasets using the county-level population distribution and a time series of new disease case counts. The time series' periodicity defines the frequency at which the updated dataset is released (e.g., every day or every week). To simulate the COVID-19 pandemic, I input time series derived from the Johns Hopkins COVID-19 tracking data[18]. The details of the simulation algorithm are presented in Section 3.2.3. The algorithm computes the re-identification risk on the patient set at each time point, according to a specified risk measure.

The fourth and final step of the framework uses the privacy risk distributions to estimate when the policy meets a privacy risk threshold. Computing the longitudinal privacy risk estimates under several data sharing policies for the same county identifies which policies likely meet the threshold at each point in the time series. The data sharer can then choose which policy to apply according to information priorities (e.g., prioritizing age granularity over sex granularity).

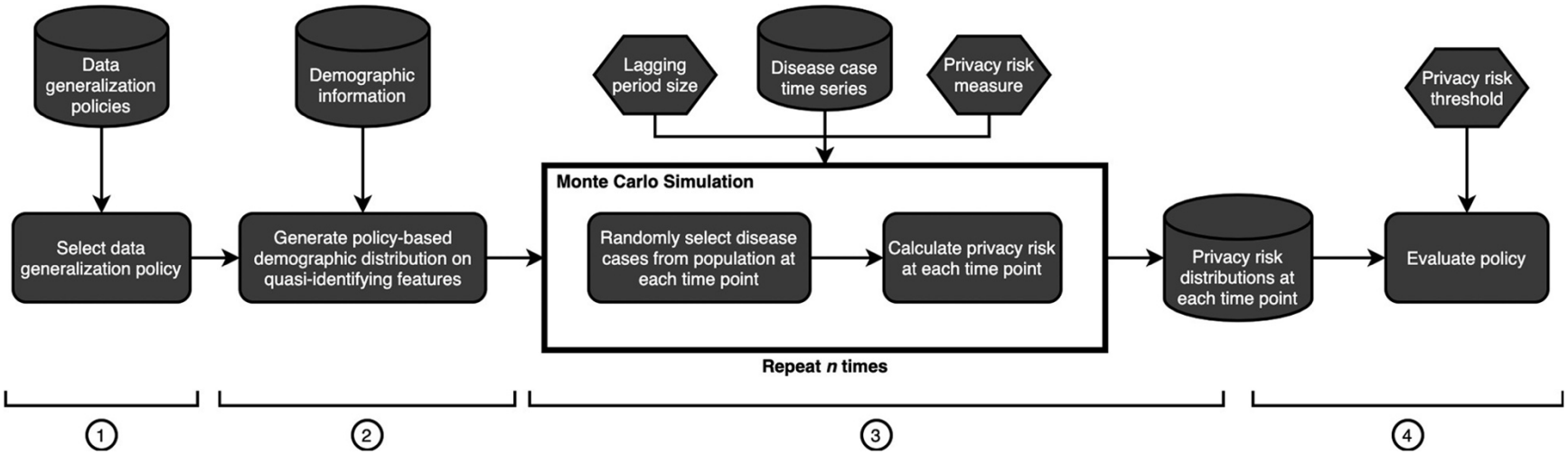

Figure 3.1. Privacy risk estimation framework. The curved rectangles represent processes, the cylinders represent data, and the hexagons represent user-defined parameters. The algorithm that performs the processes within the black box is in the core of the proposed framework and employs Monte Carlo random sampling. To obtain the privacy risk distributions, the simulation is repeated n times. The circled numbers denote the framework steps.

Table 3.1. The quasi-identifying features considered in this study. The middle column describes the generalization strategy for each feature. The third column provides an example generalization for each feature. In the case of sex and ethnicity, the information is either included or null. AIAN = American Indian/ Alaskan Native, and PI = Pacific Islander. (*These values cannot be generalized since I simulate on a county level. †This definition of a week is consistent with the one used by the CDC's COVID-19 case forecasts[138].)

Field	Generalization Strategy	Generalization Example
State of residence	None*	NA
County of residence	None*	NA
Date of diagnosis	Combine into week ranges (Sunday-Saturday†)	01/05/21 → 01/03/21 - 01/09/21
Year of birth	Convert to age ranges	1980 → 40-45 years old
Sex	Nullify value	Female → null, Male → null
Race	Combine race groups	AIAN → AIAN or PI, PI→ AIAN or PI
Ethnicity	Nullify value	Hispanic-Latino → null, Non-Hispanic → null

3.2.2 Privacy risk estimation

There are various methods for measuring the risk a recipient with certain background knowledge successfully re-identifies individual records[46,139]. In this chapter, I consider combinations of two privacy risk measures and two quasi-identifiers. The privacy risk measures are the PK risk and marketer risk (defined below). The two quasi-identifiers are {*state of residence, county of residence, year of birth, sex, race, and ethnicity}* and {*date of diagnosis, state of residence, county of residence, year of birth, sex, race, and ethnicity}*; i.e., they only differ in terms of date of diagnosis. As presented in publication [20], the risk evaluations in Sections 3.3-4 considers two attack scenarios: 1) PK risk with respect to the quasi-identifier

that includes date of diagnosis and 2) marketer risk with respect to the quasi-identifier that does not include date of diagnosis. As presented in publication [21], the utility evaluations in Sections 3.5-6 consider three attack scenarios: 1) PK risk with respect to the quasi-identifier that includes date of diagnosis, 2) PK risk with respect to the quasi-identifier that does not include date of diagnosis, and 3) marketer risk with respect to the quasi-identifier that does not include the date of diagnosis.

The PK risk is defined as the proportion of individuals in the dataset that fall into a group of size less than *k*, where each group is defined by a unique quasi-identifier value[140,141]. I evaluate this risk measure given a set of *k* values (5, 11, and 20) consistent with the standard thresholds used by public health authorities[58,60,142–144]. The PK risk assumes a data recipient knows 1) an individual is a member of the dataset and 2) the value of the target individual's quasi-identifier. In this scenario, the data recipient attempts re-identification to learn the target individual's sensitive information from additional features included in the dataset (e.g., comorbidities[145,146]). The more unique the record's representation, the more likely the data recipient can re-identify the individual[43,46]. I focus on this risk measure to follow the CDC's application of *k*-anonymization[147]. The PK risk effectively measures the proportion of records that fail to achieve *k*-anonymity for a given value of *k* and therefore allows the dynamic policy approach to identify generalization policies that are likely to meet *k*-anonymity.

In practice, obtaining such patients' quasi-identifying information is difficult[9,40]. Thus, evaluating the PK risk provides an upper bound of re-identification risk for the dataset. To demonstrate the approach's flexibility as well as to offer a different perspective on privacy risk, in which the adversary has different assumed background knowledge, the second privacy risk measure is the marketer risk[48]. The marketer risk is an amortization of the re-identification risk across all records, relaxes assumption (1), and considers the scenario in which the data recipient is motivated to re-identify as many patients as possible to learn who has the infectious disease of interest.

I highlight that, when applying the PK risk measure and assuming the adversary knows an individual's date of diagnosis (i.e., date of diagnosis is part of the quasi-identifier), I assume the adversary knows the diagnosis occurred within a lagging period of time (e.g., within one, three, or five days prior to the documented date). I allow this flexible assumption as it is unlikely a data recipient knows the targeted individual's exact diagnosis date[41], particularly when the time from a diagnostic test to case report extends beyond one day. The group corresponding to an individual contains all patients in the simulated patient set that match the individual on the demographic features, with a diagnosis date falling within the lagging period.

3.2.3 Framework algorithm inputs

The privacy risk estimation framework's Monte Carlo-driven core algorithm (denoted by the black box in Figure 3.1) calculates the privacy risk estimates from four inputs: 1) the county's demographic distribution, transformed according to the data generalization policy; 2) the time series of the number of new cases reported in the county, adjusted to match the generalization of date of diagnosis in the policy; 3) the size of the lagging period; and 4) the privacy risk measure.

The first input is the demographic distribution. The distribution defines the number of county residents that fall into each demographic group, where each group is defined by a unique quasi-identifier value (excluding the date of diagnosis). For example, assume a policy designates sharing state and county of residence, date of diagnosis, and 30-year age ranges. The input distribution is the number of people living in the county that fall into each 30-year age interval. I obtain the county distributions for the quasi-identifying features listed in Table 3.1 from the U.S. 2010 Census PCT12 tables[19].

Each PCT12 table contains joint statistics on age, sex, and county for a given Census-defined race[19]. The race values include White, Black, Asian, Native Hawaiian or Pacific Islander (NHPI), American Indian or Alaskan Native (AIAN), Mixed, and Some Other Race (referred to as "Other" hereafter). An additional table (PCT12H) provides joints statistics for Hispanic-Latino residents without race, while another (PCT12I) provides the joint statistics of non-Hispanic White residents. I calculate joint statistics for age, race, sex, ethnicity, and county of residence by first subtracting the PCT12I table from the White race table (PCT12A). The remainder is the number of White, Hispanic-Latino residents per race, sex, and county combination. I then subtract these statistics from the PCT12H table. The new remainder is the number of non-White, Hispanic-Latino residents. I distribute the non-White, Hispanic-Latino individuals among the remaining races proportional to the size of each racial group per age, sex, and county combination. For example, assume 15 people in Davidson County are non-White, 35 years old, and female. Further, assume 5 of the 15 residents are Asian and the other 10 are black or African American. Now, if there are 9 non-White, Hispanic, 35-year-old female residents in Davidson, I assign 3 of the 5 Asian residents and 6 of the 10 black or African American residents as Hispanic-Latino. Though this method may not accurately capture the true joint statistics of age, race, sex, and ethnicity per U.S. county, it provides a reasonable estimate for the framework. Distributing the Hispanic-Latino residents across all races spreads the county's demographic distribution more equally among demographic groups. Randomly sampling from a more

uniform distribution produces more conservative risk estimates as individuals are more likely to be uniquely represented in the simulated dataset[148]. The final joint statistics for age, race, sex, ethnicity, and county are used to define the demographic distributions for each county, where the counts are aggregated according to the generalization policy's specifications.

The second input is the time series, which defines the number of new disease cases reported, or the number of new records added to the dataset, per time period. The algorithm calculates the privacy risk at each time point in the time series. The time series periodicity defines the date of diagnosis generalization (e.g., date or week) and the dataset release schedule. I use the Johns Hopkins COVID-19 tracking data for COVID-19 disease case times series[18]. The Johns Hopkins data provides the cumulative number of COVID-19 cases diagnosed in each U.S. county on each day. Data preprocessing includes converting from cumulative counts to daily increases, and then setting all negative values to zero. To simulate the weekly release schedule, the preprocessed data is resampled into weekly periods (Sunday – Saturday).

The third input is the length of the lagging period. This value is a positive number that adjusts the privacy risk calculation according to the assumed knowledge of a data recipient regarding the date of diagnosis. For example, if new disease cases are not reported until five to seven days after obtaining the test sample, it is unlikely that the data recipient can know the exact date of diagnosis of an individual in the dataset. It would be more reasonable in such a case to set a 5-day lagging period, which suggests the data recipient knows at best the 5-day range in which the patient was diagnosed. A 1-day lagging period (equivalent to no lag) in this scenario would overestimate the data recipient's capabilities, inflate the privacy risk estimate, and potentially lead to unnecessary generalization of the data.

The final input is the privacy risk measure. Here, it is either the PK risk or marketer risk for a specified quasi-identifier. Different measures and quasi-identifiers consider different types of re-identification attacks. I show several variations in this chapter.

3.2.4 Privacy risk estimation framework algorithm

3.2.4.1 PK risk implementation

The algorithm follows the process described in Figure 3.2 to evaluate the PK risk. The algorithm first creates the uninfected population from the input demographic distribution, where each county resident is uniquely

represented by their demographic group (step 1). It then sums each value in *Cases* to obtain the total number of disease cases that will occur in the time series (2). The algorithm then applies Monte Carlo sampling to choose who gets "infected" from *UninfectedPop* and returns the list of individuals in random order (3). The sampling selects individuals without replacement, assuming equal weights across the entire uninfected population. Sampling one time without replacement prevents individual reinfection in the simulation. After initializing two lists (4 and 5), the algorithm enters a loop, which iterates for each value in the input time series (6). The first step within the loop removes the first *c* individuals from *InfectedPop*, counts how many of the individuals fall into each demographic group, and returns a vector of the results (7). The *NewCases* vector is added to a list of vectors from previous iterations, whose maximum size is the user-defined lag (8-11). To evaluate the PK risk under the lagging period assumption, the algorithm calculates the cell-wise sum of the vectors in *RecentCases* (12). The resulting vector, *CasesInPeriod*, represents the number of records for each unique quasi-identifier value in the dataset, whose date of diagnosis falls within the lagging period. The PK risk is then calculated on this final vector (13) and appended to the results (14) before proceeding to the next loop iteration.

The PK risk calculation is based on a formulation posed by Skinner and Elliott[141]. In the equation, let J denote the number of unique demographic groups allowed by the data generalization policy. Let f_j denote the number of records in demographic group j, for $j = 1, \dots, J$. Let $I(\cdot)$ denote the indicator function, where $I(A) = 1$ when A is true and $I(A) = 0$ otherwise. The PK risk is therefore

$$\frac{\sum_{k=1}^{K-1} \sum_{j=1}^{J} I(f_j = k) \cdot k}{n} \tag{3.1}$$

where n is the total number of records shared in the lagging period and K is the user-defined *k* value. The result is the proportion of the records shared in the lagging period that fall into a demographic group of size less than K.

Repeating the algorithm produces a distribution of risk outcomes at each point in the time series. The distribution can be analyzed for the expectation, the range, and confidence intervals of the privacy risk measure.

Algorithm 1: PK Risk Estimation

Input : *Demographics*, a list of the number of people per demographic bin in the county, where the bins are defined by the data generalization policy;
Cases, a list of the new daily or weekly disease case counts in the county;
lag, the length of the lagging period;
k, the specified k value for the PK risk calculation.

Output: *PKrisk*, a list of the PK risk values at each time point in *Cases*.

```
1  UninfectedPop ← createPopulation(Demographics)
2  nSick ← sum(Cases)
3  InfectedPop ← chooseInfected(nSick, UninfectedPop)        // This
     function Monte Carlo samples nSick individuals from
     UninfectedPop without replacement.
4  RecentCases ← [ ]
5  PKrisk ← [ ]
6  for c in Cases do
7  |  NewCases ← countPerBin(c,InfectedPop)        // This function
   |   removes the first c individuals from InfectedPop, and returns
   |   a vector of the number those individuals that fall into each
   |   demographic bin.
8  |  if length(RecentCases) = lag then
9  |  |  remove first vector from RecentCases
10 |  end if
11 |  RecentCases.append(NewCases)
12 |  CasesInPeriod ← cell-wise sum of the vectors in RecentCases
13 |  NewPKrisk ← calculatePKrisk(CasesInPeriod,k)
14 |  PKrisk.append(NewPKrisk)
15 end for
16 return PKrisk
```

Figure 3.2. PK risk estimation algorithm.

3.2.4.2 PK risk algorithm complexity

Here, I walk through the algorithm's worst-case time complexity. When each county citizen falls into their own demographic group, step 1 makes n executions, where n is the size of the county's population. Similarly, if every citizen is infected at some point in the time series, there are n Monte Carlo random sampling executions. Within the loop, when all the cases occur on the same time point, step 7 makes n executions. The remaining steps execute in constant time until the PK risk calculation in step 13. The PK risk calculation executes l times, where l equals $k - 1$ in Eqn. 3.1, for each non-empty demographic group. The value of l typically remains between 1 and 20. When the number of groups equals the number of citizens, there are ln executions made. The complexity for the loop, and subsequently the algorithm, is therefore $O(ln)$. Repeating the algorithm for m simulations increases the complexity to $O(mln)$. The

number of simulations, m, is typically on the order of 1,000. Since most US counties possess more than 1,000 residents (and may exceed 1,000,000), n dominates the time complexity.

3.2.4.3 Marketer risk implementation

The marketer risk considers a different attack scenario than the PK risk, where the data recipient attempts to re-identify as many individuals in the shared dataset as possible by matching the quasi-identifier values in the shared dataset to those in a separate, identified dataset. A common example of the latter is a voter registration list[9,46]. Not every county resident registers to vote, but for simplicity, I assume in this analysis the data recipient possesses an identified dataset containing every county resident. This assumption models the worst-case scenario, in terms of the completeness of background knowledge from an identified population dataset in the context of a marketer attack. I further assume the dataset contains all demographic information listed in Table 3.1, except for the date of diagnosis, such that the quasi-identifier is defined as {*state of residence, county of residence, year of birth, sex, race, and ethnicity}*. Excluding the date of diagnosis better approximates the information provided by a voter registration list.

Estimating the marketer risk requires a few adjustments to the PK risk estimation algorithm that considers date of diagnosis part of the quasi-identifier. First, the marketer risk is evaluated on the cumulative dataset at each time point as date of diagnosis is not considered quasi-identifying, and therefore no longer separates records into quasi-identifying windows of time. Without the date of diagnosis, the user does not specify a lagging period size. Neither does the user specify a k value, as the marketer risk measure incorporates all k values. Figure 3.3 describes the complete marketer risk estimation algorithm.

The first three steps of the marketer risk estimation algorithm are identical to the first three steps step in the PK risk estimation algorithm. The algorithm first creates uninfected population from the input demographic distribution (step 1), obtains the total number of disease cases in the time series (2), and applies Monte Carlo random sampling to select who gets "infected" and returns the list of individuals in random order (3). The sampling is performed without replacement assuming equal weights across the entire uninfected population. Individual reinfection is again prevented. The algorithm maintains the total number of disease cases, or records, per demographic group in *AllCases*. The vector is initialized to all zeros (4). After initializing the marketer risk results list (5), the algorithm enters a loop, which iterates for each value in the input time series (6). The first step in the loop removes the first c individuals from *InfectedPop* and returns of vector of the new cases' distribution across the demographic groups (7). The order of the *NewCases* vector matches the order of *AllCases*. To evaluate the marketer risk on the cumulative dataset up to the time

point corresponding to c, the algorithm adds the new cases to vector of previously reported cases (8). The resulting vector represents the number of records for each unique combination of quasi-identifier values in the cumulative dataset. The algorithm calculates the marketer risk on the updated *AllCases* vector (9).

Algorithm 2: Marketer Risk Estimation

Input : $Demographics$, a list of the number of people per demographic bin in the county, where the bins are defined by the data generalization policy;
$Cases$, a list of the new daily or weekly disease case counts in the county.

Output: $MarketerRisk$, a list of the marketer risk values at each time point in $Cases$.

```
UninfectedPop ← createPopulation(Demographics)
nSick ← sum(Cases)
InfectedPop ← chooseInfected(nSick, UninfectedPop)        // This function Monte Carlo samples nSick individuals from UninfectedPop without replacement.
AllCases ← zero vector of the same dimension as Demographics
MarketerRisk ← [ ]
for c in Cases do
    NewCases ← countPerBin(c,InfectedPop)        // This function removes the first c individuals from InfectedPop, and returns a vector of the number those individuals that fall into each demographic bin.
    AllCases ← AllCases + NewCases
    NewMarketerRisk ← calculateMarketerRisk(AllCases)
    MarketerRisk.append(NewMarketerRisk)
end for
return MarketerRisk
```

Figure 3.3. Marketer risk estimation algorithm.

The marketer risk is calculated following the formulation from Dankar and El Emam[48]. In Eqn. 3.2, J represents the number of unique demographic groups allowed by the data sharing policy. f_j represents the number of records in demographic group j in the shared dataset, for $j = 1, \dots, J$. F_j represents the number of records in demographic group j in the identified dataset, for $j = 1, \dots, J$. It follows that $\frac{f_j}{F_j}$ represents the expected proportion of correct matches between records in the shared and the identified datasets for demographic group j. n represents the total number of records in the shared dataset.

$$\frac{\sum_{j=1}^{J} \frac{f_j}{F_j}}{n} \tag{3.2}$$

The result is the expected proportion of the records in the shared dataset correctly matched to records in the identified dataset. The marketer risk value is appended to the list of marketer risk values at the end of each loop (10).

3.2.4.4 Marketer risk algorithm complexity

The marketer risk algorithm's worst case time complexity follows that of the PK risk algorithm until the marketer risk calculation in step 9. The calculation executes one time for each non-empty demographic group. When the number of groups equals the number of citizens, there are n executions made. Therefore, the complexity for m simulations of the algorithm is to $O(mn)$, where n is the size of the county's population.

3.2.5 Dynamic policy search

3.2.5.1 PK risk

To dynamically adapt policies according to an expected infection rate, I identify policies that are likely to satisfy a specific PK risk threshold at varying volumes of new case records. For the risk evaluation, I choose a k of 11, which is as a typical group size incorporated into guidance issued at the state[58,60,143,144] and federal[142] level. It is also the group size applied to CDC's COVID-19 Public Use Data with Geography[129]. I henceforth refer to the PK risk when k is equal to 11 as the PK11 risk. I search for policies that meet a PK11 threshold of 0.01; i.e., the percentage of records falling into a demographic group of size 10 or smaller should be less than or equal to 1%. I summarize the search results in Figure 3.5. Similar investigations for k of 5 and 20 (other common group size thresholds) are also provided in Figures 3.6 and 3.7, respectively.

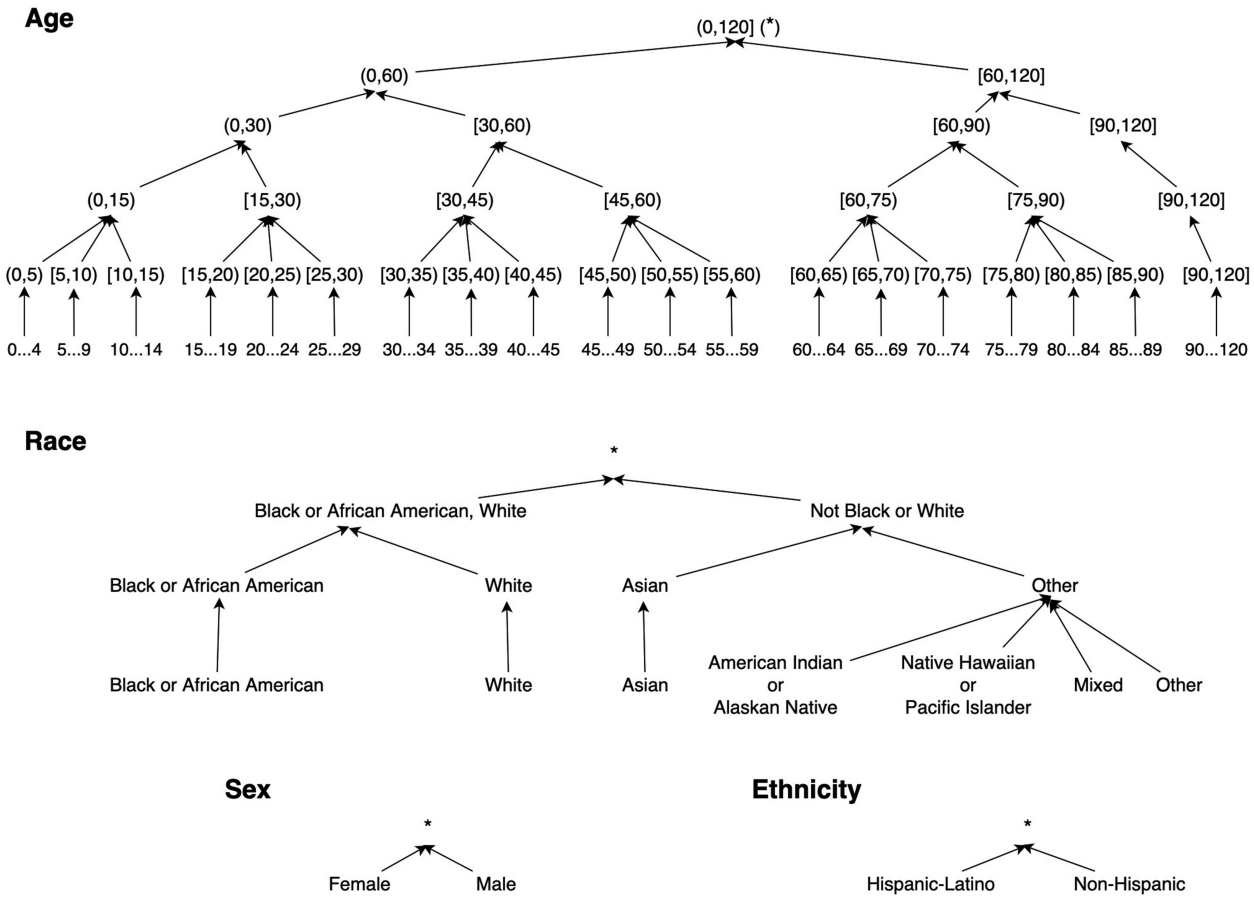

Figure 3.4. The generalization hierarchies for age, race, sex, and ethnicity used in this paper, adapted from those of Wan et al[64]. Each horizontal level is a potential generalization state for the data generalization policy. For example, the policy could specify generalizing age to 5-year age intervals to 15-year age intervals, or broader ranges. I represent year of birth as 1-year age at the bottom of the Age hierarchy. Moving up the hierarchies, the data becomes more generalized to increase privacy. An asterisk indicates the feature is generalized to a null value for all individuals, which is equivalent to suppression or non-release of the corresponding field.

The search uses the privacy risk estimation framework to evaluate 96 alternative data sharing policies for each U.S. county (with available census tract information) across a range of case count values. The policies include six potential generalizations of age, four generalizations of race, two generalizations of sex, and two generalizations of ethnicity, specified by the generalization hierarchies shown in Figure 3.4. For each policy, county, and case number combination, the framework generates 1,000 PK risk estimates. A policy meets the threshold when the upper bound of the estimates' 95% quantile range is less than or equal to 0.01. I choose to evaluate a policy in this manner to increase the likelihood supported policies meet the privacy

risk threshold in application. Note, the data sharer can adjust the size of the quantile range to modify the confidence a policy will meet a specific privacy risk threshold.

To aid in readability of the PK11 policy search results in Figure 3.5, I represent the generalization of each quasi-identifier in a policy with a four-character alphanumeric code. From left to right, the characters represent the age, race, sex, and ethnicity generalizations. I further summarize the results by categorizing US counties by population size.

Once a generalization policy meets the PK11 threshold for a given number of cases, it is unlikely records fall into a demographic group of size 10 or less. Further increasing the case volume increases the number of records in each group and decreases the PK11 value. As such, a policy is listed under the smallest case quantity at which the policy meets the PK11 threshold for every county in the category. It should also be noted there exists a parent-child relationship between policies. For example, policy 2*** is the parent of policy 3***, where the former only differs from the latter by generalizing age to a lesser degree. When a parent policy meets the PK11 threshold, all its child policies also meet the threshold.

As Figure 3.5 displays, the number of acceptable policies increases with the number of new cases. In most cases, larger counties achieve more acceptable policies than smaller counties at a given case quantity. The maximum number of acceptable policies is 73. The most granular policies across all county categories are 1C*e, 2Bse, and 3Ase. Each of these policies prioritizes different types of information. Policy 1C*e offers the most granular age information at the cost of race and sex information, while Policy 3Ase reduces age granularity to increase race and sex specificity.

The case number values are window-size agnostic, such that the policy search results hold regardless of the time period considered. For example, assume a county with fewer than 1,000 residents updates its disease surveillance dataset daily. Further, assume the county adjusts for sets a 5-day lagging period assumption. When the expected number of new cases from the current day and the previous two days sum to 50, the current day's records should be generalized according to either policy **** or **s*. The same policies are supported if, instead, the dataset is updated weekly (and diagnosis date is generalized to week of diagnosis) and 50 new cases are expected for the current week.

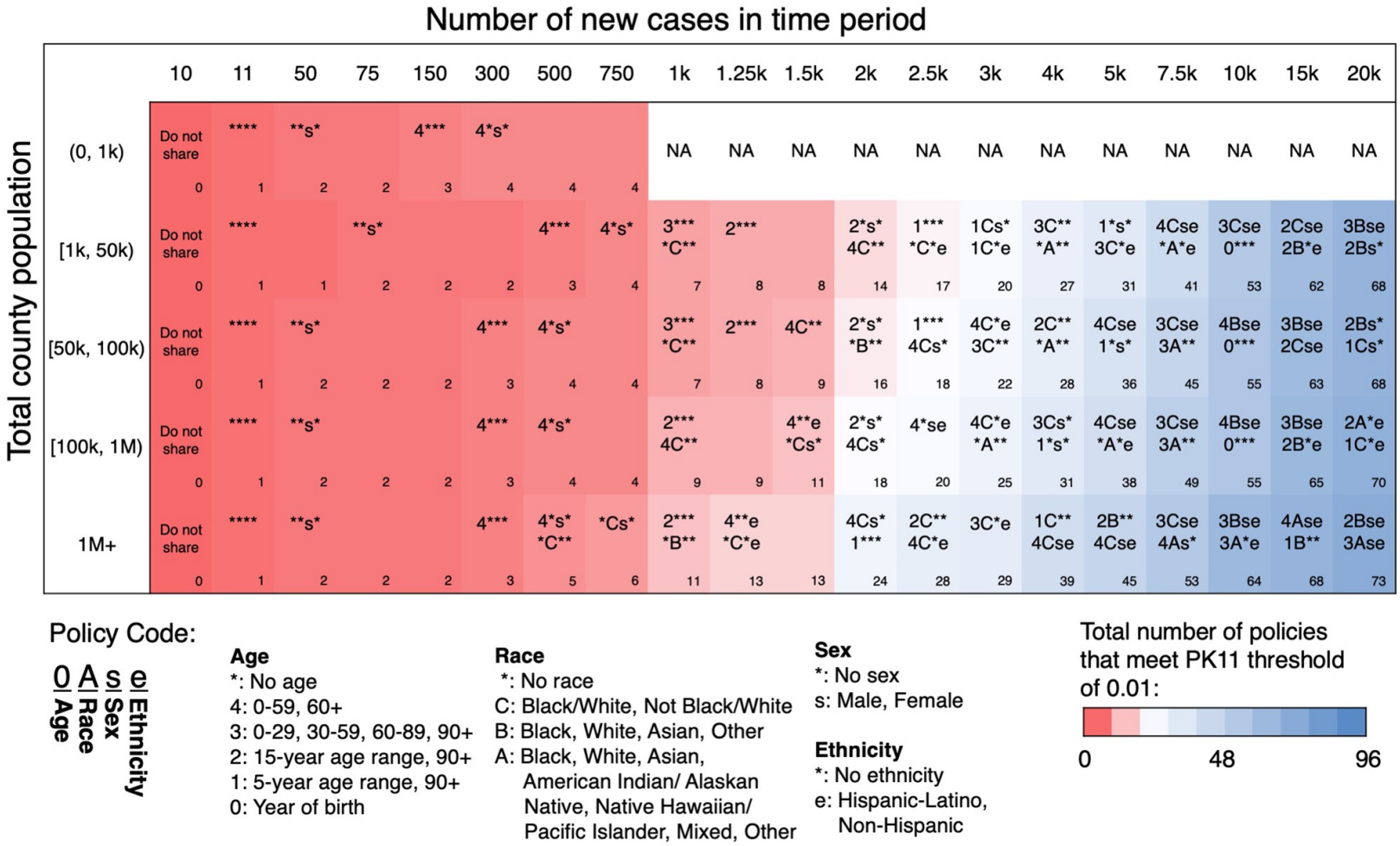

Figure 3.5. Generalization policies with a PK11 upper bound (calculated as the upper bound of the 95% quantile range of 1,000 framework simulations) less than or equal to 0.01 at varying disease case volume thresholds. A four-character alphanumeric code indicates the policy's generalization levels. All policies additionally include state and county of residence and some generalization of diagnosis date. A policy is eligible to be listed under the minimum number of new cases (table column) at which it meets the PK11 threshold for every county in the category (table row). A maximum of two policies are listed in each cell among the actual number of policies supported. The number in the bottom right-hand corner of each cell indicates how many of the 96 searched policies meet the risk threshold at the case volume.

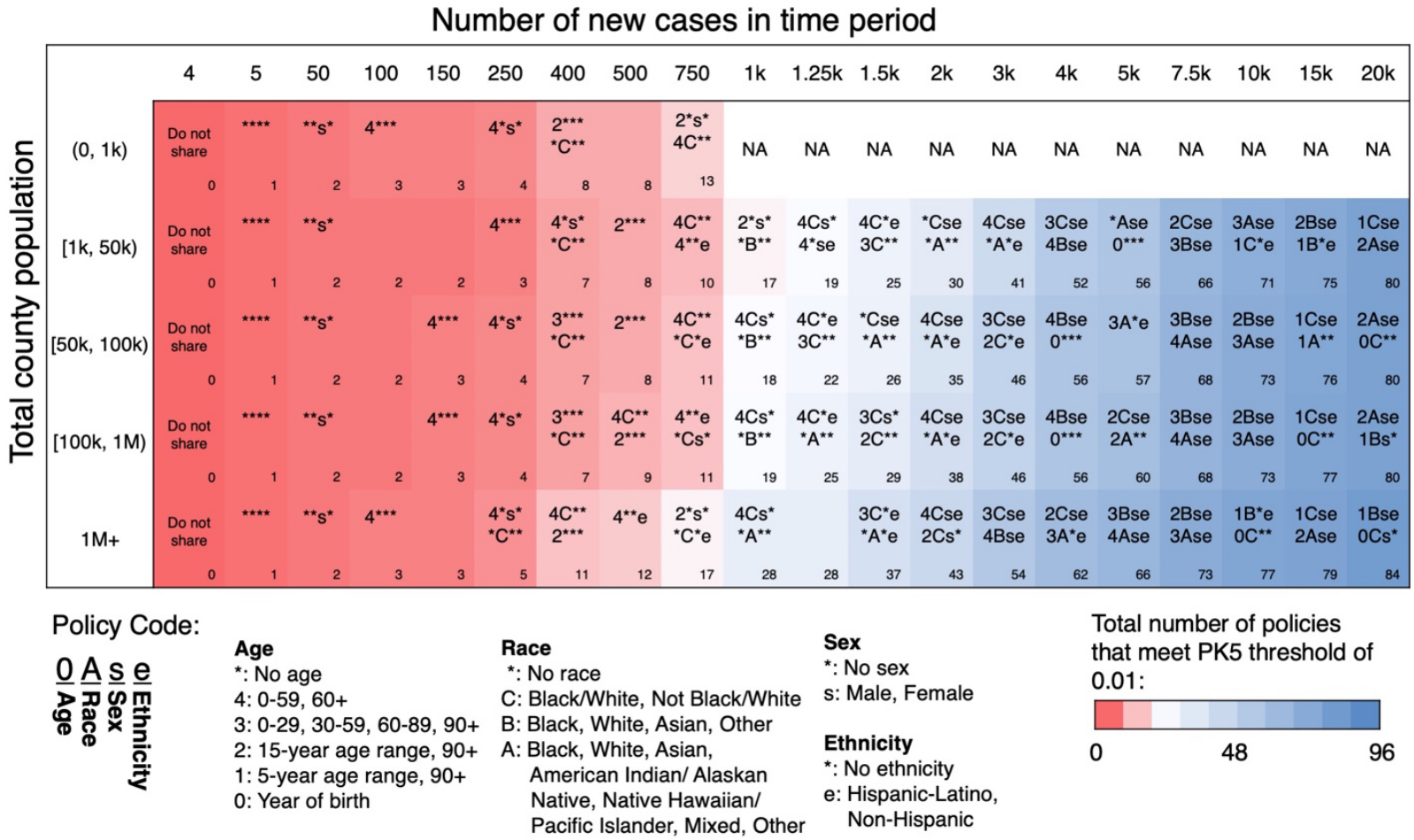

Figure 3.6. Generalization policies with a PK5 upper bound (calculated as the upper bound of the 95% quantile range of 1,000 framework simulations) less than or equal to 0.01 at varying disease case volume thresholds. A four-character alphanumeric code indicates the policy's generalization levels. All policies additionally include state and county of residence and some generalization of diagnosis date. A policy is eligible to be listed under the minimum number of new cases (table column) at which it meets the PK5 threshold for every county in the category (table row). A maximum of two policies are listed in each cell among the actual number of policies supported. The number in the bottom right-hand corner of each cell indicates how many of the 96 searched policies meet the risk threshold at the case volume.

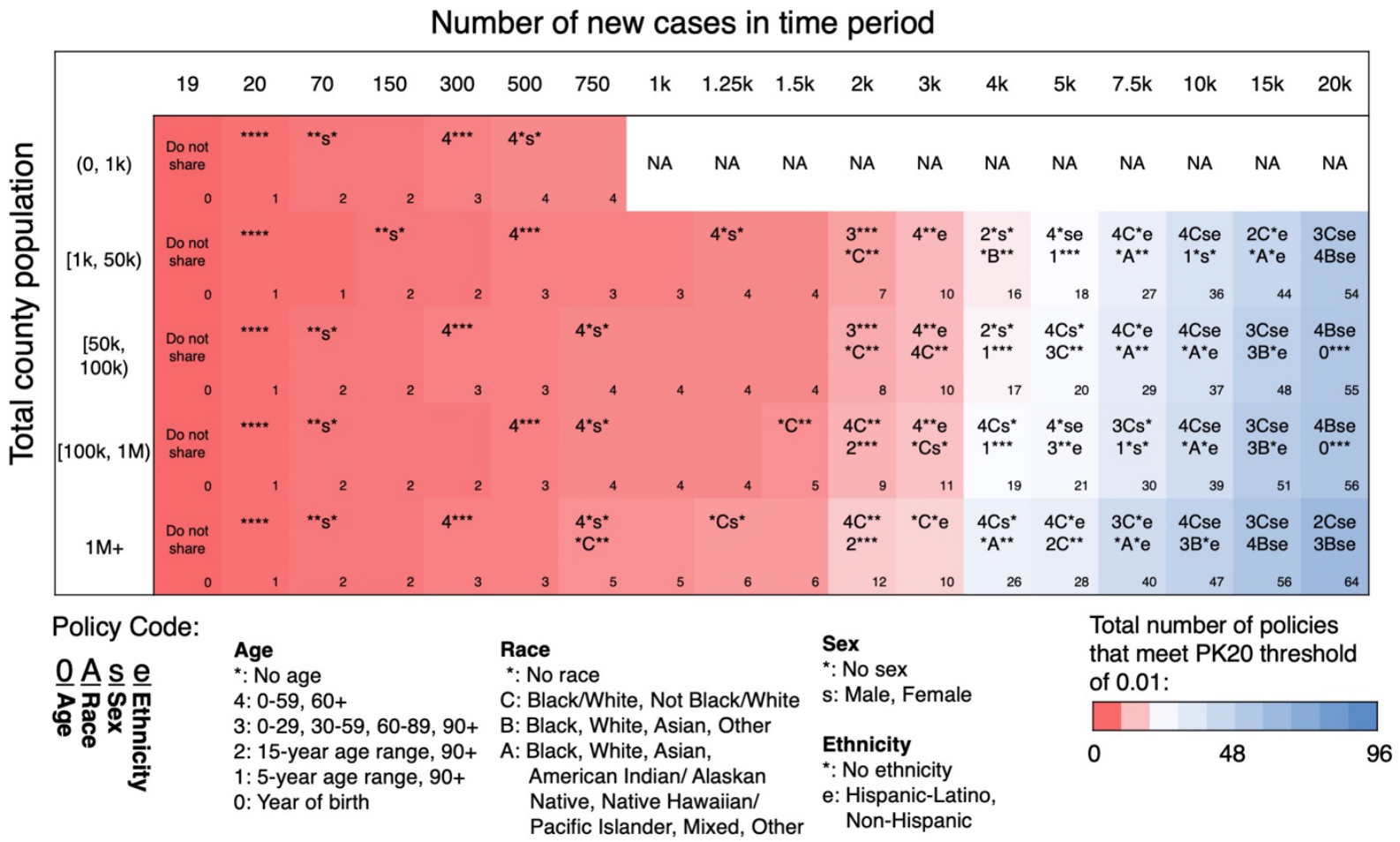

Figure 3.7. Policies with a PK20 upper bound (calculated as the upper bound of the 95% quantile range of 1,000 framework simulations) less than or equal to 0.01 at varying disease case volume thresholds. A four-character alphanumeric code indicates the policy's generalization levels. All policies additionally include state and county of residence and some generalization of diagnosis date. A policy is eligible to be listed under the minimum number of new cases (table column) at which it meets the PK20 threshold for every county in the category (table row). A maximum of two policies are listed in each cell among the actual number of policies supported. The number in the bottom right-hand corner of each cell indicates how many of the 96 searched policies meet the risk threshold at the case volume.

3.2.4.2. Marketer risk-based policy search

I apply the framework to search the same policy space as described in Figure 3.4 and identify data sharing policies that likely meet a marketer risk threshold of 0.01 at various dataset sizes. The search follows the same approach as the PK risk scenario. For each combination of U.S. county, case number, and policy I calculate the marketer risk on 1,000 independent simulations. From the 1,000 simulations, I calculate the upper bound of the 95% quantile range and compare the upper bound to a threshold of 0.01. The results indicate the minimum cumulative number of disease case records in the dataset at which a data sharing policy is supported for all counties in the population size category. I summarize the results in Figure 3.8.

Selecting a policy according to the cumulative number of records notably affects dynamic policy application. First, selecting a policy now means applying the same set of quasi-identifier generalizations to the entire dataset, including previously released records. Second, changing the generalization scheme of previously released records creates a dependency between successively applied data sharing policies. The new policy must be a parent of the current policy. If it is not, the combined information across dataset releases could expose patient identities. These differences prompt the data sharer to choose a path according to information priorities. To demonstrate, in Figure 3.8, I select a single path for each county population category and generate a corresponding results table.

Note, the nuances to the dynamic marketer risk-based policy approach would also occur in a scenario in which de-identification is guided by the PK risk and the quasi-identifier does not include date of diagnosis (see Section 3.5).

Figure 3.8 shows the number of acceptable policies increases with the cumulative number of records. For counties with more than one million residents, all 96 policies are supported when the dataset includes at least 100 records. The smallest counties achieve the fewest number of acceptable policies, with 19 equally feasible policies. The larger counties' results display a pattern where the number of supported policies at 1,000 case records remains relatively constant as the size of the dataset increases. This pattern arises from an underlying difference between the marketer risk and the PK. For a given county and policy, the PK risk fluctuates with the number of case records shared in a time window. Conversely, the marketer risk for a given county and policy converges toward a specific value as more records are accumulated. The table also displays a different pattern for the two smallest categories, because the search removes counties with a total population less than the case number threshold of interest.

To further illustrate the relationship between the marketer risk and the size of the dataset, I apply the framework to a single data sharing policy throughout the COVID-19 pandemic in Davidson County, TN. The 1Ase policy (see the key in Figure 3.8) is applied to a daily release schedule and allows for 532 potential demographic groups. Figure 3.9 shows that as the size of the disease surveillance dataset increases, the expected (mean) marketer risk remains relatively constant, and the range of risk converges toward the expectation. I note that the expected marketer risk represents the expected proportion of records correctly matched to the identified dataset. Though the proportion remains constant, the number of individuals at risk increases with the size of the dataset.

The relatively constant value for the marketer risk expectation is intuitive. Since the date of diagnosis is not considered a quasi-identifier in the attack scenario, the demographic groups increase in size as more records are added to the dataset. As the number of records in group j in the shared dataset approaches the number of records in group j in the identified dataset, the marketer risk (Eqn. 3.2) moves toward its limit, as shown in Eqn. 3.3:

$$\frac{\sum_{j=1}^{J} 1}{N} = \frac{J}{N} \tag{3.3}$$

where N is the size of the identified dataset/total population. Eqn. 3.3 approximates the expected marketer risk estimated by the framework's algorithm. The orange dotted line in Figure 3.9 was calculated using Eqn. 3.4:

$$\frac{\sum_{j=1}^{\tilde{J}} 1}{N} = \frac{\tilde{J}}{N} \tag{3.4}$$

where $\tilde{J}$ is the number of demographic groups defined by the policy for which at least one person in the population corresponds. The value of $\tilde{J}$ is obtained from the U.S. Census data. Thus, the expected marketer risk can be mathematically approximated from the framework inputs without the complete Monte Carlo simulation.

