Dynamic de-identification policies for pandemic data sharing

By

J. Thomas Brown

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTERS

in

Biomedical Informatics

August 12, 2022

Nashville, Tennessee

Approved:

Bradley A. Malin, PhD

Michael E. Matheny, MD, MS, MPH

Zhijun Yin, PhD, MS

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The novel coronavirus 2019 (COVID-19) pandemic has put a spotlight on infectious disease surveillance systems[1]. 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[2–5]. It can be used to identify the pandemic's disproportional impact on certain subpopulations and the sources of such disparities[6,7]. Moreover, it can provide the public with situational awareness of outbreaks[3,8,9]. As such, an effective data-driven pandemic response depends on the accessibility of timely 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[10,11], public data sharing on a wide scale remains limited[12]. Much of the publicly available data in the United States (U.S.) has lacked important demographic information (e.g., race or ethnicity)[8]. Data that include such information are typically limited to aggregate counts at the state level[6–8]. Moreover, most of the initiatives that have formed patient-level COVID-19 data repositories – such as the National COVID Cohort Collaborative (N3C) of the U.S. National Institutes of Health[13], the Datavant COVID-19 Research Database[14], the Centers for Disease Control and Prevention's (CDC) COVID-19 Case Surveillance datasets[15–17], and the Global.health data science initiative[18] – are not readily open to the public or do not include data shared in real time[10].

One of the primary factors limiting the public availability of surveillance data is concerns about an individual's right to privacy. In the United States, much of the infectious disease data is captured by public health authorities, hospitals, and pharmacies. Such organizations may be subject to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and related laws and policies. Under HIPAA, organizations may share patient-level data provided they establish certain privacy protections. When sharing identifiable data, for instance, HIPAA requires the organization to obtain prior consent from the individuals to which the data correspond. Obtaining consent is difficult, however, and may bias representation to those more willing to disclose their identifiable health information[19]. Alternatively, HIPAA permits organizations to share a "limited data set" of protected health information without individual authorization if "certain specified direct identifiers of individuals … have been removed" and the data recipient "enters into a data use agreement promising specified safeguards for the protected health information."[20] Since a data use agreement prohibits public access and use of the data, a limited data set cannot widely support public health research. Notably, HIPAA includes a public health exemption for data dissemination; however, the exemption is restricted to preventing an imminent threat to society and does not apply to public health research or hypothesis generation[19,21]. Nevertheless, HIPAA enables an organization to publicly share patient-level data if it is de-identified, that is, when "there is no reasonable basis to believe that the information can be used to identify an individual."[22] Even when organizations are not covered by HIPAA, they may be permitted to share data in a de-identified form as well. The California Consumer Protection Act, the Virginia Consumer Data Protection Act, the Colorado Privacy Act, and the Utah Consumer Privacy Act also provide exemptions to de-identified data sharing[23–26]. Still, the process of de-identifying data is nontrivial. Numerous demonstration attacks have shown that residual information (e.g., race, age, ZIP code of residence) can combine to uniquely represent an individual in a dataset[27–29]. With the proper background knowledge, a data recipient can leverage this information to re-identify the individuals to whom the data corresponds[28,30–32]. Concerns over such intrusions to anonymity have discouraged various organizations – from schools to public health authorities to hospitals – from sharing data[33,34], raising the importance of the question: How can organizations best comply with regulatory requirements while making surveillance data publicly available?

HIPAA allows de-identification to be satisfied through two alternative implementations. The first is Safe Harbor, which requires the suppression of eighteen direct (e.g., patient name) and quasi-identifying features (e.g., geocodes with populations smaller than 20,000 residents). However, Safe Harbor requires historical data to be shared within an uncertainty period of a year – achieved by generalizing date of event to year of event and imposing a delayed publication schedule – rendering it ineffective for continuous monitoring of infectious diseases. The alternative is Expert Determination, which indicates data is de-identified when "the risk is very small that the information could be used to identify an individual who is a subject of the information."[35] Various methods for risk assessment have been developed, including those previously developed for surveillance data[36], but provide limited guidance on adapting policies to the evolving needs of a pandemic. Rather, they retrospectively assess the privacy risk, assuming data have already been collected and are ready for dissemination. Moreover, most methods further assume the number of records in the dataset remains fixed instead of growing on a daily basis at a dynamic rate[37]. These assumptions differ from the requirements of case reporting ~~while~~ in the face of a pandemic. Waiting to publish the data hinders the ability to characterize the current state and evolution of an outbreak and appropriately respond[1,38–40]. The dynamic infection rate must also be considered as it directly influences the number of records in the dataset and subsequently the re-identification risk of each record. Furthermore, the de-identification method must consider other factors affecting the privacy risk, including the disease cases' demographics[27,28] as well as the geolocations to which the pandemic spreads[41,42]. These requirements motivate the need for methods that forecast surveillance data to design a data-sharing policy that preserves patient privacy.

In addition to preserving patient privacy, the data sharing policy must also be able to support public health research. One of the major, understudied aspects of the COVID-19 pandemic is its disproportional impact on specific subpopulations. Though data have revealed that African American, Hispanic/Latino, and Native American communities have suffered higher risks of infection[43], hospitalization[44], and mortality[6] than other racial and ethnic groups, the unavailability and inaccessibility of person-level data have stifled determinations of the such disparities' sources. Most disparity studies have had to infer disparate effects by comparing aggregated COVID-19 case counts to ZIP code or state-level demographic information[6,7]. As such, researchers have been unable to properly evaluate how socioeconomic factors and the differential incidence of pre-existing conditions, among other potential sources, may underlie disparate outcomes. A data-sharing policy that shares person-level demographic information in a timely manner is needed to support public health research and enable targeted interventions.

To support data-driven responses to current and future pandemics, this thesis aims to develop a de-identification method that preserves patient privacy while supporting public health research with near real-time data sharing. This thesis has two specific aims:

Aim 1: Develop a de-identification approach that enables near real-time person-level data publication while preserving patient privacy.

Aim 2: Evaluate how well the approach preserves the evidence of underlying disparities compared to traditional data sharing methods.

I address Aim 1 in Chapter 3 and Appendix 2, in which I introduce an approach to adaptively generate policies to publicly share de-identified patient-level epidemiological data. The approach relies on forecasting the longitudinal privacy risk of sharing the surveillance dataset at varying levels of demographic granularity. Such risk estimates allow for preemptive generalization policy selection, enabling the data sharer to de-identify new disease case records and update the surveillance dataset in near real-time.

I address Aim 2 in Chapter 4, in which I evaluate how well data shared under the dynamic policy approach supports the detection of disproportionately elevated infection rates within a specific subpopulation, where detection implies the data sharing policy preserves the evidence of underlying disparate trends. I apply an outbreak detection algorithm to measure the accuracy and timeliness at which such disparities can be detected from simulated data shared under several data sharing policies: three variations of the dynamic policy approach and two policies derived from publicly available COVID-19 datasets.

It should be recognized that this work's contributions apply to any type of epidemiological disease spread. The dynamic policy approach can also be reused to address emerging data sharing needs in which the data records continually accumulate, such as for vaccine registries[45,46]. Dynamically adapting data sharing policies can support the data-driven response to a pandemic by regularly publishing data with epidemiologically critical features in a timely and privacy-preserving manner while preserving evidence of disparate trends.

1.2 Thesis Structure

The remainder of this thesis is structured as follows. In Chapter 2, I survey the related work. I then address Aim 1 in Chapter 3, introducing a privacy risk estimation framework from which dynamic de-identification policies can be designed. I evaluate the privacy protection of such dynamic policies against an adversary with strong background knowledge. An extension of the evaluation, considering a different type of adversary, is found in Appendix 2. Next, I address Aim 2 in Chapter 4, evaluating the dynamic policy approach's ability to support disparity detection. I summarize the contributions of this work in Chapter 5 and define several avenues of future research. Finally, I specify my role in developing two manuscripts comprising this thesis in Appendix 1.

Chapter 2

Related Work

2.1 Privacy legislation in the United States

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'"[20]. The Rule also specifies "standards for individuals' privacy rights to understand and control how their health information is used."[20] 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[21].

HIPAA provides several methods 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. Individual authorization is not required for limited data sets, which, in addition to requiring the data recipient to sign a data use agreement, requires the removal of the attributes indicated in Table 1.

HIPAA permits the sharing of individual health information without individual authorization or a data use agreement when the data is de-identified. The de-identification standard may be achieved by one of two 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 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."[47] Expert Determination affords a more flexible approach to de-identification, where instead of a list of fields to be removed, a covered entity 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"[35]

HIPAA's federal regulation preempts state privacy laws if the state laws require more stringent privacy protection. Otherwise, HIPAA's standards apply[48]. In the United States, however, few states have passed comprehensive privacy laws. As of April 2022, only California, Virginia, Utah, and Colorado have signed such legislation[23–26]. These laws grant consumers more control of their personal information collected by businesses, such as the right to access and delete such data maintained by certain businesses[49]. Notably, all state laws provide exemptions for data covered by HIPAA and permit the dissemination of de-identified data.

Table 2.1. Suppressed attributes for Limited data set and Safe Harbor standards[47,50]

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:		X

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.		

2.2 De-identification models

The Safe Harbor method is designed to be a one-size-fits-all de-identification solution for every dataset; however, it may produce suboptimal results. Besides its inability to support public health research during a pandemic, studies have shown Safe Harbor's vulnerability to re-identification attacks[30,51]. Moreover, alternative data transformations may achieve the same level of anonymity as Safe Harbor (or better) while preserving more data utility[52–54]. As such, privacy researchers have developed more sophisticated anonymization models and methods to meet HIPAA's Expert Determination implementation and support biomedical research.

De-identification through Expert Determination requires that a risk assessment applying "generally accepted statistical and scientific principles" reveals "the risk is very small" an individual can be re-identified[35]. Privacy risk assessments measure the likelihood an adversary can successfully re-identify data subjects, with respect to an adversary's assumed background knowledge and how that knowledge can exploit the distinguishability of individual records. Researchers use these assessments to develop and tune anonymization models to share data with a minimal risk. 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[30,55]. In fact, Xia et al. showed how such worst-case assumptions effectively overestimate the privacy risk[56]. Thus, de-identification models need not provide perfect protection to reasonably mitigate re-identification.

One of the more well-studied and applied anonymization models is *k*-anonymity[57]. The *k*-anonymity model is designed to mitigate re-identification of individual records by ensuring that each record is indistinguishable, in terms of its 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. It has been shown that the combination of only a few quasi-identifying features can uniquely represent the majority of large population datasets[27–29]. As such, *k*-anonymity is often achieved by generalizing quasi-identifiers to coarser representations and/or suppressing quasi-identifiers corresponding to small equivalence classes[57,58].

A *k*-anonymous dataset guarantees the probability an attacker can re-identify any individual in the dataset is less than or equal to $1/k$, under a variety of scenarios. For instance, the worst-case scenario against a strong attacker assumes 1) the attacker knows a target individual's record is in the dataset, 2) the attacker knows all 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 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[57,59]. Here, each record's probability of being re-identified 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 for larger in size in the population, the probability an individual record is correctly re-identified is bounded to 1/*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[60]. 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. 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 inappropriate disclosures, models such as *l*-diversity are applied in conjunction with *k*-anonymity[60]. Even though *k*-anonymity, by itself, may be vulnerable to inappropriate disclosures, the model can protect against patient re-identification as outlined by the HIPAA Privacy Rule. It has also been broadly applied in practice, to the extent that federal and state legislation has established standard values of *k*[61–63].