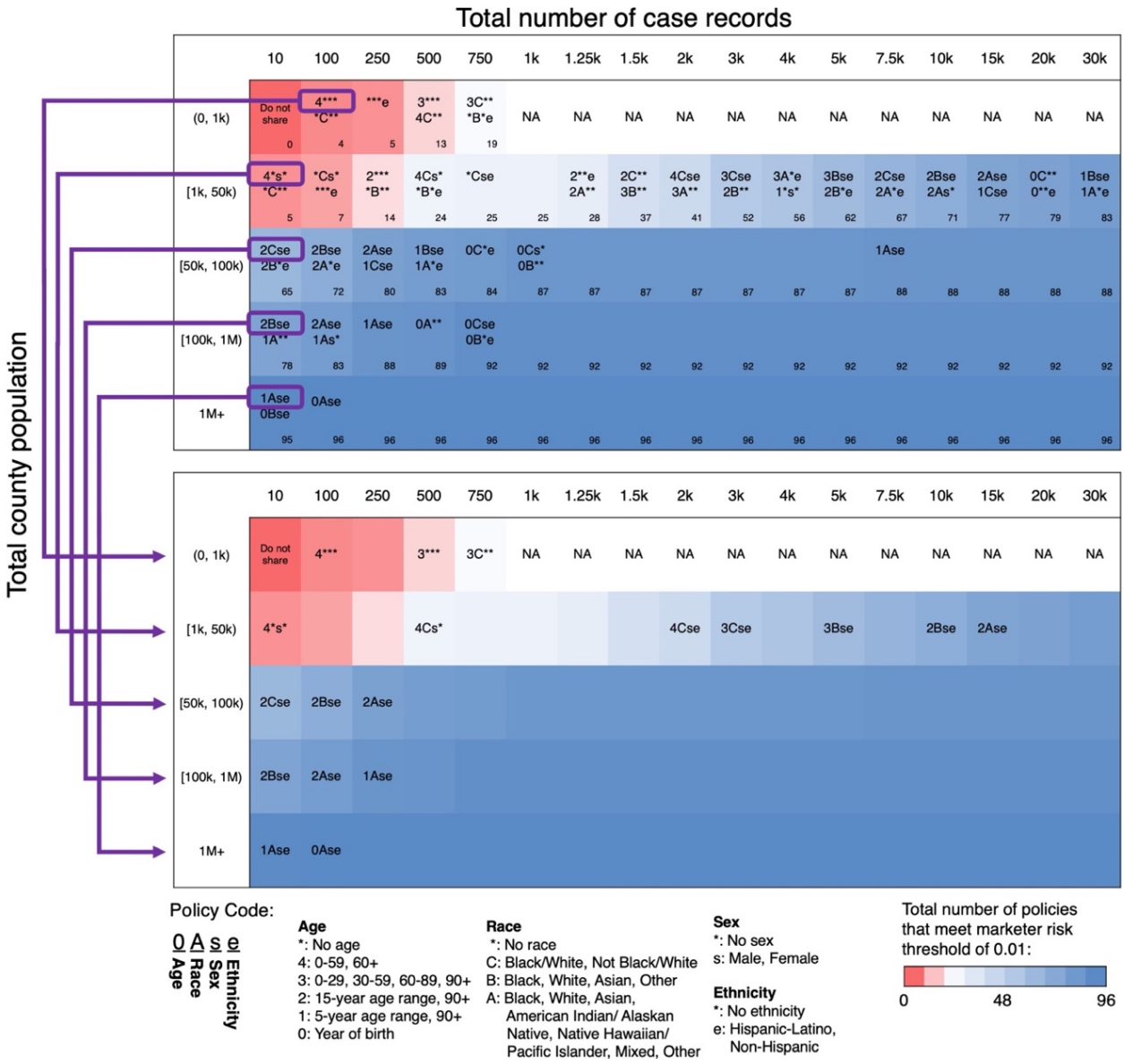

Figure 3.8. (Top) Policies with a marketer risk upper bound (calculated as the upper bound of the 95% quantile range of 1,000 framework simulations) less than or equal to 0.01 at varying disease case volume thresholds. A four-character alphanumeric code indicates the policy's generalization levels. All policies additionally include state and county of residence and some generalization of diagnosis date. A policy is eligible to be listed under the minimum number of new cases (table column) at which it meets the marketer risk threshold for every county in the category (table row). A maximum of two policies are listed in each cell among the actual number of policies supported. The number in the bottom right-hand corner of each cell indicates how many of the 96 searched policies meet the risk threshold at the case volume. The purple circles indicate the starting policy for each county population category, from which the generalization paths are generated in the table below. (Bottom) The child-parent generalization path for each category. Moving from left to right in a row, each new policy listed is a parent of those previously listed.

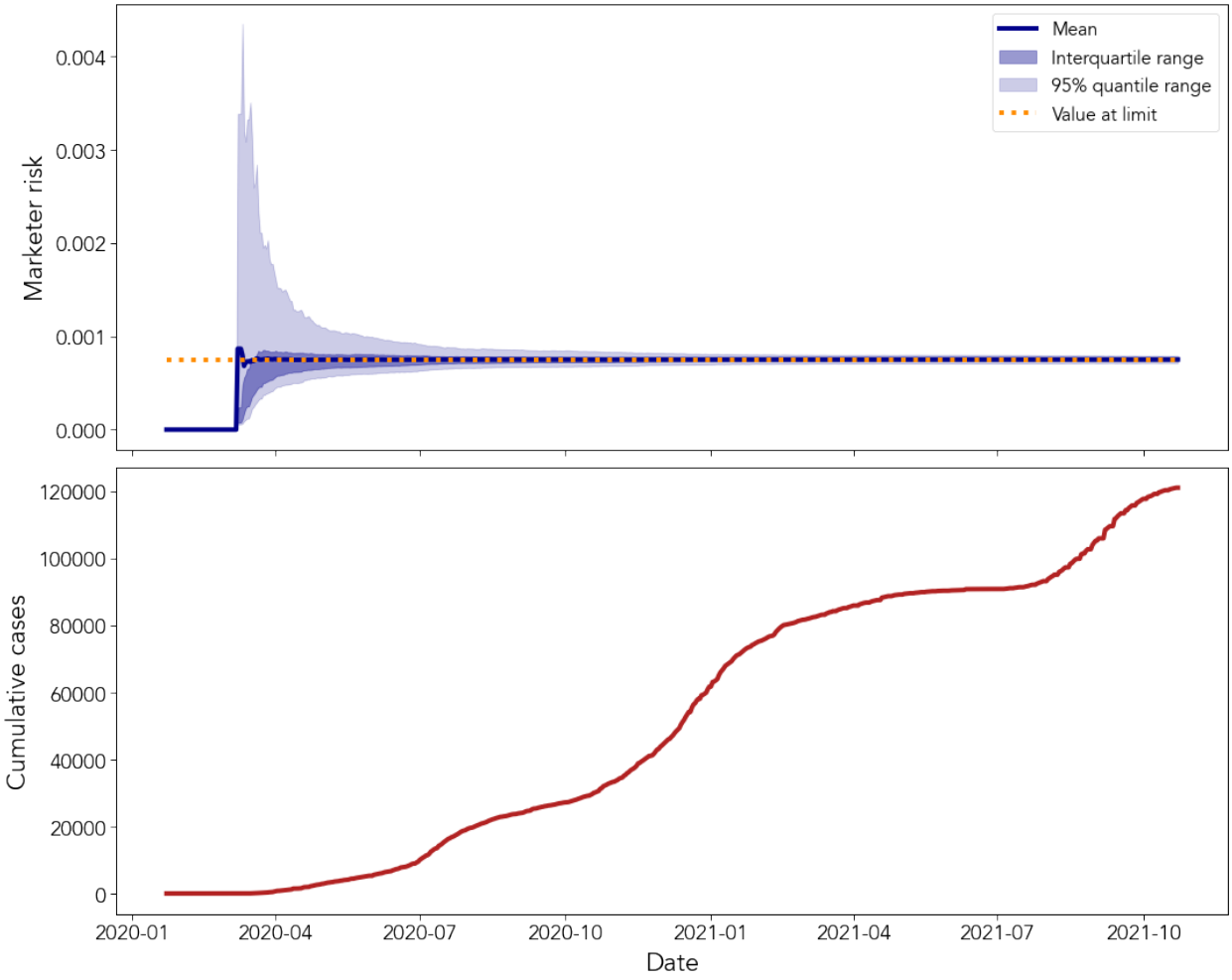

Figure 3.9. Marketer risk estimation of the 1Ase policy applied to daily releases of COVID-19 disease case surveillance data in Davidson County, TN. The expectation and quantile ranges were calculated from 1,000 independent simulations. The marketer risk is evaluated each day (Top) on the cumulative number of cases (Bottom). The orange dotted line represents the marketer risk when the size of the shared dataset is equal to the size of the population. The height of the dotted line was calculated according to Eqn. 3.4.

3.3 Risk evaluation

Now that the dynamic policy approach and privacy risk estimation framework have been described, I evaluate the dynamic policy approach's ability to maintain the re-identification risk below an established threshold. This section begins with an overview of the evaluation methods and ends with a presentation of the results for both the PK risk (with date of diagnosis included in the quasi-identifier) and the marketer risk (with date of diagnosis not included in the quasi-identifier) scenarios.

3.3.1 Evaluation overview

I use the summarized policy search results and forecasted COVID-19 disease case counts to evaluate dynamic policy selection in the context of the COVID-19 pandemic. In this evaluation, I measure the proportion of data releases in which the PK11 or marketer risk likely remains below the policy search threshold of 0.01. The dynamic policy is evaluated for two distinct alternative data sharing scenarios: 1) a daily release schedule with a 1-day lagging period assumption and 2) a weekly release schedule. The daily release schedule shares the actual date of diagnosis, prioritizing date granularity at the potential cost of demographic granularity. The weekly release schedule generalizes the date to week of diagnosis.

For every county that is both represented in the US Census PCT12 tables and Johns Hopkins COVID-19 datasets, the PK11-driven dynamic policy method selects the generalization policy from the search results at the beginning of each week according to the forecasted COVID-19 case volumes. I use the CDC COVID-19 ensemble model's county-specific, one-week forecasts for its superior accuracy over other models[138,149,150]. I specifically used the model's point estimates, calculated as the median of the point estimates of the various prediction models. For the evaluation, I collected all model predictions from August 2020 through October 2021. I obtain daily increase predictions by uniformly distributing the weekly increase point estimate. In selecting policies for the daily release schedule, I use the minimum number of predicted cases in the week. This applies the most privacy preserving policy to all new cases reported in the week. For the weekly release schedule, I use the forecasted one-week increase.

In the marketer risk scenario, I do not use the CDC's COVID-19 ensemble prediction model to inform dynamic policy selection. Since the size of the dataset monotonically increases, the minimum number of case records will always occur on the first day of the week, regardless of the predicted weekly increase in case numbers. Therefore, at the beginning of each week (Sunday, to be consistent with the prior week

definition) I use the current total number of disease case records in the dataset to select the policy for the upcoming week. This applies the most private policy to the week's new cases while allowing the policy to potentially change on a weekly basis. The policy for the weekly release schedule is chosen according to the size of the cumulative dataset at the end of each week (Saturday).

After selecting the sequence of policies for each county, I estimate the privacy risk of sharing the actual reported number of records via the privacy risk estimation framework. I define the actual number of disease cases per day or week by the Johns Hopkins COVID-19 tracking data. The PK11 or marketer risk value for each time point in each county is calculated as the upper bound of the 95% quantile range of 1,000 simulations. The evaluation measures the proportion of releases the upper bound remains below 0.01.

I additionally evaluate the static application of a policy designed with current, retrospective de-identification techniques, akin to those applied to the CDC's COVID-19 Public Use Data with Geography[129]. The policy, hereafter referred to as the *k*-anonymous policy, shares age intervals in the form (0-17, 18-49, 50-64, and 65+); race (Black or African American, White, Asian, American Indian or Alaskan Native (AIAN), Native Hawaiian or Pacific Islander (NHPI), Multiple/Other); ethnicity (Hispanic-Latino and Non-Hispanic); sex (Female and Male); state and county of residence; and date or week of diagnosis. I note the CDC's policy, from which the *k*-anonymous policy derives, was developed to meet regulatory requirements and public health standards under a different release schedule (once every two weeks to once every month) and in a retrospective manner (the actual patient records are collected, de-identified and released in a batch). The CDC's policy is designed to achieve 11-anonymity (i.e., PK11 = 0) by generalizing the date of diagnosis to month and by nulling out quasi-identifier information for small groups[44,129,131]. Thus, the *k*-anonymous policy resembles a policy developed with traditional de-identification, but notably differs in its treatment of dates of events and in its assumption of no suppression. I further note this last feature is another unique factor to sharing surveillance data in near-real time. Suppression cannot be applied with confidence because it is almost impossible to forecast exactly which records will fall into small demographic groups. Notably, the CDC's policy suppresses around 3% of each quasi-identifier to achieve 11-anonymity[147].

To provide a specific illustration of the dynamic policy approach to daily releasing updated, record-level disease surveillance data, after the primary validation experiments, I additionally provide case studies for two Tennessee counties. The first, Davidson County, is a relatively large metropolitan region with a population of approximately 630,000 residents. The second, Perry County, is a relatively rural area with

around 8,000 residents. Table 3.2 displays the counties' population demographics according to recent estimates from U.S. Census Bureau[19].

In each case study, I select a policy on a weekly basis in the same manner as the evaluation. However, to demonstrate how the framework incorporates the data recipient's potential knowledge of diagnosis date (for the PK risk scenario), and accounting for the general turnaround time of COVID-19 diagnostic tests results[151–153], I set a 5-day lagging period. Under these constraints, weekly dynamic policy selection first calculates a 5-day rolling sum of new disease case numbers through the coming week. The minimum value of the rolling sum is used to select the policy. I again estimate the privacy risk of sharing the actual number of records under the sequence of selected policies with the privacy risk estimation framework and the Johns Hopkins COVID-19 tracking data. To evaluate the dynamic policy under optimal case load forecasting, I repeat the process by replacing the forecasted case counts with the actual case numbers in policy selection.

Table 3.2. County demographics

		Davidson County, TN *n* = 626,681	**Perry County, TN** *n* = 7,915
Race	White	385,039 (61.4%)	7,584 (95.8%)
	Black	173,730 (27.7%)	119 (1.5%)
	Asian	19,027 (3.0%)	14 (0.2%)
	AIAN	2,091 (0.3%)	48 (0.6%)
	NHPI	394 (0.06%)	0 (0%)
	Other	30,757 (4.9%)	30 (0.4%)
	Mixed	15,643 (2.5%)	120 (1.5%)
Ethnicity	Hispanic/Latino	61,086 (9.7%)	117 (1.5%)
	Non-Hispanic	565,595 (90.3%)	7,798 (98.5%)
Age group	[0, 10)	82,304 (13.1%)	927 (11.7%)
	[10, 20)	72,903 (11.6%)	1,041 (13.2%)
	[20, 30)	115,876 (18.5%)	819 (10.3%)
	[30, 40)	97,154 (15.5%)	887 (11.2%)
	[40, 50)	83,472 (13.3%)	980 (12.4%)
	[50, 60)	79,768 (12.7%)	1,192 (15.1%)
	[60, 70)	49,803 (7.9%)	1,096 (13.8%)
	[70, 80)	26,901 (4.3%)	645 (8.1%)
	[80, +]	18,500 (3.0%)	328 (4.1%)
Sex	Female	323,141 (51.6%)	3,941 (49.8%)
	Male	303,540 (48.4%)	3,974 (50.2%)

*number of individuals (% of population)

3.3.2 PK risk evaluation

I summarize the PK11 risk evaluation results, categorizing counties in the same manner as the policy search, in Table 3.3. There are several major findings. First, dynamically adapting the generalization policy meets the PK11 threshold more frequently than statically applying the *k*-anonymous policy. On average, the dynamic policy meets the threshold for at least 92.8% of the 448 daily releases and 96.0% of the 64 weekly releases. The *k*-anonymous policy meets the threshold as few as 11.8% of the daily releases and 0.4% of the weekly releases. Second, I find that new cases do not occur every day or every week, particularly in counties with fewer residents. As such, there are fewer days the PK11 upper bound can potentially exceed the threshold, inflating proportions in smaller counties.

Table 3.3. Average proportion of time periods where the upper bound of the 95% quantile range of the PK11 risk is less than or equal to 0.01 in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). The average and 95% quantile range in each cell are taken across all counties in the corresponding population size category. The *k*-anonymous policy shares age intervals (0-17, 18-49, 50-64, and 65+), race (Black or African American, White, Asian, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander, Multiple/Other), ethnicity (Hispanic-Latino and Non-Hispanic), sex (Female and Male), and state and county of residency. The *k*-anonymous policy is statically applied to each release. The daily release PK11 estimates apply a 1-day lagging period, while the weekly release estimates assume the actual date of diagnosis is generalized to week of diagnosis.

County Population Size	Average proportion of daily releases that meet the PK11 threshold in the COVID-19 pandemic [95% Quantile Range] (n = 448)		Average proportion of weekly releases that meet the PK11 threshold in the COVID-19 pandemic [95% Quantile Range] (n = 64)	
	k-anonymous Policy	Dynamic Policy	*k*-anonymous Policy	Dynamic Policy
< 1,000 (n = 35)	0.900 [0.790, 0.998]	1 [1, 1]	0.605 [0.266, 0.987]	0.999 [0.984, 1]
1,000 - 50,000 (n = 2,129)	0.389 [0.118, 0.815]	0.971 [0.902, 1]	0.072 [0, 0.406]	0.960 [0.906, 1]
50,000 - 100,000 (n = 398)	0.181 [0.042, 0.532]	0.928 [0.868, 0.987]	0.004 [0, 0.031]	0.974 [0.922, 1]
100,000 - 1,000,000 (n = 538)	0.145 [0.009, 0.521]	0.947 [0.882, 0.998]	0.008 [0, 0.026]	0.982 [0.938, 1]
> 1,000,000 (n = 39)	0.118 [0.007, 0.304]	0.961 [0.874, 0.998]	0.057 [0, 0.288]	0.962 [0.906, 1]

3.3.3 PK risk case studies

Figure 3.10 shows how the forecasted case volumes do not match the weekly seasonality of the actual reported cases in Davidson County. Consequently, the CDC ensemble model tends to overestimate case loads, leading to the selection of more granular policies. Despite the rippling effects of the overestimation, the 95% quantile range of the forecast-driven PK11 remains below 0.01 throughout most of the time frame. Several days exceed the threshold, most of which occur when the selected policies disagree whether to share record-level data under the **** policy or to not share. When sharing fewer than 11 new case records in a 5-day window under the forecast-driven dynamic policy, all new records fall into a demographic group smaller than size 11, resulting in a PK11 of 1.0. Notably, the PK11 never exceeds the threshold when selecting policies according to the actual case counts. Adapting the policy according to perfect forecasts provides optimal privacy protection.

Figure 3.11 shows that case counts remain relatively small before, as well as after, infection spikes in October 2020 and August 2021. Throughout most of these intervals of low-infection rates, the selected policies from each data source indicate that record-level data should not be shared on a daily basis. However, when the 5-day rolling sums oscillate around 11 cases, the forecasted values again overestimate the weekly minimum case loads, resulting in a PK11 of 1.0. Despite the privacy leaks in the forecast-driven dynamic policy, the dynamic policy guided by the actual disease case counts again maintains the PK11 values below the threshold throughout the time frame.

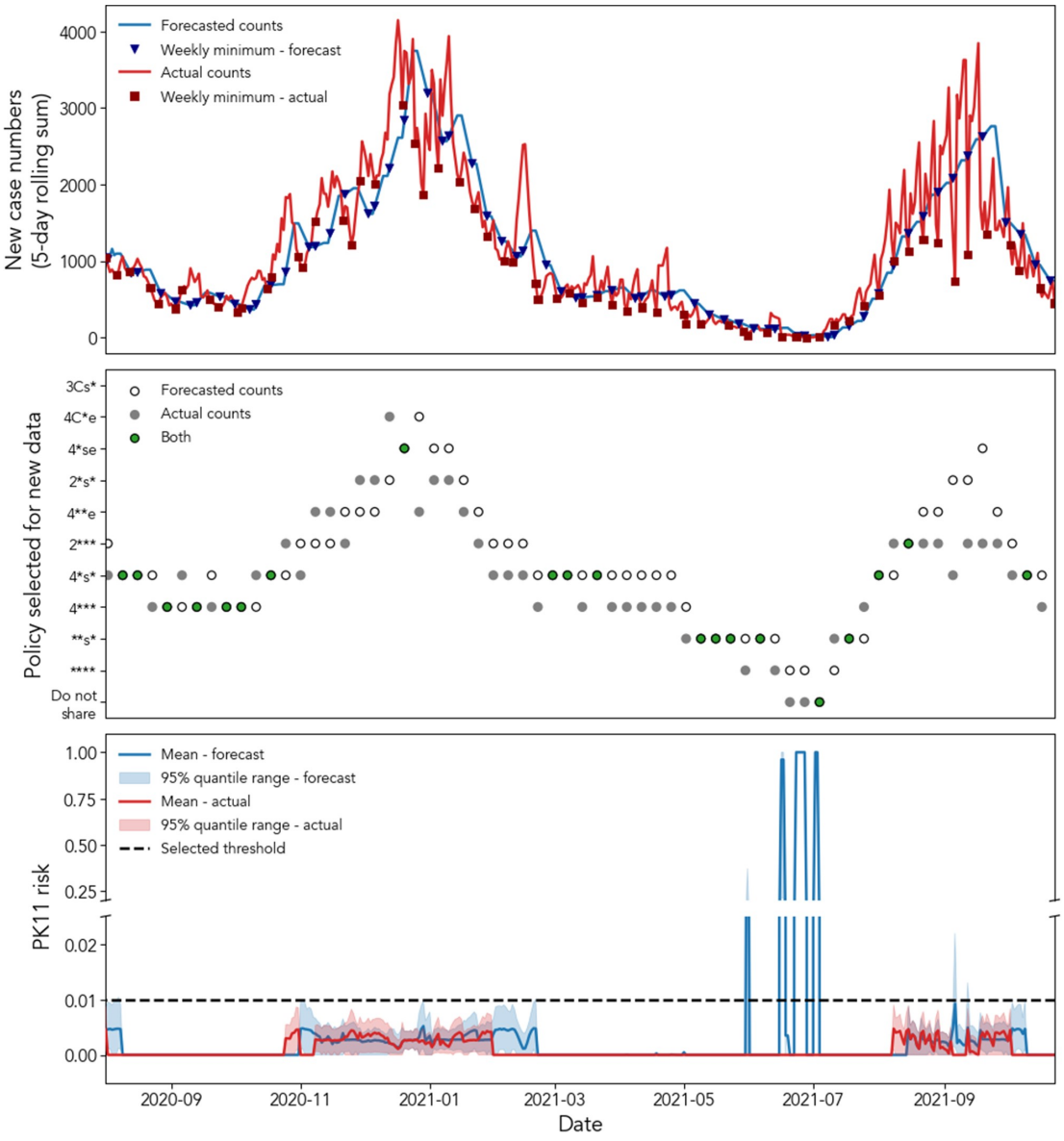

Figure 3.10. Dynamic policy selection applied to Davidson County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The 5-day rolling sum of the forecasted and actual case counts reported in Davidson County. The forecasted counts are from the CDC's COVID-19 ensemble model and the actual counts are from the Johns Hopkins surveillance data. The blue triangles and red squares denote the minimum value within each week (defined as Sunday-Saturday per the CDC model's definition). The minimum values are used to select a policy from policy search results. (Middle) The selected policy at the beginning of each week in the pandemic. Each policy is represented by a 4-character alphanumeric code following the key in Figure 3.5. The policies are ordered by increasing case count thresholds from bottom to top. Green circles indicate agreement between the policies selected from the forecasted and actual case counts. (Bottom) The PK11 from sharing the actual number of records under the two sequences of policies detailed in the middle graph. The expectation and 95% quantile range are calculated from 1,000 independent framework simulations, while applying a 5-day lagging period assumption. The horizontal dashed line marks the PK11 threshold of 0.01.

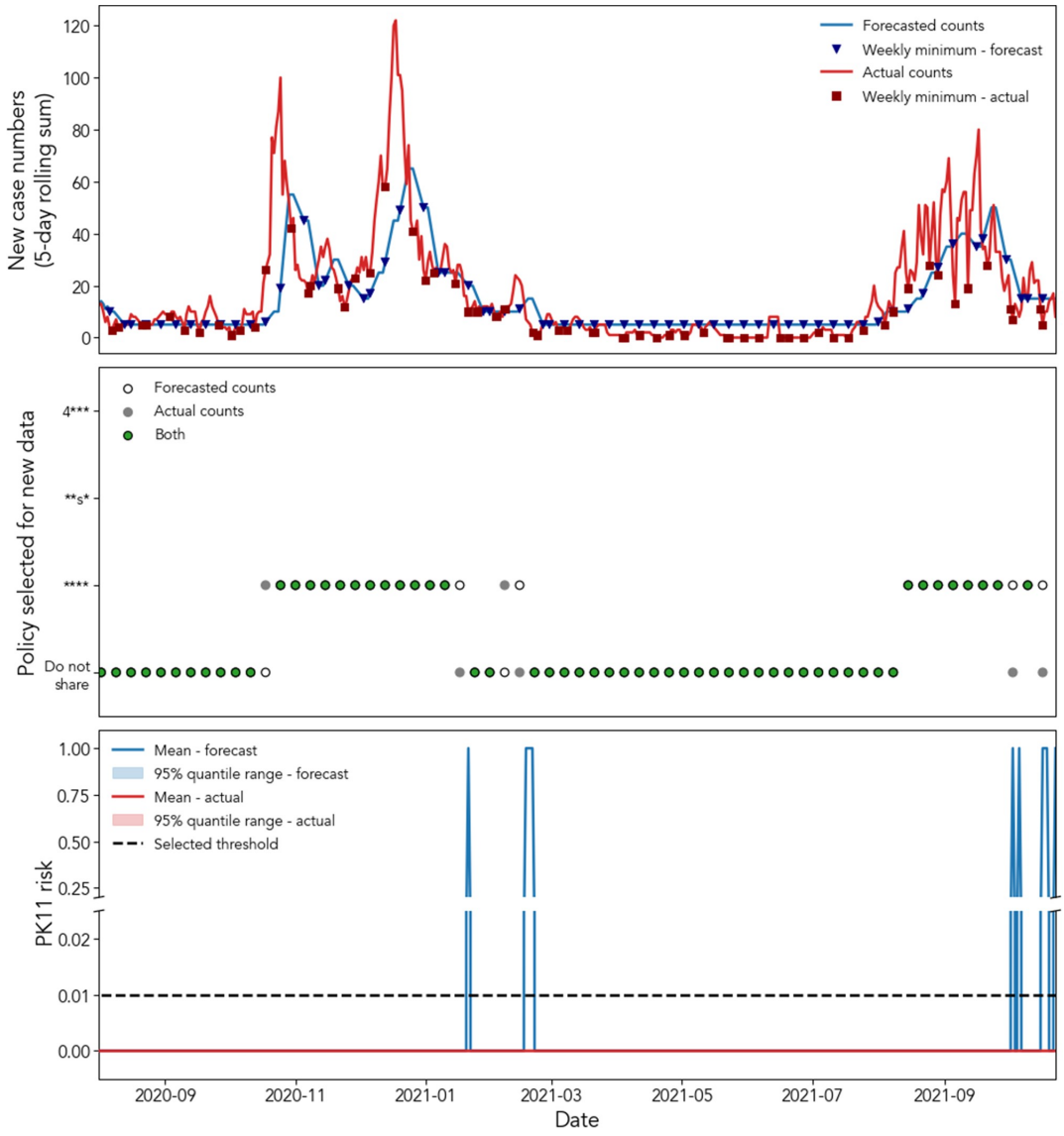

Figure 3.11. Dynamic policy selection applied to Perry County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The 5-day rolling sum of the forecasted and actual case counts reported in Davidson County. The forecasted counts are from the CDC's COVID-19 ensemble model and the actual counts are from the Johns Hopkins surveillance data. The blue triangles and red squares denote the minimum value within each week (defined as Sunday-Saturday per the CDC model's definition). The minimum values are used to select a policy from policy search results. (Middle) The selected policy at the beginning of each week in the pandemic. Each policy is represented by a 4-character alphanumeric code following the key in Figure 3.5. The policies are ordered by increasing case count thresholds from bottom to top. Green circles indicate agreement between the policies selected from the forecasted and actual case counts. (Bottom) The PK11 from sharing the actual number of records under the two sequences of policies detailed in the middle graph. The expectation and 95% quantile range are calculated from 1,000 independent framework simulations, while applying a 5-day lagging period assumption. The quantile ranges are too narrow to be seen outside the mean. The horizontal dashed line marks the PK11 threshold of 0.01.

3.3.4 Marketer risk evaluation

Table 3.4 displays risk evaluation results for the marketer-risk driven dynamic policy. Dynamic policy selection, guided by the framework's forecasts, never exceeds the marketer risk threshold of 0.01. For the smallest county size category, the total case number never reaches 100 and no data is shared. Data is shared for all other county size categories. For county's with at least 50,000 residents, the *k*-anonymous policy meets the marketer risk threshold as frequently as the dynamic policy, but with lesser data utility in terms of available demographic groups. The *k*-anonymous policy allows for 112 unique combinations of age, race, sex, and ethnicity (or 112 unique quasi-identifier values). Under the dynamic policy selection and the case loads beginning in August 2020, counties with a population between 50,000 and 100,000 residents tend to share data with at least the 2Bse policy, which also designate 112 unique demographic groups. Counties with a population between 100,000 and 1,000,000 tend to share data with at least the 2Ase policy, which allows 196 groups. And the counties with at least 1 million residents apply the 0Ase policy that allows for 2,884 groups. The dynamic policy selection tailors the data sharing policy to both case load and county population to balance privacy and utility better than the *k*-anonymous policy at the marketer risk threshold of 0.01.

Table 3.4. Average proportion of time periods where the upper bound of the 95% quantile range of the marketer risk is less than or equal to 0.01 in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). The average and 95% quantile range in each cell are taken across all counties in the corresponding population size category. The *k*-anonymous policy shares age intervals (0-17, 18-49, 50-64, and 65+), race (Black or African American, White, Asian, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander, Multiple/Other), ethnicity (Hispanic-Latino and Non-Hispanic), sex (Female and Male), and state and county of residency. The *k*-anonymous policy is statically applied to each release. The daily release estimates assume the dataset is updated on a daily basis, while the weekly releases estimates assume the dataset is updated on a weekly basis.

	Average proportion of daily releases that meet the marketer risk threshold in the COVID-19 pandemic [95% Quantile Range] (n = 161)		Average proportion of weekly releases that meet the marketer risk threshold in the COVID-19 pandemic [95% Quantile Range] (n = 23)	
County Population	*k*-anonymous Policy	Dynamic Policy	*k*-anonymous Policy	Dynamic Policy
< 1,000 (n = 35)	0.074 [0, 0.345]	1 [1, 1]	0.072 [0, 0.336]	1 [1, 1]
1,000 - 50,000 (n = 2,129)	0.689 [0, 1]	1 [1, 1]	0.691 [0, 1]	1 [1, 1]
50,000 - 100,000 (n = 398)	1 [1, 1]	1 [1, 1]	1 [1, 1]	1 [1, 1]
100,000 - 1,000,000 (n = 538)	1 [1, 1]	1 [1, 1]	1 [1, 1]	1 [1, 1]
> 1,000,000 (n = 39)	1 [1, 1]	1 [1, 1]	1 [1, 1]	1 [1, 1]

3.3.5 Marketer risk case studies

The case studies results for Davidson County and Perry County are displayed in Figures 3.12 and 3.13, respectively. As the generalization path in Figure 3.8 instructs, the 1Ase policy is applied to every data release in Davidson County, TN, as the size of the dataset remains above 250 records throughout the time interval. The mean and 95% quantile range of the marketer risk remain below the threshold of 0.01 at each time point. The 95% quantile range, in this case, is too narrow to be seen outside the expectation. The generalization policy in Perry County, TN changes from 4*s* to 4Cs* the week after the number of disease case records in the dataset surpasses 500. The expectation and 95% quantile range of the marketer risk stay below the 0.01 marketer risk threshold throughout the time interval.

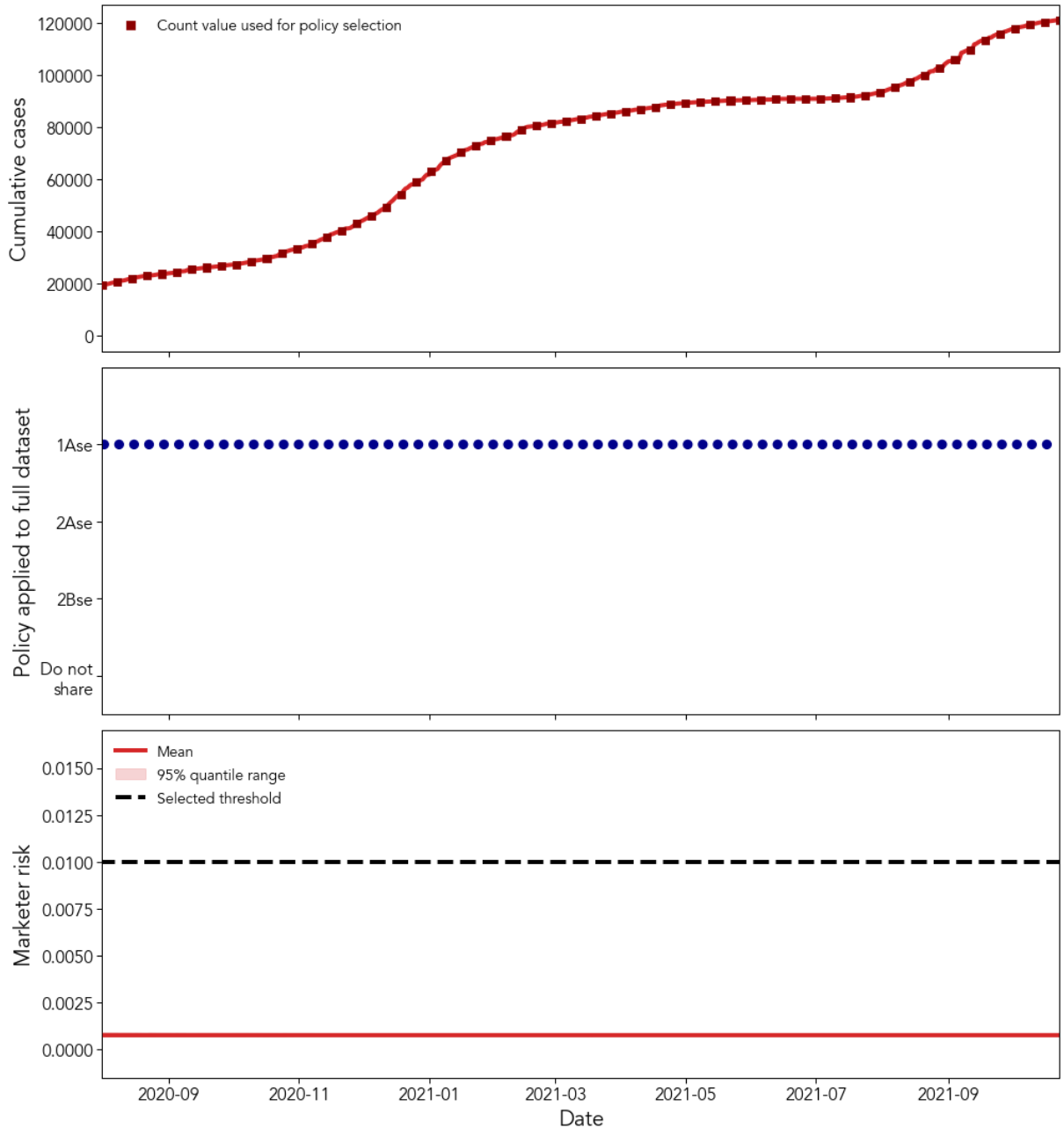

Figure 3.12. Dynamic policy selection applied to Davidson County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The cumulative sum of the case counts reported in Davidson County, according to the Johns Hopkins COVID-19 tracking data. The red squares represent the case record number value and the end of the previous week (through Saturday) used in selecting the next week's policy from Supplementary Figure E5. (Middle) The selected policy at the beginning of each week in the pandemic. Each policy is represented by a 4-character alphanumeric code following the key in Figure 3.8. (Bottom) The marketer risk from sharing the actual number of records under the sequence of policies detailed in the middle graph. The expectation and 95% quantile range are calculated from 1,000 independent simulations. The horizontal dashed line marks the marketer risk threshold of 0.01.

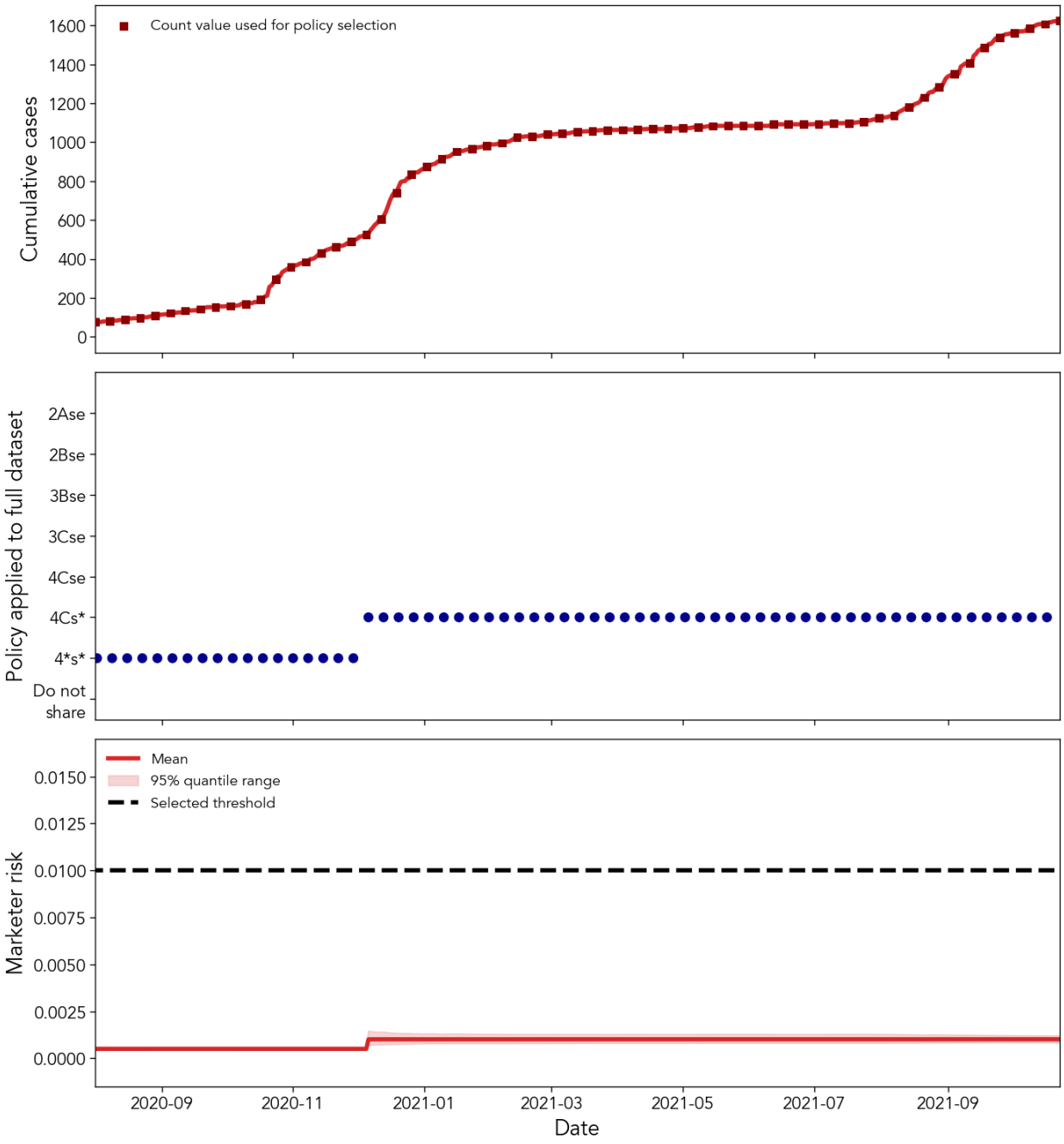

Figure 3.13. Dynamic policy selection applied to Perry County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The cumulative sum of the case counts reported in Davidson County, according to the Johns Hopkins COVID-19 tracking data. The red squares represent the case record number value and the end of the previous week (through Saturday) used in selecting the next week's policy from Supplementary Figure E5. (Middle) The selected policy at the beginning of each week in the pandemic. Each policy is represented by a 4-character alphanumeric code following the key in Figure 3.8. (Bottom) The marketer risk from sharing the actual number of records under the sequence of policies detailed in the middle graph. The expectation and 95% quantile range are calculated from 1,000 independent simulations. The horizontal dashed line marks the marketer risk threshold of 0.01.

3.4 Risk fairness evaluation

3.4.1 Distribution of PK risk

Not only do data stewards have an ethical obligation to minimize the privacy risks of patients represented in a dataset, they also have an ethical obligation to equally distribute the risks between groups of patientsl[13,71]. As such, I use the framework to investigate how a data sharing policy is likely to distribute the privacy risk across demographic groups. First, I measure how the PK risk stratifies across racial groups at k=5, k=10, and k=20. I calculate the expected proportion of the overall PK risk to which each race corresponds when applying a single policy (1Bse, according to the key in Figure 3.5) for the duration of the COVID-19 pandemic within Davidson County, TN. The case counts input into the framework are the actual counts from the Johns Hopkins data.

According to the US Census PCT12 tables, 61.4% of Davidson County residents are White, 27.7% are Black, 7.8% fall into the new Other group, and 3.0% are Asian. Figure 3.14 illustrates that the proportion of the privacy risk each racial subpopulation bears is not equal and varies over time. Around January 2021, the Asian subpopulation is expected to bear more of the overall PK5 risk than the other subpopulations, but less than the White and Black subpopulations for the overall PK11 and PK20 risk. This suggests that under the 1Bse generalization policy, Asian individuals are more likely to be more unique (i.e., fall into an equivalence class group of size 5 or less) than the other racial groups (more likely to fall into a equivalence class group between 6 and 20). The same phenomenon occurs for Black individuals around October 2020 and April 2021. Note, the “Other” group shown in Figure 3.14 represents the combination of the initial racial subpopulations American Indian/Alaskan Native, Native Hawaiian/Pacific Islander, Mixed, and Other, which were generalized into a new Other group. Therefore, these subpopulations collectively bear the proportion of the PK risk shown below, but their individual contributions may vary. Still, the results highlight the unequal distribution of the privacy risk between groups. If fairness with respect to privacy is defined, in this scenario, as each group bearing an equal proportion of the PK risk, the distribution of risk is unfair.

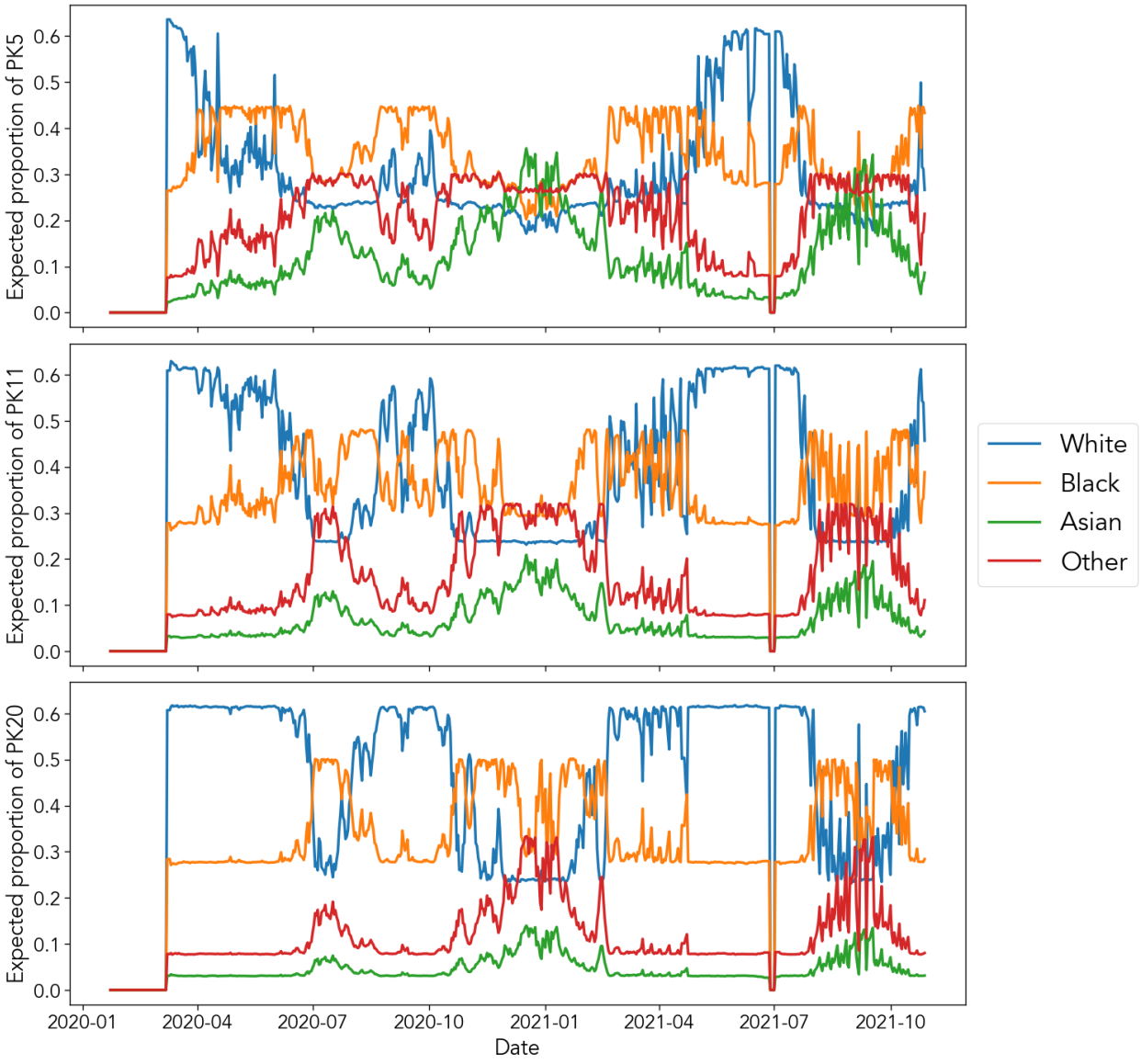

Figure 3.14. The proportion of the overall expected PK5 (Top), PK10 (Middle), and PK20 (Bottom) each racial subpopulation bears, throughout October 2021 of the COVID-19 pandemic, when applying a 1Bse policy to Davidson County, TN. Proportion of risk is reported as the average of 1,000 independent framework simulations with a 5-day lagging period assumption. The racial subpopulations are based on the U.S. Census: 1) Asian, 2) Black, 3) White, and 4) Other, which is composed of Alaskan Native/American Indian, Pacific Islander/Native Hawaiian, Two or More Races, and Some Other Race.