An alternative to *k*-anonymity is the differential privacy model. Instead of generalizing quasi-identifiers to make individual records less distinguishable, the model protects patient privacy by injecting noise into the data. Initially designed for sharing statistical aggregates, differential privacy provides formal privacy guarantees to every individual in a dataset[64]. 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. Even though differential privacy addresses the weakness of *k*-anonymity in regard to sensitive disclosures, it may not meet the de-identification standard of the HIPAA Privacy Rule in every situation[65]. Injecting noise may not be appropriate for every data sharing scenario either[66]. Moreover, surveillance datasets, such as the CDC's COVID-19 datasets, have relied on generalization and suppression instead of differential privacy techniques[17].

2.3 Privacy vs. Utility

There is an inherent tradeoff between patient privacy and data utility. Increasing patient indistinguishability requires distorting the raw data, but distortion degrades the retained information. As such, there has been a lot of research in developing algorithms to minimize the distortion necessary to achieve *k*-anonymity. The problem of finding the minimal generalization is NP-hard[67]. Therefore, optimization solutions are approximations of the global optimum. Some of the most influential algorithms include Sweeney's original Datafly algorithm[68], Sweeney's theoretical MinGen algorithm[58], Bayardo and Aggarwal's heuristic-based search algorithm[69], and the LeFevre's Mondrian algorithm[70].

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 then 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[58], or the divergence between the original data and the generalized data[54]. Third, the algorithms 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. 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. Notably, several generalization methods have been developed to de-identify datasets that sequentially add data features[71] or continuously add new records[72]. Some of these methods achieve both *k*-anonymity and *l*-diversity in dynamic environments[73]. However, such methods still rely on retrospective, deterministic risk assessments to develop a data-sharing policy. They do not consider the inherent uncertainty of an evolving pandemic, nor can they design policies in the absence of actual data.

k-anonymity algorithms search for an optimal generalization by minimizing the distortion to the data. However, the utility of that data needs to be evaluated in the context of downstream applications. When designing data sharing policies for general use cases, divergence measures and information loss metrics, such as the Kullback-Leibler divergence, are often used to measure the original information retained in the transformed dataset[53,70,73]. When downstream applications are known, the data utility evaluation can be more specific. For example, Jeffery et al. measured how the statistical power to detect outbreaks in spatial surveillance data varied with the level of geographic aggregation applied to disease cases' geolocation[74]. The evaluation applied a global scan statistic to various types of synthetic data. The results showed that strong outbreak signals were detected with the greatest power from the most specific data. On the other hand, the weaker outbreak signals were detected with the greatest power when the level of aggregation was similar in size to the outbreak. Therefore, depending on the application, data utility can be more nuanced than the assumption that the most specific data supports the best performance. When more specific utility functions are available, targeted evaluations provide important insights into the utility achieved by de-identification methods.

2.4 COVID-19 Disparities

The novel coronavirus 2019 (COVID-19) pandemic has disproportionately affected the population. McLaren found racial and ethnic minorities to have disproportionately high COVID-19 mortality rates in Spring 2020[6]. Rossen et al. found 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. They 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[75]. Levin et al. 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[76]. Other studies found disparities in infection[43] and hospitalization rates[44,77] 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[78,79]. 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%[80].

Yet, the differing impact of COVID-19 remains poorly understood due to data limitations. 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[6]. The dearth of publicly available COVID-19 data with racial and ethnic information is widespread. Gross et al. 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[81]. 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.

2.5 Outbreak detection algorithms

At the turn of the century, government agencies and researchers became increasingly interested in developing methods to detect bioterrorist attacks. It motivated the Defense Advanced Research Project Agency (DARPA) to sponsor the Bio-event Advanced Leading Indicator Recognition Technology (BioALIRT) project, which began in October 2001. The project funded the development of novel biosurveillance methods to detect outbreaks in various types of data. Where traditional algorithms solely relied on univariate count data, these new algorithms improved detection simultaneously incorporating data from several sources, spatiotemporal information, and/or covariate information[82]. Unique among these algorithms is the What's Strange About Recent Events (WSARE) algorithm[83]. 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[84]. The result is an algorithm that both detects subpopulation outbreaks and explains the features (e.g., race, ZIP code, etc.) describing the outbreak group. These features 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[82,85,86]. 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.

Chapter 3

Dynamically Adjusting Case Reporting Policy to Maximize Privacy and Public Health Utility in the Face of a Pandemic

3.1 Introduction

The novel coronavirus 2019 (COVID-19) pandemic has put a spotlight on infectious disease surveillance systems[1] and the importance of making such information widely accessible[38]. Sharing surveillance data in a timely manner can support a wide variety of public health research endeavors (e.g., from modeling disease transmissibility to simulating interventions[2–5]) and provide the public with situational awareness of outbreaks[3,8,9]. In recognition of such benefits, over the past year and a half, various organizations have worked to broaden access to large epidemiological datasets. Recent instantiations of COVID-19 initiatives include the National COVID Cohort Collaborative (N3C) of the U.S. National Institutes of Health[13], the Datavant COVID-19 Research Database[14], the Centers for Disease Control and Prevention's (CDC) COVID-19 Case Surveillance datasets[15–17], and the Global.health data science initiative[18], among others.

While advances in surveillance have spurred rapid growth in the volume and diversity of epidemiological resources, public data sharing on a wide scale remains limited[87]. This is due to numerous social and political factors, but it is evident that privacy is a core driving factor. In the United States, for instance, infectious disease data is captured by a variety of organizations, such as public health authorities, hospitals, and pharmacies. In regard to public data dissemination, such organizations may be subject to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and related laws and policies. Under HIPAA, an organization is permitted to publicly share patient-level data only when it is de-identified, that is, when "there is no reasonable basis to believe that the information can be used to identify an individual."[22] Even when organizations are not covered by HIPAA, they may be permitted to share data in a de-identified form as well. For example, the California Consumer Protection Act, the Virginia Consumer Data Protection Act (VCDPA), and the Colorado Privacy Act provide exemptions to de-identified data sharing[23–25]. However, transforming data into a de-identified form is a non-trivial endeavor. Numerous demonstration attacks have shown that, with the right background knowledge, a data recipient can leverage residual information in the records to re-identify the individuals to whom the data corresponds[27–32]. Concerns over such intrusions to anonymity have discouraged organizations from sharing data[33,34], which raises the importance of the question: How can organizations best comply with regulatory requirements while making surveillance data publicly available?

Under HIPAA, de-identification can be satisfied through two alternative implementations. The first is Safe Harbor, which requires the suppression of eighteen direct (e.g., patient name) and quasi-identifying features (e.g., geocodes with populations smaller than 20,000 residents). However, Safe Harbor requires hiding epidemiologically critical factors, such as reducing the granularity of dates of events to their year, which renders such a policy useless for characterizing infectious disease transmission. The alternative is Expert Determination, which indicates data is de-identified when "the risk is very small that the information could be used to identify an individual who is a subject of the information."[35] Various methods for risk assessment have been developed, including those previously developed for surveillance data[36], but provide limited guidance on adapting policies to the needs of the moment. Rather, they are retrospective in nature in that they assume data have already been collected and are ready for dissemination. Moreover, most methods further assume the number of records in the dataset remains fixed[37]. These assumptions differ from the requirements of case reporting while in the face of a pandemic. Waiting to publish the data will hinder the ability to characterize the current state and evolution of an outbreak[1,38–40]. The infection rate must also be considered in the de-identification approach as it directly and dynamically influences the number of records in the dataset. Furthermore, several factors affect the privacy risk, including the demographics of the people infected[27,28] and the geolocations to which the pandemic spreads[41,42]. These requirements motivate the need for methods that forecast surveillance data.

In this paper, we introduce an approach to adaptively generate policies to publicly share de-identified patient-level epidemiological data. The framework simulates disease cases to estimate the longitudinal privacy risk of sharing infected individuals' quasi-identifier information at different levels of granularity in the absence of actual patient data. Periodically adjusting the policy allows the data sharer to adapt data granularity according to the influx of new patient records, while simultaneously allowing periods of consistent quasi-identifier representation. We specifically apply the framework to illustrate how policies could be developed to share COVID-19 patient health information and compare such policies to a more traditional de-identification approach relying on retrospective risk assessment. Furthermore, to be consistent with the CDC's current practice of using generalization and suppression for privacy[17], we use the framework to explore a wide range of data generalization policies.

It should be recognized the framework applies to any type of epidemiological disease spread, adjusts for the demographic diversity of individual US counties, and relies on public data sources. The framework can also be reused to address emerging data sharing needs, such as for vaccine registries[45,46]. Dynamically adapting data sharing policies holds the potential to consistently share more data with the public in a timely and privacy-preserving manner, fueling our data-driven response to infectious disease[9].

3.2 Methods

Due to the challenge of predicting exactly who will be infected, prospectively fixing a data sharing policy requires probabilistic risk assessment. Our 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 research, we 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[15,16]. In this section, we summarize the framework's features and its application to contextualize the results. Specific technical details are provided in Appendix 2.

3.2.1 Privacy risk estimation framework

Figure 3.1 summarizes the framework. In the first step, we select a data generalization policy, which defines the generalization of each quasi-identifying feature considered. In this paper, we consider basic demographic features and the date of diagnosis as quasi-identifiers, as they are typical features organizations have been requested to share (Table 3.1). The second step generates the county-level population across the quasi-identifying features per the selected policy. We 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[88], where each group is defined by a unique combination of quasi-identifier values, excluding date of diagnosis.

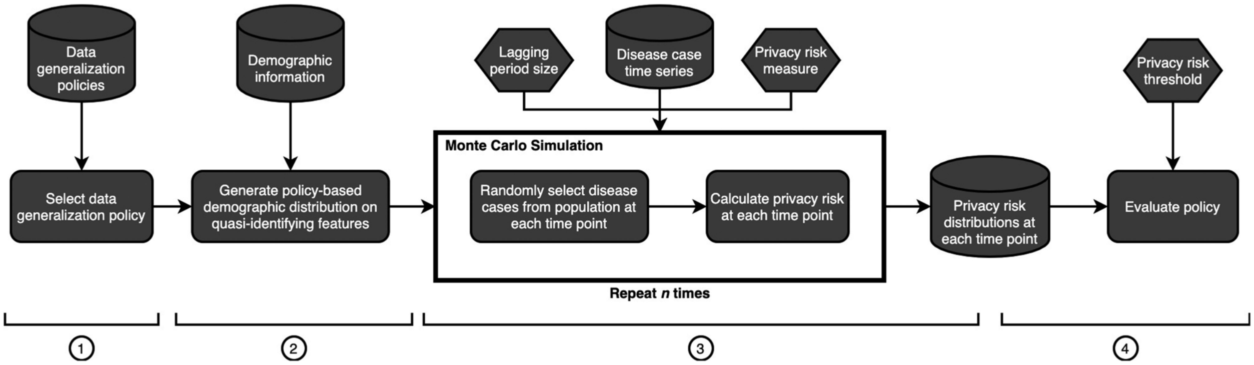

Data generalization policies
Demographic information
Lagging period size
Disease case time series
Privacy risk measure
Privacy risk threshold
Monte Carlo Simulation
Select data generalization policy
Generate policy-based demographic distribution on quasi-identifying features
Randomly select disease cases from population at each time point
Calculate privacy risk at each time point
Privacy risk distributions at each time point
Evaluate policy
Repeat n times
1
2
3
4

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, employs Monte Carlo random sampling, and is presented in greater detail in the Methods section. To obtain the privacy risk distributions, the simulation is repeated *n* times. The circled numbers denote the framework steps.

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, we input time series derived from the Johns Hopkins COVID-19 tracking data[89]. The simulation algorithm (details of which are in Appendix 2) initially assumes that the no one in the county is infected. Then, for each time point, we randomly sample the number of disease cases (without replacement) from the uninfected population to form the newly reported patient dataset. The framework assumes individuals are not re-infected (for simplicity, considering a potentially negligible COVID-19 reinfection rate[90]) and assumes equal weighting across all individuals when sampling (to model the general uncertainty of disease spread, particularly in pandemics[91]).