3.4.2 Distribution of marketer risk

I repeat the same experiment as above, this time calculating each racial subpopulation's expected proportion of the overall marketer risk. Figure 3.15 shows that the two largest racial subpopulations, Black and White, essentially bear an equal amount of the overall amortized re-identification risk. The two smallest subpopulations do the same.

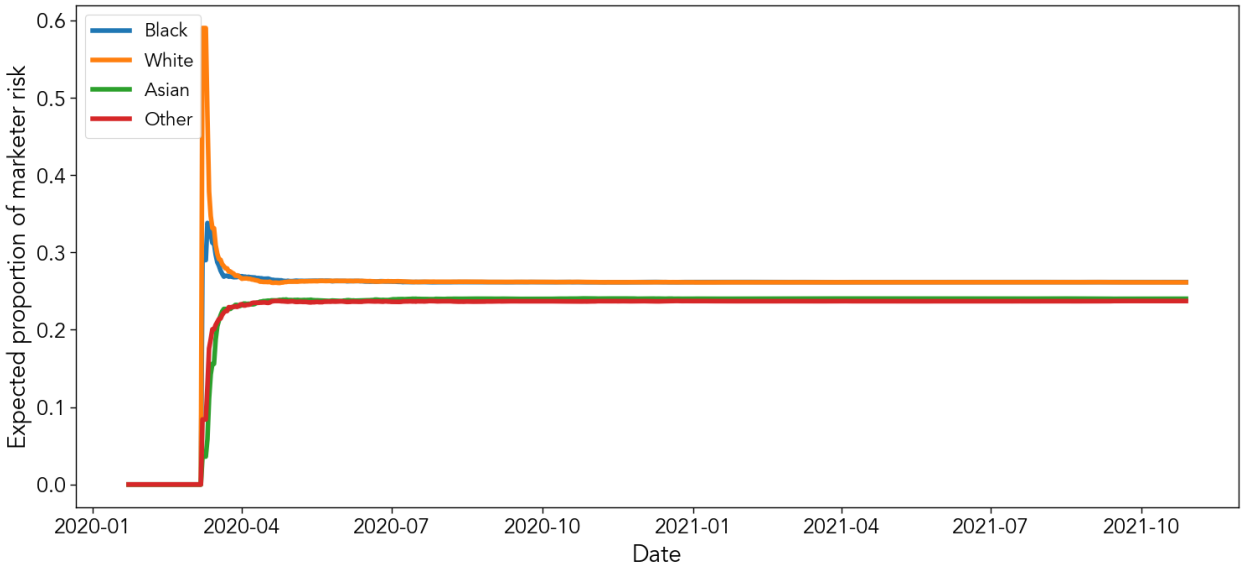

Figure 3.15. The proportion of the overall expected marketer risk each racial subpopulation bears, throughout October 2021 of the COVID-19 pandemic, when applying a 1Bse policy to Davidson County, TN. Proportion of risk is reported as the average of 1,000 independent framework simulations. The racial subpopulations are based on the U.S. Census: 1) Asian, 2) Black, 3) White, and 4) Other, which is composed of Alaskan Native/American Indian, Pacific Islander/Native Hawaiian, Two or More Races, and Some Other Race.

3.5 Utility evaluation

To complement the privacy risk evaluation of the dynamic policy approach to de-identifying a COVID-19 pandemic registry in Section 3.3, in this section, I evaluate the utility of such data. Specifically, I evaluate the ability to detect simulated infection disparities within the data.

This section begins with a description of the different types of de-identification methods considered in the evaluation, including variations of the dynamic policy approach and methods derived from two real-world COVID-19 datasets. I then describe how I simulate disparities in infectious disease surveillance data, provide details regarding how I detect disparities with an outbreak detection algorithm for each de-identification method, and review the experimental design and performance evaluation measures. Finally, I present the results of the experiments.

3.5.1 Data sharing policies and assumptions

In this utility evaluation, I compare the ability to detect infection disparities between three variations of the dynamic policy approach, each considering an adversary with different background knowledge, as well as two de-identification policies derived from real-world COVID-19 datasets. In this section and the next (Section 3.6), a data sharing policy refers to the (static or dynamic) generalization applied to the quasi-identifier values as well as the rate at which the dataset is update and released. This differs from the generalization policy defined in the previous sections that defined a single set of generalization specifications.

All three dynamic policies include date of diagnosis and county of residence and are updated on a daily basis. The first dynamic policy, hereafter referred to as the strong adversary policy (**SAP**), is the same as the PK11 policy in Section 3.3. Similar to the PK risk case studies, I assume the adversary knows the date of diagnosis within a five-day period, accounting for the separation between diagnostic test date and date of confirmed diagnosis, and search for generalization strategies that are likely to meet a PK11 risk threshold of 0.01. To evaluate the optimal SAP implementation, in terms of forecasting the influx of new disease case records, I also assume the data sharer can estimate the number of daily cases that will accrue in the coming week within ± 5 cases.

The reasonable adversary policy (**RAP**) protects against an adversary who knows a target individual's demographic information, but not their diagnosis date. Therefore, the quasi-identifier does not include date of diagnosis. This is likely a more reasonable assumption due to the difficulty of ascertaining a patient's exact date of diagnosis[41]. Since date of diagnosis is not included in the quasi-identifier, similar to the marketer risk scenario, the data sharer updates the generalization strategy of all records in the dataset at the end of each week, according to the cumulative number of records. The generalization policy adaptation is also constrained to represent demographic quasi-identifiers with equal or greater granularity than previous strategies determine, following a generalization path.

The marketer adversary policy (**MAP**) protects against the marketer attack in the same manner as in Section 3.3. Again, I estimate the marketer risk under the assumption the adversary has an identified dataset that covers every population resident - a worst-case scenario. Figure 3.16 displays the dynamic policy search results guiding the three dynamic policies for both Davidson and Perry counties. For Davidson county, I prioritize generalization strategies that preserve race and ethnicity granularity. Due to Perry county's racial and ethnic homogeneity (Table 3.2), for that county, I prioritize strategies that preserves age and sex granularity.

The first of the two real-world COVID-19 dataset policies is the ***k*-anonymous** policy used in the risk evaluation in Section 3.3, originally derived CDC's COVID-19 Case Surveillance Public Use Data with Geography[129]. In the utility evaluation, however, the *k*-anonymous policy is more similar to the CDC's original policy by sharing month of diagnosis, instead of date or week. Due to the generalized month of diagnosis, I assume the dataset is updated on the first day of each month. Also, for simplicity and to match the dynamic policy implementation, the *k*-anonymous policy again differs from the CDC's policy in that it does not strategically suppress quasi-identifiers to achieve 11-anonymity[42].

The **Marginal Counts** policy resembles the non-person-level data displayed in state COVID-19 dashboards[125] that have been used in several disparity investigations[107]. Though most racial data have been shared at the state level, for consistency with the other policies, I assume it shares county-level marginal counts for each race, ethnicity, age, and sex value, without preserving joint statistics. For example, the marginal counts for African Americans would be the daily counts of all African American cases, independent of ethnicity, age, and sex variation. I assume the dataset shared under this policy is updated on a daily basis. Table 3.5 summarizes the five de-identification policies' details.

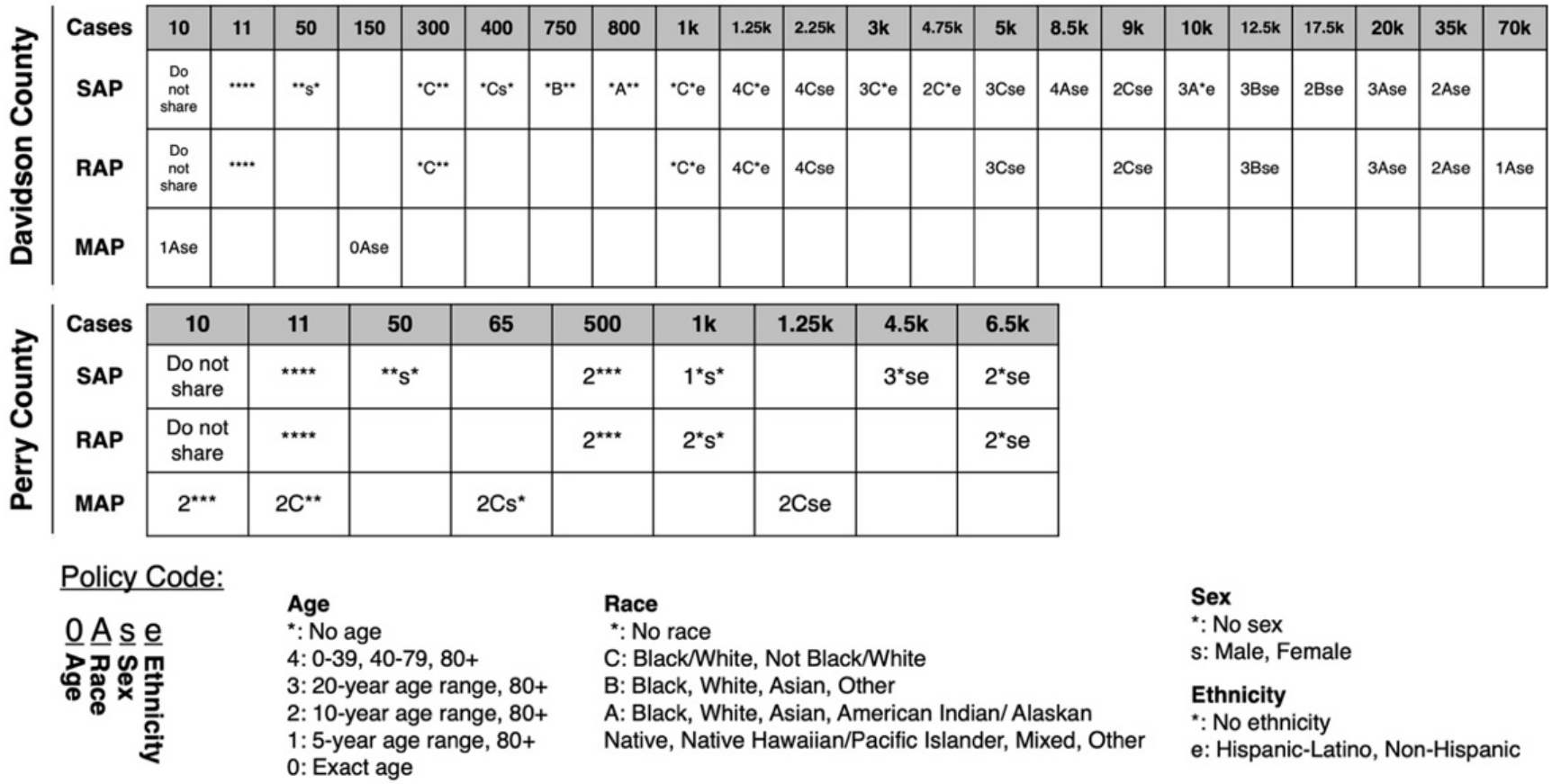

Figure 3.16. Dynamic policy search results for SAP, RAP, and MAP. The SAP and RAP strategies meet a PK11 threshold of 0.01, and the MAP strategies meet a marketer risk threshold of 0.01.

Table 3.5. Details of the de-identification policy assessed in this study.

	Strong Adversary Policy (SAP)	**Reasonable Adversary Policy (RAP)**	**Marketer Adversary Policy (MAP)**	***k*-anonymous**	**Marginal Counts**
Diagnosis date granularity	Date	Date	Date	Month	Date
Publication schedule	Daily	Daily	Daily	Monthly	Daily
Demographic generalization	Varies between time periods	Updated over time	Updated over time	Fixed	Fixed, single feature
Format	Row-level	Row-level	Row-level	Row-level	Daily counts by feature value
Includes comorbidity information	Yes	Yes	Optional	Yes	No
Assumed worst-case adversarial knowledge	Target individual's demographics and date of diagnosis	Target individual's demographics	Identified dataset of population residents	Target individual's demographics and date of diagnosis	NA

3.5.2 Simulating surveillance data

Labelling real world surveillance data for disparities can be both time consuming and arbitrary, such that outbreak detection is normally evaluated on simulated data[154]. For this evaluation, I generate partially synthetic data through constrained random sampling. It is partially synthetic in that the number of daily case records is informed by the Johns Hopkins University COVID-19 county-level tracking data[18], but how the records distribute across demographic subpopulations is simulated. Since a disparity manifests as an anomalous increase in the number of cases corresponding to a specific demographic subpopulation relative to the subpopulation's size[23,24], the baseline distribution is generated by randomly sampling individuals from the population, without replacement. To simulate a disparity, I disproportionately sample from the affected subpopulation.

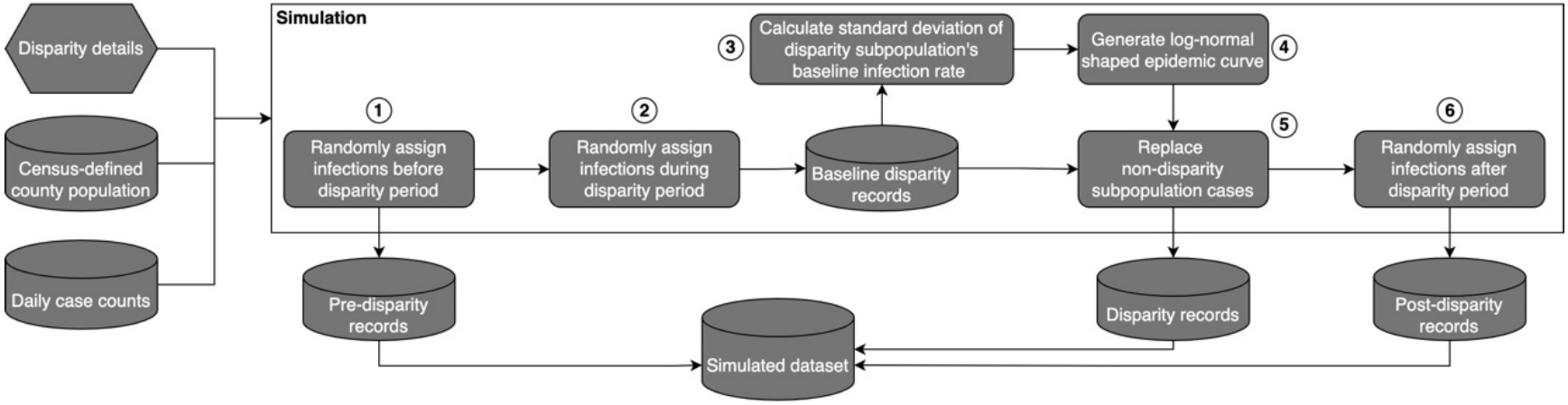

Figure 3.17. The pipeline for simulating disparity data in this study.

Figure 3.17 depicts the complete simulation process. A disparity is defined by a start date, peak date, duration, and subpopulation affected. In the simulation, all records are randomly sampled without replacement from the representative county population I generated from U.S. Census PCT12 tables[19] (see Section 3.2). I generate the baseline demographic distribution by randomly assigning which county residents are infected on each day leading up to (step 1) and throughout the disparity period (2). To simulate a disparity in the specified subpopulation, I first calculate the standard deviation of the subpopulation's baseline infection rate during the disparity period (3). I then generate a log-normal shaped epidemic curve[155] (4), whose values define the additional proportion of daily cases that need to correspond to the disparity subpopulation. For example, if the curve has a value of 0.2 on a given day, then an additional 20% of the day's records need to correspond to the disparity subpopulation. I rely upon a log-normal shaped curve, following the standard practice in the literature, to approximate real world epidemic curves[154,156]. The curve reaches its apex on the peak date, at a value set to four times the standard deviation of the baseline infection rate. This induces a disparity proportional to the subpopulations' baseline rate, peaking at a 99.9% significance level. In scenarios where no baseline cases correspond to the disparity subpopulation, and the standard deviation is zero, the peak value is set to a proportion value of 0.5. I then randomly replace records within the disparity period that do not belong to the disparity subpopulation with those that do, according to the proportion values defined by the epidemic curve (5). Finally, I continue baseline sampling for the remainder of the time series (6).

All simulated disparities are 45 days in duration, as the evaluation emphasizes early disparity detection, with an epidemic curve increasing rapidly to a peak on day 10 before decreasing slowly[155]. The affected subpopulation is defined as a combination of demographic values the Census provides for race, ethnicity, sex, and age. The definition includes up to one value for each of these four features. Since a disparity typically affects a range of ages instead of an exact age, I transform age into age groups ([0, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), [80, +]) when simulating and detecting disparities.

3.5.3 Disparity detection

Many outbreak detection algorithms have been developed as a consequence of the Defense Advanced Research Project Agency (DARPA) sponsoring the Bio-event Advanced Leading Indicator Recognition Technology (BioALIRT) project[157]. Unique among these algorithms is the What's Strange About Recent Events (WSARE) algorithm[158]. Designed for multivariate categorical data that includes both spatial and temporal information, such as that available in Limited data sets, WSARE combines association rule

mining, hypothesis testing, and randomization to detect significant patterns in surveillance data[159]. The result is an algorithm that both detects subpopulation outbreaks and explains the features (e.g., race, ZIP code, etc.) describing the outbreak group. WSARE's characteristics are unique as most outbreak detection algorithms, even state-of-the-art machine learning algorithms, either do not detect significant patterns in multivariate categorical data or do not explain the reason an alert was raised[157,160,161]. Moreover, WSARE has been implemented in several real world settings, including American and Israeli outbreak detection monitoring systems[158]. As such, WSARE provides the opportunity to detect disparate emerging disparities within COVID-19 surveillance data, and, therefore, to evaluate how well data sharing policies enable disparity detection.

For each time period in the dataset, WSARE searches for the most statistically significant increase in case records using a set of rules. The rule consists of a single value for one or more covariates. For instance, WSARE may return an alert indicating an unusually high number of records from October 10, 2020, that correspond to 20-30-year-old males. WSARE uses a greedy search to identify the most anomalous rule through a series of Fisher Exact Tests[162], comparing the current time period's records to baseline records at a user-defined statistical significance threshold. False positives due to multiple hypothesis testing are mitigated via randomization tests. Variations of the WSARE algorithm (namely, 2.0, 2.5, 3.0) apply different methods for defining baseline records[158]. Here, I employ WSARE 2.0 because it does not require extensive historical data (which are likely unavailable in novel pandemics). WSARE 2.0 generates a baseline from dataset records 35, 42, 49, and 56 days prior to the date of evaluation. I apply WSARE 2.0 to each de-identification policy. To further evaluate SAP, where the quasi-identifier generalization varies within the dataset, I additionally apply a variation of WSARE 3.0. This variation generates a baseline by randomly sampling up to 10,000 county residents from the U.S. Census population statistics.

I apply WSARE to the de-identification policies in the following manner. On each day in the WSARE 2.0 application to SAP, referred to as SAP 2.0, the quasi-identifiers in the current day's records and the baseline days' records are transformed to the most coarse version specified by the set of generalization strategies applied to those records. In the WSARE 3.0 application to SAP, referred to as SAP 3.0, the current day's generalized records are compared to the census-derived baseline. For both RAP and MAP, the records in the full dataset are transformed according to the current day's generalization strategy. To standardize this comparison between policies, I convert the *k*-anonymous policy's month of diagnosis to date of diagnosis by randomly assigning a date within the month to each record. I generate assignments by randomly sampling the date with replacement, where each date within the month is equally weighted. For the Marginal

Counts policy, I consider a single covariate that includes all race, ethnicity, age group, and sex values. Finally, for comparison, I apply WSARE 2.0 to the raw data.

3.5.4 Experimental design

I broadly evaluate how well each of the de-identification policies enables disparity detection at different significance thresholds in both Davidson and Perry County, TN. I simulate 50 datasets, each with the same two-component disparity starting on a different day – every 10 days from 5/10/2020 to 9/12/2021. For Davidson county, the two components are Black or African American race and age group [30, 40). Likely due to the racial and ethnic homogeneity of the county residents and the constraints of the simulation method, I was unable to simulate detectable disparities with a racial or ethnic component in Perry county. Therefore, for Perry county, the disparity components are Female sex and age group [30, 40). I apply WSARE at five different statistical significance thresholds (0.1, 0.05, 0.01, 0.005, 0.001) to each dataset, under each de-identification policy. I then measure the proportion of the datasets in which the disparity is detected. I consider the disparity detected if WSARE raises an alert within the disparity period, and the alert's feature value exactly matches or contains the true value. For instance, if the simulated disparity occurs in the [30, 40) age group and the data is shared under the *k*-anonymous policy, an alert for age group [18, 50) raised within the disparity period is considered an accurate detection. I also measure the time to detection, defined as the number of days since the start of the simulated disparity to the first date an alert is raised with correct demographic features. Note, the detection time considers the date at which the data is made available by the data sharing policy. If the disparity is not detected, I assign a detection time of 90 days, or twice the disparity duration. Finally, I measure how many false positives are generated. False positives are defined as an alert raised during the disparity period that does not have any of the correct features and any alert raised outside the period. Since WSARE 2.0 generates a baseline from records occurring up to 56 days prior to the evaluation date, I do not count false positives (for any WSARE implementation) prior to day 56 or during the first 56 days following the simulated disparity. This is done because a representative baseline cannot be acquired.

3.5.5 Evaluation results

I first measure the proportion of the 50 experiment datasets in which the disparity is accurately detected, at each statistical significance threshold. Figures 3.18 and 3.19 present the results for Davidson and Perry County, respectively.

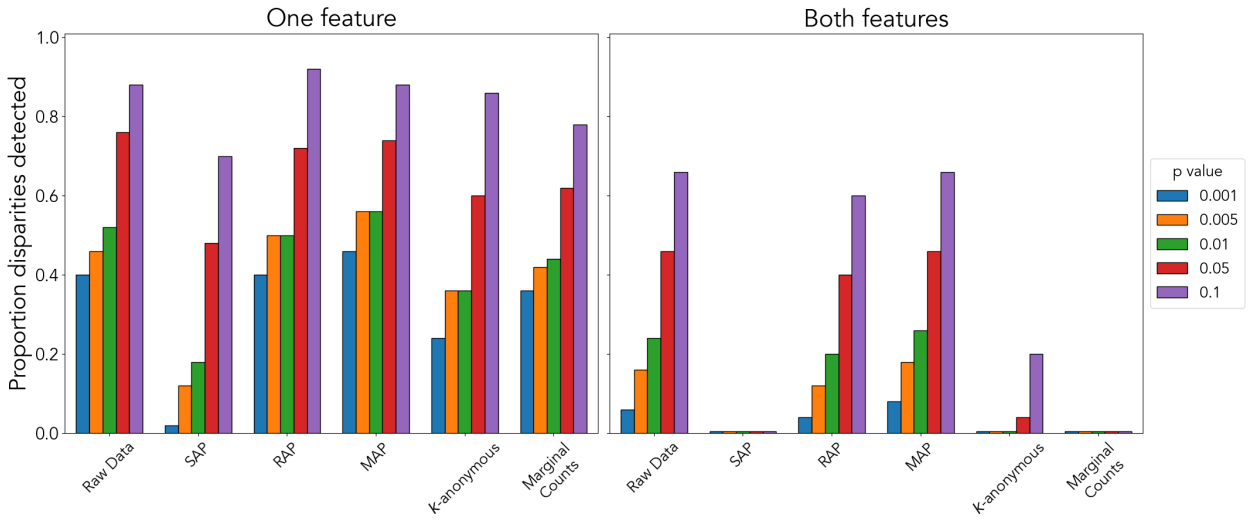

Figure 3.18. Proportion of detected disparities for Davidson County, TN, in which at least one of the simulated disparity features (left) and both features (right) are detected. The proportion is out of 50 different experiment datasets.

In Davidson County, RAP and MAP detect the greatest proportion of the simulated disparities. In some cases, RAP and MAP detect more disparities than the raw data. When detecting at least one of the features defining the demographic subpopulation within which the disparate infection rate occurs (30–39-year-old African Americans), the *k*-anonymous and Marginal Counts policies also detect a large proportion of the disparities across the significance thresholds. However, the *k*-anonymous policy detects both demographic features only 20% of the time at a 0.1 significance level, and the Marginal Counts policy's lack of joint statistics prevents the detection of both features entirely. The SAP 3.0 implementation detects one of the disparity features more often than the SAP 2.0. Yet, both implementations detect fewer disparities than the other policies, and neither detect both features.

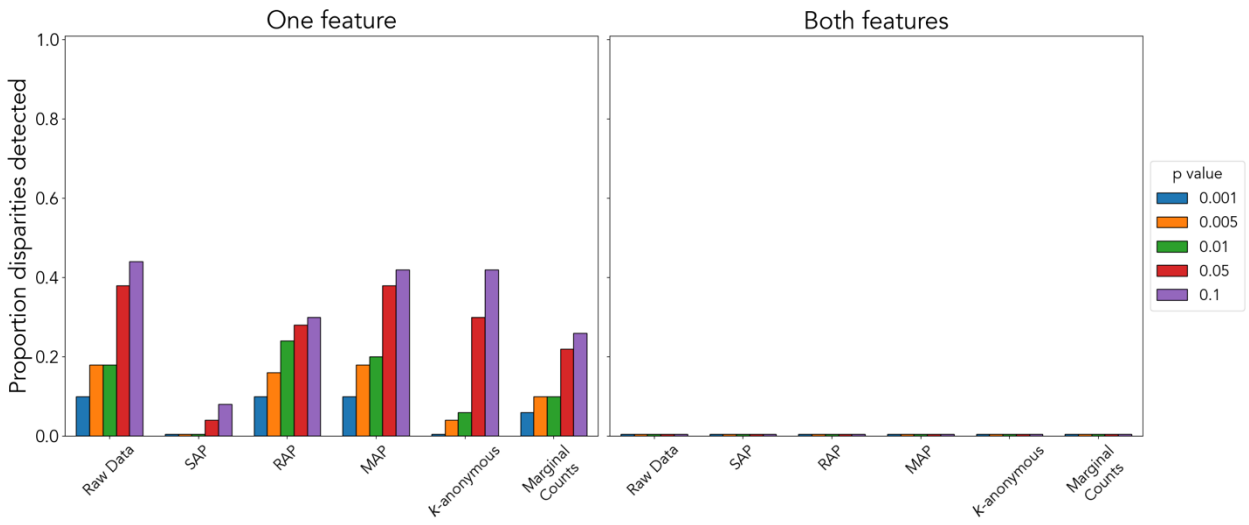

Figure 3.19. Proportion of detected disparities for Perry County, TN, in which at least one of the simulated disparity features (left) and both features (right) are detected. The proportion is out of 50 different experiment datasets.

In Perry County, MAP detects one of the disparity features (either Female or 30-39 years old) nearly as often as the raw data. The *k*-anonymous policy detects one of the features more often than RAP at statistical significance thresholds of 0.1 and 0.05 and less often at the other thresholds. SAP 2.0 did not detect any disparities, where SAP 3.0 detected one feature of less than 10% of the disparities at thresholds of 0.1 and 0.05. None of the de-identification policies, nor the raw data, enabled both disparity features to be detected in Perry County.

I next consider the detection times and false positives generated by each data sharing policy. I create Activity Monitoring Operating Characteristic (AMOC) curves by averaging the detection times and false positives for each policy at each significance threshold. A larger p-value threshold tends to decrease the detection time while increasing the false positive rate. A more significant threshold has the opposite effect. Thus, the results generate curves where the optimal value is a detection time of 1 day (1 day after the disparate infection rate began) with no false positives. Figures 3.20 and 3.21 present the AMOC curves for Davidson and Perry County, respectively.

For Davidson County, the RAP and MAP policies enable the shortest times to detect at least one and both simulated disparity features. The Marginal Counts policy provides comparable detection times for detecting only one disparity feature. The SAP 2.0 and SAP 3.0 implementations do not support detection of both features either. However, SAP 2.0 and SAP 3.0 enable, on average, similar detection times to the *k*-

anonymous policy while generating fewer false positives. This is because the *k*-anonymous policy's monthly publication schedule delays the time to detection, even though the *k*-anonymous policy detects more disparities than either SAP implementation.

For Perry County, MAP enables the earliest detection of at least one disparity feature, followed by RAP and the *k*-anonymous policy. Again, no policy enabled the detection of both disparity features, producing average detection times of 90 days.

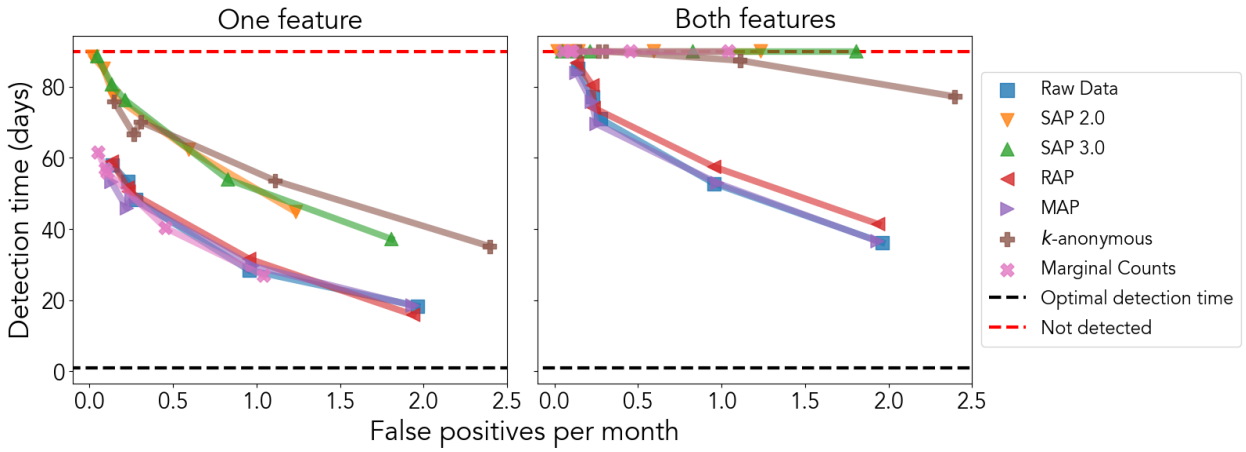

Figure 3.20. AMOC curves for Davidson County, TN, for detecting at least one of the simulated disparity features (left) and both features (right). Each point is the average of 50 different experiment datasets.

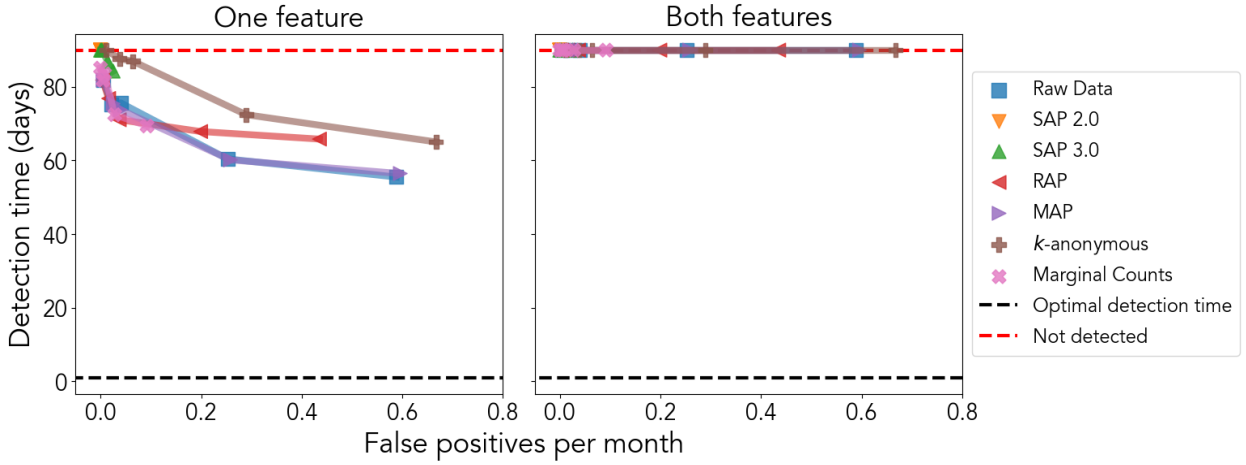

Figure 3.21. AMOC curves for Perry County, TN, for detecting at least one of the simulated disparity features (left) and both features (right). Each point is the average of 50 different experiment datasets.

3.6 Utility fairness evaluation

3.6.1 Evaluation overview

I evaluate fairness with respect to data utility in the context of detecting infection disparities. That is, I consider a de-identification method to fairly preserve data utility if the resulting data supports equal disparity detection performance across subpopulations. The smaller the difference between the proportion of disparities detected and the smaller the difference between the disparity detection times, the fairer the performance is considered.

In this experiment, I simulate 10 datasets with a single-component disparity for each of the race, ethnicity, age group, and sex values. There is one dataset for each of 10 dates spread across COVID-19's multiple waves. I apply WSARE to search for the best single component increase at a significance threshold of 0.05. I measure bias, or the lack of fairness, between subpopulations data utility by calculating the standard deviation across subpopulations' average proportion of disparities detected and average detection times. A smaller standard deviation indicates more fair disparity detection. I calculate both feature-specific standard deviations (e.g., race-specific deviations to measure racial bias) and standard deviations across all subpopulations. Additionally, I test for statistically significant differences in the detection performance, across all subpopulations, when sharing the data under the de-identification policies vs sharing the raw data. I do so with McNemar test and two-sided paired t-tests for the proportion of disparities detected and the average detection time, respectively. In each case, the null hypothesis is that the detection performance supported by the de-identification policies and the raw data is the same.

3.6.2 Evaluation results

3.6.2.1 Fairness of detection rates

Table 3.6 displays the McNemar test results for any statistically significant differences in the proportion of disparities detected between the de-identification policies and the raw data. The subpopulation-specific results are shown in Table 3.7.

Table 3.6. McNemar test results for the proportion disparities detected (p-values).

	Davidson	**Perry**
SAP 2.0	5.16×10^{-32}	3.31×10^{-24}
SAP 3.0	3.02×10^{-6}	1.32×10^{-23}
RAP	1	3.81×10^{-6}
MAP	1	0.774
k-anonymous	1.52×10^{-5}	0.0755
Marginal Counts	3.05×10^{-5}	1.53×10^{-5}

* Compared to raw data. Across all 200 simulated datasets.

In Davidson County, RAP, MAP, and the raw data enable detection of 90% of all the disparities. The *k*-anonymous and Marginal Counts enable the detection of 80% of all disparities. The SAP 3.0 implementation outperforms the SAP 2.0 implementation, detecting 70% of the disparities to SAP 2.0's 30%. The McNemar tests suggest there is insufficient evidence to reject the null hypothesis that the proportion of disparities detected under the RAP and MAP policies are similar to that of the raw data. Regarding the other de-identification policies, however, there is sufficient evidence to reject the null hypothesis, where the SAP 2.0 implementation produces the most significant p-value. In terms of supporting relatively similar detection rates across racial groups in Davidson County, SAP is the fairest with a standard deviation of the proportion of disparities detected across racial groups of 0.1. However, SAP does not detect as many age group disparities. The SAP implementations' differential performance between race and age group disparities reflects the dynamic policies' prioritization of racial and ethnic granularity in Davidson County. Across all subpopulations, SAP 3.0, RAP, MAP, and the *k*-anonymous are the fairest, with a standard deviation of 0.2.

In Perry County, only the MAP and *k*-anonymous policies produced p-values greater than 0.05 in the McNemar tests. The SAP implementations produced the most significant p-values. In fact, the SAP policy does not support disparity detection for almost any group. This is because SAP does not share many records due to excessively high privacy risks in the context of a strong adversary. The RAP and MAP's differential performance between racial disparities and age group disparities reflect the dynamic policies' prioritization for age group and sex granularity in Perry County. Though it detects fewer disparities overall, the *k*-anonymous policy enables the fairest detection rate across all subpopulations, with a standard deviation of 0.2.

Table 3.7. Proportion of disparities detected in each single-feature subpopulation.

		Davidson							Perry						
		Raw Data	SAP 2.0	SAP 3.0	RAP	MAP	*k*-anonymous	Marginal Counts	Raw Data	SAP 2.0	SAP 3.0	RAP	MAP	*k*-anonymous	Marginal Counts
Race	**Asian**	0.9	0.6	0.9	0.9	0.9	0.9	0.9	0.1	0	0	0	0.1	0	0.1
	American Indian/ Alaskan Native	0.6	0.6	0.8	0.6	0.7	0.7	0.6	0.2	0	0	0	0.2	0.2	0
	Black	1	0.7	1	1	1	0.9	1	0.2	0	0	0	0	0.3	0.1
	Mixed	1	0.9	1	1	1	0.7	1	0.2	0	0	0	0.3	0.2	0
	Native Hawaiian/ Pacific Islander	0.2	0.5	0.7	0.2	0.2	0.4	0.2	0	0	0	0	0	0	0
	Other	0.9	0.7	1	0.9	0.9	0.9	0.9	0	0	0	0	0.3	0.2	0
	White	0.9	0.7	1	0.9	0.9	1	0.8	0	0	0	0	0	0	0
	Average	0.8	0.7	0.9	0.8	0.8	0.8	0.8	0.1	0.0	0.0	0.0	0.1	0.1	0.0
	Standard deviation	0.3	0.1	0.1	0.3	0.3	0.2	0.3	0.1	0.0	0.0	0.0	0.1	0.1	0.0
Ethnicity	**Hispanic-Latino**	0.9	0.1	0.8	0.9	0.9	1	0.9	0.4	0	0	0	0	0.4	0.3
	Non-Hispanic	1	0.2	0.8	1	1	0.9	0	0	0	0	0	0	0	0
	Average	1.0	0.2	0.8	1.0	1.0	1.0	0.5	0.2	0.0	0.0	0.0	0.0	0.2	0.2
	Standard deviation	0.1	0.1	0.0	0.1	0.1	0.1	0.6	0.3	0.0	0.0	0.0	0.0	0.3	0.2
Age group	**[0, 10)**	0.9	0	0.6	0.9	0.9	0.9	0.9	0.7	0	0	0.7	0.7	0.5	0.6
	[10, 20)	0.9	0.1	0.6	0.9	0.9	0.5	0.9	0.6	0	0	0.6	0.6	0.7	0.6
	[20, 30)	0.9	0	0.4	0.9	0.9	0.8	0.9	0.7	0	0	0.7	0.7	0.3	0.6
	[30, 40)	1	0	0.4	1	1	0.4	1	0.5	0	0	0.5	0.5	0.5	0.5
	[40, 50)	1	0	0.5	1	1	0.3	0.9	0.6	0	0	0.6	0.6	0.6	0.6
	[50, 60)	1	0	0.6	1	1	0.8	0.9	0.7	0	0	0.6	0.7	0.5	0.6
	[60, 70)	1	0	0.4	1	1	0.6	1	0.7	0	0	0.7	0.7	0.7	0.6
	[70, 80)	1	0	0.5	1	1	0.5	1	0.6	0	0	0.6	0.6	0.6	0.6
	[80, 120)	0.9	0	0.5	0.9	0.9	0.8	0.8	0.4	0	0	0.4	0.4	0.4	0.4
	Average	1.0	0.0	0.5	1.0	1.0	0.6	0.9	0.6	0.0	0.0	0.6	0.6	0.5	0.6
	Standard deviation	0.1	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.1
Sex	**Female**	1	0.2	0.9	1	1	1	0.8	0.7	0	0.1	0.3	0.7	0.4	0.3
	Male	1	0.1	0.9	1	1	1	1	0.6	0	0.1	0.3	0.6	0.4	0.3
	Average	1.0	0.2	0.9	1.0	1.0	1.0	0.9	0.7	0.0	0.1	0.3	0.7	0.4	0.3
	Standard deviation	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0
All fields	*Average*	0.9	0.3	0.7	0.9	0.9	0.8	0.8	0.4	0.0	0.0	0.3	0.4	0.3	0.3
	Standard deviation	0.2	0.3	0.2	0.2	0.2	0.2	0.3	0.3	0.0	0.0	0.3	0.3	0.2	0.3

* Proportion is out of 10 experiment datasets

3.6.2.1 Fairness of detection times

Table 3.8 displays the paired t-test results for any statistically significant differences in the average disparity detection times between the de-identification policies and the raw data. The subpopulation-specific results are shown in Table 3.9.

In Davidson County, RAP and MAP enable the most similar detection times to the raw data. The paired t-tests comparing detection times between the de-identification policies and the raw data, produced p-values of 0.350 and 0.271 for RAP and MAP, respectively. All the other policies generated p-values < 0.0001. In terms of supporting relatively similar detection times across subpopulations, the *k*-anonymous policy is, on average, the fairest, with a standard deviation of 14.2 days. However, the detection times are longer than those for RAP and MAP. Notably, RAP and MAP support relatively fair detection times across subpopulations, except for AIAN and NHPI, the two smallest subpopulations in Davidson County.

In Perry County, the SAP implementations have the smallest standard deviations in average detection times. However, that is due to SAP broadly preventing disparity detection. Of the policies that generally detect disparities, MAP produces the most similar results to the raw data, with a p-value of 0.666, while the *k*-anonymous policy is the fairest. Across all subpopulations, the *k*-anonymous policy's standard deviation in detection time is 11.2 days. Regarding age group disparities, specifically, RAP and MAP support the fairest detection times.

Table 3.8. Paired t-test results for average detection times (p-values).

	Davidson	**Perry**
SAP 2.0	1.33×10^{-43}	2.43×10^{-23}
SAP 3.0	4.07×10^{-9}	5.68×10^{-23}
RAP	0.350	2.39×10^{-6}
MAP	0.271	0.666
k-anonymous	5.57×10^{-27}	1.17×10^{-8}
Marginal Counts	1.46×10^{-5}	2.63×10^{-6}

† Compared to raw data. Across all 200 simulated datasets.

Table 3.9. Average time to detect, in days, disparities in each single-feature subpopulation.