The algorithm computes the re-identification risk on the patient set at each time point, according to a specified risk measure. There are various methods for measuring privacy risk[37]. In this work, we measure risk 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 set of quasi-identifier values[92,93]. We refer to this measure as the *PK risk* and evaluate it given a set of *k* values (as defined below) consistent with the standard thresholds used by public health authorities[61,63,94–96]. The PK risk assumes a data recipient knows 1) an individual is a member of the dataset, 2) the individual's name and quasi-identifying information, and 3) the individual's relative date of diagnosis for the disease of interest. 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[97,98]). The more unique the record's representation, the more likely the data recipient can re-identify the individual[27,28]. In this research, we focus on this risk measure to follow the CDC's application of *k*-anonymization[99]. The PK risk effectively measures the proportion of records that fail to achieve *k*-anonymity.

In practice, obtaining such patient information is difficult[30,55]. Thus, evaluating the PK risk provides an upper bound of re-identification risk for the dataset. To demonstrate the approach's flexibility and to offer a different perspective on privacy risk, we further analyze the amortized re-identification risk[59] in Appendix 2. The amortized re-identification risk relaxes assumptions (1) and (3) 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.

Table 3.1. The quasi-identifiers considered in this study. The middle column describes the generalization strategy for each quasi-identifier. The third column provides an example generalization for each quasi-identifier. In the case of sex and ethnicity, the information is either included or null. AIAN = American Indian/ Alaskan Native, PI = Pacific Islander. *These values cannot be generalized since we simulate on a county level. †This definition of a week is consistent with the one used by the CDC's COVID-19 case forecasts[100].

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

We highlight that, when applying the PK risk measure, we assume the attacker knows the diagnosis occurred within a lagging period of time (e.g., within one, three, or five days prior to the documented date). We allow this flexible assumption as it is unlikely a data recipient knows the targeted individual's exact diagnosis date[56], 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.

The 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).

3.2.2 Dynamic policy search

To dynamically adapt policies according to an expected infection rate, we identify policies that are likely to satisfy a specific PK risk threshold at varying volumes of new case records. For this policy search, we choose a *k* of 11, which is as a typical group size incorporated into guidance issued at the state[61,63,95,96] and federal[94] level. It is also the group size applied to CDC's COVID-19 Public Use Data with Geography[15]. We henceforth refer to the PK risk when *k* equal to 11 as the PK11 risk. In this paper, we 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%. Similar investigations for *k* of 5 and 20 (other common group size thresholds) are provided in Appendix 2.

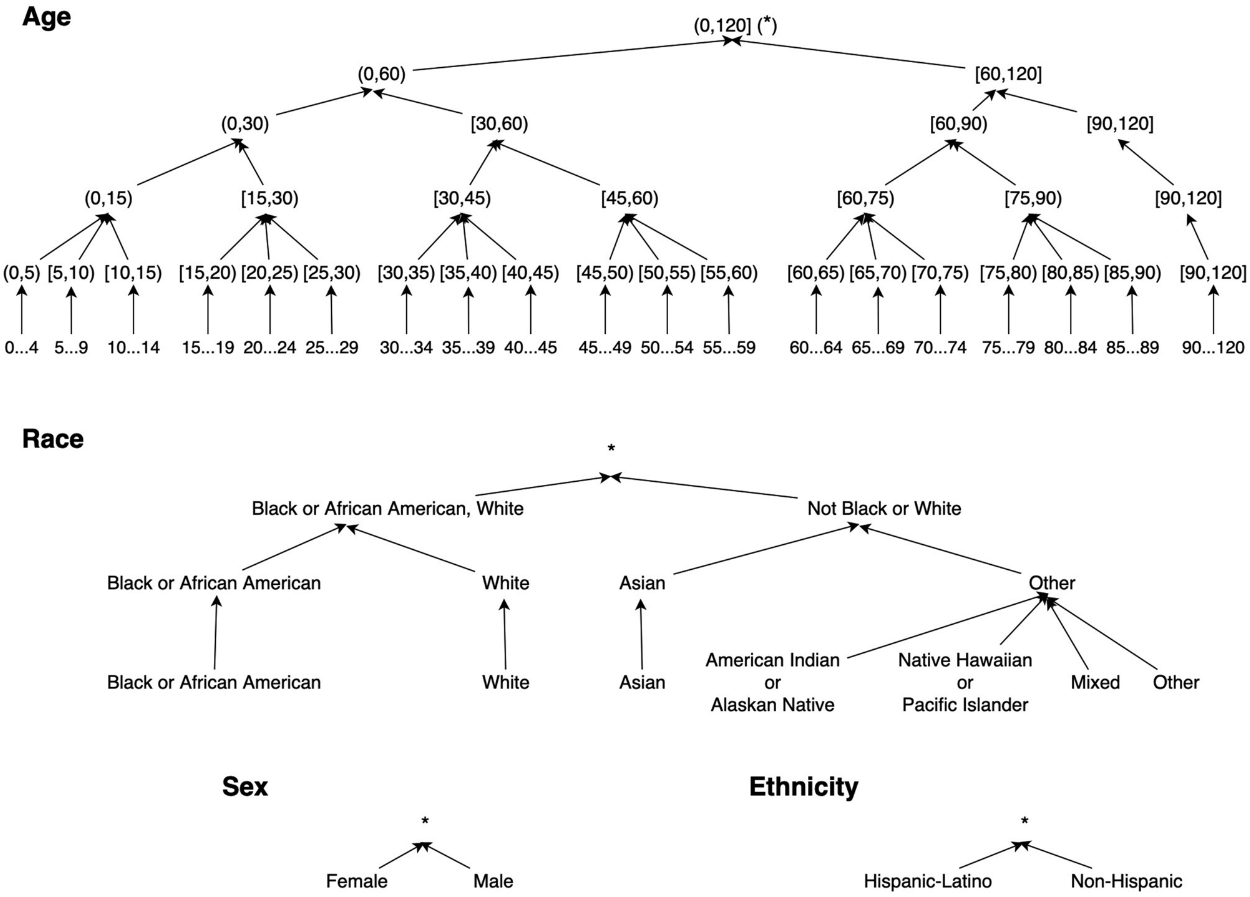


Figure 3.2. The generalization hierarchies for age, race, sex, and ethnicity used in this paper, adapted from those of Wan et al[53]. Each horizontal level is a potential generalization state for the data generalization policy. For example, the policy could specify generalizing age to 5-year age intervals to 15-year age intervals, or broader ranges. We 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. The generalization options follow a hierarchical structure (see Figure 3.2), where moving up the hierarchy generalizes the information to increase privacy at the cost of utility[58]. For each policy, county, and case number combination, the framework generates 1,000 PK11 estimates. A policy meets the threshold when the upper bound of the estimates' 95% quantile range is less than or equal to 0.01. We 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.