		Davidson							Perry						
		Raw Data	SAP 2.0	SAP 3.0	RAP	MAP	*k*-anonymous	Marginal Counts	Raw Data	SAP 2.0	SAP 3.0	RAP	MAP	*k*-anonymous	Marginal Counts
Race	Asian	14.8 [4.0, 54.0]	43.1 [3.9, 90.0]	14.9 [4.4, 54.0]	14.8 [4.0, 54.0]	14.8 [4.0, 54.0]	32.6 [13.8, 67.5]	15.3 [4.0, 54.9]	81.6 [43.8, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	82.0 [46.0, 90.0]	90 [90.0, 90.0]	81.6 [43.8, 90.0]
	American Indian/ Alaskan Native	42.5 [4.4, 90.0]	52.2 [9.2, 90.0]	29.3 [5.9, 90.0]	42.7 [4.4, 90.0]	33.2 [4.4, 90.0]	47.6 [20.2, 90.0]	40.9 [4.4, 90.0]	74.0 [9.9, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	73.7 [8.2, 90.0]	77.8 [28.4, 90.0]	90.0 [90.0, 90.0]
	Black	7.1 [3.4, 12.9]	33.4 [1.9, 90.0]	10.1 [6.4, 17.6]	7.1 [3.4, 12.9]	7.1 [3.4, 12.9]	34.9 [19.8, 67.5]	8.8 [4.4, 18.6]	74.5 [12.2, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	71.1 [23.0, 90.0]	82.0 [46.0, 90.0]
	Mixed	9.4 [3.0, 23.3]	19.3 [3.9, 61.6]	6.5 [3.4, 10.5]	11.3 [3.0, 23.3]	9.8 [3.0, 23.3]	49.1 [23.9, 90.0]	8.6 [3.0, 21.0]	74.3 [11.2, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	67.0 [10.0, 90.0]	77.8 [28.4, 90.0]	90.0 [90.0, 90.0]
	Native Hawaiian/ Pacific Islander	73.5 [7.2, 90.0]	54.4 [7.7, 90.0]	43.4 [7.7, 90.0]	73.5 [7.2, 90.0]	73.6 [7.8, 90.0]	67.5 [28.4, 90.0]	73.5 [7.2, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90 [90.0, 90.0]	90.0 [90.0, 90.0]
	Other	15.3 [2.4, 55.3]	30.5 [2.4, 90.0]	6.3 [1.9, 10.1]	14.5 [2.0, 54.9]	14.6 [2.0, 54.0]	32.9 [13.0, 66.1]	14.9 [2.0, 55.3]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	65.7 [7.8, 90.0]	77.8 [28.4, 90.0]	90.0 [90.0, 90.0]
	White	14.6 [4.4, 54.0]	31.9 [4.4, 90.0]	7.7 [4.9, 10.1]	14.8 [4.4, 54.0]	14.6 [4.4, 54.0]	28.4 [19.8, 38.6]	25.9 [5.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90 [90.0, 90.0]	90.0 [90.0, 90.0]
	Average	25.3	37.8	16.9	25.5	24.0	41.9	26.8	82.1	90.0	90.0	90.0	79.8	82.1	87.7
	Standard deviation	24.2	12.7	14.2	24.1	23.4	13.8	23.5	7.9	0.0	0.0	0.0	10.9	7.8	4.0
Ethnicity	Hispanic-Latino	15.1 [3.4, 54.9]	81.7 [44.4, 90.0]	23.6 [4.0, 90.0]	14.8 [3.4, 54.9]	15.1 [3.4, 54.9]	28.4 [19.8, 38.6]	15.7 [4.4, 54.9]	61.4 [10.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	65.8 [23.0, 90.0]	68.7 [10.0, 90.0]
	Non-Hispanic	5.8 [2.4, 9.6]	76.0 [18.7, 90.0]	23.3 [2.9, 90.0]	5.8 [2.4, 9.6]	5.8 [2.4, 9.6]	34.3 [19.8, 67.5]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	90 [90.0, 90.0]	90.0 [90.0, 90.0]
	Average	10.5	78.9	23.5	10.3	10.5	31.4	52.9	75.7	90.0	90.0	90.0	90.0	77.9	79.4
	Standard deviation	6.6	4.0	0.2	6.4	6.6	4.2	52.5	20.2	0.0	0.0	0.0	0.0	17.1	15.1
Age group	[0, 10)	14.7 [4.4, 54.9]	90.0 [90.0, 90.0]	40.9 [2.4, 90.0]	15.5 [4.4, 54.9[	14.7 [4.4, 54.9]	38.2 [22.4, 71.5]	14.9 [5.0, 54.9]	32.2 [4.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	33.7 [4.4, 90.0]	32.2 [4.4, 90.0]	52.6 [20.2, 90.0]	41.7 [5.4, 90.0]
	[10, 20)	15.0 [5.0 , 54.9]	84.2 [58.1, 90.0]	45.6 [8.8, 90.0]	15.3 [4.4, 54.9]	15.4 [5.0, 54.9]	57.4 [20.2, 90.0]	15.4 [5.0, 54.9]	40.0 [4.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	41.5 [3.9, 90.0]	40.0 [4.4, 90.0]	64.6 [10.3, 90.0]	42.5 [5.4, 90.0]
	[20, 30)	16.0 [4.4, 58.0]	90.0 [90.0, 90.0]	63.5 [7.9, 90.0]	15.3 [4.4, 54.9]	16.0 [4.4, 58.0]	41.5 [22.4, 90.0]	16.2 [5.0, 58.0]	35.4 [5.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	35.2 [5.4, 90.0]	33.9 [5.4, 90.0]	90 [32.2, 90.0]	44.0 [6.0, 90.0]
	[30, 40)	7.9 [5.4, 14.2]	90.0 [90.0, 90.0]	58.7 [8.4, 90.0]	8.2 [3.9, 14.6]	7.9 [5.4, 14.2]	66.5 [26.2, 90.0]	7.9 [5.4, 14.2]	48.7 [4.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	50.8 [3.4, 90.0]	48.7 [4.4, 90.0]	77.8 [23.0, 90.0]	48.9 [5.4, 90.0]
	[40, 50)	6.9 [4.4, 10.5]	90.0 [90.0, 90.0]	48.8 [2.9, 90.0]	7.4 [4.4, 11.1]	7.3 [4.4, 10.5]	70.5 [22.4, 90.0]	15.5 [4.4, 54.9]	40.9 [3.9, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	42.7 [3.9, 90.0]	40.9 [3.9, 90.0]	71.1 [23.4, 90.0]	41.4 [3.9, 90.0]
	[50, 60)	7.3 [5.4, 10.1]	90.0 [90.0, 90.0]	44.5 [6.9, 90.0]	7.5 [5.4, 10.1]	7.2 [5.4, 10.1]	41.9 [20.2, 90.0]	15.7 [5.4, 54.4]	32.0 [3.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	40.4 [3.4, 90.0]	32.6 [4.9, 90.0]	65.8 [20.2, 90.0]	41.7 [4.9, 90.0]
	[60, 70)	5.6 [2.9, 7.0]	90.0 [90.0, 90.0]	59.9 [3.8, 90.0]	5.8 [2.9, 7.5]	5.8 [2.9, 7.5]	53.8 [22.4, 90.0]	6.2 [2.9, 9.1]	32.1 [5.0, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	33.8 [5.0, 90.0]	32.1 [5.0, 90.0]	77.8 [19.8, 90.0]	41.4 [5.4, 90.0]
	[70, 80)	8.9 [3.8, 18.9]	90.0 [90.0, 90.0]	51.4 [6.0, 90.0]	8.2 [4.4, 18.0]	8.9 [3.8, 18.9]	60.9 [26.2, 90.0]	8.2 [3.8, 18.0]	39.7 [4.4, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	41.3 [4.4, 90.0]	39.7 [4.4, 90.0]	90 [20.2, 90.0]	39.9 [4.4, 90.0]
	[80, 120)	17.0 [3.4, 63.4]	90.0 [90.0, 90.0]	49.0 [2.9, 90.0]	17.0 [3.4, 63.4]	17.0 [3.4, 63.4]	40.5 [20.2, 90.0]	25.3 [3.4, 90.0]	57.3 [4.9, 90.0]	90.0 [90.0, 90.0]	90.0 [90.0, 90.0]	58.4 [4.9, 90.0]	57.0 [4.9, 90.0]	90 [20.2, 90.0]	58.3 [6.0, 90.0]
	Average	11.0	89.4	51.4	11.1	11.1	52.4	13.9	39.8	90.0	90.0	42.0	39.7	75.5	44.4
	Standard deviation	4.5	1.9	7.7	4.5	4.5	12.2	5.8	8.5	0.0	0.0	8.2	8.5	13.2	5.8
Sex	Female	7.4 [4.4, 14.0]	74.1 [10.2, 90.0]	25.6 [5.4, 67.9]	7.5 [4.4, 14.0]	7.4 [4.4, 14.0]	28.4 [19.8, 38.6]	24.3 [4.9, 90.0]	32.1 [3.2, 90.0]	90.0 [90.0, 90.0]	84.2 [58.1, 90.0]	64.8 [3.7, 90.0]	32.1 [3.2, 90.0]	69.1 [32.2, 90.0]	67.3 [6.4, 90.0]
	Male	6.8 [3.9, 13.7]	82.4 [48.2, 90.0]	29.3 [5.4, 68.8]	6.8 [3.9, 13.7]	6.8 [3.9, 13.7]	28.4 [19.8, 38.6]	10.2 [5.4, 18.2]	39.6 [4.0, 90.0]	90.0 [90.0, 90.0]	84.2 [58.1, 90.0]	65.2 [6.0, 90.0]	39.8 [4.9, 90.0]	69.1 [32.2, 90.0]	66.4 [7.0, 90.0]
	Average	7.1	78.3	27.5	7.2	7.1	28.4	17.3	35.9	90.0	84.2	65.0	36.0	69.1	66.9
	Standard deviation	0.4	5.9	2.6	0.5	0.4	0.0	10.0	5.3	0.0	0.0	0.3	5.4	0.0	0.6
All fields	*Average*	15.6	69.2	34.1	15.7	15.2	44.2	22.7	57.8	90.0	89.4	65.9	58.4	77.4	65.3
	Standard deviation	15.9	25.1	18.9	15.8	15.2	14.2	21.9	23.0	0.0	1.8	24.0	23.7	11.2	21.0

[†]Mean [95% quantile range]

3.7 Discussion

3.7.1 Dynamic policy approach

The aim of this work was to develop a de-identification method that supported timely publication of a de-identified pandemic registry while preserving the data's utility for public health research.

In this chapter, I introduced a framework to dynamically adjust data sharing policies to publicly share infectious disease surveillance data in a timely manner. The framework forecasts privacy risk according to the expected volume of new cases, enabling data sharers to prospectively adapt policies before seeing caseloads while incorporating the uncertainty of who will be infected in the future. In Section 3.3, I demonstrated how dynamically changing the policy per the framework's recommendations can maintain the privacy risk below the specified privacy risk threshold more frequently than statically applying a policy developed through retrospective de-identification methods, for both the PK and marketer risk-based approaches. In Section 3.5, I showed how dynamic policies designed with reasonable adversaries enable more timely and accurate detection of underlying disparities than data sharing policies derived from current, published COVID-19 datasets.

The dynamic policy approach is designed to maximize the utility of surveillance data for public health research and disease surveillance use cases. It does so by fluctuating data generalization with the infection rate to avoid the potential identity exposures or the loss of utility inevitably imposed by fixed data sharing policies applied to dynamic datasets. The dynamic policy approach also bypasses the delay of accumulating patient records before performing a risk assessment and shares dates of events. I showed how these last two features are crucial for effective disease monitoring[119,163], as they reduced the time to disparity detection. Furthermore, forecasting the privacy risk from population estimates enables greater consistency in quasi-identifier representation, as the policy can be maintained throughout the forecasted interval of time, and enables the data sharer to design a data sharing policy in the absence of the actual data. Moreover, predicting which policies provide sufficient privacy protection could potentially automate patient de-identification.

I demonstrated three approaches to dynamic policy adaptation. In the PK risk-based approach where it is assumed a strong adversary knows the target individual's diagnosis date within a window of time (the PK11 policy or SAP), I fixed county of residence and date of diagnosis granularity while increasing or decreasing the demographic granularity with the influx of new disease case records. I made this tradeoff to support

consistent data updates but acknowledge that it may induce certain data utility constraints. For instance, if an application requires uniform demographic granularity, the demographic values may need to be further generalized. An alternative dynamic policy approach could preserve the demographic granularity over time by using the privacy risk estimation framework's predictions to generalize the date of diagnosis into variably sized time windows. Still, this would impose a utility constraint on date information and cause the data publication schedule to vary. In the PK-risk based approach where it is assumed a reasonable adversary does not know the target individual's diagnosis date (RAP) and in the marketer risk-based approach (MAP), I showed how the dynamic policy can preserve date of diagnosis granularity while monotonically increasing the demographic granularity of the entire dataset over time. The weaker adversary increased the data sharer's ability to share more granular information over time. This, intuitively, follows the privacy-utility tradeoff underlying data sharing.

The disparity detection evaluation's results suggest that both in large, urban populations and small, rural populations, RAP and MAP can support better disparity detection performance than the data sharing policies derived from current, publicly available COVID-19 datasets. RAP and MAP detected a larger proportion of both single and double-feature disparities than the other policies, and with lower detection times. The *k*-anonymous policy's (in Sections 3.5-6) generalization of date of diagnosis induced uncertainty with respect to intramonth demographic variation in the dataset, broadly preventing the detection of more specific, multi-feature disparities. Its monthly data publication schedule also increased detection times. The Marginal Counts policy detected disparities in a timely manner, but its removal of joint distributions prevented the complete characterization of multi-feature disparities. Though SAP 3.0 outperformed SAP 2.0, it still provided suboptimal detection performance for both counties.

In this chapter, I evaluated several dynamic policies, each designed to meet a privacy risk threshold against adversaries with different types of background knowledge. I do not, however, advocate for which policy should be implemented in every case. This investigation showed how the privacy risk estimation framework's flexibility can inform different approaches to dynamic policy adjustment. Furthermore, the results highlight the importance of adversarial modelling in data sharing policy development and selection. If the adversary does not know (or cannot know) the COVID-19 diagnosis date of a target individual, the data sharer has the potential to share more granular information under RAP or MAP. If the adversary can reasonably obtain such information, SAP and the *k*-anonymous policies provide better privacy protection. The difference in disparity detection performance between these two groups highlights the need to investigate the likelihood an adversary can know the date if diagnosis, if they even know the complete demographic information[40,41].

3.7.2 Fairness in de-identification

The fairness evaluations, with respect to risk (Section 3.4) and utility (Section 3.6), highlight minority subpopulations' disadvantages in de-identified data. First, minority subpopulations may remain disproportionately exposed to re-identification compared to the majority group. When caseloads were high Davidson County, TN, around October 2021 and September 2021 (see Figure 3.10), the majority of records that fell into an equivalence class smaller than 5, 10, or 20 belonged to the Black, Asian, and (generalized) Other individuals (see Figure 3.14). This is because the majority of new disease cases corresponded to the White race, broadly increasing the racial subpopulations' equivalence class sizes and reducing their distinguishability.

Second, de-identification transformations can mask the evidence of health disparities. While the dynamic and real-world policies varied in their ability to support fair infection disparity performance, a consistent trend appeared: disparities were less frequently detected in smaller subpopulations, regardless of de-identification method. For example, disparities in the NHPI population in Davidson County were detected less frequently than disparities in other racial subpopulations, for each policy. And, across all subpopulations, fewer disparities were detected in rural Perry County than in urban Davidson County.

These disadvantages are a consequence of the fact that de-identification methods, by design, target the quasi-identifiers of the most unique individual records and, as the experiments in this chapter showed, the most unique records tend to correspond to minority subpopulations. Therefore, data stewards face an ethical dilemma when aiming to support public health research and an effective data-driven response to a pandemic: should the records that fall into the smaller equivalence class sizes be further distorted to preserve their privacy or should their granularity be maintained to preserve their representation? The former disproportionately removes records from minority subpopulations, who, in reality, suffered disparate outcomes during the COVID-19 pandemic[101]. The latter disproportionately exposes minority subpopulations to re-identification, whose communities, in reality, suffered discrimination during the COVID-19 pandemic[164,165]. Indeed, each subpopulation inherits its own privacy-utility tradeoff, and the tradeoffs are not equal[71]. And while the COVID-19 pandemic makes the implications of the tradeoff between fair privacy and fair utility more salient, the impact of this tradeoff extends, in our increasingly data-driven world, to our pursuit of health equity. Further complicating the matter is the fact that the fair privacy-fair utility tradeoff has received limited attention in the privacy community and therefore remains poorly understood[14,63,71].

3.8 Limitations and future directions

Despite the merits of this investigation, I wish to highlight several limitations to guide future extensions and transition into application. First, the dynamic, forecast-driven approach did not always meet the privacy risk threshold in the SAP, PK risk-based scenario. However, the framework's policy search results remained relatively robust. Policies chosen from forecasted counts were typically similar or close to those chosen from actual case counts. And when overestimating the number of cases, the privacy risk did not always dramatically exceed the threshold. Furthermore, I selected policies according to a 95% empirical confidence interval, but the policy search can readily incorporate larger confidence intervals as organizations deem desirable. Expanding the intervals further increases the likelihood the dynamic policy will meet the threshold in application. Moreover, when adjusting policies according to the actual case counts, the privacy risk never exceeded the threshold. Thus, the dynamic policy approach can be improved through more accurate forecasts and a model that accounts for potential case load overestimation.

Second, my approach did not incorporate suppression to protect the most unique patient records in the dataset. This is because it is nearly impossible to accurately forecast the exact records which will fall into small demographic groups. It is possible, however, during the enforcement of a selected policy (using the framework) to suppress actual patient records that need to be published and fall into population demographic bins corresponding to very few individuals, such as patient records that are population uniques, or patient records that correspond to population groups with fewer than k individuals (for PK risk). Such records with certainty would not meet the k-anonymity requirement. Additional risk analysis can be performed to estimate the risk of actual records in not meeting the k-anonymity requirement in a data release and suppress fields in records that are associated with a high estimated risk. Still, the framework's policy search and the policy selection approach depend on many adjustable parameters (e.g., the number of performed simulations, the expected number of new disease cases, the specific bins randomly selected to simulate new cases, the size of the quantile range used for the confidence a policy will meet a given risk threshold), which can be adjusted to mitigate the need for suppression.

Third, the privacy risk estimation framework depends on random sampling methods that may not realistically simulate the pandemic spread of disease. I assigned an equal likelihood of infection to all uninfected county residents at any given time in the simulations, and did not allow reinfections. In reality, the actual likelihood varies according to contact patterns of infectious individuals (i.e., through households or at work)[166,167], and reinfections are possible, though not likely in the case of COVID-19[168]. Still, I believe that Monte Carlo simulations, constrained to run within the relatively contained geographic region of a

county, provide a reasonable range and estimate of infection outcomes, as they have shown to be adept at simulating complex, high-dimensional patterns[169]. Further framework refinement should address the possibility of reinfection for diseases for which reinfection is more likely.

Fourth, the framework does not compute the re-identification risk of sharing a specific record. Rather, it estimates the range and expectation of privacy risk for a population. Future work should evaluate how well the framework's estimates compare to the re-identification risk of sharing actual disease surveillance data. Fifth, the utility evaluations in Sections 3.5-6 measured the ability to detect a disparity without quantifying how accurately the disparity was represented by the data sharing policy. Though data representation may be sufficient for accurate detection, implying the data sharing policy sufficiently preserves the representation of the underlying disparate trends, it is likely the data sharing policies still distort disparity features (e.g., severity or duration). Moreover, the simulated surveillance data did not consider potential simultaneous disparities in multiple subpopulations. Future work should consider more complex disparities and quantify how well data sharing policies preserve their features.

Fifth, my experiments using simulated data did not consider the effect of suppressing values and missing data on disparity detection. As *k*-anonymity is often achieved in practice through suppression[42] and real-world data is rarely complete, future work should quantify the robustness the policies' performance under suppression and varying levels of missingness.

Sixth, the data utility evaluations in Sections 3.5 and 3.6 relied on a single outbreak detection algorithm. It is possible that other outbreak detection algorithms improve performance and fairness. Notably, however, as discussed in Section 3.5, most outbreak detection algorithms were not designed to detect disparities in categorical data. Anomaly detection algorithms, from the statistical process control-based methods commonly applied by public health agencies to the state-of-the-art deep learning methods, often rely on univariate count data. Of the outbreak detection algorithms that take advantage of multivariate count data, most focus on monitoring disease spread in time and space with granular geolocation information[161,170]. Outbreak detection algorithms designed to detect changes in demographic subpopulations within categorical data are few, and even fewer are those that indicate which subpopulation experiences the outbreak[160]. In fact, to the extent of my knowledge, the only algorithm, other than WSARE, that combines association rule mining, hypothesis testing, and explainable disease surveillance is Neill and Kumar's Multidimensional Subset Scan (MD-Scan)[171]. Alternatively, different statistical methods, such as regression[24], could be used to identify temporal disparities. Future work should apply alternative algorithms and methods to more broadly evaluate the data share policies' ability to preserve underlying disparities.

Finally, I focused my evaluation on disparities within counties while only briefly comparing performance between two counties. The difference in Davidson and Perry county performance suggests all five data sharing policies were unfair in terms of providing similar disparity detection performance between counties. Future work should analyze performance differences between all counties in a state or country.

3.9 Conclusion

Disease surveillance data is variable, between geographic areas and over time. As such, data must be frequently updated and de-identified in a manner that incorporates such dynamics. To support disease monitoring and disparity investigations by public health researchers and the general public, the data must also contain granular date information. The privacy risk estimation framework I introduced enables a prospective approach to surveillance data de-identification. In contrast to traditional methods, prospective policy selection offers increased flexibility to support near-real time data dissemination. I showed that forecast-driven de-identification offers better privacy protection than the static data sharing policy application when applied to a pandemic registry that increases in size at a variable rate. Moreover, I showed that when protecting against a potential adversary of reasonable strength – an adversary who, at most, knows a target individual's complete demographic information – dynamic policy de-identification enables timely publication of person-level data that preserves evidence of underlying disparities better than current public datasets. As such, dynamic policy de-identification has the potential to support the detection and characterization of disparities, and the investigation of their sources, in current and future pandemics.

Furthermore, the results of this work highlight a tradeoff between fairly distributing the privacy risk and fairly distributing the data utility when sharing de-identified data. The lack of understanding of this tradeoff and its significant societal implications serve as the motivation for the remainder of this dissertation.

Chapter 4

Data-based constraints to fairness in de-identified data

4.1 Introduction

The COVID-19 pandemic was a single example of the modern data economy's demands for ever growing amounts of data. The current enthusiasm around artificial intelligence requires larger and more diverse datasets to develop algorithms that are high performing, just and fair[172,173]. Data is also more frequently being recognized as a public good[174], such that US National Institutes of Health's recently updated their Data Management and Sharing Policy in a manner that "establishes the expectation for maximizing the appropriate sharing of scientific data generated from NIH-funded or conducted research".[175] The push for more data sharing, and the reliance on de-identification to support large-scale data sharing, makes the need to understand the potential privacy risk and data utility inequities of de-identification even more urgent.

However, as described in Section 2.6 and despite the shared understanding among privacy researchers that more distinguishable populations have a less favorable privacy-utility tradeoff[71], the fairness of de-identification is not well understood[13]. Only recently has the privacy community begun to seriously investigate the interplay of fairness and de-identification, where most investigations empirically measure the extent to which de-identification data transformations degrade algorithmic fairness[71]. Several investigations have sought to formalize the tradeoff between privacy and fairness; however, they treat privacy as a binary outcome instead of a varying level of protection[16]. This limits the ability to define the relationship between fair privacy risks and fair data utility, and how one may have to be sacrificed for the other. Moreover, the investigations have missed the larger question: is it even possible to simultaneously equalize privacy risk and data utility, and, if not, how do we respond?

In this chapter, I formalize the relationship between achieving fair privacy risk and fair data utility between records when de-identifying a dataset via generalization and suppression. I show that it is, in fact, impossible to achieve equal privacy and utility across groups in nearly all scenarios. I further illustrate the constraints and consequences defined by the impossibility theorem in the context of the *k*-anonymity privacy model. Finally, I discuss the ethical implications of the impossibility theorem and propose a data sharing model that supports sharing more representative data without sacrificing privacy protections.

4.2 The fairness tradeoff theorem

In this section, I formally prove that it is not possible for generalization and suppression to simultaneously equalize the re-identification risk and data utility across records in a dataset that has unequal re-identification risks prior to the implementation of these techniques. The proof is the first to explicitly consider both fairness with respect to privacy and fairness with respect to utility, and to formalize the tradeoff between the two. I begin the formalization by establishing several definitions.

Definition. Dataset.

Let D be a dataset of n records $\{d_1, d_2, \dots, d_n\}$ where each record corresponds to a single individual. Each record contains values for m attributes, $\{A_1, A_2, \dots, A_m\}$. Suppose there are no missing values for any $d \in D$.

As described in Section 2.2, quasi-identifying attributes contribute to the re-identification risk. A record's quasi-identifier is defined as its set of quasi-identifying attributes.

Definition. Quasi-identifier.

The quasi-identifier of D, Q_D, is the subset of attributes in D, $\{A_i, \dots, A_j\} \subseteq \{A_1, \dots, A_m\}$, that can distinguish individual data subjects in a manner that can support re-identification.

A record's re-identification risk is inversely proportional to its distinguishability with respect to the quasi-identifier. The more records with the same quasi-identifier value, the less distinguishable the records, and the lesser their re-identification risk.

Definition. Equivalence class.

An equivalence class, e, is the set of records that share the same quasi-identifier value.

Let $\mathcal{E}(D)$ be the set of equivalence classes of a dataset D. I assume that all records $d \in D$ that are suppressed share an equivalence class $e_* \in \mathcal{E}(D)$. Finally, I define $|e_*| \equiv |D| + 1$.

Definition. Re-identification risk.

Let e be the equivalence class of d and $|e|$ be the size of (number of records pertaining to) e. The re-identification risk of d is $\frac{1}{|e|}$.

To reduce a record's re-identification risk, the quasi-identifier values of a subset of records in the dataset can be generalized to coarser representations, such that more records share its quasi-identifier value and therefore the record belongs to a larger equivalence class. Entire records can also be suppressed such that they are not present in the de-identified dataset. Generalization and suppression effectively reduce records' distinguishability; however, this comes at the cost of data granularity. Hence, the privacy-utility tradeoff.

A standard assumption in current generalization and suppression strategies is that of deterministic transformations, such that records with the same quasi-identifier value in the raw data (i.e., records that belong to the same equivalence class prior to de-identification) are transformed in the same manner by the de-identification function. This makes it so that records that belong to the same equivalence class before de-identification also belong to the same equivalence class after de-identification.

Definition. Data transformation.

Let $\phi(d)$ be a function that transforms attributes of a record $d \in D$. I say that ϕ is *determinisitic* if for all $d, d' \in D$ with $d = d'$, $\phi(d) = \phi(d')$. I let $\phi(D)$ denote a dataset created by applying ϕ to each record d_i

Let $F(d)$ return the size of the equivalence class of a record d. Similarly, $F\big(\phi(d)\big)$ will return the size of the equivalence class of the record d transformed by ϕ, i.e., $\phi(d)$. Thus, if d is suppressed, $F(d) = |D| + 1$ (and note that $|\phi(D)| = |D|$).

Lemma 1.

Suppose that ϕ is a deterministic function. Then for any $e \in E(D)$ there is $e' \in E\big(\phi(D)\big)$ such that $e \subseteq e'$. Moreover, $F(d) \leq F\big(\phi(d)\big)$ for all $d \in D$.

Proof:

Suppose that there is $e \in E(D)$ and $d, d' \in e$, and $\phi(d) \in e'$ but $\phi(d') \notin e'$ for some $e' \in \mathcal{E}\big(\phi(D)\big)$. Then $\phi(d) \neq \phi(d')$. However, since $d, d' \in e$, $d = d'$, which means that ϕ is not deterministic, a contradiction. Since $e \subseteq e'$ for every $e \in E(D)$ and some $e' \in E\big(\phi(D)\big)$, it follows immediately that $F(d) \leq F\big(\phi(d)\big)$ for all $d \in D$.

■

Definition. De-identification.

A deterministic function ϕ is a de-identification function if $\exists\ d \in D$ such that $F(d) \leq F\big(\phi(d)\big)$.

By its very nature, applying a de-identification function leads to utility loss. There are many ways to measure the utility loss. Here, I apply a domain-agnostic definition of the utility loss to a record $d \in D$ from applying a de-identification function ϕ as

$$UL(d_i; \phi) = F\big(\phi(d)\big) - F(d) \tag{4.1}$$

In other words, a record's utility loss is defined as the difference between the size of a record's equivalence class after and before being transformed by a deterministic de-identification function. Note that by Lemma 1 $UL(d_i; \phi) \geq 0$ for any deterministic ϕ.

I define utility loss in this manner for several reasons. First, the difference in group size is the same as measuring the information-theoretic entropy of the de-identified data compared to the raw data, except that this measure does not include the log transform. Measuring utility loss as entropy is commonly applied in practice when optimizing de-identification transformations, as it is monotonic (i.e., the entropy increases with increased generalization and suppression) and it considers the differential utility loss between groups within a non-uniform distribution[6,74] – both critical for this fairness evaluation. Second, it measures utility loss in terms of the data's intrinsic utility, which estimates the data's global utility independent of a specified use case. Third, it intuitively measures the extent to which de-identification transformations have diluted a record's presence in its equivalence class. For example, assume 2 records pertain to 20-year olds and 10 records pertain to 21-year olds in the raw dataset. If age is generalized by ϕ such that all 12 records become part of the same equivalence class defined as 20-21 year olds, most of those records actually pertain to 21-year olds. The presence of the 2 records pertaining to 20-year olds has been diluted by de-identification more than the presence of the 10 records pertaining to 21-year olds. Finally, this definition of utility loss generalizes to any utility loss measure that monotonically increases with the difference in a record's equivalence class size before and after transformation, such that the theorems also generalize to such measures.

Next is the key impossibility result.

Theorem 1. Impossibility of simultaneous fair risk and fair utility.

Let ϕ be a deterministic de-identification function and suppose that $\exists\ d, d' \in D$ such that $F(d) \neq F(d')$. Then either $F\big(\phi(d)\big) \neq F\big(\phi(d')\big)$ or $UL(d;\phi) \neq UL(d';\phi)$.

Proof:

Suppose $F(d) \neq F(d')$, and both $F\big(\phi(d)\big) = F\big(\phi(d')\big)$ and $UL(d;\phi) = UL(d';\phi)$. Then $UL(d;\phi) = F\big(\phi(d)\big) - F(d) = UL(d';\phi) = F\big(\phi(d')\big) - F(d')$. Since $F\big(\phi(d)\big) = F\big(\phi(d')\big)$, this implies that $F(d) = F(d')$, a contradiction.

■

I further strengthen this impossibility result by showing that it holds even up to approximations. This is formalized next.

Theorem 2. Fairness tradeoff.

Let $\gamma \geq 0$ and $\epsilon \geq 0$. Suppose that $\exists d, d' \in D$ with $|F(d) - F(d')| \leq \gamma$, and let ϕ be a deterministic de-identification function such that $\big|F\big(\phi(d)\big) - F\big(\phi(d')\big)\big| \geq \epsilon$. Then $|UL(d;\phi) - UL(d';\phi)| \geq \epsilon - \gamma$.

Proof:

$$
\begin{aligned}
|UL(d;\phi) - UL(d';\phi)| &= \Big|\big(F\big(\phi(d)\big) - F(d)\big) - \big(F\big(\phi(d')\big) - F(d')\big)\Big| \\
&= \Big|\big(F\big(\phi(d)\big) - F\big(\phi(d')\big)\big) - \big(F(d) - F(d')\big)\Big| \\
&\geq \Big|\big(F\big(\phi(d)\big) - F\big(\phi(d')\big)\big)\Big| - |F(d) - F(d')| \\
&\geq \epsilon - \gamma
\end{aligned}
$$

■

I visually graph the fairness tradeoff theorem in Figure 4.1. The only time the risk and utility can be simultaneously equalized between two records in a dataset is when they start with the same re-identification risk (i.e., when γ=0). Otherwise, either the records' risk can be equalized (when ϵ=0) or the records' utility can be equalized (when $|UL(d;\phi) - UL(d';\phi)|$=0). Furthermore, the greater the difference is between their initial privacy risks, the greater the potential disparity in risk or utility after de-identification.

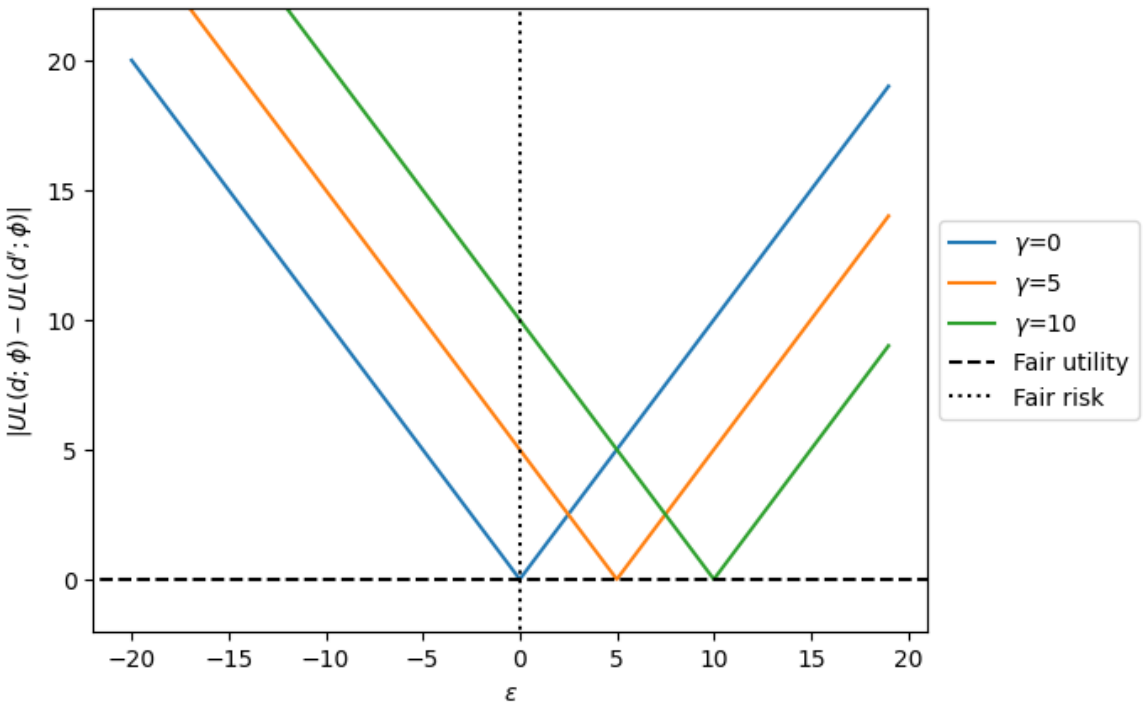

Figure 4.1. Visualization of fairness tradeoff theorem. Here, I assume $|F(d) - F(d')| = \gamma$ and $\left|F\left(\phi(d)\right) - F\left(\phi(d')\right)\right| = \epsilon$.

4.3 Empirical illustration of the fairness tradeoff theorem

I further illustrate the constraints to simultaneously equalizing privacy risk and data utility using the *k*-anonymity privacy model[42] (see Section 2.3). I use the *k*-anonymity model for its legal precedent, intuitive structure, and its natural implementation via generalization and suppression. Moreover, *k*-anonymity applies an equal privacy risk upper bound to all records, facilitating the demonstration of the differential utility loss that occurs when prioritizing fair privacy protections. Nevertheless, I highlight that the fairness constraints defined by the fairness tradeoff theorem extend beyond a single privacy model.

This evaluation measures disparities in utility loss and privacy risk between racial subpopulations when *k*-anonymizing the United States population. The analysis begins by identifying several ways generalization and suppression can disproportionately degrade the minorities' data utility in the context of the *k*-anonymity model. Specifically, I evaluate the overall utility loss and distribution of utility loss across racial subgroups while varying the value of k for *k*-anonymization, while varying the proportion of records that can be suppressed, when *k*-anonymizing a uniformly distributed population (i.e., when all records start with the

same re-identification risk), and when varying the manner in which the race variable is generalized. I then evaluate the differential risks each racial subgroup must assume to achieve equal data utility after de-identification.

4.3.1 Experiment parameters

I again represent the demographic distribution of the US population according to the 2010 Decennial Census PCT 12 tables[19] (see Section 3.2.3). Each table contains joint statistics by age, sex (Male/Female), and 5-digit ZCTA (used as a proxy for 5-digit ZIP code in our analysis) for each of seven different race values defined by the Census. Several tables also include joint statistics by ethnicity, but, here, I restrict the quasi-identifier to {*age, sex, race, and ZIP code*}. The race values include American Indian or Alaskan Native (AIAN), Asian, Black, Native Hawaiian or Pacific Islander (NHPI), White, Mixed, and Other. I combine all race-specific counts for all 50 states plus Puerto Rico to create the final dataset.

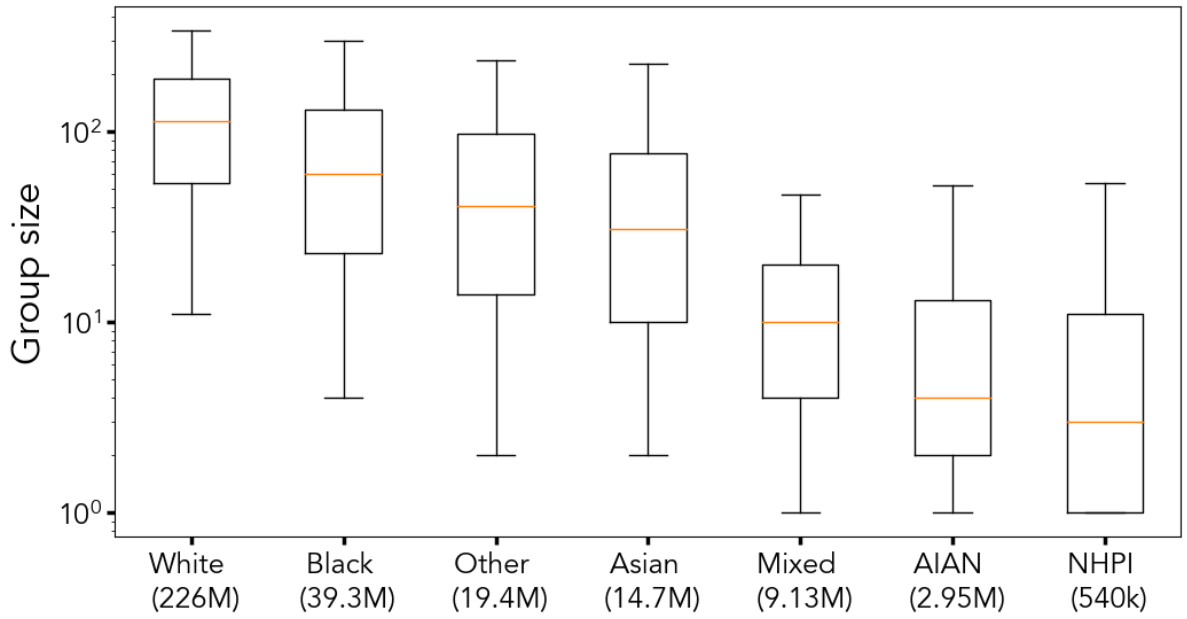

Figure 4.2. Distribution of group sizes for each race in the United States population, per the 2010 Decennial Census. Group size is defined as the number of individuals with the same set of values for race, age, sex, and ZIP5. Re-identification risk is inversely proportional to the group size. The numbers in parentheses indicate the number of United States residents corresponding to each race. For each distribution, brackets denote 95% confidence interval, boxes denote inter-quartile range, and orange line denotes median value. AIAN = American Indian or Alaskan Native; NHPI = Native Hawaiian or Pacific Islander.

The fairness tradeoff theorem shows that risk and utility cannot be equalized when records start with different privacy risks. Different starting points are likely to occur in nearly all populations, including between racial subgroups in the US population, as shown in Figure 4.2.

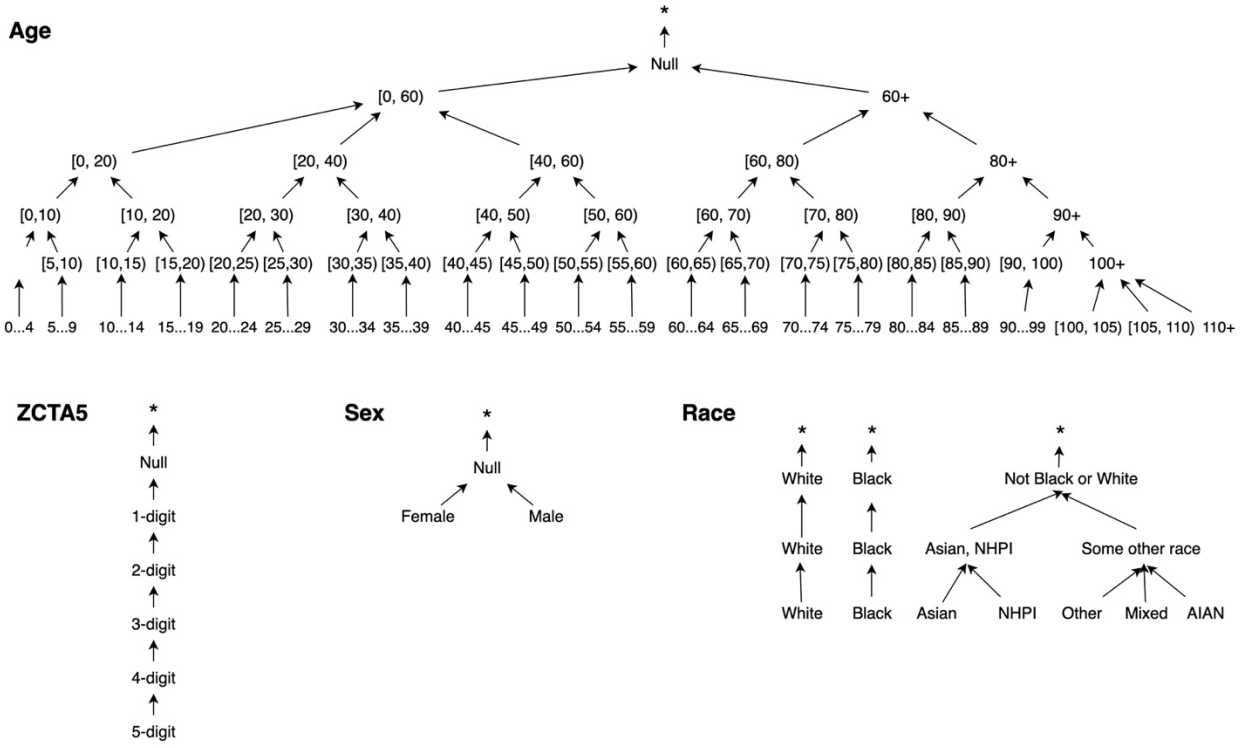

Figure 4.3. Generalization hierarchies for the four quasi-identifying attributes in this chapter, where the race and age generalization hierarchies vary slightly from those used in Chapter 3 (see Figure 3.4). Differing from Figure 3.4, the "*" symbol denotes suppressing the record entirely from the dataset. Note, the Race hierarchy does not allow the Race attribute to be generalized to a null value.

To implement *k*-anonymity, I defined generalization hierarchies for each of the quasi-identifying attributes, displayed in Figure 4.3. I searched for optimal generalizations, following the hierarchies, with three different algorithmic approaches that are standards in practice and the data privacy literature[139]. The first implements the OLA algorithm without suppression[6] (see Section 2.5). Recall, the OLA algorithm applies global recoding, which consistently generalizes all records to the same levels in the hierarchy, where local recoding may vary generalization levels between subgroups of records (e.g., some records have 5-year age intervals while others have 10-year age intervals). OLA identifies the globally optimal global recoding

according to a monotonic utility loss measure. The utility loss measure is normalized entropy as defined in Eqn. 4.2, below. The second *k*-anonymization approach also implements OLA, but this time allowing 1% of all records to be entirely suppressed. The third approach implements the Mondrian algorithm, which uses a greedy search to approximate the optimal local recoding for the dataset[80]. The combination of implementations allows for the comparison between suppression and no suppression as well as between global and local recoding.

I measure utility loss according to a normalized version of entropy as defined by Gionis and Tassa[6,74]. This is the same utility loss measure as defined in theorems above; however, we apply the log transform here to be consistent with the original entropy measure's definition. Let dataset D contain n records $\{d_1, d_2, \dots, d_n\}$. Let $F(d_i)$ denote the size of d_i's equivalence class prior to de-identification and $F\left(\phi(d_i)\right)$ denote the size of d_i's equivalence class after de-identification by a deterministic de-identification function ϕ. The utility loss is defined as:

$$\frac{\sum_{d_i \in D} -\log_2\left(\frac{F(d_i)}{F\left(\phi(d_i)\right)}\right)}{n} \tag{4.2}$$

I normalize the measure by the number of records to be able to compare values between racial subgroups of different sizes.

4.3.2 Evaluating the effect of k

I first measure 1) the overall utility loss and 2) the inequality in utility loss between racial subgroups as the Gini coefficient of each group's respective entropy, while varying *k.* Figure 4.4 shows that increasing the value of *k* increases the overall utility lost in the dataset for all three algorithmic approaches. Increasing *k* also decreases the inequality in race-specific utility loss. This result is intuitive, as increasing the value of *k* requires more records from the less distinguishable subgroups to be distorted, equalizing with the distortion experienced by the smaller, more distinguishable subgroups. The results also highlight a tradeoff between overall utility and fairness with respect to utility when anonymizing a dataset, as the *k*-anonymizations with the fairest distribution in utility also induced the greatest utility loss. The Mondrian algorithm induces the least amount of utility loss, via local recoding, and also the greatest inequality in

utility loss. Global recoding with suppression induces less utility loss and more utility loss inequality than global recoding without suppression.

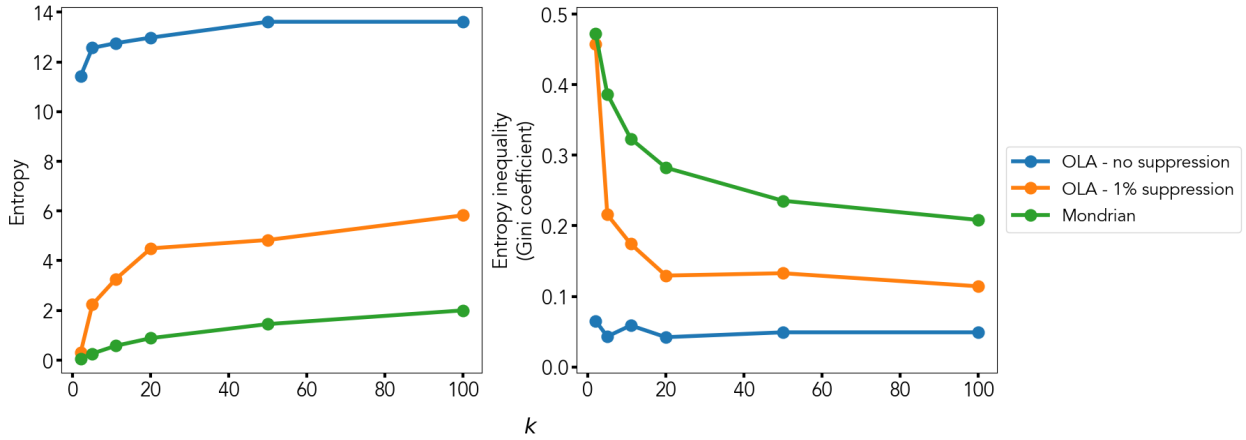

Figure 4.4. (Left) Overall utility loss, measured as entropy, when applying each *k*-anonymization implementation at *k* values of {2, 5, 11, 20, 50, 100}. (Right) Inequality in utility loss between racial subgroups measured as the Gini coefficient of the race-specific utility loss values (one for each of the seven racial subgroups defined in the US Census). The results show a tradeoff between minimizing the overall utility loss and minimizing the inequality of utility loss.

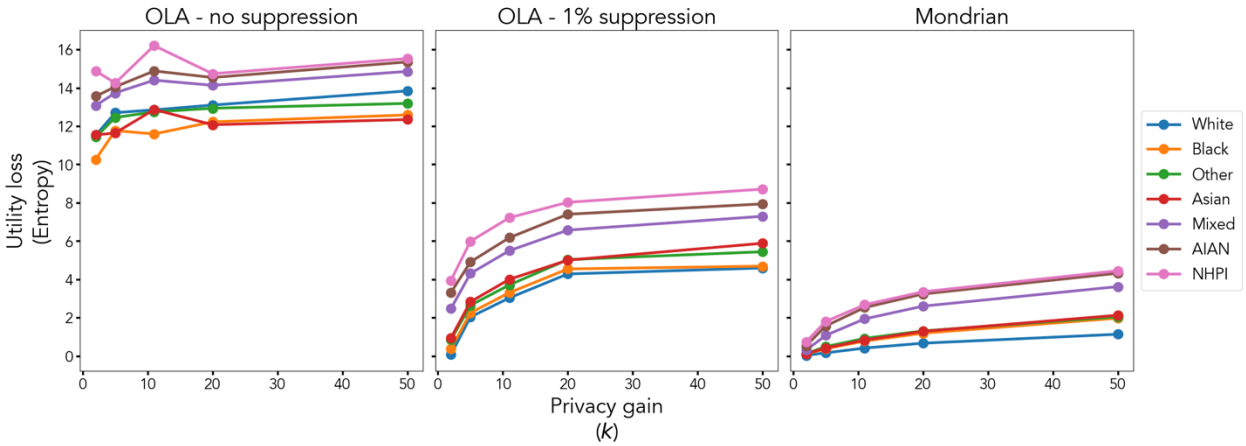

Figure 4.5. Race-specific privacy-utility curves when k-anonymizing the United States population on the features: race, age, sex, and ZIP code. Utility loss is measured as entropy (Eqn. 4.2), which measures the divergence between the original data and the transformed data[6]. Privacy is gained at increasing values of *k*. Points correspond to *k* values {2, 5, 11, 20, 50} – thresholds found in current state and federal guidance. (Left) OLA k-anonymization algorithm with no records suppressed. (Center) OLA algorithm with up to 1% of all records suppressed. (Right) Mondrian *k*-anonymization algorithm. AIAN = American Indian or Alaskan Native; NHPI = Native Hawaiian or Pacific Islander.

Figure 4.5 shows how each racial subgroup experiences a different privacy-utility tradeoff, represented as curves, when *k*-anonymizing the full US population. Each point on the curve indicates the average utility

loss records within a racial subgroup incur when being de-identified to a particular value of *k.* To simultaneously achieve fair privacy and fair utility, all race-specific curves must cross at the same point. Instead, they almost never cross; and when they do, only two or three curves cross at a time. In terms of utility distribution, the smaller and more distinguishable racial minorities lose more utility than the majority subgroup, across *k* values and algorithmic implementations. Again, since *k*-anonymity equalizes privacy risk, the racial subgroups' utility loss cannot be equalized.

4.3.3 Evaluating the effect of suppression

I next evaluate how suppression affects different racial subpopulations. Here, I apply OLA at *k*=11 (CMS' standard[59]) while varying the proportion of records within the dataset that can be suppressed by the algorithm. The suppression thresholds vary between 0 and 0.1. Figure 4.6 shows that increasing the amount of suppression generally increases the data utility (or decreases entropy/utility loss). Utility loss increases at a suppression threshold of 0.1 because at that point the algorithm can achieve *11*-anonymity without any generalization – all records in equivalence classes smaller than 11 are suppressed – and the suppression induces greater utility loss. However, the benefit in utility generally comes at the cost of fairness in utility. Figure 4.7 provides a more detailed description of the racial subgroup-specific effects when increasing the proportion of total records suppressed. The minority racial subgroups are more likely to be suppressed than the majority and subsequently lose more utility than the majority.

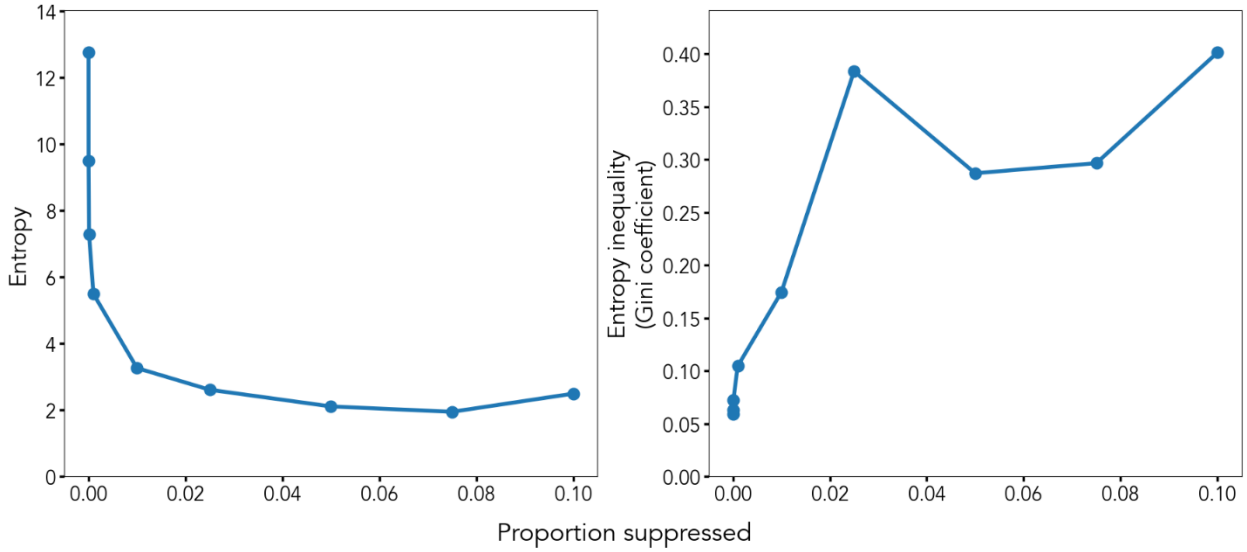

Figure 4.6. Overall utility loss measured as entropy (left) and inequality in utility loss between racial subgroups (right) when varying the proportion of records suppressed. Each de-identification applies the OLA algorithm at k=11 to the US population.

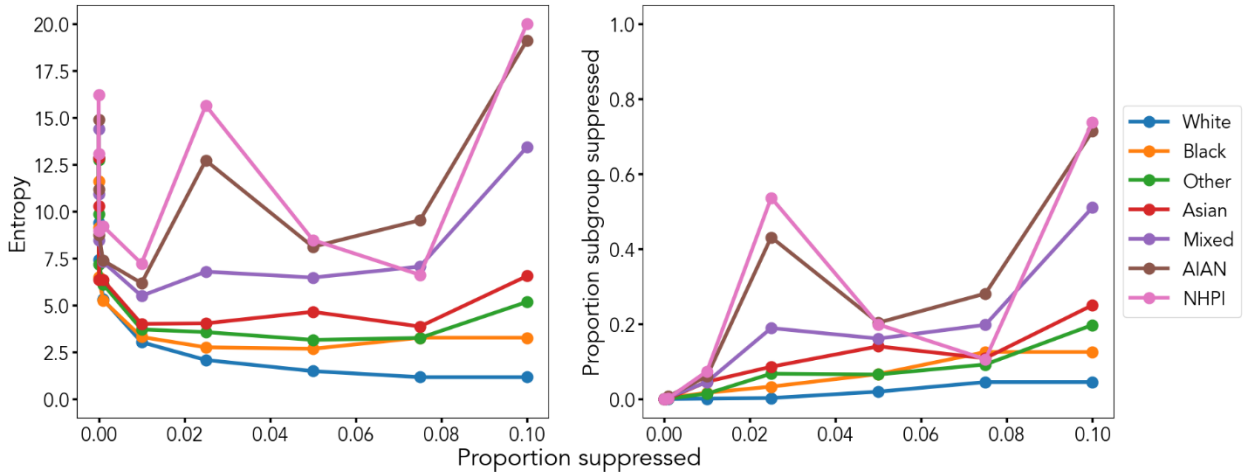

Figure 4.7. Race-specific utility loss (left) and proportion of racial subgroup's records (right) when varying the overall proportion of records suppressed. Each de-identification applies OLA algorithm at k=11 to the US population.

4.3.4 Evaluating inequalities in a uniform distribution

The fairness tradeoff theorem states that the risk and utility of records can only be equalized when they start with equal re-identification risk. Here, I measure utility loss inequalities when *k*-anonymizing a uniformly distributed/distinguishable population.

I simulate a uniform distribution by first creating a dataset with all the race, age, and sex values available in PCT12 tables and 1,000 distinct ZIP code values (0-999). I then assign 3 individuals to each combination of race, age, sex, and ZIP code. Finally, I *k*-anonymize the uniform population in the same manner as before. Figure 4.8 shows the race-specific utility loss values at varying levels of *k.*

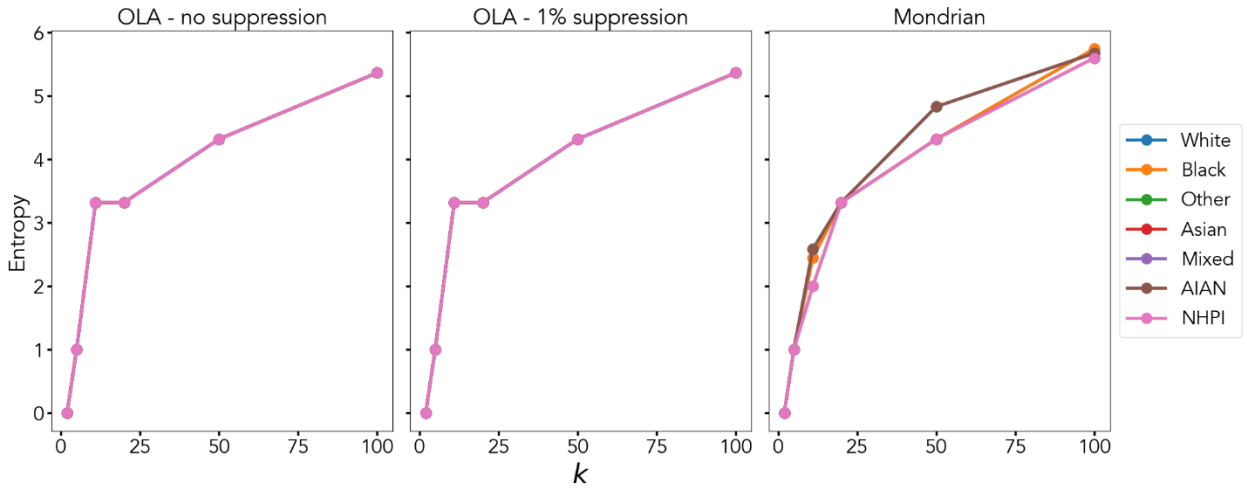

Figure 4.8. Race-specific privacy-utility curves when *k*-anonymizing a uniformly distributed population on the features: race, age, sex, and ZIP code. Points correspond to k values {2, 5, 11, 20, 50}.

The OLA implementations with and without suppression do not induce unequal utility loss between racial subgroups. They are also identical, as the records that are suppressed by the algorithm are those that are anomalously distinguishable. Since all records start with the same re-identification risk in this population, OLA does not suppress any records. Notably, however, the Mondrian algorithm unequally distributes data utility at some values of *k.* Specifically, Asian and NHPI subgroups have different utility than the Black and White subgroups, who have a different utility from the AIAN, Mixed, and Other subgroups. This is due to a combination of the Mondrian algorithm's greediness[80] and it locally recoding the dataset according to the imbalanced generalization hierarchy for race (Figure 4.3). The dataset is partitioned on the race attribute according to generalization level second from the bottom. This separates Black and White into their

individual races while grouping Asian and NHPI together and AIAN, Mixed, and Other together. This result highlights the potential for generalization hierarchies, which may be defined semantically or in a data-driven manner[176], to potentially bias the utility distribution in a de-identified dataset.

4.3.5 Evaluating the effect of race generalization hierarchies

Given the results of uniform distribution experiment, I investigate how race generalization hierarchy structure can influence the distribution of utility across racial subgroups when *k*-anonymizing the US population. Here, I define four different race generalization hierarchies, shown in Figure 4.9. The first, applied in all experiments thus far, attempts to preserve the semantics across race generalizations. For example, NHPI individuals may be more likely to have similar ancestry with Asian individuals than other groups. However, I note that the semantics are difficult to define and may differ between use cases, such that this hierarchy enforces one type of semantically driven generalization. Second, I test what happens when combining the larger and smaller subgroups via generalization. The most extreme combination generalizes White race, the largest and least distinguishable subgroup, with NHPI, the smallest and most distinguishable subgroup, via generalization. Third, I test how utility fairness is affected by removing the race values entirely, mimicking the concept of fairness via unawareness[88]. Finally, I test how the algorithms respond when enforcing no race generalization; records must contain the most specific values or be suppressed entirely.

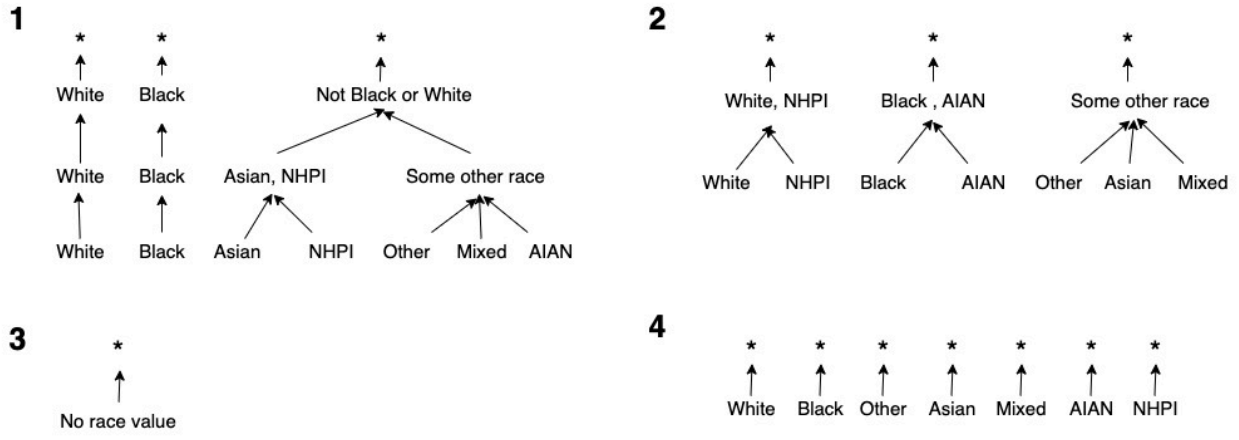

Figure 4.9. Four distinct race generalization hierarchies considered in our experiment. "*" symbol denotes suppressing the record entirely from the dataset.

As shown in Figure 4.10, the results are varied. With global recoding and no suppression (OLA – no suppression), forcing the algorithm to preserve the most specific race values produces the most equal distribution of data utility while removing the race variable produces the most inequality. However, the results are mixed across *k*-anonymity implementations and *k* values, and in no scenario is perfect utility equality achieved. Nevertheless, removing the race variable does not remove utility inequality among racial subgroups, highlighting the limitation of fairness via unawareness in the context of de-identification.

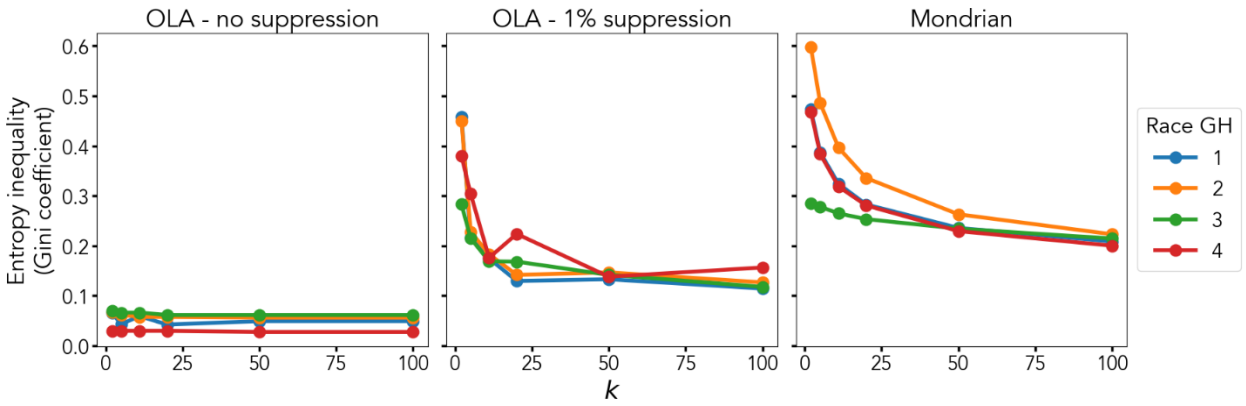

Figure 4.10. Inequality in utility loss between racial subgroups when applying different race generalization hierarchies (shown in Figure S7) to each *k*-anonymization.

4.3.6 Evaluating risk inequality when equalizing utility loss

Finally, I take a different perspective to de-identification in which I equalize utility loss between racial subgroups while allowing their re-identification risks to vary. To do so, I *k*-anonymize each racial subgroup in the US population dataset independently on the quasi-identifying variables {*age, ZIP code, sex*}. To establish a utility loss upper bound, I first *100*-anonymize ($k = 100$) the majority subgroup, or the White race subgroup. I then use a binary search to find the maximum value of *k* at which each of the other racial subgroups meets that utility loss upper bound. Figure 4.11 displays the maximum re-id risk ratios for each group, where the ratio is calculated as one over the subgroup-specific *k* value divided by 1/100 (the risk of the White subgroup). In most cases, the non-White racial subgroups assume about ten times as much risk as the White subgroup. In some cases, the racial minorities assume 100 times as much risk as the White subgroup, meaning that they are *k*-anonymized to a *k* value of 1. In other words, in such cases, there is no generalization that achieves the same utility as the White *100*-anonymized subgroup.

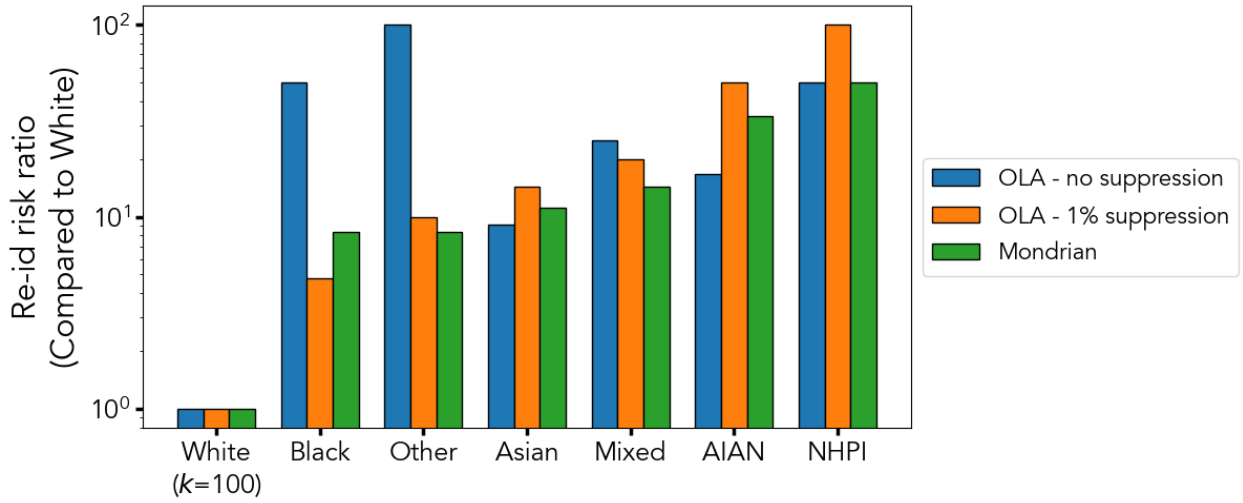

Figure 4.11. Re-identification risk ratios when independently *k*-anonymizing each racial subgroup such that they retain at least as much utility as the *100*-anonymized White subgroup. Re-id risk ratio is calculated as one over the *k* for the subgroup divided by 1/100 (the *k* for the White subgroup).

4.4 Ethical implications of the fairness tradeoff theorem

More distinguishable populations – such as racial and ethnic minorities, members of the LGBTQ community, those who live in rural areas, and other underrepresented groups – have a less favorable privacy-utility tradeoff than the majority population – a fact that cannot be remedied with better optimizations. Such populations must either retain high privacy risk to be represented in the dataset or have their representation significantly distorted if they are included at all. Given that individual privacy rights are a prominent component of US and international law and that no comparable regulations protect an individual's right to data utility, practitioners repeatedly prioritize privacy. As a consequence, minority's lack of representation can lead to inequitable benefits and interventions. For example, France's prohibition of collecting data on individual's race and ethnicity limited the country's ability to identify vulnerable populations during the COVID-19 pandemic[177]. Data transformations can mask health disparities in underrepresented minorities, as shown in Chapter 3 and by Xu et al.[14,15] in two different studies. Furthermore, even though the formalization of the problem in this chapter focuses on generalization and suppression, there exists growing empirical evidence that alternative data transformation strategies are similarly constrained. For instance, differential privacy has been shown to also mask health disparities[14] as well as distort minorities' representation in the 2020 Decennial Census[16] in a manner that could lead to

inequities in Census-guided intitiatives[17]. Sharing a fully synthetic replica of a dataset can also exacerbate inequities in artificial intelligence performance[25].

While current privacy legislation motivates practitioners to prioritize equal privacy protections over equal data utility, navigating the ethical implications is more complex. Let us analyze it in the context of the guiding research principles outlined by the Belmont Report: respect for persons, beneficence, and justice[178].

First, the principle of respect for persons includes recognizing individuals' personal dignity and autonomy. In terms of data sharing, this principle is regulated by requiring individuals' informed consent to share their identified data. Their de-identified data, however, can be shared without their consent, as de-identified data is legally not considered personal data, evidently on the ground that the privacy risks are sufficiently small that consent is not required. But were a data steward to prioritize equal representation when de-identifying a dataset, which necessarily allows the privacy risk to fluctuate between groups, at what point are minority groups' records still considered "de-identified"? If the overall privacy risk of a dataset meets standard de-identification thresholds but the subgroup-specific privacy risks do not, should those patients' consent be required to share their data?

Second, the principle of beneficence denotes the obligation to maximize benefits while minimizing harm. In terms of privacy-preserving data sharing, preserving minorities' representation in the data set maximizes their potential benefit while preserving their privacy minimizes their potential harm from privacy intrusions. However, prioritizing minorities' privacy protections over their data utility can also lead to potential harms. For example, when de-identification prevents health disparities from being detected, the disparities are not addressed. As argued by Faden et al.[179], there is an ethical obligation to optimize health care for all individuals through meaningful research. The inherent tradeoff between fair privacy and fair utility in de-identified data makes it difficult to maximize benefits while minimizing harms.

Third, the principle of justice, as defined by the Belmont Report, is achieved by fairly distributing the risk and benefits across groups. The fairness tradeoff theorem clearly states that both cannot be concurrently equalized.

4.5 Rethinking privacy-preserving data sharing

Alternative data sharing methods are required to resolve the ethical dilemma imposed by the data-based constraints to fair de-identification. Here, I discuss several.

First, minority populations could be financially compensated for their privacy or utility loss. Steed et al.[17] recently analyzed how the application of differential privacy to the 2020 Census can lead to inequities in Census data-guided educational funding allocations. They showed that differential privacy is likely to distort the data in a way that school districts with more racial and ethnic minority students would receive less funding. To resolve this problem, they proposed, among other solutions, to financially compensate school districts that are expected to lose funding as the result of differential privacy transformations. While this may alleviate some of the inequities of de-identification, an economic solution is unlikely to accommodate all data sharing scenarios as proper compensation may not be feasible to define or fulfill.

Another solution is to engage communities in the data sharing process[174]. For example, representatives from minority populations could be involved in determining an acceptable privacy-utility tradeoff for their communities in the de-identified data[71]. Respecting data sovereignty in this manner has proven beneficial for several data sharing initiatives, such as the All of Us Research Program's consulting with Tribal nations[180]. Nevertheless, it remains possible for data transformations to become so unfavorable, either in risk or benefit, that communities would prefer not to be included in the dataset at all, further biasing representation. Indeed, relying solely on data transformations to protect individuals' privacy severely constrains the ability to share representative data. Without additional privacy protection measures that do not require data distortion, it may not be possible to do so.

To mitigate the unfairness of data transformation strategies, I propose that privacy practitioners, policy makers, and data sharing initiatives should supplement data transformations with additional sociotechnical mechanisms that deter users from invading individuals' privacy in the first place. If no intrusion is attempted, patients' privacy is equally protected, even when distinguishability varies across groups. Examples of such mechanisms include requiring users to a sign a data use agreement and constraining users' access to within a monitored environment. While implementing sociotechnical deterrents has legal[27] and practical[180] precedents, the additional privacy protections they provide come at the expense of data accessibility. Obtaining access becomes more cumbersome and fewer users will gain access compared to sharing the data in the public domain. The challenge then becomes how to facilitate data access to trustworthy users while sufficiently disincentivizing bad actors from misusing the data.

To improve data accessibility while implementing sociotechnical deterrents to privacy intrusions, we propose a "passport-visa" credentialling model, an extension of the passport models proposed by Dyke et al.[181] and implemented by the All of Us Research Program[180]. In the passport-visa model, each potential data user must first obtain a passport from a sponsoring institution. For example, this could be a researchers' home institution. The sponsoring institution provides the individual with a passport according to their credibility. Credibility can be established through various sociotechnical mechanisms, such as requiring training on ethical research principles, requiring sponsorship from previously credentialed users, and/or contractual obligations. After obtaining a passport, the individual can then apply for a visa from the organization possessing the data of interest. The visa is granted according to the individual's and sponsoring institution's credibility, which may involve additional deterrents to data misuse. Upon receiving the visa, the individual then obtains access to the data resource. To standardize appropriate thresholds for credibility, we further propose that a generally agreed upon agency establish minimum standards for passport-visa approvals. The standards would serve as starting point, upon which passport- and visa-granting organizations could build upon as they see fit, to prevent negligence in granting data access. The standards would also adjust to the sensitivity of different data types, such that passports and visas are granted according to tiered levels of access (similar to different levels of security clearance or drivers' licenses).

To effectively limit data access to users who have a visa, the passport-visa model necessarily limits data access to controlled environments and frameworks. Otherwise, as is true with the public domain model, the data could be untraceably shared with anyone. Examples of controlled environments and frameworks include a centralized access model, where users can only access the transformed data within a monitored environment[180], and blockchain solutions to track data usage[182]. Controlling the analytics environment also enables the set of potential data operations a user can perform to be restricted, which could wholly prevent certain types of misuse.

While the passport-visa model limits data access compared to a public domain model, it is not as cumbersome as traditional access-control models. Traditionally, potential users must apply for access to each data resource, individually. Stewards of the resource must also define their own requirements for access and determine each potential user's credibility, often with limited background information[181]. By contrast, a generally recognized passport, granted according to general standards, would establish baseline credibility to any visa-granting organization. This would reduce the visa-granting organization's administrative burden and risk when granting a visa. The data sharing model may also limit the

administrative burden for the data user and the passport-granting institution, where the passport can be applied for and renewed on a regular or semi-regular basis.

The passport-visa model also addresses several key components for effective deterrence in terms of implementing certain, severe, and swift consequences for data misuse[183]. First, by restricting access to controlled environments and frameworks, the model makes it easier for data misuse to be detected. Being able to reliably detect data misuse validates the threat of recourse in the event of a violation.

Second, the passport-visa model can support the imposition of diverse penalties for privacy violations. For one, a violator can lose their visa and passport. Without a visa a user cannot access a single resource, but without a passport a user could not access any resource implementing the passport-visa model. The broader the adoption and reach of the passport, the greater the penalty a violator experiences for losing their passport. Further, it may be desirable to allow the person whose privacy was violated to seek legal redress from the visa-granting institution, the data user, and the data user's sponsoring institution. The visa-granting institution could also seek redress from the user and their sponsoring institution, and the sponsoring institution could seek redress from the user. The levels of potential recourse increase the magnitude of penalties levied against the organizations and especially against the user. This increases the passport-visa model's leverage to deterring data misuse, as the potential benefit a user could gain from an intrusion is unlikely to outweigh the penalty.

4.6 Discussion

The aim of this work was to define the constraints to simultaneously achieving the fair distribution of privacy risk and data utility in a de-identified data set and then to discuss the implications of and potential solutions to such constraints.

In this chapter, I formally proved that it is impossible to concurrently equalize risk and utility between records that start with differing re-identification risks in the original data – a condition that is likely to be true in nearly all real-world datasets. I empirically illustrated the fairness constraints by applying the *k*-anonymity model to the United States population and measuring race-specific disparities in data utility and privacy risk. The smaller the racial subpopulation, the greater was their utility loss (when equalizing privacy risk) or their privacy risk (when equalizing utility loss) compared to the majority population. While I focused on differences between racial subpopulations for their prevalence in health outcomes, the fairness

constraints apply to any quasi-identifying attribute that is unequally distributed within a population. The fairness constraints also apply to different methods of de-identification, as empirical investigations for differential privacy and synthetic data suggest.

Navigating the fair privacy-fair utility tradeoff becomes more complex when realizing that the subpopulations most disadvantaged by the fairness constraints are also those who generally suffer worse health outcomes and discrimination. And while legal obligations to protect individuals' privacy currently govern how de-identified data is transformed, competing ethical obligations demand we reevaluate how to share data for biomedical research. The solution I proposed is to combine de-identification transformations with sociotechnical deterrents to privacy intrusions, through a passport-visa credentialling model, such that data utility can be more fairly distributed without sacrificing privacy protections. The passport-visa credentialling model is more scalable, more generalizable, and facilitates access to trustworthy data recipients better than traditional controlled-access models.

4.7 Limitations and future directions

I highlight several limitations to this work. First, the fairness tradeoff theorem defined privacy risk as re-identification risk and utility loss as a derivative of entropy. I defined privacy risk in this manner because of its prominence in privacy regulations: HIPAA and other privacy regulations do not apply to de-identified data, or data for which the re-identification risk is low. Even though fairness investigations related to differential privacy and synthetic data have illustrated similar constraints[16,25] and these data transformation strategies consider different types of privacy intrusions, it is possible that other types of privacy risks are not constrained in the same manner as formalized by the impossibility theorem. Furthermore, it is possible other types of utility measures can be equalized with the privacy risk. One example is Latanya Sweeney's precision metric (PREC)[47,184], which measures the extent to which records are generalized with respect to the levels in each generalization hierarchy. It is trivial for every record in a data set to be generalized to the same level via global recoding (supporting equal utility as defined by PREC) and be *k*-anonymized to the same level of *k* (supporting equal privacy risk). However, PREC was initially designed for *k*-anonymity optimization and neglects the changes in the distribution of the data caused by generalization and suppression[6]. As such, I used a derivative of entropy. Still, it may be possible to equalize privacy risk and utility among records when utility is measured with respect to a particular use case, such as ML prediction. Future work should consider different types of privacy risks and specific use cases.

Second, the passport-visa credentialling model, like any controlled access model, introduces additional constraints and potential inequities to privacy-preserving data sharing. For instance, the passport-vias model may be expensive to implement. Data accessibility and data user privacy (as users' actions and credentials would be tracked) are also limited by the passport-visa model. Moreover, obtaining a passport and visa may be easier for investigators from some subpopulations than others, which could create inequities in data access[71]. I argue that the constraints defined by the impossibility theorem and their ethical implications may require that accessibility and *user* privacy be reduced to obtain fairer *data subject* privacy and utility. However, accessibility and user privacy have their own tradeoffs. The public domain data sharing model represents one end of the spectrum, where anyone can access a published dataset and user's actions and identities are never stored. The passport-visa model represents a preliminary solution that seeks to balance the data subjects' needs and the data users' needs. Future work should develop sociotechnical mechanisms and data sharing models that optimize the many tradeoffs inherent to data sharing, while reducing inequities in data access and prioritizing data subjects' utility and privacy.

Third, the passport-visa credentialling model cannot guarantee privacy protection. As with any credentialling system or deterrence mechanism, there is a risk that users defraud the system. There is also a risk that users will act irrationally[183]. There have been several theoretical investigations into the privacy protection afforded by sociotechnical systems[63–65]. There are also several real-world examples of successful implementations of sociotechnical deterrents, such as MIMIC[67], which only requires users to sign a data-use agreement to gain access. Nevertheless, future work should develop and validate sociotechnical mechanisms' ability to deter privacy intrusions.

Finally, the fairness tradeoff theorem assumed de-identification transformations are deterministic. This assumption is pervasive across generalization and suppression algorithms and models for the past decades. However, differential privacy and the randomized response method[185] (originally developed to preserve survey respondent's privacy by randomly adding noise to their answers) could be considered non-deterministic transformation methods. While, again, empirical evidence suggests differential privacy is similarly constrained[16], non-deterministic transformation methods offer a potential opportunity to relax such constraints if they explicitly consider the fairness of data utility. In Chapter 5, I investigate how more representative de-identified data can be shared by using non-deterministic transformations to allow cooperative privacy protections between groups.

4.8 Conclusion

Even though controlling data access imposes additional challenges – such as passports and visas being unequally available among groups, the potential for fraud, the limitations of deterrence methods, and preventing truly "open science" – these challenges can be navigated and to some extent mitigated. The impossibility of achieving fairness with respect to both privacy risk and data utility via data transformations cannot. Our pursuit of a more equitable, data-informed society demands a different approach to data sharing; one in which data transformations are combined with external deterrents to simultaneously protect privacy and preserve the representation of the full population. At the same time, the fairness tradeoff theorem highlights potential avenues to develop de-identification transformation methods that relax, but perhaps not entirely resolve, the fairness constraints. This opportunity is the motivation for the next chapter.

Chapter 5

Altruistic Masking: a method to improve fairness in de-identified data

5.1 Introduction

Records in a dataset have varying distinguishability prior to de-identification. The more unique the record is, the greater its initial privacy risk and the less favorable its privacy-utility tradeoff. As shown in Chapter 4, it is the variability in starting points that imposeses a tradeoff between fairly distributing the privacy risks and fairly distributing data utility between subgroups of records in a de-identified dataset. While sharing data according to the passport-visa model can alleviate the fairness constraints, doing so reduces data accessibility and may create access inequities[71]. It may also be expensive to implement such a model, making it more feasible for large data sharing consortia to adopt compared to smaller organizations. Moreover, the passport-visa model still requires the data to be de-identified to some extent, such that minority subgroups' representation may still be disproportionately distorted. Broadly supporting more equitable privacy-preserving data sharing needs additional methods that relax the data-based constraints to fair de-identification.

While the fairness tradeoff theorem presented in Chapter 4 defines the constraints to achieving fairness in de-identified data via generalization and suppression, it also highlights a potential opportunity to alleviate such constraints. Namely, the theorem assumes de-identification transformations are deterministic. In other words, every record with the same quasi-identifier value (i.e., in the same equivalence class) prior to de-identification will also have the same transformed quasi-identifier value after de-identification. This rigid assumption, that is derived from standard practice and appears to be ubiquitous among generalization and suppression methods, restricts the manner in which records can gain privacy – they can be generalized further or be suppressed. However, breaking such an assumption with non-deterministic transformations could allow for cooperation and contributions between subgroups. That is, if a subset of records in a large equivalence class is transformed in a manner that gives records in a smaller equivalence classes greater privacy protections, the records in the smaller equivalence classes could theoretically retain greater data utility while still gaining privacy. The altruistic contribution of privacy protection from the large equivalence class still comes at a cost, as transforming data always comes at the expense of utility; however, such a method would take utility from the group with a more favorable privacy-utility tradeoff instead of

taking from the group with a less favorable one. This, in turn, could allow for a more equal distribution of privacy risk and data utility in a de-identified dataset.

In this chapter, I investigate the potential to leverage non-deterministic de-identification transformations to allow the majority subgroups' records to contribute to the minority subgroups' privacy, with the goal of improving minorities' utility. I develop and validate such a de-identification method, called Altruistic Masking (AM). I first describe AM and the cooperative privacy protections it supports. Notably, AM allows subgroups of records to contribute to each other's privacy protections in some respect, but it cannot provide the same privacy guarantees as *k*-anonymity. The nuance in privacy protections highlights yet another constraint to fairness in de-identified data, as it is the more distinguishable populations that are disadvantaged. Nevertheless, as I show, relaxing fairness with respect to such privacy guarantees provides the flexibility to more equally distribute data utility. After describing the AM method, I present an algorithm to implement it. Finally, I test the overall utility and distribution of utility of data de-identified using AM and compare it to that of standard *k*-anonymization methods. I measure utility both intrinsically, according to the entropy utility loss measure (Eqn. 4.2), as well as in the context of disparity detection.

5.2 Conceptual description of Altruistic Masking

AM supports cooperative privacy protections by mimicing the effect of unknown or missing data on patient distinguishability. Figure 5.1 displays variations of the same example dataset to illustrate this effect. Figure 5.1A displays the ground truth dataset of which Figure 5.1B is an abstraction. In this example dataset, there are two equivalence classes that only differ in their race values: White (corresponding to the majority equivalence class with 7 records) and American Indian or Alaskan Native (AIAN) (corresponding to the minority equivalence class with 3 records). Assume that both the data steward and the adversary do not know record 8's race value, as shown in the dataset in Figure 5.1B. This omission could be due to a variety of reasons, including clerical error or the patient declining to answer. If the adversary does not know and cannot infer the ground truth distribution shown in Figure 5.1A – an example of such an inference would be if there are only three individuals in the population with a race value of AIAN then record 8's value must be White – then the adversary cannot confidently determine patient 8's race value as either White or AIAN. Therefore, the uncertainty induced by the unknown value has a privacy protective effect for both records with race value White and records with race value AIAN, as the adversary would have to consider record 8 as potentially belonging to both equivalence classes.

AM attemps to mimic the effect of unknown data by strategically masking values in records' quasi-identifiers to create uncertainty around their original equivalence class. The value that is masked is the quasi-identifying attribute for which the steward is optimizing fairness. For example, to improve the fairness of representation with respect to race, as shown here, the steward masks the race values in certain records. Improving the fairness of representation with respect to age would involve masking age values of certain records. While the number of records masked within each equivalence class depends on the privacy protection threshold the data steward defines, the records that are ultimately masked are chosen at random. The process of determining the number of records masked per equivalence class is described in Section 5.3. To protect against certain inference attacks, AM masks values under the assumption the adversary *does* know the ground truth distribution. Therefore, instead of only masking values from the majority equivalence class' records, as shown in Figure 5.1C, it additionally masks records from the minority equivalence class' records, as shown in Figure 5.1D. As at least one masked record comes from the minority equivalence class, the adversary is less certain which masked record belongs to each equivalence class. Sections 5.3 and 5.4 describe how AM tries to mask only one record from each minority equivalence class to maximize the minorities' retained utility. However, the sections also describe how certain scenarios may demand more than one record from the minority equivalence class to be masked in order to meet the privacy risk threshold the data steward defines.

A) Ground truth

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	White
9	21	AIAN
10	21	AIAN
11	21	AIAN

B) Missing value

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	Unknown
9	21	AIAN
10	21	AIAN
11	21	AIAN

C) Mask majority record

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	?
9	21	AIAN
10	21	AIAN
11	21	AIAN

D) Altruistic masking

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	?
9	21	?
10	21	AIAN
11	21	AIAN

Figure 5.1. Varying representations of the same dataset.

Before deriving how many records must be masked to achieve specific privacy protection thresholds, I would like to highlight the differences between AM and standard de-identification methods. Figure 5.2 shows the original data table shown in Figure 5.1A after additional generalization (Figure 5.2A) or suppression (Figure 5.2B). Figure 5.2C also shows an alternative to cooperative privacy protections where record 8's race value is changed from AIAN to White.

First, generalization (Figure 5.2A) of the race value may bring all records to the same granularity. It could be argued that such a transformation equally degrades utility between the two equivalence classes, as all records now have the same quasi-identifier. However, the majority of the generalized records still correspond to patients that originally had a race value of White. As such, and as captured by the entropy utility loss measure used in Chapter 4, the signal of the records originally having a race value of AIAN is diluted more than that of the records originally having a race value of White. AM, on the other hand, sacrifices a few records' race values to preserve a more granular representation of the rest. And as I show later, most of the records that are masked derive from the majority equivalence classes.

Second, suppressing (Figure 5.2B) the records entirely such that they are not shared with the adversary removes the minority equivalence class' representation entirely. AM does not remove records from the dataset. AM does mask some values in some records, but only the values pertaining to the attribute for which fairness is being optimized. The remainder of the quasi-identifier is not modified.

Third, replacing the race value in a majority equivalence class' record with that of the minority equivalence class effectively decreases the distinguishability of records in the minority equivalence class. However, similar to generalization, the more that majority equivalence class records are mixed in, the more diluted the minority equivalence class' signal becomes. AM instead creates an additional equivalence class where the race value is masked. The masked records vary in their original representation – in this example, some originally had a race value of White; others AIAN – but the unmasked records retain more accurate representation.

AM aims to preserve minority representation better than these de-identification alternatives. It does so under the assumption that fewer, more granular, and more truthful records provide better representation and better preserves their signal for diverse applications than having less granualar (generalization), unshared (suppression), and untruthful records (flipping race values).

A) Generalization

ID	Age	Race
1	21	Any
2	21	Any
3	21	Any
4	21	Any
5	21	Any
6	21	Any
7	21	Any
8	21	Any
9	21	Any
10	21	Any
11	21	Any

B) Suppression

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	White

C) Flip value

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	AIAN
9	21	AIAN
10	21	AIAN
11	21	AIAN

Figure 5.2. Alternative transformations to AM (Figure 5.1D) of the dataset shown in Figure 5.1A.

5.3 Privacy protections of Altruistic Masking

In this section, I define the cooperative privacy protections AM supports. I show how AM provides similar privacy protections as *k*-anonymity, but it cannot provide the same privacy guarantees.

5.3.1 Preliminaries

Following the definitions presented thus far in this dissertation, I define a **quasi-identifier** as the set of features in the dataset that can enable re-identification. Quasi-identifying features are those that can be both known by an adversary and used to distinguish individual records, such as demographic features. I define an **equivalence class** as the set of records that share the same the same quasi-identifier value. I define the **fairness attribute** as the attribute for which the steward aims to optimize the fairness of de-identification trainsformations. I assume the fairness attribute is also part of the quasi-identifier such that the fairness attribute is masked by AM. I define a **masking class** as the set of records that share the same set of quasi-identifying features minus the fairness attribute.

Figure 5.3 visually depicts these concepts. The quasi-identifier consists of two attributes: age and race. There are three equivalence classes. The first includes 8 records corresponding to 21-years-old White individuals, the second includes 3 records corresponding to 21-years-old AIAN individuals, and the third equivalence class includes 3 records corresponding to 30-years-old AIAN individuals. Since the first two equivalence classes have the same quasi-identifier value but for the fairness attribute (the race attribute, in this example), they belong to the same masking class. The third equivalence class has a quasi-identifier value that differs with respect to both the age and fairness(race) attribute, and therefore belongs to a different masking class. To create cooperative privacy protections, the fairness attribute is the value that is masked by AM. Therefore, AM involves replacing certain records' race values with "?", or some sort of null value.

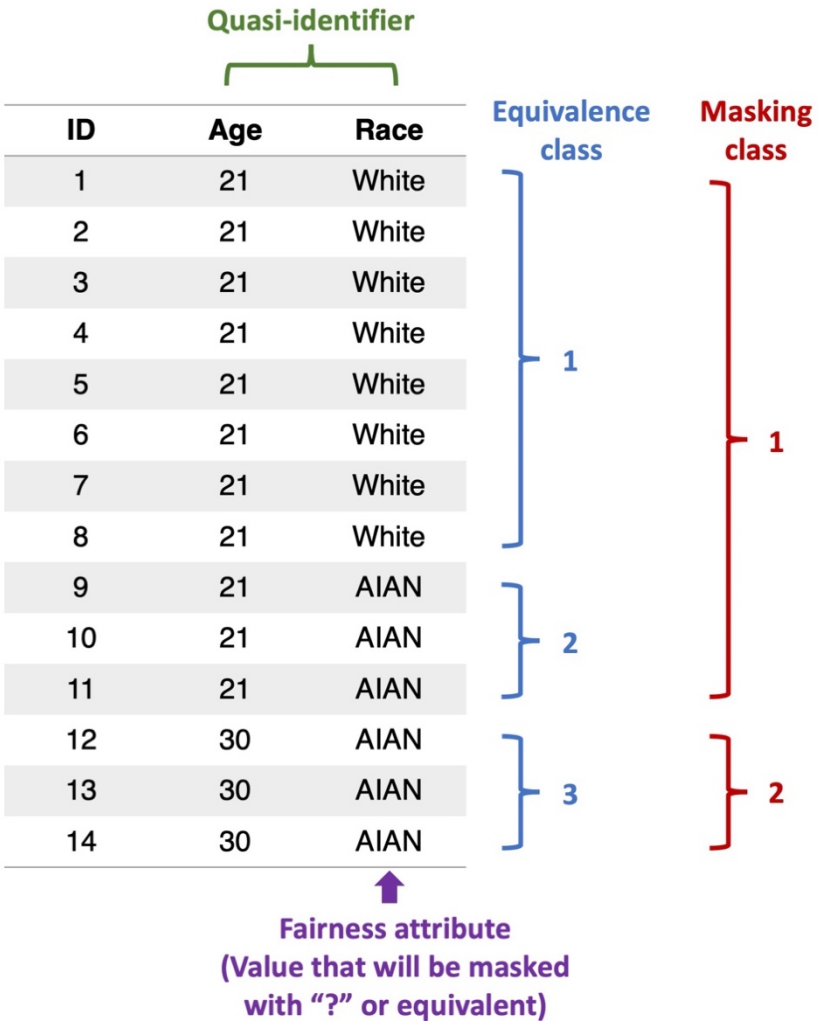

Figure 5.3. Preliminary concepts underlying AM.