3.2.3 Dynamic policy evaluation

We 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 experiment, we measure the proportion of data releases in which the PK11 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 each county, the 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. We use the CDC COVID-19 ensemble model's county-specific, one-week forecasts for its superior accuracy over other models[100–102]. For the evaluation, we collected all model predictions from August 2020 through April 2021. We obtain daily increase predictions by uniformly distributing the weekly increase point estimate. In selecting policies for the daily release schedule, we 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, we use the forecasted one-week increase.

After selecting the sequence of policies for each county, we estimate the privacy risk of sharing the actual reported number of records via the privacy risk estimation framework. We define the actual number of disease cases per day or week by the Johns Hopkins COVID-19 tracking data. The PK11 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.

We 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[15]. The policy, hereafter referred to as the *k*-anonymous policy, shares age intervals in the form (0-17, 18-49, 50-64, and 65+); nearly fully specified race; fully specified ethnicity, sex, and state and county of residence; and date or week of diagnosis. We 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[15,17,57]. 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. We 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.

3.2.4 Case studies

To provide a specific illustration of the dynamic policy approach to daily releasing updated, record-level disease surveillance data, we consider 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.

In each case study, we 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, and accounting for the general turnaround time of COVID-19 diagnostic tests results[103–105], we 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. We 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, we repeat the process by replacing the forecasted case counts with the actual case numbers in policy selection.

3.2.5 Code

All experiments are performed using Python (version 3.8). The code, and walkthroughs corresponding to each experiment, can be found at:
https://github.com/vanderbiltheads/PandemicDataPrivacy

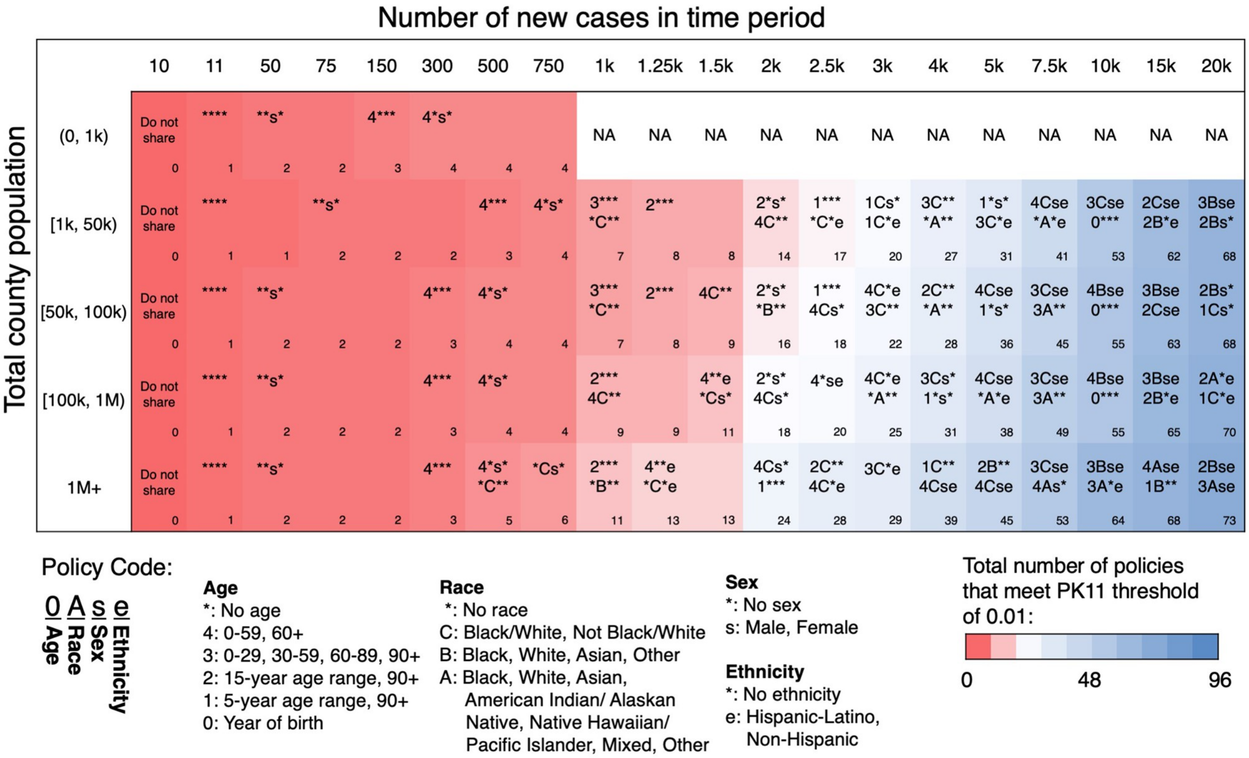
3.3 Results

3.3.1 Dynamic policy search

We summarize the policy search results in Figure 3.3. To aid in readability, we 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. We 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 3.3 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.



Total county population	10	11	50	75	150	300	500	750	1k	1.25k	1.5k	2k	2.5k	3k	4k	5k	7.5k	10k	15k	20k
(0, 1k)	Do not share 0	**** 1	**s* 2	2	4*** 3	4*s* 4	4	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[1k, 50k)	Do not share 0	**** 1	1	**s* 2	2	2	4*** 3	4*s* 4	3*** *C** 7	2*** 8	8	2*s* 4C** 14	1*** *C*e 17	1Cs* 1C*e 20	3C** *A** 27	1*s* 3C*e 31	4Cse *A*e 41	3Cse 0*** 53	2Cse 2B*e 62	3Bse 2Bs* 68
[50k, 100k)	Do not share 0	**** 1	**s* 2	2	2	4*** 3	4*s* 4	4	3*** *C** 7	2*** 8	4C** 9	2*s* *B** 16	1*** 4Cs* 18	4C*e 3C** 22	2C** *A** 28	4Cse 1*s* 36	3Cse 3A** 45	4Bse 0*** 55	3Bse 2Cse 63	2Bs* 1Cs* 68
[100k, 1M)	Do not share 0	**** 1	**s* 2	2	2	4*** 3	4*s* 4	4	2*** 4C** 9	9	4**e *Cs* 11	2*s* 4Cs* 18	4*se 20	4C*e *A** 25	3Cs* 1*s* 31	4Cse *A*e 38	3Cse 3A** 49	4Bse 0*** 55	3Bse 2B*e 65	2A*e 1C*e 70
1M+	Do not share 0	**** 1	**s* 2	2	2	4*** 3	4*s* *C** 5	*Cs* 6	2*** *B** 11	4**e *C*e 13	13	4Cs* 1*** 24	2C** 4C*e 28	3C*e 29	1C** 4Cse 39	2B** 4Cse 45	3Cse 4As* 53	3Bse 3A*e 64	4Ase 1B** 68	2Bse 3Ase 73

Figure 3.3. 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.

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.

3.3.2 Dynamic policy evaluation

We summarize the evaluation results, categorizing counties in the same manner as the policy search, in Table 3.2. 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, we 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 3.2. 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.

	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)	
County Population Size	*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 Case Study: Davidson County, TN

Figure 3.4 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.

3.3.4 Case Study: Perry County, TN

Figure 3.5 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.

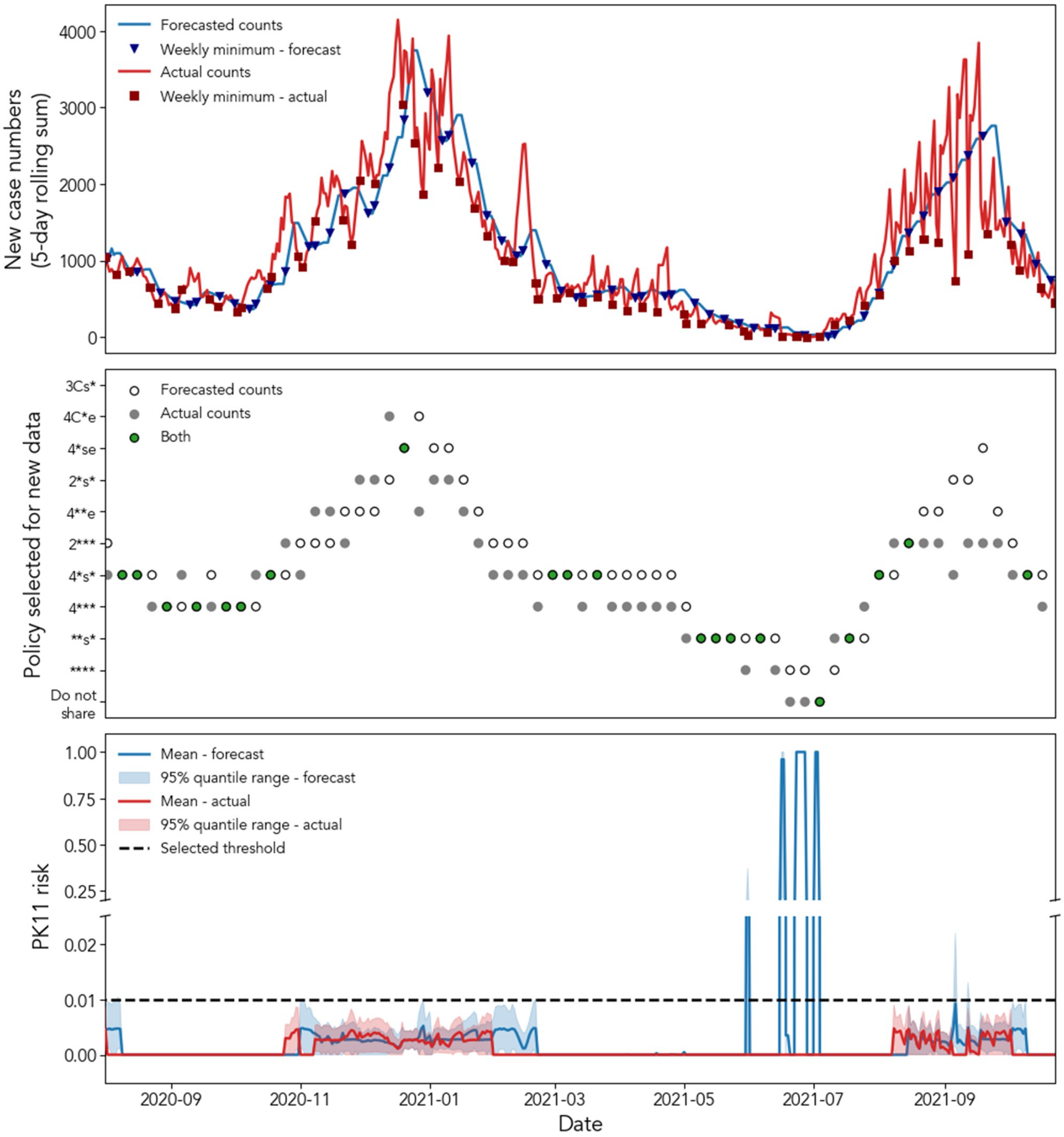


Figure 3.4. Dynamic policy selection applied to Davidson County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The 5-day rolling sum of the forecasted and actual case counts reported in Davidson County. The forecasted counts are from the CDC's COVID-19 ensemble model and the 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. 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.