Table 5.1 summarizes additional notation used to facilitate the derivation and description of the privacy protections. A_i represents the size of equivalence class i prior to masking. B_i and C_i represent the number of records from equivalence class i that are not masked and masked, respectively, by AM. Therefore, $A_i = B_i + C_i$. The symbol D_{ij} represents the number of records masked within masking class j that do not belong to equivalence class i. Finally, E_j represents the total number of records masked in masking class j, such that $E_j = C_i + D_{ij}$ for every equivalence class i belonging to masking class j. Figure 5.4 provides an illustrative example describing this notation.

Table 5.1. Preliminary notation.

A_i	Size of equivalence class i before masking.
B_i	Number of records from i that are not masked.
C_i	Number of records from i that are masked.
D_{ij}	Number of records masked within masking class j that do not belong to equivalence class i.
E_j	Total number of records masked within masking class j.

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	White
8	21	White
9	21	AIAN
10	21	AIAN
11	21	AIAN

ID	Age	Race
1	21	White
2	21	White
3	21	White
4	21	White
5	21	White
6	21	White
7	21	?
8	21	?
9	21	?
10	21	AIAN
11	21	AIAN

	White equivalence class	AIAN equivalence class
A_i	8	3
B_i	6	2
C_i	2	1
D_{ij}	1	2
E_j	3	3

Figure 5.4. Masking example and the corresponding notation values.

5.3.2 Adversarial assumptions

When modeling the privacy protections of AM, I assume the adversary attempts to re-identify a target individual in the dataset, instead of attempting to re-identify as many individuals as possible by linking the shared dataset to a population register (i.e., a marketer attack[48]). I further assume the adversary knows the target individual's record is in the patient population, the target individual's complete quasi-identifier, and the distribution of equivalence classes in the patient population. Finally, I assume the data steward does not know which individual the adversary is targeting such that similar privacy protections should be applied to every record in the dataset.

Under these assumptions, I model three different attack strategies the adversary could take. Each strategy varies in how the adversary prioritizes attacking unmasked vs. masked records. Table 5.2 summarizes the

three strategies. The first attack strategy prioritizes attacking the unmasked records. The adversary iteratively attacks all unmasked records first. If the target individual cannot be re-identified from the unmasked records, the adversary then attacks the masked records. The second attack strategy is the converse of the first, where the adversary prioritizes attacking the masked records before attacking the unmasked records. In the third attack strategy, the adversary does not prioritize either masked or unmasked records. They attack either masked or unmasked records with equal probability.

Table 5.2. Adversary's potential attack strategies against a dataset with AM.

Strategy	With only one re-identification attempt	With multiple re-identification attempts
1	Randomly attack an unmasked record.	Attack all unmasked records first, then attack masked records.
2	Randomly attack a masked record.	Attack all masked records first, then attack unmasked records.
3	Randomly attack any masked or unmasked record.	Iterate through all masked and unmasked records with equal probability.

Ultimately, I assume that an attacker will act rationally such that they attack records in a manner that maximizes their rate of success. AM must then protect against the most potent attack, which I derive in the following sections.

5.3.3 Re-identification risk on the first attempt

I first define the expected probability an adversary re-identifies a targetted individual on the first attempt when the dataset is de-identified using AM. I define the re-identification risk at expectation as masked records are chosen randomly and the adversary does not know whether or not the target individual's record has been masked. I define the expected re-identification risk on the first attempt for each of the three attack strategies. For all derivations, let target individual t belong to equivalence class i and masking class j.

Against the first attack strategy (see Eqn. 5.1), the target individual's expected re-identification risk is equal to the probability the target individual's record is not masked by AM multiplied by the probability the adversary attacks the correct record given the target is not masked. The expected re-identification risks

against the second and third attack strategies are defined in Eqn. 5.2 and 5.3, respectively. To compare the re-identification risks of AM to that of *k*-anonymity, Eqn. 5.4 displays the expected re-identification risk in a *k*-anonymous dataset. Without masking, there is only one attack strategy: randomly attack a record in the target individual's equivalence class.

Strategy 1:

$$E(reid_t|AM, Strategy\ 1)$$
$$= P(t\ is\ unmasked) * P(re - id_t|t\ is\ unmasked, attacks\ unmasked)$$
$$= \left(\frac{B_i}{A_i}\right) * \left(\frac{1}{B_i}\right) = \frac{1}{A_i} \quad (5.1)$$

Strategy 2:

$$E(reid_t|AM, Strategy\ 2)$$
$$= P(t\ is\ masked) * P(re - id_t|t\ is\ masked, attacks\ masked)$$
$$= \left(\frac{C_i}{A_i}\right) * \left(\frac{1}{E_j}\right) = \frac{C_i}{A_i E_j} \quad (5.2)$$

Strategy 3:

$$E(reid_t|AM, Strategy\ 3)$$
$$= P(re - id_t|attacks\ masked\ and\ unmasked)$$
$$= \frac{1}{B_i + E_j} \quad (5.3)$$

Attack against *k*-anonymous dataset:

$$E(reid_t|k - anonymity)$$
$$= \frac{1}{A_i}\ where\ A_i \geq k \quad (5.4)$$

Against an adversary who knows the original distribution of equivalence classes prior to AM, at least one record must be masked both within and outside the target individual's equivalence class. Thus, $C_i \geq 1$ and $D_i \geq 1$ such that $C_i + D_i = E_j > C_i$. Therefore, the adversary maximizes their probability of re-identifying t by taking attack strategy 1.

Notably, AM does not decrease the expected probability of re-identification on the first attempt against attack strategy 1. Without any masking, the expected probability would also be $\frac{1}{A_i}$. The expected risk is also equal to that of a k-anonymous dataset in the scenario that $A_i = k$. However, were a steward to choose between applying AM or k-anonymizing the dataset to a higher value of k, AM would not decrease the expected re-identification risk on the first attempt while k-anonymization would. Moreover, Eqn. 5.1 shows that AM does not provide the same privacy guarantees as k-anonymity, in that records are not guaranteed to remain in an equivalence class of a certain size. While masked records will be re-identified with a probability of 0, as attack strategy 1 involves attacking only masked records, unmasked records will be correctly re-identified with probability $\frac{1}{B_i}$, which is greater than $\frac{1}{A_i}$. Hence, AM increases the risk of some of unmasked records while decreasing the risk of the masked records. Nevertheless, from the perspective of the adversary, who does not know if the target individual is masked, the expected probability of successful re-identification of a target individual in equivalence class i remains to be $\frac{1}{A_i}$.

To validate this mathematical derivation, I simulate re-identification attacks against target individual t who resides in the dataset summarized in Table 5.3. There are 27 records unequally distributed between three equivalence classes. All equivalence classes belong to the same masking class. Record t belongs to equivalence class 3.

On each iteration of the simulation, one record is randomly masked from each equivalence class. The adversary then randomly chooses to attack either an unmasked record belonging to equivalence class 3 or one of the masked records, according to the attack strategy. Figure 5.5 shows the rate of re-identification calculated from 10,000 simulations.

Table 5.3. Summarized description of dataset used for re-identification simulations. All records belong to the same masking class.

Equivalence class	# records	# masked
1	14	1
2	9	1
3*	4	1
Overall	**27**	**3**

* Target individual t resides in this equivalence class.

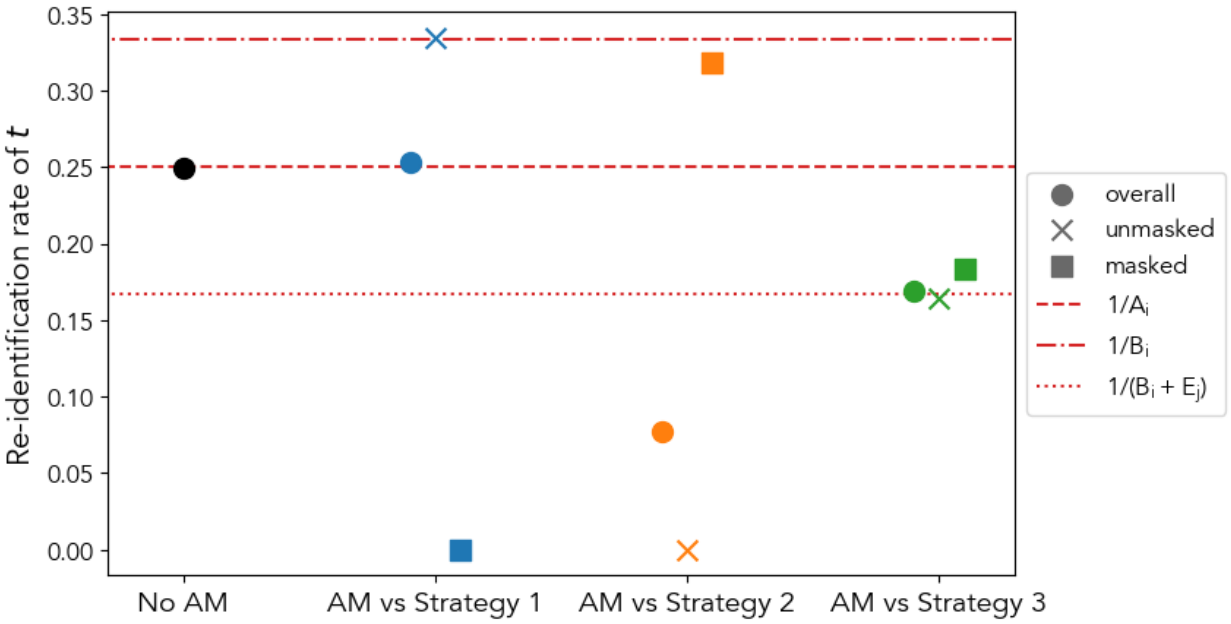

Figure 5.5. Re-identification rate of target individual t, when sharing the dataset described in Table 5.3 against the attack strategies described in Table 5.2. Overall rates are defined as proportion of correct re-identifications across 10,000 independent simulations. Overall = re-identification rate across all simulations. Unmasked = re-identification rate when t is not masked via AM. Masked = re-identification rate when t is masked via AM.

Following the mathematical derivation, the adversary maximizes their expected re-identification rate when applying attack strategy 1. The expected rate when applying AM against attack strategy 1 is equal to that of not applying AM, or $\frac{1}{A_i} = \frac{1}{4}$. However, when applying AM to the dataset, t is actually more likely to be re-identified (compared to the dataset without AM) when t is not randomly chosen to be masked, denoted by the blue "X" at $\frac{1}{B_i} = \frac{1}{3}$.

Given these results, a rational adversary would attempt to re-identify a target individual using attack strategy 1, where the adversary only attacks unmasked records. The results highlight that AM does not decrease a target individual's re-identification risk at expectation and may, in fact, randomly increase their re-identification risk. Nevertheless, as I show in the next section, AM protects records from re-identification by increasing the effort required to achieve correct re-identification. That is, it increases the expected number of attacks an adversary must attempt to correctly re-identify a target individual.

5.3.4 Effort of re-identification

De-identification methods that leverage generalization and suppression, and the privacy models that underly them, generally model an adversary that makes a single re-identification attempt. Against such an attack, the re-identification risk is estimated to be one over the records' equivalence class size. However, as shown by Xia et al.[49], a rational adversary (i.e., an adversary that considers both the cost and benefit of attempting re-identification and attacks in a manner that maximizes their payoff) may attack more than one record in the target individual's equivalence class. The authors also show that an adversary's motivation to attack, and ultimately a record's risk of re-identification, depends on the amount of effort the adversary must exert to re-identify the individual. The greater the effort, the less likely it is that the adversary continues attacking records in search of the target individual. In some cases, the expected effort may be so great that a rational adversary would not attempt re-identification at all. Following this premise, I show how AM can increase the effort of re-identification to that of *k*-anonymity for a specified value of *k.*

I define the effort an adversary must exert to re-identify a target individual as the expected number of attempts until correct re-identification. I use a hypergeometric distribution to model the attack scenario, such that the expected number of attacks until re-identification is defined as $\frac{N+1}{2}$, where N is the size of the pool of records the adversary attacks. The expected number of attempts until re-identification when applying AM against the three attack strategies outlined in Table 5.2 and when *k*-anonyming a dataset, are

defined below. Again, for all derivations, let target individual t belong to equivalence class i and masking class j.

Strategy 1:

$$E(\#attempts\ until\ reid_t|AM, Strategy\ 1)$$
$$= P(t\ is\ unmasked)E(\#attempts|t\ is\ unmasked, attacks\ unmasked) +$$
$$P(t\ is\ masked)E(\#attempts|t\ is\ masked, attacks\ masked)$$
$$= \left(\frac{B_i}{A_i}\right)\left(\frac{B_i + 1}{2}\right) + \left(\frac{C_i}{A_i}\right)\left(B_i + \frac{E_j + 1}{2}\right) \quad (5.5)$$

Strategy 2:

$$E(\#attempts\ until\ reid_t|AM, Strategy\ 2)$$
$$= P(t\ is\ masked)E(\#attempts|t\ is\ masked, attacks\ masked) +$$
$$P(t\ is\ unmasked)E(\#attempts|t\ is\ unmasked, attacks\ unmasked)$$
$$= \left(\frac{C_i}{A_i}\right)\left(\frac{E_j + 1}{2}\right) + \left(\frac{B_i}{A_i}\right)\left(E_j + \frac{B_i + 1}{2}\right) \quad (5.6)$$

Strategy 3:

$$E(\#attempts\ until\ reid_t|AM, Strategy\ 3)$$
$$= E(\#attempts|attacks\ masked\ an\ unmasked\ )$$
$$= \frac{B_i + E_j + 1}{2} \quad (5.7)$$

Attack against k-anonymous dataset:

$$E(\#attempts\ until\ reid_t | k-anonymity)$$
$$= \frac{A_i + 1}{2}\ where\ A_i \geq k \tag{5.8}$$

To determine the adversary's most potent attack strategy, I again simulate re-identification attacks against target individual t in the dataset described in Table 5.3. Figure 5.6 displays the average number of attacks required to re-identify t across 10,000 simulations.

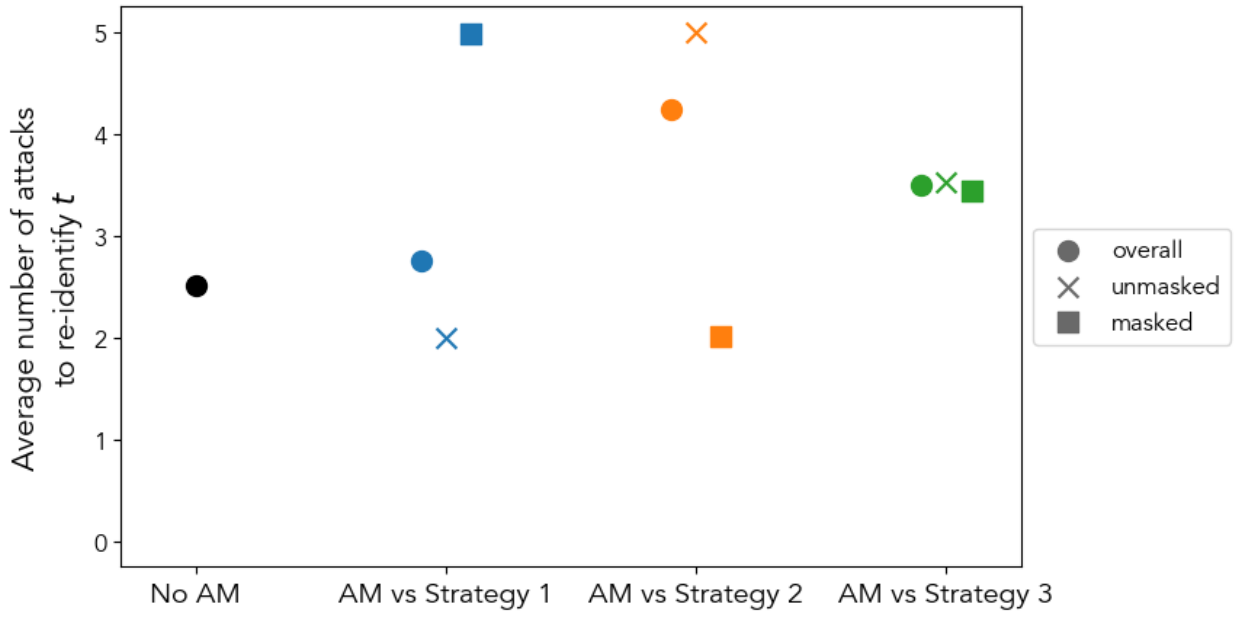

Figure 5.6. Average number of attacks until correctly re-identifying individual t, when sharing the dataset described in Table 5.3 against the attack strategies described in Table 5.2. The "overall" values are calculated as the average across 10,000 independent simulations. "unmasked" and "masked" are the overall results stratified into when when t is not masked and masked by AM, respectively.

The results are similar to those in Figure 5.5, in that the adversary is expected to minimize their effort in re-identifying t when taking attack strategy 1. Therefore, I assume a rational adversary will again always use attack strategy 1 such that AM should protect against such a strategy. Individual t again receives different privacy protections against strategy 1 depending on whether or not their record is ultimately

masked by AM. However, differing from the previous section, the overall expected number of attempts required to re-identify t is slightly higher when AM is applied than when it is not. In fact, Eqn. 5.5 shows that increasing the number of records masked both within and outside t's equivalence class increases the adversary's expected effort to achieve re-identification. Therefore, I derive the number of records that must be masked in order for AM to require the same amount of effort at expectation as *k*-anonymity, for a specified value of *k*, shown below.

$$E(\#attempts\ until\ reid_t | k-anonymity) = E(\#attempts\ until\ reid_t | AM, Strategy\ 1)$$

$$\frac{k+1}{2} = \left(\frac{B_i}{A_i}\right)\left(\frac{B_i + 1}{2}\right) + \left(\frac{C_i}{A_i}\right)\left(B_i + \frac{E_j + 1}{2}\right)$$

$$k = A_i + \frac{C_i D_i}{A_i}$$

$$k - A_i = \frac{C_i D_i}{A_i}$$

$$\Delta_i = \frac{C_i D_i}{A_i} \tag{5.9}$$

Eqn. 5.9 shows that the privacy gain of AM – equal to the increase in the expected effort required to re-identify a target individual – is proportional to both the number of records masked within t's equivalence class, denoted by C_i, and the number of records masked outside t's equivalence class but within t's masking class, denoted by D_i. The latter, in particular, allows the majority groups' records to contribute to the minority groups' privacy protections. Figure 5.7 displays simulation results that show that AM can increase the adversary's effort to re-identify t, on average, to that of *k*-anonymity at specified *k*-values. AM increases protections while only masking one record within t's equivalence class (i.e., $C_i = 1$); the rest of the masked records come from the majority equivalence classes. With respect to Eqn. 5.9, $C_i = 1$ for all iterations and values of D_i, and $A_i = 4$ (see Table 5.3).

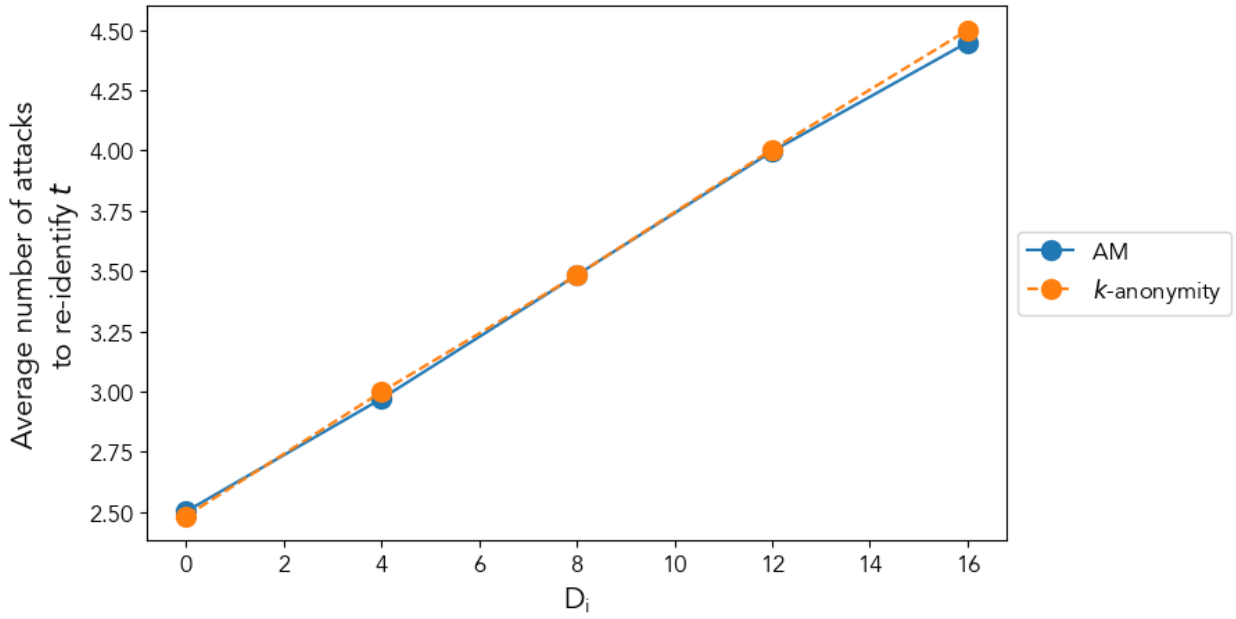

Figure 5.7. Average number of attacks until correctly re-identifying individual t, when sharing the dataset described in Table 5.3 using AM against the attack strategy 1 (described in Table 5.2) and using k-anonymity. The average number of attacks is calculated across 10,000 independent simulations. The k value for each value of D_i was defined using Eqn. 5.9, where $C_i = 1$ and $A_i = 4$ (see Table 5.3).

5.3.5 Summary of Altruistic Masking's privacy protections

AM leverages non-determinstic transformations to allow cooperation in privacy protections such that the majority groups' records can contribute privacy protections to the minority groups' records. In this section, I showed how AM can reduce the re-identification risk by increasing an adversary's expected effort to re-identify a target individual. Given U.S. privacy legislation's focus on reducing individual's re-identification risk without mandating the correct manner in doing so, I argue that such a method could meet HIPAA's Expert Determination standard.

However, AM does not provide the same privacy guarantees as k-anonymity. Table 5.4 summarizes the differences. k-anonymity guarantees that every record in the dataset will fall into an equivalence class of size k or larger. Such a threshold applies an equal, or fair, upper bound to every record's privacy risk. AM cannot guarantee that records fall into an equivalence class of a certain size. In fact, AM may increase the re-identification risk of records that are not randomly masked (as the size of their equivalence class is actually reduced) while decreasing the risk of those that are. Instead of ensuring that records belong to sufficiently large equivalence classes, AM focuses on increasing an adversary's perceived cost of re-

identification. As long as the adversary does not know which records are masked, AM imposes an equal, or fair, lower bound to the adversary's expected effort to attempt re-identification. I argue that this serves as a relaxation of providing fair privacy protections to every record in a dataset, which, as I will show in the next sections, allow for more equal retention of data utility. Still, this represents another tradeoff to privacy-preserving data sharing: between providing more ideal privacy guarantees and more equal data utility.

Table 5.4. Summary of AM and *k*-anonymity's privacy protections.

	AM	***k*-anonymity**
Re-identification risk on the first attempt	Does not change the risk at expectation. Unmasked records are exposed to greater risks than unmasked records, but all records have the same expected risk from the perspective of the adversary.	Expected risk is bound to $\frac{1}{k}$.
Number of attempts to re-identify target individual	Increases the number of attempts at expectation. Unmasked records require fewer attempts than masked records, but all records' expected effort can be bound from the perspective of the adversary.	Bounds the expected value to $\frac{k+1}{2}$.
Method to reduce re-identification risk	Create a new equivalence class of masked records to increase the record's equivalence class size at expectation.	Increase record's equivalence class size.
Fairness	Applies a minimum threshold to an adversary's expected effort to re-identification.	Applies a minimum equivalence class size threshold.

5.4.1 Altruistic Masking pipeline

Figure 5.8 describes the pipeline for implementing AM. As described by Eqn. 5.9, cooperative privacy protections require that minority equivalence classes contribute at least one record to be masked. Therefore, AM cannot protect population uniques and may not protect records in very small equivalence classes. As such, the first step of the pipeline is to *k*-anonymize the raw dataset to a large enough value of *k* to support masking. The user specifies the value of this initial *k* value, $k_{initial}$, as well as the generalization hierarchies that guide *k*-anonymization. The most straightforward method to increase the size of equivalence classes and masking classes is via global recoding. In this chapter, I *k*-anonymize the raw dataset using the OLA algorithm without suppression. I do not allow OLA to suppress records to avoid the disproportionate utility loss imposed by suppression (see Section 4.3.3).

The second step of the pipeline applies AM. Guided by Eqn. 5.9, AM increases an adversary's expected effort to re-identify any record to at least that of a *k*-anonymous dataset, where *k* is equal to the user-defined k_{target}. The user also specifies the fairness attribute as well as thresholds that limit the extent to which the minority and majority equivalence classes can be masked. The dataset is then masked according to the AM algorithm, producing the final de-identified dataset.

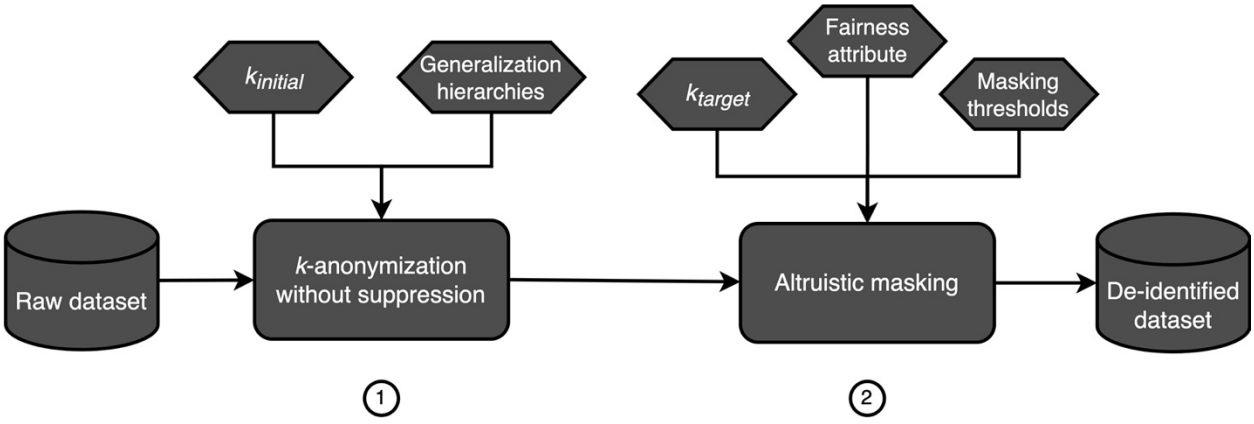

Figure 5.8. De-identification pipeline for AM.

5.4.2 Altruistic Masking algorithm

Details of the algorithm implementing AM are shown in Table 5.5 and Figure 5.9. In the first step (line 1), the algorithm makes a copy of the dataset that has been *k*-anonymized to a *k* value of $k_{initial}$. The algorithm then partitions the dataset into masking classes (2). The algorithm then enters a loop (3-27), iterating across each masking class class (3). The algorithm first identifies the majority (4) and minority (5) equivalence classes. The majority equivalence classes are those that already have at least k_{target} records. The minority equivalence classes are those that have less than k_{target} records. The algorithm then counts the total number of records belonging to the majority equivalence classes (6) and initializes the *feasible* variable to *True*. The AM algorithm is not guaranteed to achieve the cooperative privacy protections to the level of k_{target} while meeting the majority and minority masking thresholds. The larger the difference is between $k_{initial}$ and k_{target}, the more records must be masked. In some instances, there are not enough that can be masked. As such, in addition to the transformed dataset, the algorithm returns whether the desired level of masking was feasible.

The algorithm then enters a second loop (8-26), iterating across the minority equivalence classes within the current masking class (8). The algorithm counts the number of records in the current equivalence class (9), or A_i as defined in Table 5.1; initializes the number of records to be masked within that equivalence class to 1 (10), or C_i; and counts the number of records that have already been masked within the masking class and outside the equivalence class (11), or D_{ij}. The algorithm enters a while loop (12-21) in which it determines the minimal number of records that must be masked within the current minority equivalence class, C_i, to meet the privacy protections threshold of k_{target}. If C_i and D_{ij} can remain below the minority and majority equivalence class masking thresholds – or if the masking is feasible – the corresponding number of records are randomly masked in the dataset (15-16). If it is not feasible, the algorithm terminates and returns the current version of the masked dataset and the *feasible* variable set to *False* (22-25). If masking is feasible for every equivalence class, the algorithm returns the transformed dataset and the *feasible* variable set to *True* (28).

Table 5.5. Description of pseudocode functions in Figure 5.9.

Function	Description
getMaskingClasses(*Data'*, $Attr_{fair}$)	Partitions *Data'* into unique masking classes with respect to $Attr_{fair}$.
getMajorityEquivalenceClasses(*Data', j,* k_{target})	Returns the indices of the records corresponding to equivalence classes within masking class *j* that have at least k_{target} records.
getMinorityEquivalenceClasses(*Data', j,* k_{target})	Returns the indices of the records corresponding to equivalence classes within masking class *j* that have less than k_{target} records. The indices are partitioned into individual equivalence classes.
sizeEquivalenceClass(*i*)	Returns the size of equivalence class *i* before masking.
alreadyMasked(*Data*, j, i*)	Returns the number of records within masking class *j* and outside equivalence class *i* that have already been masked in $Data^*$.
numToMask(A_i, C_i, k_{target})	Using Eqn. 5.9, returns the number of records that must be masked to achieve the privacy protections at k_{target}.
mask(*number to mask, indices of records that can be masked*, *Data**, $Attr_{fair}$)	Randomly chooses, without replacement, which records are masked within the specified indices in *Data**. Records that were previously masked are ignored. Masking involves changing $Attr_{fair}$ values to '?' symbol, or equivalent.

Algorithm 3: Altruistic Masking

Input : $Data'$, Dataset k-anonymized to $k_{initial}$;
k_{target}, k value to which masking should protect;
$Attr_{fair}$, fairness attribute.
$Threshold_{Majority}$, maximum proportion of majority equivalence class records that can be masked;
$Threshold_{Minority}$, maximum proportion of minority equivalence class records that can be masked.

Output: $Data^*$, masked Dataset;
$feasible$, boolean indicating whether masking to k_{target} was feasible.

$Data^* \leftarrow Data'$
$J \leftarrow$ getMaskingClasses($Data'$, $Attr_{fair}$)
for j **in** J **do**
  $I_{majority} \leftarrow$ getMajorityEquivalenceClasses($Data'$, j, k_{target})
  $I_{minority} \leftarrow$ getMinorityEquivalenceClasses($Data'$, j, k_{target})
  $N_{majority} \leftarrow$ numberMajorityRecords($Data'$, $I_{majority}$)
  $feasible \leftarrow True$
  for i **in** $I_{minority}$ **do**
    $A_i \leftarrow$ sizeEquivalenceClass(i)
    $C_i \leftarrow 1$
    $D_{masked} \leftarrow$ alreadyMasked($Data^*$, j, i)
    while $C_i \leq (A_i * Threshold_{minority})$ **do**
      $D_{ij} \leftarrow$ numToMask(A_i, C_i, k_{target}) $- D_{masked}$
      if $D_{ij} \leq (N_{majority} * Threshold_{majority})$ **then**
        $mask(C_i, i, Data^*, Attr_{fair})$
        $mask(D_{ij}, I_{majority}, Data^*, Attr_{fair})$
        break
      else
        $C_i \leftarrow C_i + 1$
      end if
    end while
    if $C_i > (A_i * Threshold_{minority})$ **then**
      $feasible \leftarrow False$
      return $Data^*$, $feasible$
    end if
  end for
end for
return $Data^*$, $feasible$

Figure 5.9. AM algorithm.

5.5 Utility evaluation of Altruistic Masking

In this section, I compare the overall utility and fairness with respect to utility that AM supports to that of standard *k*-anonymization methods. I measure utility in two ways. First, I measure intrinsic utility using the normalized entropy measure described in Eqn. 4.2. Second, I measure utility in the context of detecting disparate outcomes with respect to a binary attribute. Similar to Chapter 4, I define fairness with respect to utility in terms as equal data utility between racial subpopulations.

The section begins by describing the datasets used in the evaluations and the de-identification methods being compared to AM. The intrinsic utility results are then presented, followed by the disparity detection utility results.

5.5.1 Datasets and generalization hierarchies

Two datasets are used in both the intrinsic utility and disparity detection evaluations: a simulated dataset and the Adult dataset from UC Irvine[186]. I use simulated data to evaluate how well the de-identification methods preserve the utility of racial subpopulations of varying proportions. The Adult dataset serves as an example of a real-world dataset upon which many de-identification methods have been tested.

The simulated dataset contains four attributes: age, sex, race, and a binary outcome (used for the disparity detection). I define the quasi-identifier as {*age, sex, race*}. I define the generalization hierarchies for each quasi-identifying attribute as shown in Figure 5.10. The demographic feature values are selected by randomly sampling with replacement according to a feature-specific probability distribution. The age probability distribution is derived from the Adult dataset. The sex distribution is equally distributed between Female and Male. The race distribution includes three possible values: majority, minority, and super-minority. The probability distribution of these values varies in the experiments. The binary outcome is randomly assigned to records according to a race-specific rate. Each iteration of the simulated dataset contains 100,000 records. Multiple iterations are performed to estimate each de-identification method's expected performance in each experiment.

The Adult dataset, also known as the Adult income dataset, is a sample of U.S. Census data containing several demographic and socio-economic features. I define the quasi-identifer as {*age, gender* (equivalent to biological sex in this case), *race, native-country, educational-num, workclass, marital-status,*

occupation}. I use the same generalization hierarchies for age and gender/sex as shown in Figure 5.10. The hierarchies for the other quasi-identifying attributes, including a different race hierarchy, are shown in Figure 5.11. As indicated by the hierarchies, I consider the "raw" Adult dataset to contain generalized forms of {*native-country, workclass, marital-status, occupation}*. I perform this initial generalization because of the attributes' impact on *k*-anonymization algorithms. Each of these four attributes contain many values with skewed distributions. Since the *k*-anonymization algorithms minimize utility loss with respect to entropy, they frequently sacrifice age, race, and sex information to preserve these more entropic features. Therefore, to balance the scales, I initialize *native-country, workclass, marital-status,* and *occupation* to a more generalized starting point. Preprocessing also includes dropping duplicate records and records with null values, producing a final dataset of 45,175 records. The re-identification risk distribution for each racial subgroup is illustrated in Figure 5.12.

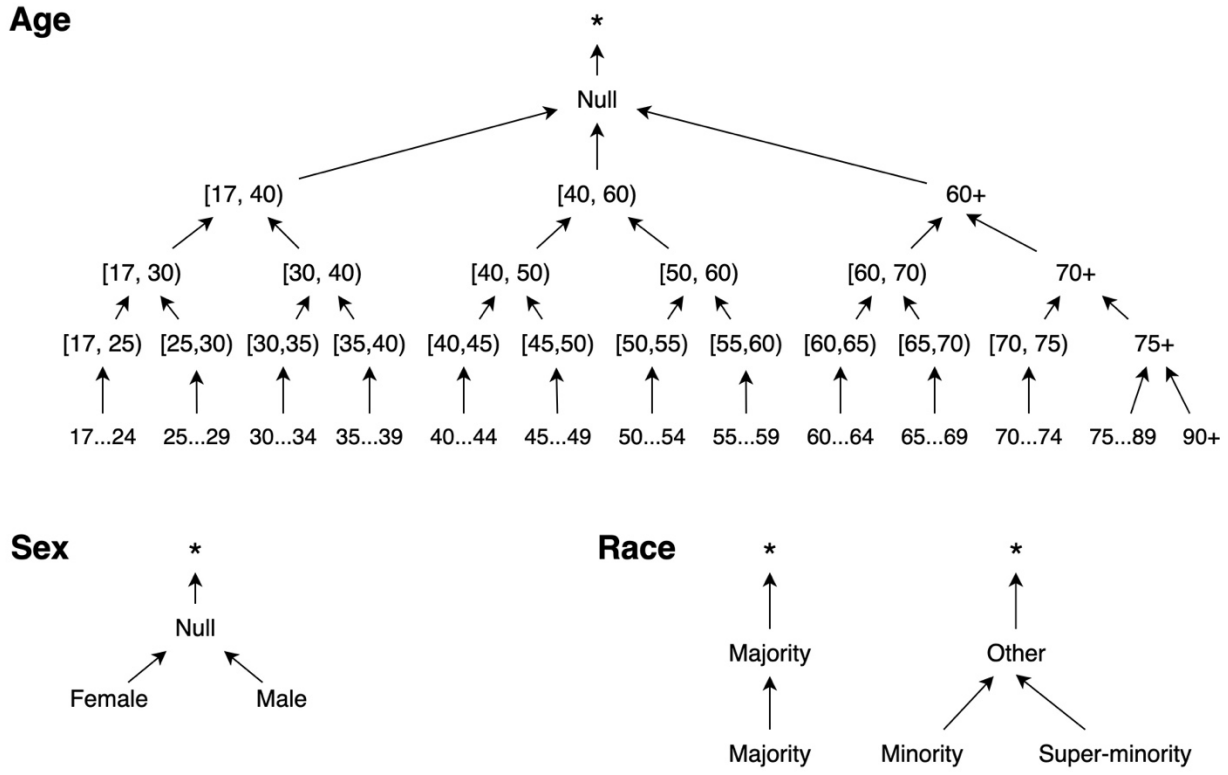

Figure 5.10. Generalization hierarchies for the simulated dataset. The '*' symbol denotes suppression of the complete record.

5.5.2 De-identification methods

I compare AM's ability to preserve overall data utility and support racial fairness with respect to utility to that of two variations of standard *k*-anonymization. Similar to Chapter 4, the variations apply global recoding using the OLA algorithm[6], where the first implementation allows up to 1% of records to be suppressed and the second implementation does not apply suppression. As described in Section 5.4, de-identifying the dataset via AM involves first *k*-anonymizing the dataset with the OLA algorithm (without suppression) to a *k* value of $k_{initial}$. AM is then applied to the k_{target} value, where k_{target} is equal to the k applied to the standard *k*-anonymization implementations. For example, AM with $k_{initial} = 3$ and $k_{target} = 11$ would be compared to standard *k*-anonymization where $k = 11$. All OLA implementations are optimized according to the entropy utility loss measure defined in Eqn. 4.2.

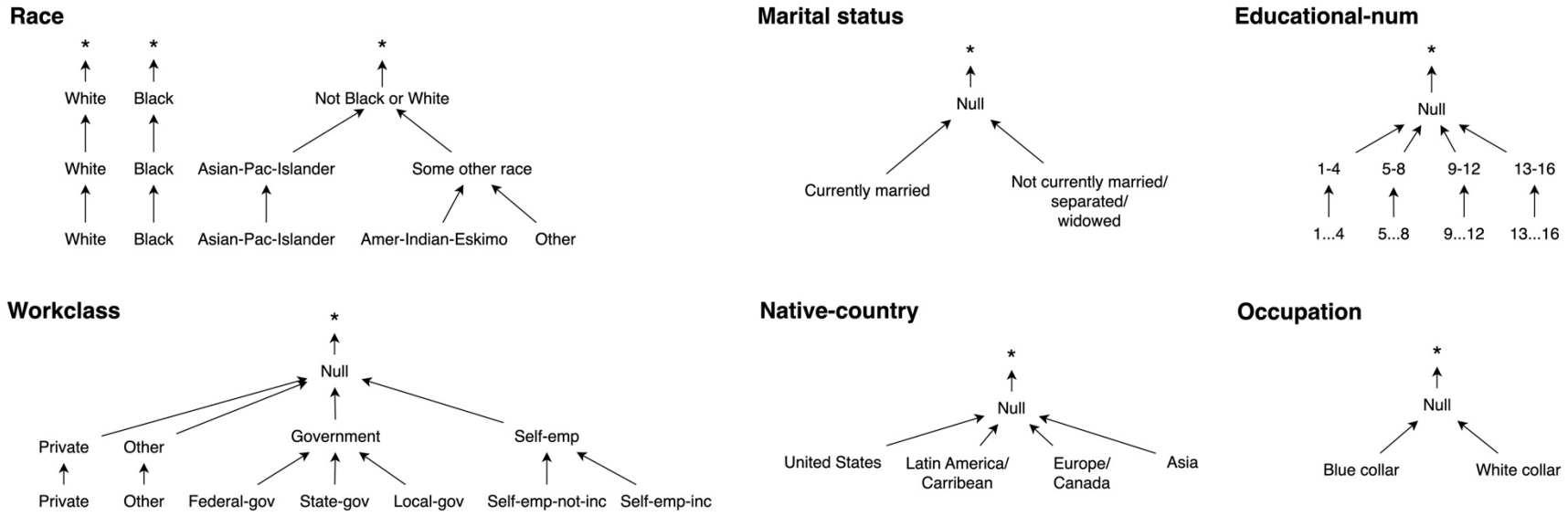

Figure 5.11. The generalization hierarchies guiding de-identification of the Adult dataset include those presented here and the age and sex hierarchies shown in Figure 5.10. The * symbol denotes suppression of the complete record.

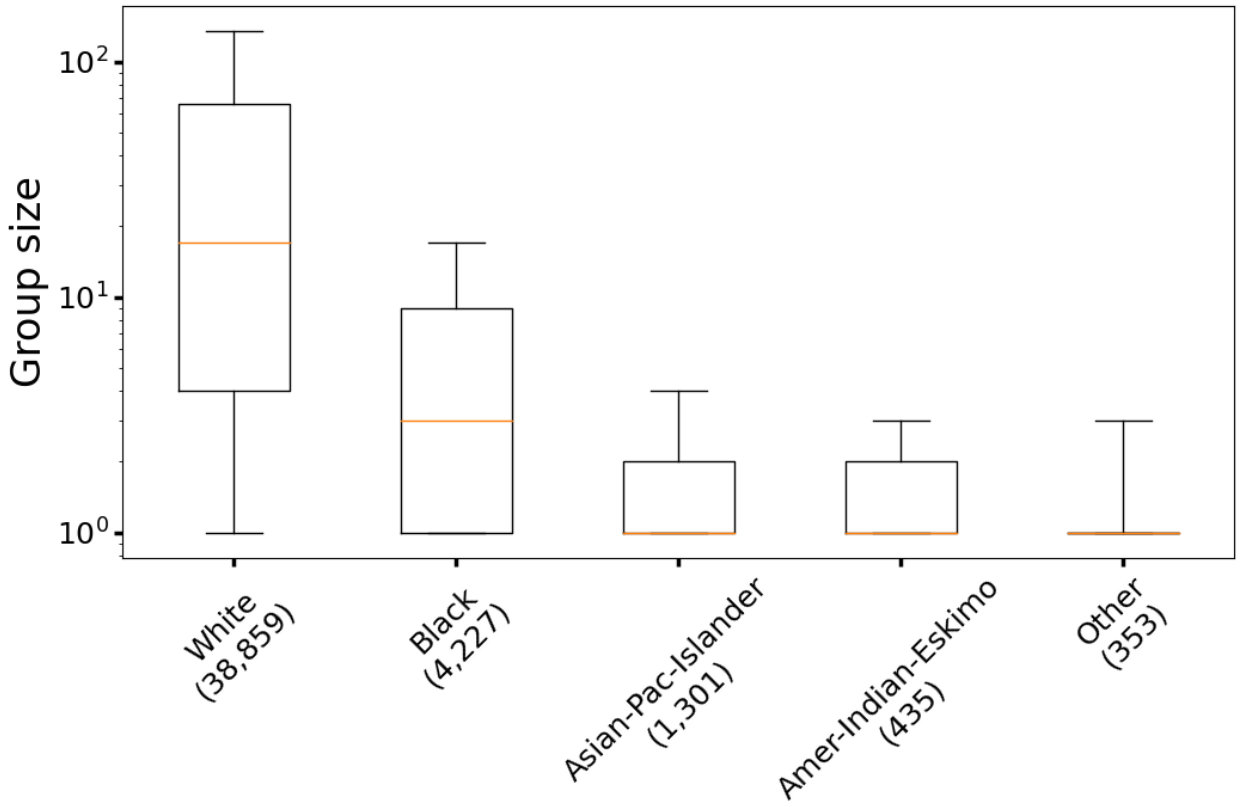

Figure 5.12. Distribution of group sizes for each race in the Adult dataset. Group size is defined as the number of individuals with the same quasi-identifier value. Re-identification risk is inversely proportional to the group size. The numbers in parentheses indicate the number of records corresponding to each race. For each distribution, brackets denote 95% confidence interval, boxes denote inter-quartile range, and orange line denotes median value.

5.5.3 Intrinsic utility evaluation

5.5.3.1 Simulated data

I first measure intrinsic data utility loss when applying each de-identification method to simulated data at various values of *k*. The probability distribution for the race values is defined as: {majority=0.9, minority=0.09, super-minority=0.01}. Figure 5.13 displays the overall utility loss in the dataset and the inequality, or unfairness, of the distribution of the utility loss across the racial subpopulations. Inequality is calculated as the Gini coefficient of the race-specific utility loss values. Figure 5.14 displays the race-specific utility loss curves. AM is implemented with $k_{initial} = 3$ and k (k_{target}) values ranging from 3 to 50. The expected values and 95% quantile intervals are calculated from 100 simulations.

Intuitively, AM's utility loss at k =3 is equal to that of k-anonymity without suppression, as no masking is applied when $k_{initial} = k_{target}$. As the value of k increases, AM retains greater data overall data utility than k-anonymization without suppression and more equally distributes data utility between racial subpopulations than both k-anonymization implementations. While k-anonymization with 1% suppression minimizes overall utility loss, it disproportionately reduces the utilty of the super-minority racial subpopulation, particularly at higher k values. AM, however, retains relatively consistent data utility for all racial subpopulations across k values. It also preserves more data utility for the super-minority group than the k-implementations for k values greather than and equal to 11.

Figures 5.15 and 5.16 display the overall rates and race-specific rates, respectively, of suppression (for k-anonymity) and masking (for AM) at the various k values. At k values greater than 30, AM masks more than 1% of all records, on average. The k-anonymization implementation with suppression limits suppression to below 1% of all records. Note that even though AM may mask more records than the k-anonymity implementation is suppressing, masking only changes the race values of records while suppression removes the records from the dataset entirely. Figure 5.16 highlights how suppression targets the smaller groups in the dataset. The minority and super-minority racial subpopulations are suppressed more frequently than the majority. In some cases, more than 30% of super-minority records are expected to be suppressed. AM, on the other hand, targets the majority subpopulations' records. At k = 50, the majority racial subpopulation is masked at a higher rate than the minority, which itself is masked at a higher rate than the super-minority.

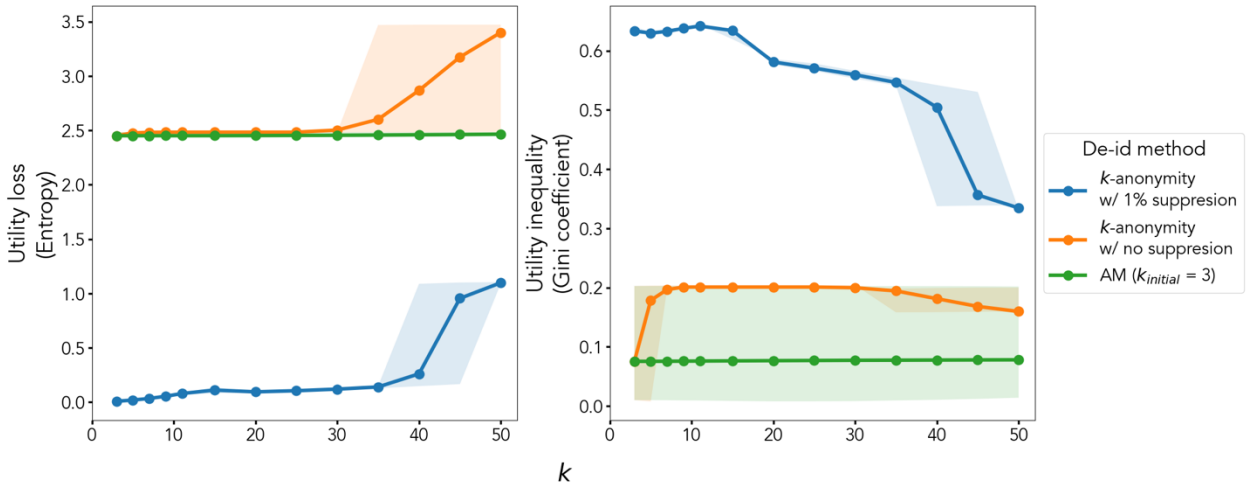

Figure 5.13. (Left) Overall utility loss, measured as entropy (see Eqn. 4.2), when applying each de-identification method at k values ranging from 3 to 50 to simulated data. (Right) Inequality in utility loss between racial subgroups measured as the Gini coefficient of the race-specific utility loss values. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. The race value probability distribution when simulating the data was defined as {majority=0.9, minority=0.09, super-minority=0.01}. Each simulated dataset contained 100,000 records.

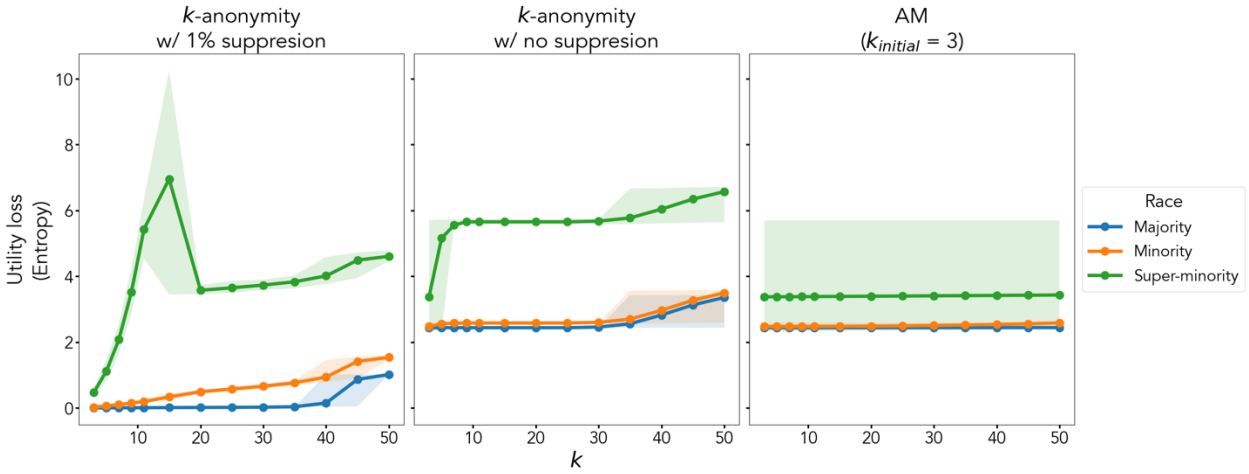

Figure 5.14. Race-specific utility loss curves when de-identifying simulated data at varying levels of k. (Left) OLA k-anonymization algorithm with up to 1% of all records suppressed. (Center) OLA algorithm with no records suppressed. (Right) Altruistic Masking where $k_{initial} = 3$. Utility loss is measured as entropy according to Eqn. 4.2. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. The race value probability distribution when simulating the data was defined as {majority=0.9, minority=0.09, super-minority=0.01}. Each simulated dataset contained 100,000 records.

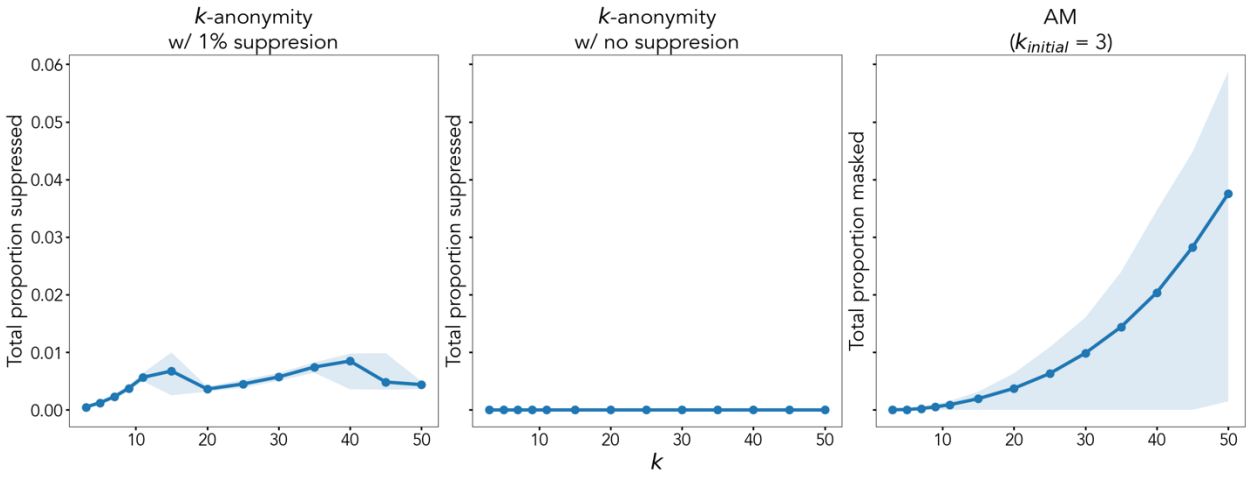

Figure 5.15. Total proportion of records either suppressed or masked when de-identifying simulated datasets at varying levels of *k*. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. The race value probability distribution when simulating the data was defined as {majority=0.9, minority=0.09, super-minority=0.01}. Each simulated dataset contained 100,000 records.

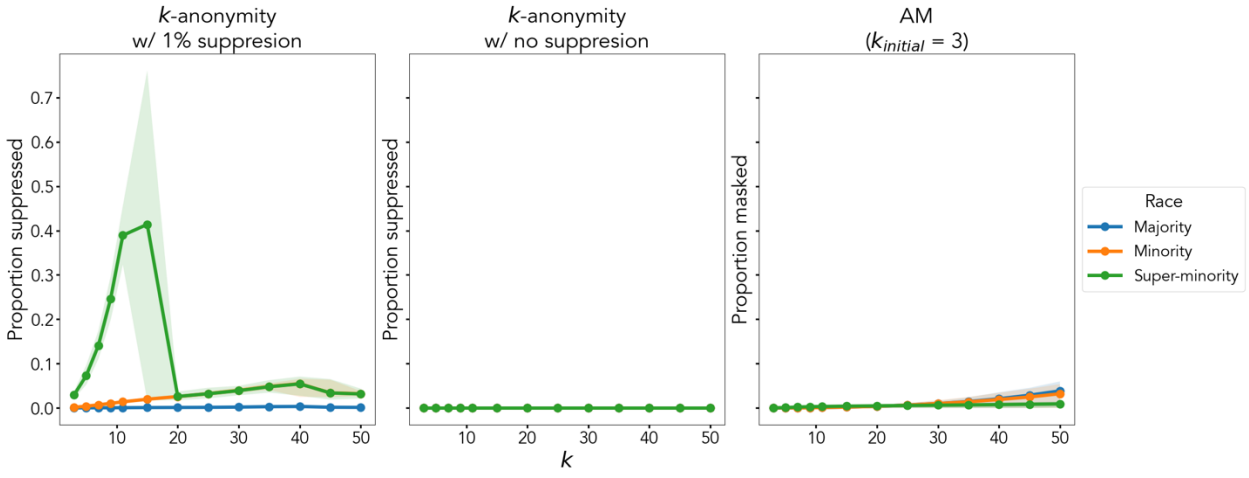

Figure 5.16. Race-specific suppression and masking rates when de-identifying simulated data at varying levels of *k*. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. The race value probability distribution when simulating the data was defined as {majority=0.9, minority=0.09, super-minority=0.01}. Each simulated dataset contained 100,000 records.

I next evalute each de-identification method's ability to minimize utility loss for the super-minority population, while varying the relative size of the super-minority population. The probability distribution for the race values is defined as: {majority=0.9 – x, minority=0.09, super-minority=x}. For the k-anonymization implementations, k = 30. For AM, $k_{initial}$ = 3 and k_{target} = 30. The expected values and 95% quantile intervals are again calculated from 100 simulations. Figure 5.17 displays the overall utility loss results, Figure 5.18 displays the race-specific utility loss, Figure 5.19 displays the overall suppression and masking rates, and Figure 5.20 displays the race-specific suppression and masking rates.

Figure 5.17 shows that k-anonymity with suppression minimizes the overall utility loss compared to the other de-identification methods. AM produces less overall utility loss than k-anonymity without suppression, while also reducing utility loss inequality. As shown in Figure 5.18, AM reduces the disparity in data utility between the super-minority race and the other racial subpopulations more often than the other de-identification methods. k-anonymization with suppression produces the greatest utility loss for the super-minority group when the super-minority group is very small. This is because when the super-minority group makes up a sufficiently small proportion of the overall population, k-anonymization with suppression actually suppresses all of the super-minority's records (Figure 5.20).

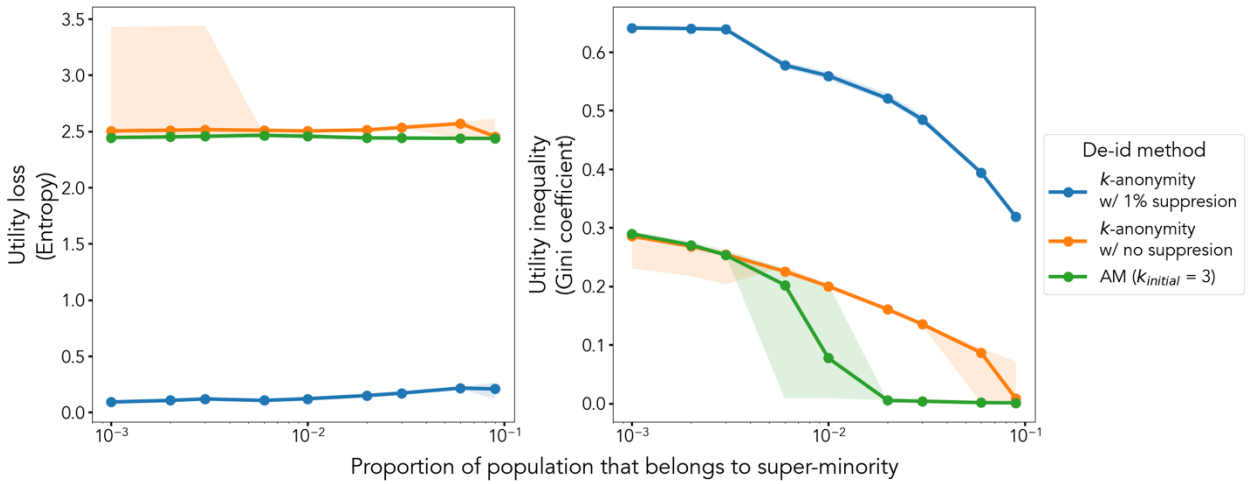

Figure 5.17. (Left) Overall utility loss, measured as entropy (see Eqn. 4.2), when applying each de-identification method to simulated data within which the super-minority racial subpopulation makes up a different proportion of the overall population. (Right) Inequality in utility loss between racial subgroups measured as the Gini coefficient of the race-specific utility loss values. All de-identification methods are applied with a k value of 30. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations.

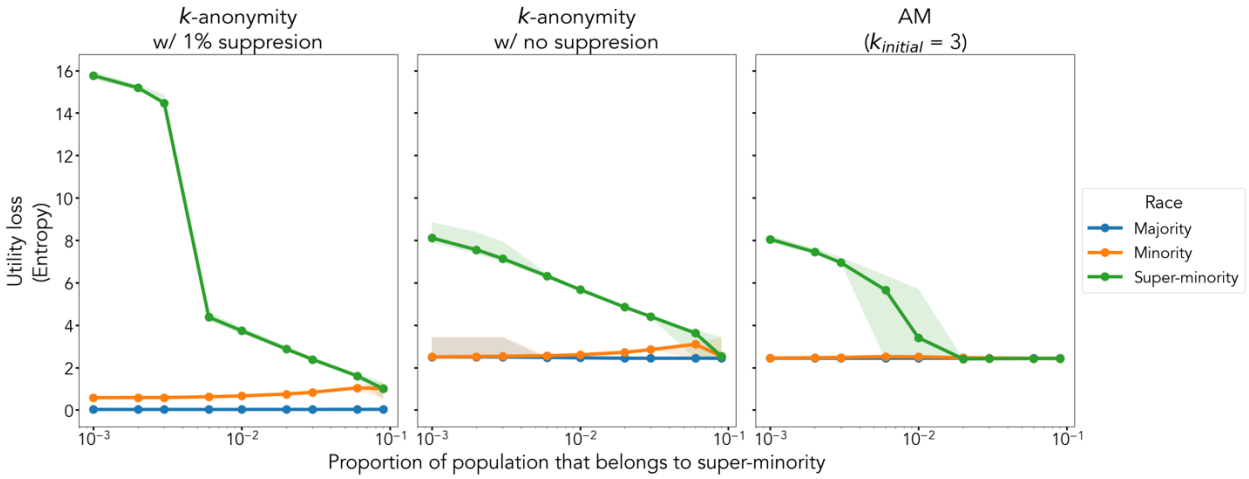

Figure 5.18. Race-specific utility loss curves when de-identifying the simulated data with varying proportions of records corresponding to the super-minority race. (Left) OLA k-anonymization algorithm with up to 1% of all records suppressed. **(**Center) OLA algorithm with no records suppressed. (Right) Altruistic Masking where $k_{initial} = 3$. All de-identification methods are applied with a k value of 30. Utility loss is measured as entropy according to Eqn. 4.2. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations.

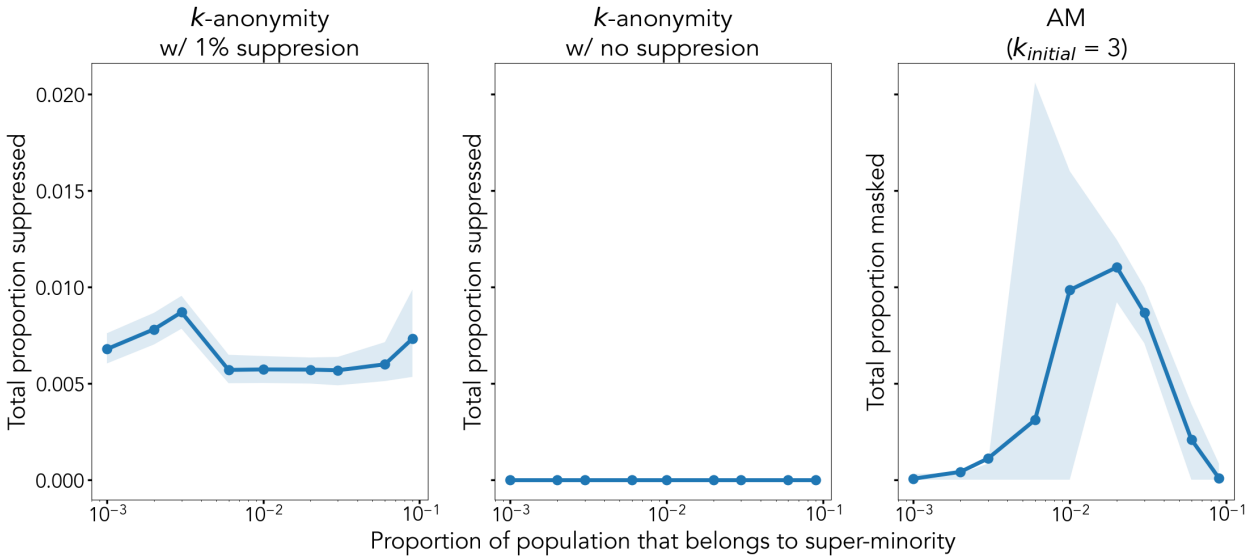

Figure 5.19. Total proportion of records either suppressed or masked in the simulated data by the de-identification methods while varying the relative size of the super-minority population. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. All de-identification methods are applied with a *k* value of 30.

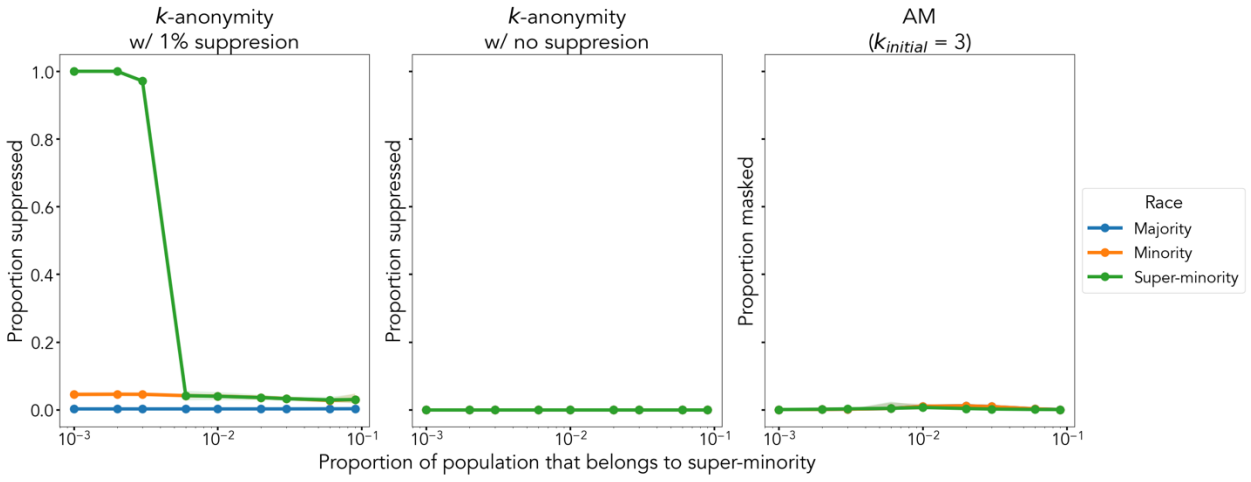

Figure 5.20. Race-specific suppression and masking rates when de-identifying simulated data while varying the relative size of the super-minority population. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. All de-identification methods are applied with a *k* value of 30.

5.5.3.2 Adult dataset

I repeat the intrinsic utility evaluation on the Adult dataset. AM is implemented with $k_{initial} = 11$ and k (k_{target}) values ranging from 11 to 40. Figures 5.21-24 display the results in the same format as that of the simulated data.

The results are similar to those of the simulated data. *k*-anonymity with suppression again minimizes overall utility loss (Figure 5.21). It also disproportionately suppresses records corresponding to the minority racial subpopulations, while AM masks more records from the majority subpopulation (White race) than from the minorities (Figure 5.23). In contrast to the simulated data, AM does not produce the least utility inequality for all *k* values (Figure 5.21). It does, however, consistently produce relatively low utility inequality. It also consistently retains greater overall data utility compared to *k*-anonymity without suppression. Furthermore, masking minimally degrades each racial subgroup's utility after the initial *k*-anonymization to $k_{initial} = 11$ (Figure 5.22).

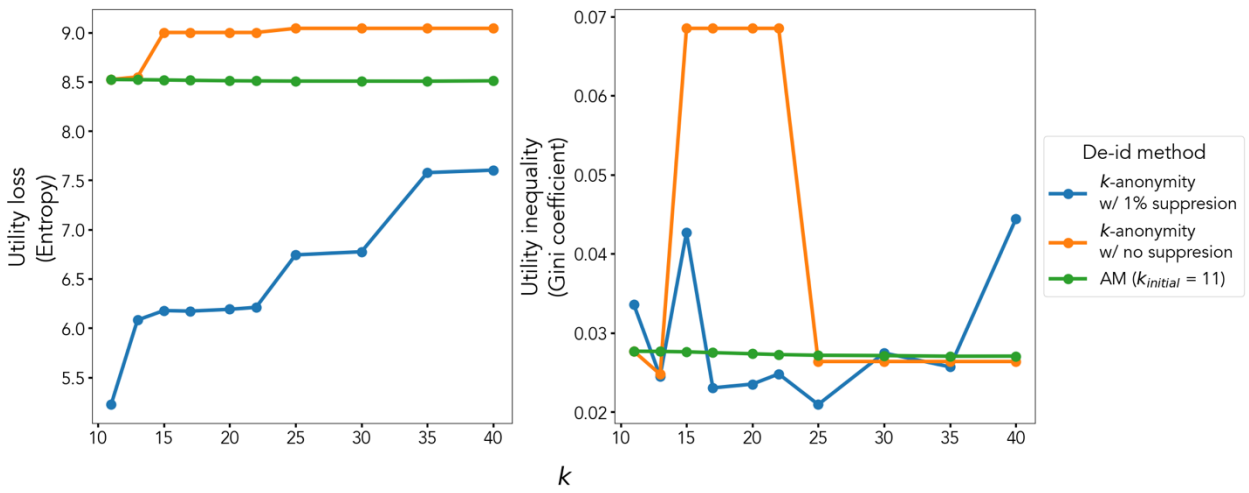

Figure 5.21. (Left) Overall utility loss, measured as entropy (see Eqn. 4.2), when applying each de-identification method at k values ranging from 3 to 50 to the Adult dataset. (Right) Inequality in utility loss between racial subgroups measured as the Gini coefficient of the race-specific utility loss values.

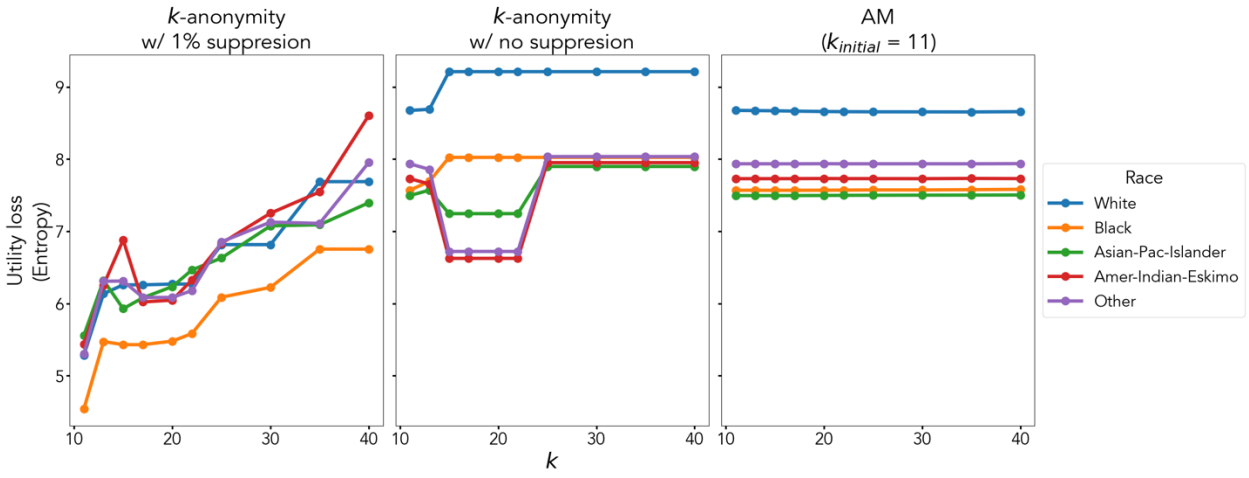

Figure 5.22. Race-specific utility loss curves when de-identifying the Adult dataset at varying levels of k. (Left) OLA k-anonymization algorithm with up to 1% of all records suppressed. (Center) OLA k-anonymization algorithm with no records suppressed. (Right) Altruistic Masking where $k_{initial} = 11$.

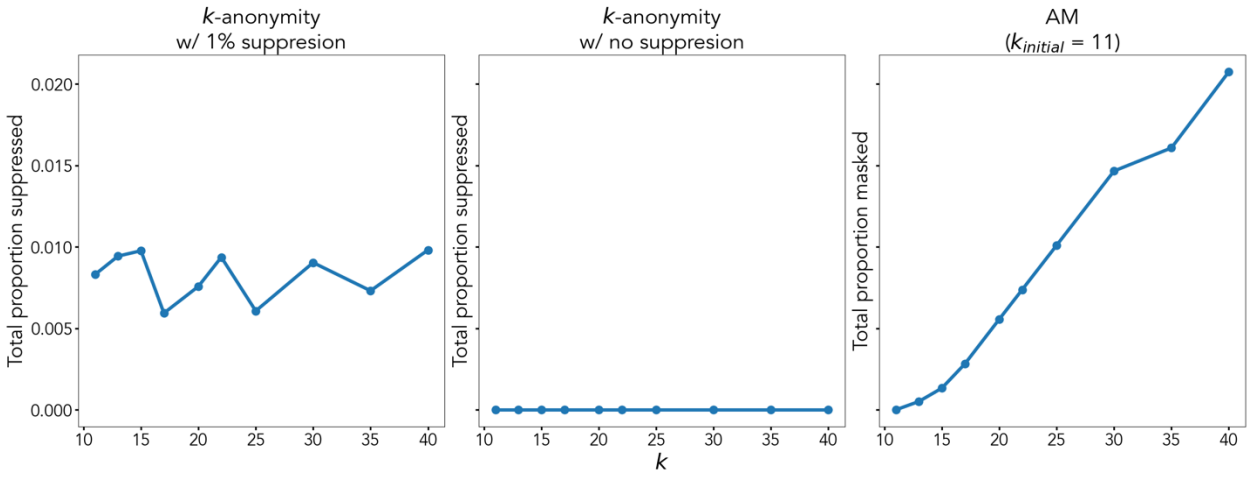

Figure 5.23. Total proportion of records either suppressed or masked when de-identifying the Adult dataset at varying levels of *k*.

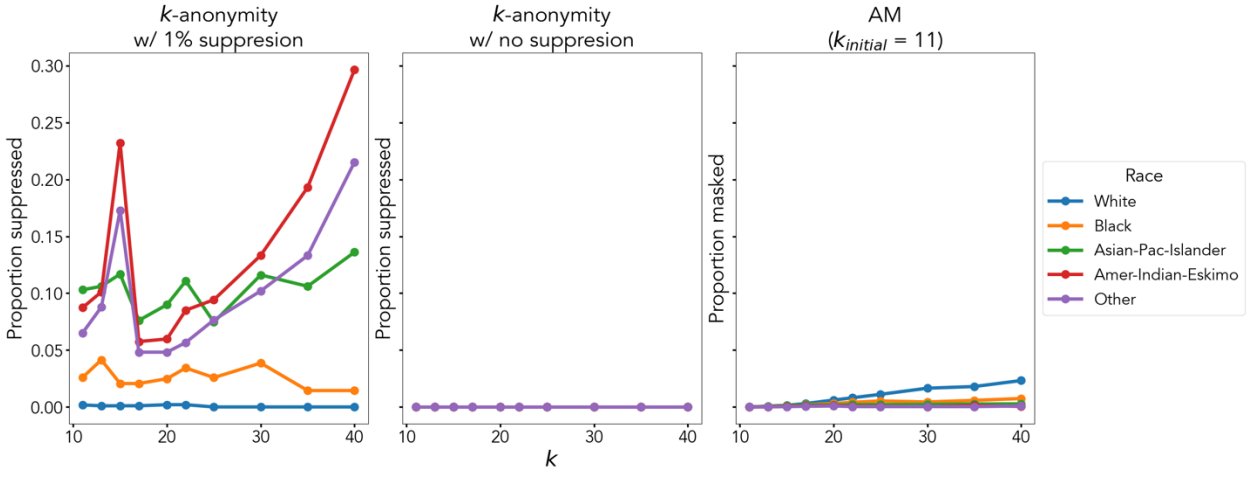

Figure 5.24. Race-specific suppression and masking rates when de-identifying the Adult dataset at varying levels of *k*.

5.5.4 Disparity detection utility evaluation

Since de-identification methods can mask the evidence of disparities[14,15], as shown in Chapter 3, in this section, I evaluate AM's ability to preserve such evidence compared to the standard *k*-anonymization methods. In the simulated data, I randomly assign a value for a binary outcome according to a race-specific rate. The rates are {majority: 0.1, minority: 0.2, super-minority:0.5}. In the Adult dataset, the outcome is defined as whether or not the individual has an annual salary of more than $50,000 dollars.

I estimate disparities in the binary outcomes between racial subpopulations by applying a logisitic regression model. For the simulated dataset, the dependent variables include the three quasi-identifying features. The baseline values for categorical variables are {*race_White, gender_Male*}. *Age* is treated as a continuous variable, regardless of generalization. For the Adult dataset, the dependent variables include the seven quasi-identifying attributes and the *hours-per-week* variable. The baseline values for categorical variables are {*race_White, gender_Male, marital-status_Married, native-country_US, workclass_Private, occupation_white_collar*}. *Age, educational-num*, and *hours-per-week* are treated as continuous variables. For each dataset, the model includes only first order terms. In scenarios in which race values are generalized into coarser representations (i.e., minority and super-minority combined into "Other"), I assign the same estimated odds ratio to each racial subpopulation corresponding to the generalized value. Where an odds ratio cannot be estimated – for example, when records for a particular racial subgroup are not present in the de-identified dataset – I assign an odds ratio of 1. The more accurate the estimated odds ratio is, the better I consider the de-identification method to support disparity detection.

5.5.4.1 Simulated data

For the simulated data, I first evaluate disparity detection performance when varying the level of *k*. Similar to the intrinsic utility evaluation, the probability distribution for the race values is defined as: {majority=0.9, minority=0.09, super-minority=0.01}.

Figure 5.25 displays the odds ratio estimates for the minority and super-minority populations across 100 independent simulations. *k*-anonymity with suppression supports the most accurate odds ratio estimates up to *k*=ll. Afterward, the minority and super-minority race values are generalized to "Other" and the same odds ratio is estimated for both groups. *k*-anonymity without suppression runs into the same problem at *k*=5. The expected odds ratios estimates from the dataset sharing AM are lesser than the true value for the super-minority race and greater than the true value for the minority race. This is because initializing the

dataset to $k_{initial} = 3$ involves generalizing the minority and super-minority race value to "Other" in a fraction of the simulations, as indicated by the 95% quantile ranges. Nonetheless, increasing the number of records masked (i.e., as k_{target} increases while k_{intial} remains constant) does not change the odds ratios estimates at expectation. Therefore, on average, AM supports more accurate odds ratio estimation, for both the minority and super-minority populations, than the *k*-anonymity implementations at higher *k* values.

Figure 5.26 displays the percentage of simulations in which the odds ratio estimate is statistically significant to a *p*-value of 0.05. Disparities for both the minority and super-minority subpopulations are detected in every simulation for each de-identification method.

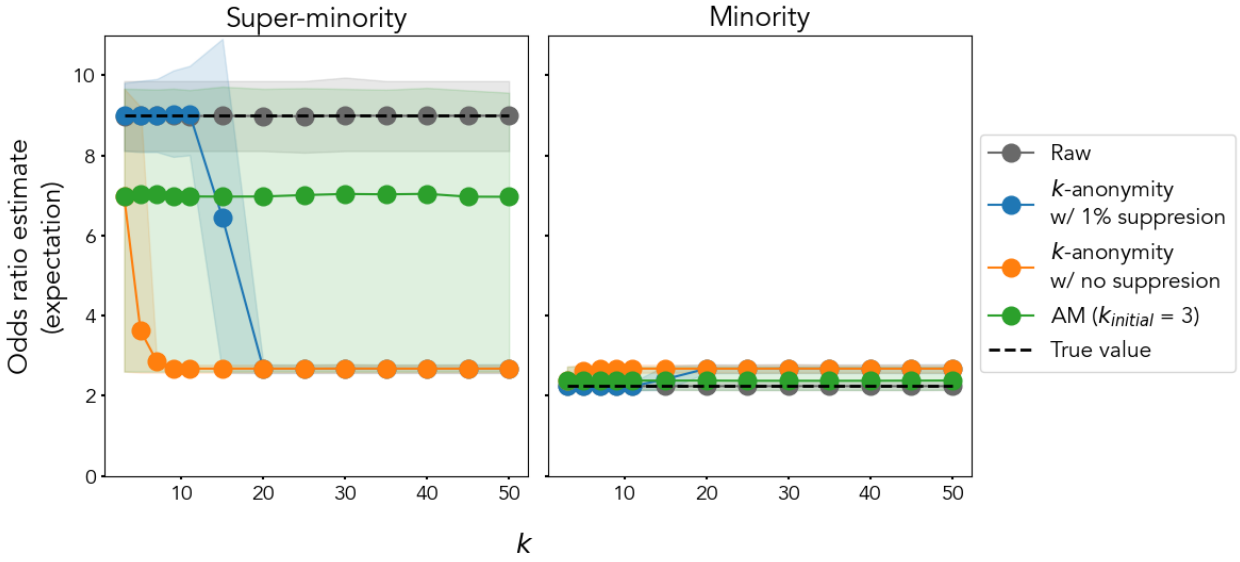

Figure 5.25. Odds ratio estimates for racial disparities in simulated data that has been transformed by different de-identification methods, when varying the level of *k*. Race probability distribution is defined as {majority=0.9, minority=0.09, super-minority=0.01}. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations.

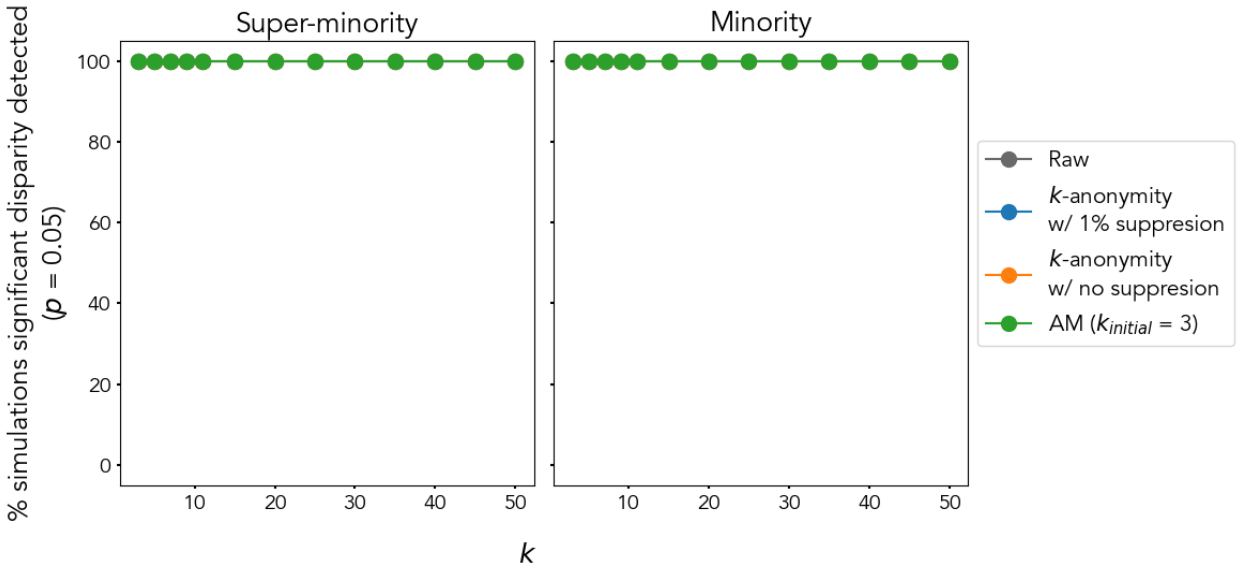

Figure 5.26. Percentage of 100 simulations in which odds ratio estimate (displayed in Figure 5.25) has a *p*-value less than 0.05. Race probability distribution is defined as {majority=0.9, minority=0.09, super-minority=0.01}.

I repeat the experiment, varying the race value probability distribution in the data simulation process in the same manner as the intrinsic utility evalution. For the *k*-anonymization implementations, k = 30. For AM, $k_{initial}$ = 3 and k_{target} = 30. Figure 5.27 displays the average odds ratio estimates, calculated across 100 simulations. Figure 5.28 displays the percentage of simulations the odds ratios meet the statistical significance threshold of 0.05.

Figure 5.27 shows that AM supports the most accurate odds ratio estimates of the de-identification implementations. When the proportion of the population corresponding to the super-minority race is 0.001, AM and *k*-anonymization without suppression provide essentially identical estimates. However, performance begins to diverge when the super-minority's proportion is 0.002, with AM's performance equaling that of the raw data starting at the super-minority proportion value of 0.02. *k*-anonymity without suppression's performance never equals that of the raw data, but consistently improves as the proportion of the population corresponding to the super-minority groups increases. Notably, *k*-anonymization with suppression supports the worst disparity detection performance, despite it minimizing the data's overall intrinsic utility loss (Figure 5.13). In fact, the disproportionate suppression rates of minority and super-minority records (see Figure 5.20) consistently inhibit disparity detection. When the proportion all of population corresponding to the super-minority group is < 0.006, this *k*-anonymity implementation wholly

suppresses the super-minority population's representation in the de-identified dataset such that the disparity cannot be detected (see Figures 5.27 and 5.28).

Figure 5.27 Odds ratio estimates for racial disparities in simulated data that has been transformed by different de-identification methods, when varying proportion of the dataset corresponding to the super-minority race. Expected values (lines) and 95% quantile ranges (shaded areas) are calculated from 100 independent simulations. k=30 for all de-identification methods.

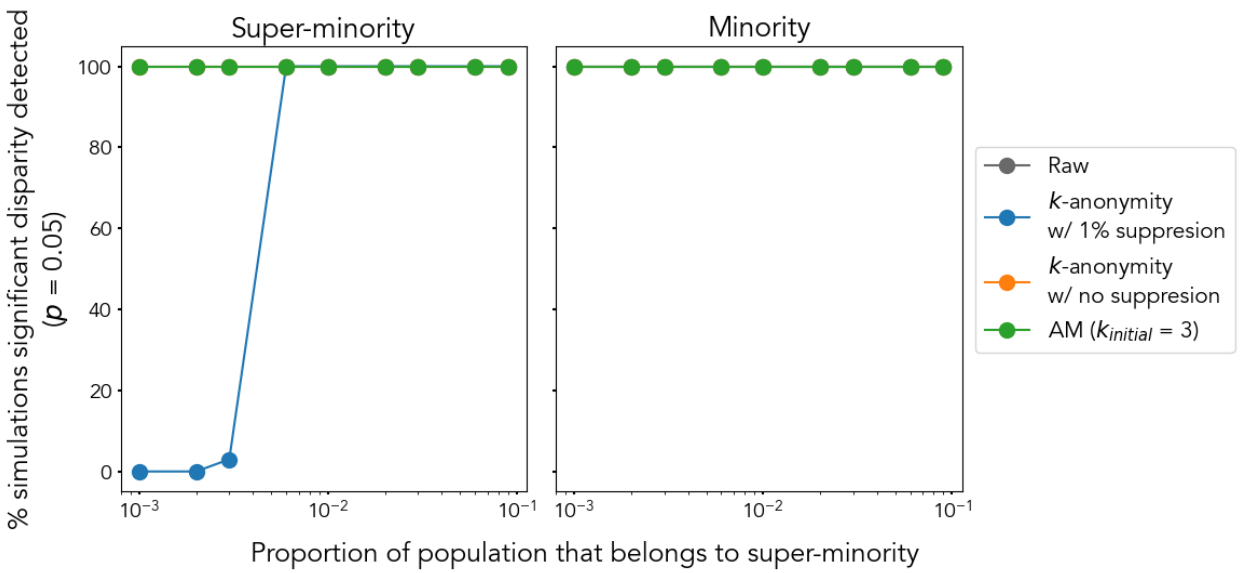

Figure 5.28 Percentage of 100 simulations in which odds ratio estimate (displayed in Figure 5.27) has a *p*-value less than 0.05. k=30 for all de-identification methods.

5.5.4.2 Adult dataset

The race-specific odds ratio estimates and *p*-values for the Adult dataset, when varying the value of k, are displayed in Figure 5.29 and 5.30, respectively. In this case, the true odds ratios for racial disparities in having a salary above $50,000 dollars are not known. As such, the results highlight how the odds ratio estimates change when using the raw data vs. the de-identified versions.

Figure 5.29 shows that k-anonymization implementations have an inconsistent effect on the odds ratio estimates, particularly at smaller k values. k-anonymization with suppression is the least stable, where sometimes it flips the direction of the disparity. For example, the odds ratio for Asian-Pac-Islander changes from <1 to > 1. k-anonymization without suppression does not flip the direction of the disparity, but it does change the estimated magnitude of the disparity. AM provides consistent odds ratio estimates across k-values. Indeed, the odds ratio estimates depend more on the generalization applied to the dataset when initializing the dataset to $k_{initial}$ than by the increasing number of records masked. The initial generalization combines Amer-Indian-Eskimo, Asian-Pac-Islander, and Other to "Not Black or White". As such, the same odds ratio is estimated for the three original racial subpopulations. Whereas the Amer-Pac-Islander and Other estimates from AM are very similar to those from the raw data, the Amer-Indian-Eskimo odds ratio from AM underestimates the disparity compared to the raw data.

In terms of the de-identification methods' affect on the statistical significance of the estimated odds ratios, as shown in Figure 5.30, the *k*-anonymity implementations are again unstable. At some *k* values, *k*-anonymity with and without suppression produces statistically significant odds ratio when the raw data does not, and vice versa. *k*-anonymity without suppression also produces. Notably, whereas masking has little effect on the magnitude of the odds ratio estimates (Figure 5.29), it can decrease the *p*-value of the estimated coefficients (Figure 5.30). As such, AM at higher masking rates (or greater difference between k_{target} and $k_{initial}$) can produce statistically significant coefficients when the raw data does not.

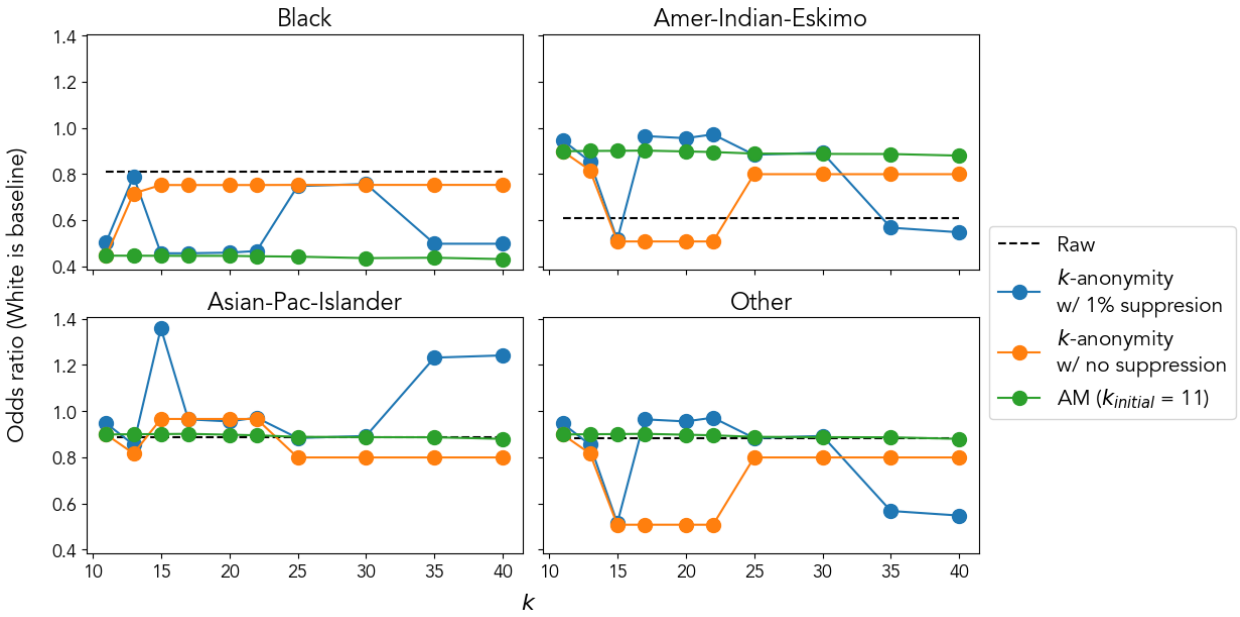

Figure 5.29 Odds ratio estimates for racial disparities in the Adult dataset with respect to the binary outcome of having an annual salary greater than $50,000.

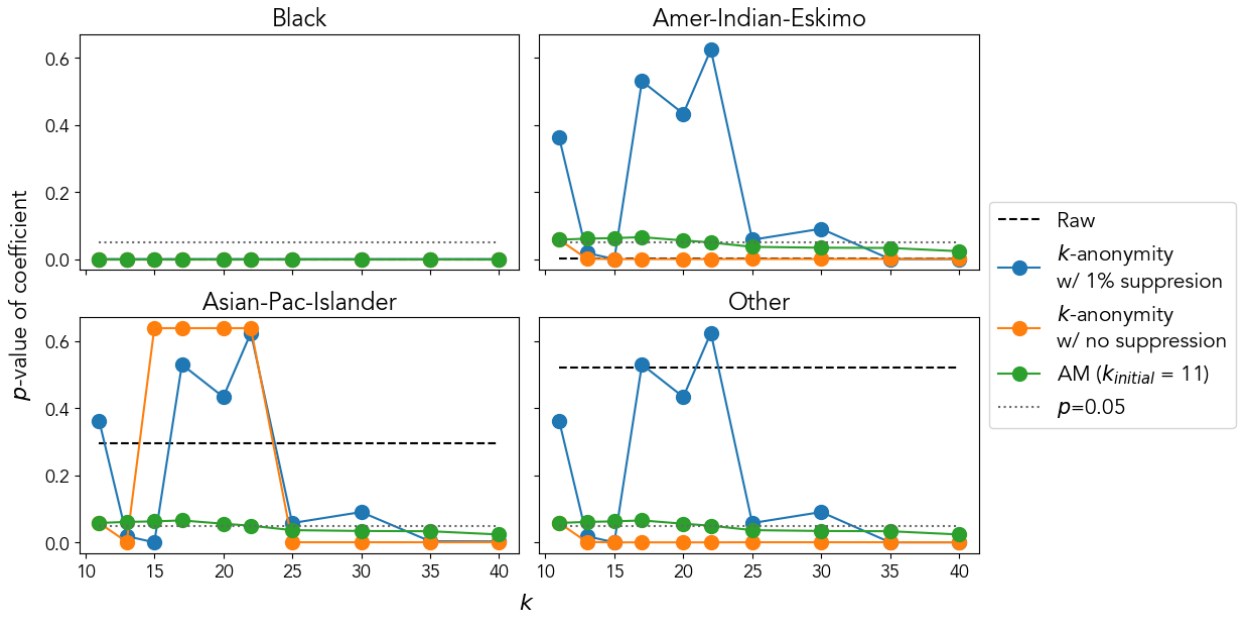

Figure 5.30 *p*-values for the estimated odds ratio coefficients in Figure 5.29.

5.6 Discussion

The aim of this work was to develop a non-deterministic de-identification method that relaxes the fairness constraints defined by the fairness tradeoff theorem, and identify any additional constraints to equailizing privacy risk and data utility in a de-identified dataset.

In this chapter, I developed a de-identification method that leverages non-determinisitic data transformations to mimic the protective effect of missing data such that privacy protections can be cooperatively distributed between subgroups of records. Such cooperation allows larger equivalence classes to altruistically reduce the distinguishabilty of the smaller equivalence classes, increasing the data utility the smaller equivalence classes would otherwise retain. I derived the privacy protections Altruistic Masking can and cannot provide as well as developed an algorithm that implements a relaxed form of fair privacy protections while reducing the distortion applied to minority groups' records. The resulting data preserved minorities' data intrinsic utility better than standard *k*-anonymization, which was critical for supporting disparity detection. This was true in both simulated and real-world data. Furthermore, while I defined the fairness evaluation in terms of equal performance between racial subpopulations, AM can generalize to any quasi-identifying variable to support specific fairness criteria.

Nonetheless, the improvement in fairness with respect to data utility comes at a price. AM cannot provide the same privacy guarantees as *k*-anonymity and other group-based privacy protection methods. AM instead provides fair privacy protections creating an equal floor to the adversary's expected effort to re-identify a target individual. This highlights another fairness tradeoff in de-identification: more equal data utility can be provided at the expense of certain privacy guarantees, and vice versa. Moreover, the utility evaluations showed that AM cannot entirely equalize the distribution of data utility. The utility distribution of a masked dataset is still constrained to that of the initial *k*-anonymization, where $k = k_{initial}$.

5.7 Limitations and future directions

I acknowledge several limitations to guide future directions. First, AM's privacy protections depend on the adversary's rationality. AM assumes that increasing an adversary's expected effort to re-identify a target individual will deter the adversary from attempting re-identification. While the effectiveness of such deterrence has been shown theoretically[49,65], and arguably implied by the scant evidence of real-world re-identifications[187,188], it is still possible an adversary will repeatedly attack the dataset to re-identify a target

individual. Given that AM's privacy protections are only fair at expectation, more distinguishable records that are not randomly masked could remain susceptible to re-identification. There are two potential solutions to this problem. First, privacy protections against motivated adversary's can be enhanced by sharing the dataset in a monitored environment in which the re-identification attempts could be detected. However, as discussed in Chapter 4, this comes at the cost of data accessibility. Second, a data steward could reapply the AM algorithm such that a different version of the dataset is shared with each data recipient. This would make it so that the same record does not remain unmasked against every potential adversary, making it less likely the adversary's target individual is unmasked. However, this would create potential for collusion. Were data recipients to combine several versions of the dataset, they may be able to reverse the masked values. Nevertheless, data sharing frameworks that prevent collusion could mitigate this risk. Future work should evaluate the privacy protections of AM, with and without additional deterrents and AM applications, against real-world attacks.

Second, I developed AM to protect a re-identification attack in which a single individual is targeted. AM could provide better or worse privacy protections against other re-identification attacks, particularly those in which the adversary attempts to re-identify more than one individual[40,48]. The fairness of the re-identification protections may also vary against diverse attacks. Furthermore, I did not consider how to incorporate AM into privacy models that protect against other types of privacy disclosures, such as *l*-diversity protecting against sensitive attribute disclosures[39]. Future work should develop masking methods that protect against diverse attack methods and types.

Third, AM may be susceptible to imputation and reverse-engineering. As imputation methods continuously improve, it may be possible that the masked values could be imputed from the residual information[189]. It may also be possible for the adversary to reverse-engineer the original dataset with knowledge of the AM algorithm, the masked dataset, and certain background knowledge (i.e., knowledge of the dataset's original distribution). Future work should investigate how well masked values could be imputed and reverse-engineered.

Fourth, the utility evaluation results for the Adult dataset revealed there are more nuances to fairness in de-identification. While the simulated data clearly showed that standard de-identification methods reduced racial subgroup's data utility inversely proportional to their relative size in the population, the Adult dataset did not. In some cases, the largest racial subpopulation – the White race group – incurred the greatest utility loss. This is likely due to presence of the additional quasi-identifying attributes with skewed distributions. For example, if the White group had a more skewed distribution in terms of *workclass* than the other racial

groups, then generalizing that variable could lead to greater utility loss (as defined by entropy[6,74]) for the White group. Nevertheless, even with the additional variables, the results showed that standard *k*-anoymization algorithms suppressed records corresponding to racial minorities more often than the White group. AM did not. Future work should evaluate AM's ability to support detection of more complex disparities as well as support more diverse applications.

Finally, I only considered fairness with respect to a single attribute. Ideally, de-identified data could fairly distribute the privacy risk and utility with respect to several attributes. Future work should investigate how to mask data in a way that supports fairness with respect to multiple attributes, and whether optimizing fairness with respect to one variable sacrifices the fairness with respect to others.

5.8 Conclusion

Non-deterministic de-identification methods can preserve the representation of minority subpopulations better than deterministic methods. I showed that such representation was critical for identifying underlying disparities. However, increasing fairness with respect to utility may require reducing fairness with respect to certain privacy guarantees. Once again, the smaller and more distinguishable populations are disadvantaged. While no silver bullet, Altruistic Masking represents another method by which fairness constraints can be relaxed in privacy-preserving data sharing. It also highlights nuances between privacy protections and privacy guarantees that may illuminate innovative solutions in the future.

Chapter 6

Conclusion

6.1 Summary

As the demand for data increases, so does the need to develop data sharing methods that both protect individuals' right to privacy and preserve equitable representation. This dissertation focused on developing de-identification and data sharing methods to address such needs, both within the context of a pandemic as well as for biomedical research more generally. It also identified privacy, utility, and fairness constraints to guide the development of privacy practice and regulation moving forward. The specific contributions are as follows.

First, I developed a prospective and dynamic de-identification method, driven by a privacy risk estimation framework, to support pandemic data sharing. The framework enables a dataset that accumulates additional records at variable rates to be de-identified and shared in near-real time. The framework can also generalize to consider different types of quasi-identifier features, particular information priorities, and adversaries with varying background knowledge. I showed how data that is dynamically generalized according to the framework's forecasts can both decrease patient distinguishability and support early and accurate disparity detection. The framework can also support the development of large data sharing consortia by informing how many subjects must be recruited from particular populations to meet privacy and utility thresholds. While these methods can optimize the privacy-utility tradeoff with respect to sharing dynamic datasets, the experiments highlighted minority groups' disadvantage in de-identified data. The fairness experiments showed that more distinguishable groups are disproportionately exposed to re-identification and/or disproportionately distorted by de-identification transformations. This finding added to growing evidence of the privacy risk and data utility disparaties of de-identification[14,15,71], which extend beyond pandemic data sharing.

Second, I formalized and empirically illustrated the constraints to concurrently equalizing the distribution of risk and utility in a de-identified dataset. The formalization included the fairness tradeoff theorem, which states that when records start with different re-identification risks in the raw data, deterministic generalization and suppression transformations can either equalize risk or utility. They cannot equalize both simultaneously. As nearly all real-world datasets are expected to have records with variable

distinguishability prior to de-identification, the fairness tradeoff theorem defines an ethical dilemma that data stewards, data users, and policy makers must face. That is, they must decide whether to prioritize equal privacy or equal utility in de-identified data. I discussed several implications of the theorem in the context of current privacy legislation and several foundational research principles, highlighting regulations' current prioritization of privacy over utility. I also discussed how the fairness constraints of de-identification transformations can be alleviated by supplementing de-identification with sociotechnical deterrents. However, doing so comes at the cost of data accessibility. To alleviate the constraints to fair privacy risk, fair data utility, and data accessibility, I proposed a scaleable controlled-access framework to data sharing, called the passport-visa model. The passport-visa model distributes the burden of verifying user trustworthiness and deterring data misuse between sponsoring insitutions (those who provide the passport) and data sharing organizations (those who grant the visa and share the data). Nevertheless, the passport-visa model involves monitoring data users, to some extent, such that data user privacy must be considered.

Finally, I developed a non-deterministic de-identification method to alleviate the constraints defined by the fairness tradeoff theorem. The method, called Altruistic Masking, breaks the conventions of standard de-identification transformations to allow privacy protections to support cooperation in privacy protections. AM leverages such cooperation to transform the majority groups' records in a manner that increases the minority groups' privacy protections, enabling minorites to retain more granular representation in the de-identified dataset while still meeting specified privacy risk thresholds. I showed how AM reduces re-identification risk by increasing an adversary's expected effort to re-identify a target individual. However, while AM can provide an equal floor to the adversary's expected effort, I also showed how it cannot equally provide the same privacy guarantees as *k*-anonymity and other group-based privacy protection methods. Nevertheless, the utility evaluation demonstrates how sacrificing equality in certain privacy guarantees can improve minority subpopulations' representation in a de-identified dataset beyond that afforded by deterministic transformation methods, enabling more accurate disparity detection and characterization.

Collectively, this dissertation identifies several constraints to de-identification and data sharing beyond the privacy-utility tradeoff – the focus in the privacy research community for decades. The fairness tradeoff theorem defines a tradeoff between fair privacy risk and fair data utility. Sharing data using the passport-visa model, and other controlled access models, imposes a tradeoff between relaxing the constraints of the fairness tradeoff theorem and data accessibility. Implementing a controlled access model may also impose a tradeoff between data user privacy and data subject privacy. Finally, the Altruistic Masking method defines a tradeoff between equally providing certain privacy guarantees and more equally preserving data utility. While this dissertation does not identify a silver bullet to address these tradeoffs – if one even exists

– the constraints reframe the problem of privacy-preserving data sharing. Indeed, if data is to be shared in a way that both preserves privacy and supports society's pursuit of health equity, solutions must consider the constraints to privacy, utility, fairness, and accessibility.

6.2 Future investigations

Beyond those already stated at the conclusion of Chapters 3-5, there are several limitations to this work that should guide future investigations. First, the fairness investigations of Chapters 4 and 5 consider datasets retrospectively; they do not consider dynamic datasets such as those in Chapter 3. While the constraints defined by the fairness tradeoff theorem apply to any deterministic application of generalization and suppression, future work should consider how to adapt non-deterministic transformations, such as AM, to dynamic datasets.

Second, the investigations do not consider how patient recruitment can impact the distribution of privacy risk in the raw dataset and, subsequently, the constraints to fair privacy and fair utility. Even though the empirical results using Census data in Chapter 4 indicate there are population constraints equalizing group size in the raw dataset, patient recruitment may be optimized to alleviate the fairness constraints[190]. Future work should investigate recruitment strategies that can prospectively optimize for fairness in de-identified data.

Finally, I did not consider how de-identified data can support the development of artificial intelligence. De-identification could satisfy AI's growing demand for data while preserving patients' privacy[173]; however, there is still limited understanding how generalization and suppression may impact the ability to develop algorithms that are high performing, just, and fair[172,173]. As shown in Chapter 3, generalization can dampen the noise to improve model performance. It could also be argued that the more granular the data, or the greater the data's intrinsic utility, the more flexibility a developer has for feature engineering and signal detection. Nevertheless, the correlation between de-identified data that has been generalized and suppressed in a manner that optimizes intrinsic utility measures and model performance has been shown to be inconsistent[91,92,191]. Future work should investigate how de-identified data may affect the development of fair AI technologies.

References

1. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary Use of Clinical Data: the Vanderbilt Approach. J Biomed Inform. 2014 Dec;52:28–35.

2. Warren S, Brandeis L. The right to privacy. In: Killing the Messenger. Columbia University Press; 1989. p. 1–21.

3. Office for Human Research Protections US Department of Health and Human Services. Federal Policy for the Protection of Human Subjects ('Common Rule') [Internet]. HHS.gov. 2009. Available from: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

4. Office for Civil Rights US Department of Health and Human Services. The HIPAA Privacy Rule [Internet]. HHS.gov. 2008. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

5. California Consumer Privacy Act (CCPA) [Internet]. State of California - Department of Justice - Office of the Attorney General. 2018. Available from: https://oag.ca.gov/privacy/ccpa

6. El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, et al. A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. J Am Med Inform Assoc. 2009;16(5):670–82.

7. Emam KE, Jonker E, Moher E, Arbuckle L. A Review of Evidence on Consent Bias in Research. The American Journal of Bioethics. 2013 Apr 1;13(4):42–4.

8. Office for Civil Rights US Department of Health and Human Services. Summary of the HIPAA Privacy Rule [Internet]. HHS.gov. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

9. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar 1;17(2):169–77.

10. Standards for Privacy of Individually Identifiable Health Information [Internet]. Federal Register. 2000. Available from: https://www.federalregister.gov/documents/2000/12/28/00-32678/standards-for-privacy-of-individually-identifiable-health-information

11. Garfinkel SL. De-identification of personal information. National Institute of Standards and Technology; 2015 Oct p. NIST IR 8053. Report No.: NIST IR 8053.

12. Garfinkel S, Near J, Dajani A, Singer P, Guttman B. De-Identifying Government Datasets: Techniques and Governance. US Department of Commerce, National Institute of Standards and Technology; 2023.

13. Ekstrand MD, Joshaghani R, Mehrpouyan H. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In: Conference on Fairness, Accountability and Transparency. PMLR; 2018. p. 35–47.

14. Xu H, Zhang N. Implications of data anonymization on the statistical evidence of disparity. Management Science. 2021;68(4):2600–18.

15. Xu H, Zhang N. Privacy in Health Disparity Research. Medical Care. 2019 Jun;57:S172.

16. Fioretto F, Tran C, Van Hentenryck P, Zhu K. Differential privacy and fairness in decisions and learning tasks: A survey. arXiv preprint arXiv:220208187. 2022;

17. Steed R, Liu T, Wu ZS, Acquisti A. Policy impacts of statistical uncertainty and privacy. Science. 2022;377(6609):928–31.

18. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 May;20(5):533–4.

19. The United States Census Bureau. Population Census Tables [Internet]. Available from: https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

20. Brown JT, Yan C, Xia W, Yin Z, Wan Z, Gkoulalas-Divanis A, et al. Dynamically adjusting case reporting policy to maximize privacy and public health utility in the face of a pandemic. JAMIA. 2022 Feb 19;29(5):853–63.

21. Brown JT, Wan Z, Gkoulalas-Divanis A, Kantarcioglu M, Malin BA. Supporting COVID-19 disparity investigations with dynamically adjusting case reporting policies. In: AMIA Annual Symposium Proceedings. 2022.

22. Webb Hooper M, Nápoles AM, Pérez-Stable EJ. COVID-19 and Racial/Ethnic Disparities. JAMA. 2020 Jun 23;323(24):2466–7.

23. Romano SD, Blackstock AJ, Taylor EV, El Burai Felix S, Adjei S, Singleton CM, et al. Trends in racial and ethnic disparities in COVID-19 hospitalizations, by region — United States, March–December 2020. MMWR Morb Mortal Wkly Rep. 2021 Apr 16;70(15):560–5.

24. McLaren J. Racial disparity in COVID-19 deaths: seeking economic roots with census data. The BE Journal of Economic Analysis & Policy. 2021 Jul 1;21(3):897–919.

25. Cheng V, Suriyakumar VM, Dullerud N, Joshi S, Ghassemi M. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2021. p. 149–60. (FAccT '21).

26. Goldstein MM, Pewen WF. The Hipaa Omnibus Rule: Implications for Public Health Policy and Practice. Public Health Rep. 2013 Nov 1;128(6):554–8.

27. Office for Civil Rights US Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [Internet]. HHS.gov. 2012. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

28. Office for Civil Rights US Department of Health and Human Services. Standards for privacy of individually identifiable health information. Final rule. Fed Regist. 2002 Aug 14;67(157):53181–273.

29. Virginia Consumer Data Protection Act Signed Into Law | Lerman Senter [Internet]. Available from: https://www.lermansenter.com/internet-e-commerce/2021/03/08/virginia-consumer-data-protection-act/

30. And Now There are Three …. The Colorado Privacy Act [Internet]. The National Law Review. Available from: https://www.natlawreview.com/article/and-now-there-are-three-colorado-privacy-act

31. The Utah Consumer Privacy Act: Utah Becomes Fourth US State with Comprehensive Privacy Law [Internet]. JD Supra. Available from: https://www.jdsupra.com/legalnews/the-utah-consumer-privacy-act-utah-2977882/

32. US State Privacy Legislation Tracker [Internet]. Available from: https://iapp.org/resources/article/us-state-privacy-legislation-tracker/

33. Utah Consumer Privacy Act Newest State Privacy Act Signed into Law [Internet]. The National Law Review. Available from: https://www.natlawreview.com/article/utah-consumer-privacy-act-newest-state-privacy-act-signed-law

34. Brothers KB, Clayton EW. "Human Non-Subjects Research": Privacy and Compliance. The American Journal of Bioethics. 2010 Sep 9;10(9):15–7.

35. What is GDPR, the EU's new data protection law? [Internet]. GDPR.eu. 2018. Available from: https://gdpr.eu/what-is-gdpr/

36. Anonymization and GDPR compliance; an overview - GDPR Summary [Internet]. Available from: https://www.gdprsummary.com/anonymization-and-gdpr/

37. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res. 2019 May 31;21(5):e13484.

38. Office for Civil Rights US Department of Health and Human Services. Disclosures for Emergency Preparedness - A Decision Tool: Limited Data Set (LDS) [Internet]. HHS.gov. 2007. Available from: https://www.hhs.gov/hipaa/for-professionals/special-topics/emergency-preparedness/limited-data-set/index.html

39. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE'06). 2006. p. 24–24.

40. Barth-Jones D. The "Re-Identification" of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. Rochester, NY: Social Science Research Network; 2012 Jul. Report No.: ID 2076397.

41. Xia W, Liu Y, Wan Z, Vorobeychik Y, Kantacioglu M, Nyemba S, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. Journal of the American Medical Informatics Association [Internet]. 2021 Jan 15;(ocaa327). Available from: https://doi.org/10.1093/jamia/ocaa327

42. Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst. 2002 Oct 1;10(05):557–70.

43. Golle P. Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of the 5th ACM workshop on Privacy in electronic society. New York, NY, USA: Association for Computing Machinery; 2006. p. 77–80. (WPES '06).

44. Samarati P, Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. In: Proceedings of the IEEE Symposium on Research in Security and Privacy (S&P). Oakland, CA; 1998.

45. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications. 2019 Jul 23;10(1):3069.

46. Sweeney L. Simple demographics often identify people uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. 2000;

47. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Unc Fuzz Knowl Based Syst. 2002 Oct;10(05):571–88.

48. Dankar FK, El Emam K. A method for evaluating marketer re-identification risk. In: Proceedings of the 2010 EDBT/ICDT Workshops. 2010. p. 1–10.

49. Xia W, Kantarcioglu M, Wan Z, Heatherly R, Vorobeychik Y, Malin B. Process-Driven Data Privacy. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia: Association for Computing Machinery; 2015. p. 1021–30. (CIKM '15).

50. Dwork C. Differential privacy. In: Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II. Berlin, Heidelberg: Springer-Verlag; 2006. p. 1–12. (ICALP'06).

51. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing. 2022;493:28–45.

52. El Emam K, Dankar FK. Protecting Privacy Using k-Anonymity. JAMIA. 2008;15(5):627–37.

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

54. Domingo-Ferrer J, Sanchez D, Blanco-Justicia A. The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning) [Internet]. Available from: https://cacm.acm.org/magazines/2021/7/253460-the-limits-of-differential-privacy-and-its-misuse-in-data-release-and-machine-learning/fulltext

55. Elliot M. Final report on the disclosure risk associated with the synthetic data produced by the sylls team. Report 2015. 2015;2.

56. El Emam K, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. Journal of medical Internet research. 2020;22(11):e23139.

57. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, et al. A Multifaceted benchmarking of synthetic electronic health record generation models. Nat Commun. 2022 Dec 9;13(1):7609.

58. Missouri Department of Health & Senior Services. Data Release Policy | HIV/AIDS Disease Surveillance [Internet]. Available from: https://health.mo.gov/data/hivstdaids/datareleasepolicy.php

59. Centers for Medicare & Medicaid Services. Cell Size Suppression Policy [Internet]. [cited 2022 Jan 10]. Available from: https://resdac.org/articles/cms-cell-size-suppression-policy

60. Washington Department of Health Agency Standards for Reporting Data with Small Numbers [Internet]. [cited 2021 Jan 25]. Available from: https://www.doh.wa.gov/Portals/1/Documents/1500/SmallNumbers.pdf

61. Wood A, Altman M, Bembenek A, Bun M, Gaboardi M, Honaker J, et al. Differential Privacy: A Primer for a Non-Technical Audience. SSRN Journal. 2018;

62. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet. 2022 Jul;23(7):429–45.

63. Wan Z, Vorobeychik Y, Xia W, Liu Y, Wooders M, Guo J, et al. Using game theory to thwart multistage privacy intrusions when sharing data. Science Advances. 2021;7(50).

64. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Ganta R, et al. A Game Theoretic Framework for Analyzing Re-Identification Risk. PLoS One. 2015 Mar 25;10(3).

65. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Malin B. Expanding Access to Large-Scale Genomic Data While Promoting Privacy: A Game Theoretic Approach. The American Journal of Human Genetics. 2017 Feb 2;100(2):316–22.

66. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014 Jun;15(6):409–21.

67. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific data. 2023;10(1):1.

68. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C ying, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019 Dec 12;6(1):317.

69. Johnson A, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV-Note: Deidentified free-text clinical notes. PhysioNet; 2023.

70. Investigators A of URP. The "All of Us" research program. New England Journal of Medicine. 2019;381(7):668–76.

71. Bowen C, Snoke J. Do No Harm Guide: Applying Equity Awareness In Data Privacy Methods. 2023;

72. Data Access Tiers – All of Us Research Hub [Internet]. [cited 2023 Oct 25]. Available from: https://www.researchallofus.org/data-tools/data-access/

73. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. J Am Med Inform Assoc. 2015 Sep 1;22(5):1029–41.

74. Gionis A, Tassa T. k-Anonymization with Minimal Loss of Information. IEEE Transactions on Knowledge and Data Engineering. 2009;21(2):206–19.

75. Maybank A. Why racial and ethnic data on COVID-19's impact is badly needed. American Medical Association [Internet]. 2020 Apr 8; Available from: https://www.ama-assn.org/about/leadership/why-racial-and-ethnic-data-covid-19-s-impact-badly-needed

76. Rivers C, Chretien JP, Riley S, Pavlin JA, Woodward A, Brett-Major D, et al. Using "outbreak science" to strengthen the use of models during epidemics. Nature Communications. 2019 Jul 15;10(1):3102.

77. Executive Order on Ensuring a Data-Driven Response to COVID-19 and Future High-Consequence Public Health Threats [Internet]. The White House. 2021. Available from: https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/21/executive-order-ensuring-a-data-driven-response-to-covid-19-and-future-high-consequence-public-health-threats/

78. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. In: Proceedings of the AMIA Annual Fall Symposium. 1997. p. 51–5.

79. Bayardo RJ, Agrawal R. Data Privacy through Optimal k-Anonymization. In: 21st International Conference on Data Engineering (ICDE'05) [Internet]. Tokyo, Japan: IEEE; 2005 [cited 2021 Oct 14]. p. 217–28. Available from: http://ieeexplore.ieee.org/document/1410124/

80. LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian Multidimensional K-Anonymity. In: 22nd International Conference on Data Engineering (ICDE'06). Atlanta, GA, USA: IEEE; 2006. p. 25–25.

81. Byun J won, Sohn Y, Bertino E, Li N. Secure Anonymization for Incremental Datasets. In: in SDM, 2006. 2006. p. 48–63.

82. Pei J, Xu J, Wang Z, Wang W, Wang K. Maintaining k-anonymity against incremental updates. In: 19th International Conference on Scientific and Statistical Database Management (SSDBM 2007). IEEE; 2007. p. 5–5.

83. Xiao X, Tao Y. M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. 2007. p. 689–700.

84. Li F, Zhou S. Challenging more updates: Towards anonymous re-publication of fully dynamic datasets. arXiv preprint arXiv:08064703. 2008;

85. Anjum A, Raschia G, Gelgon M, Khan A, Ahmad N, Ahmed M, et al. τ-safety: A privacy model for sequential publication with arbitrary updates. computers & security. 2017;66:20–39.

86. Wang K, Fung BCM. Anonymizing sequential releases. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06. Philadelphia, PA, USA: ACM Press; 2006. p. 414.

87. Shmueli E, Tassa T, Wasserstein R, Shapira B, Rokach L. Limiting disclosure of sensitive data in sequential releases of databases. Information Sciences. 2012 May 15;191:98–127.

88. Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, et al. Algorithmic fairness in computational medicine. eBioMedicine. 2022 Oct 1;84.

89. Chen J, Kallus N, Mao X, Svacha G, Udell M. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2019. p. 339–48. (FAT* '19).

90. Kleinberg J, Ludwig J, Mullainathan S, Rambachan A. Algorithmic fairness. In: Aea papers and proceedings. 2018. p. 22–7.

91. Chester A, Koh YS, Wicker J, Sun Q, Lee J. Balancing Utility and Fairness against Privacy in Medical Data. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI). 2020. p. 1226–33.

92. Chester A, Koh YS, Lee J. Understanding the Effects of Mitigation on De-identified Data. In: Advances and Trends in Artificial Intelligence Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I. 2021. p. 133–44.

93. Zhu K, Van Hentenryck P, Fioretto F. Bias and Variance of Post-processing in Differential Privacy. AAAI. 2021 May 18;35(12):11177–84.

94. McGlinchey A, Mason O. Observations on the bias of nonnegative mechanisms for differential privacy. FoDS. 2020 Nov 30;2(4):429–42.

95. Pujol D, McKenna R, Kuppam S, Hay M, Machanavajjhala A, Miklau G. Fair decision making using privacy-protected data. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2020. p. 189–99. (FAT* '20).

96. Xu D, Du W, Wu X. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. p. 1924–32.

97. Tran C, Dinh M, Fioretto F. Differentially private empirical risk minimization under the fairness lens. Advances in Neural Information Processing Systems. 2021;34:27555–65.

98. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The Problem of Fairness in Synthetic Healthcare Data. Entropy. 2021 Sep;23(9):1165.

99. van Breugel B, Kyono T, Berrevoets J, van der Schaar M. Decaf: Generating fair synthetic data using causally-aware generative networks. Advances in Neural Information Processing Systems. 2021;34:22221–33.

100. Xu D, Yuan S, Zhang L, Wu X. FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets. In: 2019 IEEE International Conference on Big Data (Big Data). 2019. p. 1401–6.

101. Rossen LM, Branum AM, Ahmad FB, Sutton P, Anderson RN. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity — United States, January 26–October 3, 2020. MMWR Morb Mortal Wkly Rep. 2020 Oct 23;69(42):1522–7.

102. Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. Eur J Epidemiol. 2020 Dec 1;35(12):1123–38.

103. Karaca-Mandic P, Georgiou A, Sen S. Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States. JAMA Intern Med. 2021 Jan;181(1):131–4.

104. Parpia AS, Martinez I, El-Sayed AM, Wells CR, Myers L, Duncan J, et al. Racial disparities in COVID-19 mortality across Michigan, United States. eClinicalMedicine [Internet]. 2021 Mar 1 [cited 2022 Feb 4];33. Available from: http://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00041-9/fulltext

105. Zelner J, Trangucci R, Naraharisetti R, Cao A, Malosh R, Broen K, et al. Racial Disparities in Coronavirus Disease 2019 (COVID-19) Mortality Are Driven by Unequal Infection Risks. Clinical Infectious Diseases. 2021 Mar 1;72(5):e88–95.

106. Keating D, Cha AE, Florit G. 'I just pray God will help me': Racial, ethnic minorities reel from higher covid-19 death rates. Washington Post. 2020 Nov 20;

107. Gross CP, Essien UR, Pasha S, Gross JR, Wang S yi, Nunez-Smith M. Racial and Ethnic Disparities in Population-Level Covid-19 Mortality. J Gen Intern Med. 2020 Oct;35(10):3097–9.

108. Hussain-Gambles M, Leese B, Atkin K, Brown J, Mason S, Tovey P. Involving South Asian patients in clinical trials. Health Technol Assess. 2004 Oct;8(42):iii, 1–109.

109. Mills EJ, Seely D, Rachlis B, Griffith L, Wu P, Wilson K, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. The Lancet Oncology. 2006 Feb 1;7(2):141–8.

110. Paskett ED, Reeves KW, McLaughlin JM, Katz ML, McAlearney AS, Ruffin MT, et al. Recruitment of minority and underserved populations in the United States: The centers for population health and health disparities experience. Contemporary Clinical Trials. 2008 Nov 1;29(6):847–61.

111. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. New England Journal of Medicine. 2016 Aug 18;375(7):655–65.

112. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 Oct;538(7624):161–4.

113. 2016 National Healthcare Quality and Disparities Report [Internet]. Available from: https://admin.ahrq.gov/research/findings/nhqrdr/nhqdr16/index.html

114. Adler NE, Rehkopf DH. US disparities in health: descriptions, causes, and mechanisms. Annu Rev Public Health. 2008;29:235–52.

115. Braveman P. What Are Health Disparities and Health Equity? We Need to Be Clear. Public Health Rep. 2014;129(Suppl 2):5–8.

116. Braveman P, Arkin E, Orleans T, Proctor D, Acker J, Plough A. What is health equity? Behavioral science & policy. 2018;4(1):1–14.

117. Cruz TM, Smith SA. Health Equity Beyond Data: Health Care Worker Perceptions of Race, Ethnicity, and Language Data Collection in Electronic Health Records. Medical Care. 2021 May;59(5):379–85.

118. Ibrahim NK. Epidemiologic surveillance for controlling Covid-19 pandemic: types, challenges and implications. J Infect Public Health. 2020 Nov;13(11):1630–8.

119. Thacker SB, Qualters JR, Lee LM. Public Health Surveillance in the United States: Evolution and Challenges* [Internet]. Available from: https://www.cdc.gov/MMWR/preview/mmwrhtml/su6103a2.htm

120. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big Data for Infectious Disease Surveillance and Modeling. The Journal of Infectious Diseases. 2016 Dec 1;214(suppl_4):S375–9.

121. Woolhouse MEJ, Rambaut A, Kellam P. Lessons from Ebola: Improving infectious disease surveillance to inform outbreak management. Science Translational Medicine. 2015 Sep 30;7(307):307rv5-307rv5.

122. Fang Y, Nie Y, Penny M. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. J Med Virol. 2020 Mar 16;

123. Benitez J, Courtemanche C, Yelowitz A. Racial and ethnic disparities in COVID-19: evidence from six large cities. J Econ Race Policy. 2020 Dec 1;3(4):243–61.

124. Madhavan S, Bastarache L, Brown JS, Butte AJ, Dorr DA, Embi PJ, et al. Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States: a perspective from 15 academic medical centers. Journal of the American Medical Informatics Association. 2021 Feb 1;28(2):393–401.

125. Dixon BE, Grannis SJ, McAndrews C, Broyles AA, Mikels-Carrasco W, Wiensch A, et al. Leveraging data visualization and a statewide health information exchange to support COVID-19 surveillance and response: Application of public health informatics. JAMIA. 2021 Jul 1;28(7):1363–73.

126. Gardner L, Ratcliff J, Dong E, Katz A. A need for open public data standards and sharing in light of COVID-19. The Lancet Infectious Diseases. 2021 Apr 1;21(4):e80.

127. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. JAMIA. 2020 Aug 17;28(3):427–43.

128. Datavant. COVID-19 Research Database [Internet]. Available from: https://covid19researchdatabase.org/

129. Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Public Use Data with Geography (dataset access date: August 1, 2021)

[Internet]. Available from: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

130. Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Restricted Data Access, Summary, and Limitations (dataset access date: August 1, 2021) [Internet]. Available from: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t

131. Lee B, Dupervil B, Deputy NP, Duck W, Soroka S, Bottichio L, et al. Protecting Privacy and Transforming COVID-19 Case Surveillance Datasets for Public Use. Public Health Rep. 2021 Jun 17;00333549211026817.

132. Maxmen A. Massive Google-funded COVID database will track variants and immunity. Nature. 2021 Feb 24;

133. Ebert J. State won't collect, release data on coronavirus cases in Tennessee schools. The Tennessean.

134. Maxmen A. Why the United States is having a coronavirus data crisis. Nature. 2020 Aug 25;585(7823):13–4.

135. N3C Data Overview [Internet]. National Center for Advancing Translational Sciences. 2020. Available from: https://ncats.nih.gov/n3c/about/data-overview

136. Hauser C. Is Your Vaccine Card Selfie a Gift for Scammers? Maybe. The New York Times. 2021 Feb 6;

137. Kempe A, Beaty BL, Steiner JF, Pearson KA, Lowery NE, Daley MF, et al. The Regional Immunization Registry as a Public Health Tool for Improving Clinical Practice and Guiding Immunization Delivery Policy. Am J Public Health. 2004 Jun;94(6):967–72.

138. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, et al. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. medRxiv. 2020 Aug 22;2020.08.19.20177493.

139. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of Biomedical Informatics. 2014 Aug 1;50:4–19.

140. Skinner CJ, Holmes DJ. Estimating the Re-identification Risk Per Record in Microdata. Journal of Official Statistics. 1998 Dec;14(4):361.

141. Skinner CJ, Elliot MJ. A Measure of Disclosure Risk for Microdata. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2002;64(4):855–67.

142. CMS Cell Size Suppression Policy | ResDAC [Internet]. Available from: https://www.resdac.org/articles/cms-cell-size-suppression-policy

143. California Department of Health Data De-identification Guidelines (DDG) [Internet]. [cited 2021 Jan 25]. Available from: https://www.dhcs.ca.gov/dataandstats/Documents/DHCS-DDG-V2.0-120116.pdf

144. Utah Department of Health Data Suppression/Data Aggregation Guidelines Summary [Internet]. [cited 2021 Jan 25]. Available from: https://ibis.health.utah.gov/ibisph-view/pdf/resource/DataSuppressionSummary.pdf

145. Sanyaolu A, Okorie C, Marinkovic A, Patidar R, Younis K, Desai P, et al. Comorbidity and its Impact on Patients with COVID-19. SN Compr Clin Med. 2020 Jun 25;1–8.

146. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010;17(3):322–7.

147. Lee B, Dupervil B, Deputy NP, Duck W, Soroka S, Bottichio L, et al. Protecting Privacy and Transforming COVID-19 Case Surveillance Datasets for Public Use. arXiv:210105093 [cs] [Internet]. 2021 Jan 13 [cited 2021 May 31]; Available from: http://arxiv.org/abs/2101.05093

148. Berenbrink P, Friedetzky T, Hu Z, Martin R. On weighted balls-into-bins games. Theoretical Computer Science. 2008 Dec;409(3):511–20.

149. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLOS Computational Biology. 2018 Feb 20;14(2):e1005910.

150. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLOS Computational Biology. 2019 Nov 22;15(11):e1007486.

151. Tennessee Department of Health. TDH Announces Testing Schedule Change [Internet]. Available from: https://www.tn.gov/health/news/2020/12/14/tdh-announces-testing-schedule-change.html

152. Virginia Department of Health. COVID-19 FAQ [Internet]. Available from: https://www.vdh.virginia.gov/covid-19-faq/, https://www.vdh.virginia.gov/covid-19-faq/

153. County of Los Angeles. COVID-19: Frequently asked questions about testing [Internet]. COUNTY OF LOS ANGELES. 2020. Available from: https://covid19.lacounty.gov/testing-faq/

154. Zhou H, Burkom H, Winston CA, Dey A, Ajani U. Practical comparison of aberration detection algorithms for biosurveillance systems. JBI. 2015 Oct 1;57:446–55.

155. Lotze T, Shmueli G, Yahav I. Simulating multivariate syndromic time series and outbreak signatures. Robert H Smith School Research Paper No RHS-06-054. 2007 May 1;

156. Sartwell PE. The distribution of incubation periods of infectious disease. Am J Epidemiol. 1995 Mar 1;141(5):386–94.

157. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics. 2005 Apr;38(2):99–113.

158. Wong WK, Moore A, Cooper G, Wagner M. What's Strange About Recent Events (WSARE): an algorithm for the early detection of disease outbreaks. JMLR. 2005;6(66):1961–98.

159. Wong WK. Data mining for early disease outbreak detection [Ph.D.]. [United States -- Pennsylvania]: Carnegie Mellon University;

160. Fanaee-T H, Gama J. EigenEvent: An algorithm for event detection from complex data streams in syndromic surveillance. Intelligent Data Analysis. 2015 Jan 1;19(3):597–616.

161. Yuan M, Boston-Fisher N, Luo Y, Verma A, Buckeridge DL. A systematic review of aberration detection algorithms used in public health surveillance. Journal of Biomedical Informatics. 2019 Jun;94:103181.

162. Good PI. Permutation tests : a practical guide to resampling methods for testing hypotheses. 2nd ed. New York: Springer; 2000. (Springer series in statistics).

163. Hope K, Durrheim DN, d'Espaignet ET, Dalton C. Syndromic surveillance: is it a useful tool for local outbreak detection? J Epidemiol Community Health. 2006 May;60(5):374–5.

164. Ellen Wright Clayton, Bradley Malin, Consuelo H. Wilkins. Do not disclose the identity of coronavirus patients and contacts to law enforcement. The Tennessean.

165. Findling MG, Blendon RJ, Benson J, Koh H. COVID-19 Has Driven Racism And Violence Against Asian Americans: Perspectives From 12 National Polls. Health Affairs Forefront.

166. Xie G. A novel Monte Carlo simulation procedure for modelling COVID-19 spread over time. Scientific Reports. 2020 Aug 4;10(1):13120.

167. Schneider KA, Ngwa GA, Schwehm M, Eichner L, Eichner M. The COVID-19 pandemic preparedness simulation tool: CovidSIM. BMC Infectious Diseases. 2020 Nov 19;20(1):859.

168. Hall V, Foulkes S, Charlett A, Atti A, Monk EJM, Simmons R, et al. Do antibody positive healthcare workers have lower SARS-CoV-2 infection rates than antibody negative healthcare workers? Large multi-centre prospective cohort study (the SIREN study), England: June to November 2020. medRxiv. 2021 Jan 15;2021.01.13.21249642.

169. Metropolis N, Ulam S. The Monte Carlo Method. Journal of the American Statistical Association. 1949 Sep 1;44(247):335–41.

170. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics. 2005 Apr;38(2):99–113.

171. Neill DB, Kumar T. Fast Multidimensional Subset Scan for Outbreak Detection and Characterization. Online J Public Health Inform. 2013 Apr 4;5(1):e91.

172. Burnell R, Schellaert W, Burden J, Ullman TD, Martinez-Plumed F, Tenenbaum JB, et al. Rethink reporting of evaluation results in AI. Science. 2023 Apr 14;380(6641):136–8.

173. The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [Internet]. The White House. 2023. Available from: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

174. Sabatello M, Martschenko DO, Cho MK, Brothers KB. Data sharing and community-engaged research. Science. 2022 Oct 14;378(6616):141–3.

175. Health NI of. Final NIH policy for data management and sharing. 2020.

176. Ayala-Rivera V, McDonagh P, Cerqueus T, Murphy L, Thorpe C. Enhancing the Utility of Anonymized Data by Improving the Quality of Generalization Hierarchies. Trans Data Priv. 2017;10(1):27–59.

177. McAuley J. How France's aversion to collecting data on race affects its coronavirus response. Washington Post. 2020 Jun 30;

178. Office for Human Research Protections US Department of Health and Human Services. The Belmont Report [Internet]. HHS.gov. 2010. Available from: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

179. Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL. An Ethics Framework for a Learning Health Care System: A Departure from Traditional Research Ethics and Clinical Ethics. Hastings Center Report. 2013;43(s1):S16–27.

180. Xia W, Basford M, Carroll R, Clayton EW, Harris P, Kantacioglu M, et al. Managing re-identification risks while providing access to the All of Us research program. Journal of the American Medical Informatics Association. 2023;30(5):907–14.

181. Dyke SOM, Linden M, Lappalainen I, De Argila JR, Carey K, Lloyd D, et al. Registered access: authorizing data access. Eur J Hum Genet. 2018 Dec;26(12):1721–31.

182. Ramachandran A, Kantarcioglu M. SmartProvenance: A Distributed, Blockchain Based DataProvenance System. In: Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy. New York, NY, USA: Association for Computing Machinery; 2018. p. 35–42. (CODASPY '18).

183. D'arcy J, Herath T. A review and analysis of deterrence theory in the IS security literature: making sense of the disparate findings. European journal of information systems. 2011;20(6):643–58.

184. Sweeney L. Computational disclosure control: A primer on data privacy protection. Massachusetts Institute of Technology; 2001.

185. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association. 1965;60(309):63–9.

186. Becker B, Kohavi R. Adult [Internet]. UC Irvine; 1996 [cited 2024 Mar 7]. Available from: https://archive.ics.uci.edu/dataset/2

187. Seastedt KP, Schwab P, O'Brien Z, Wakida E, Herrera K, Marcelo PGF, et al. Global healthcare fairness: We should be sharing more, not less, data. PLOS Digital Health. 2022 Oct 6;1(10):e0000102.

188. Emam KE, Jonker E, Arbuckle L, Malin B. A Systematic Review of Re-Identification Attacks on Health Data. PLOS ONE. 2011 Dec 2;6(12):e28071.

189. Zou J, Gichoya JW, Ho DE, Obermeyer Z. Implications of predicting race variables from medical images. Science. 2023 Jul 14;381(6654):149–50.

190. Borza VA, Clayton EW, Kantarcioglu M, Vorobeychik Y, Malin BA. A Representativeness-informed Model for Research Record Selection from Electronic Medical Record Systems. AMIA Annual Symposium Proceedings. 2022;2022:259.

191. Drechsler J. Challenges in Measuring Utility for Fully Synthetic Data. In: International Conference on Privacy in Statistical Databases. Springer; 2022. p. 220–33.