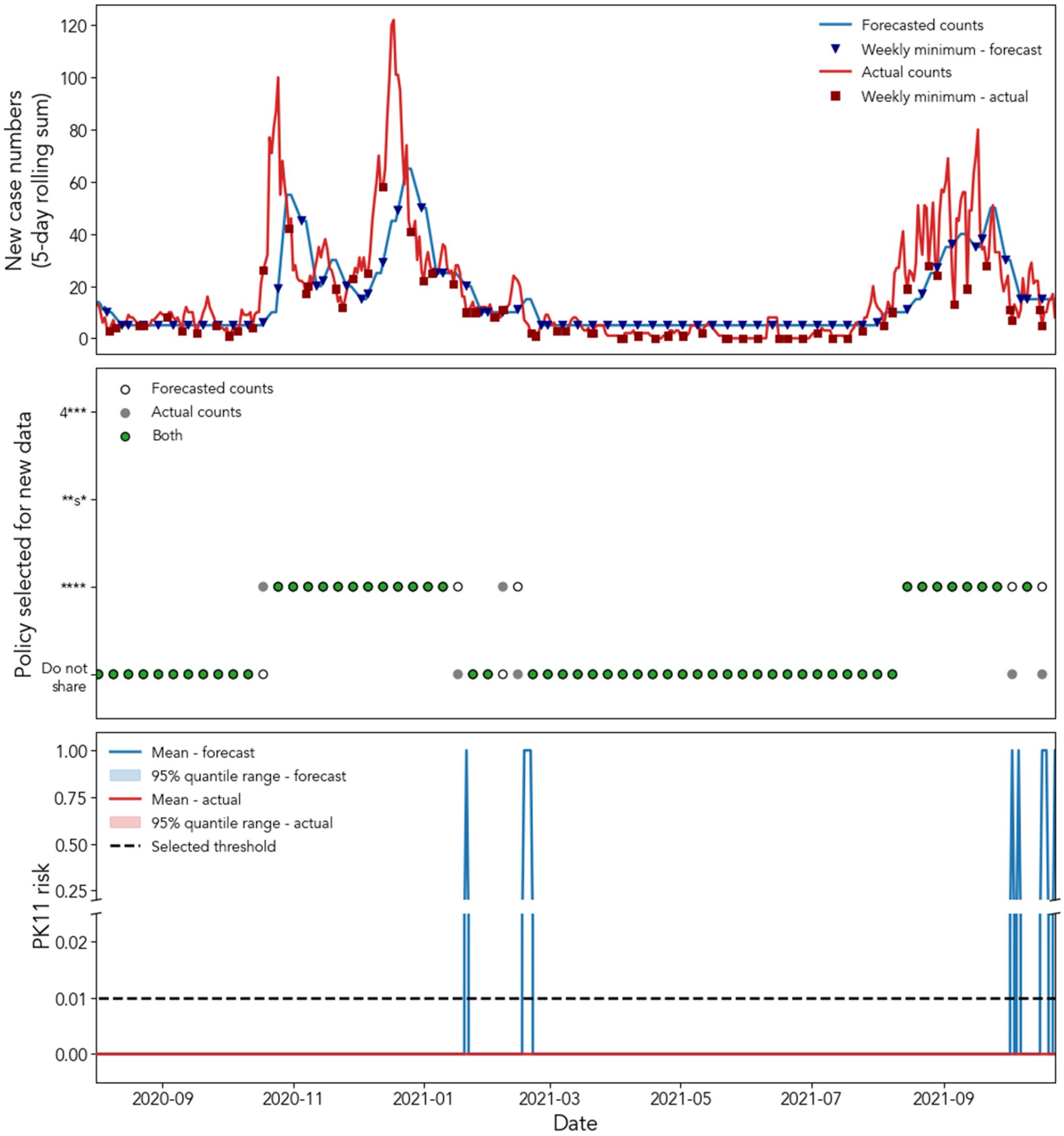


Figure 3.5. Dynamic policy selection applied to Perry County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The 5-day rolling sum of the forecasted and actual case counts reported in Davidson County. The forecasted counts are from the CDC's COVID-19 ensemble model and the actual 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. 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.

3.4 Discussion

This paper introduces a framework to dynamically adjust data sharing policies to publicly share infectious disease surveillance data. The framework forecasts privacy risk according to the expected volume of new cases, enabling data sharers to prospectively adapt policies before seeing case loads. We demonstrate how dynamically changing the policy per the framework's recommendations maintains 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. The dynamic policy also enhances surveillance utility by fluctuating data generalization with the infection rate, allowing the data sharer to prioritize sharing certain patient information; bypassing the delay of accumulating patient records before performing a risk assessment; and sharing dates of events. These last two features are crucial for characterizing disease transmission[38,39]. Forecasting also enables greater consistency in quasi-identifier representation, as the policy can be maintained throughout the forecasted interval of time. Moreover, predicting which policies provide sufficient privacy protection could potentially automate patient de-identification.

We demonstrate two approaches to dynamic policy adaptation. In the PK risk-based approach, we fix county of residence and date of diagnosis granularity while varying the demographic granularity. We make 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 marketer risk-based approach (see Appendix 2), we show that when the potential attacker has less background knowledge, the dynamic policy can preserve date of diagnosis granularity while monotonically increasing the demographic granularity of the entire dataset over time.

We do not advocate for which measure provides the best privacy protection, nor do we specify which applications each approach best supports; rather, this investigation shows how the privacy risk estimation framework's flexibility can inform different approaches to dynamic policy adjustment.

Despite the merits of this work, we 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 PK risk-based scenario. However, the framework's policy search results remain relatively robust. Policies chosen from forecasted counts are typically similar or close to those chosen from actual case counts. And when overestimating the number of new cases, the privacy risk does not always dramatically exceed the threshold. That is, except when the overestimates indicate there will be marginally sufficient records to share under the **** policy, which sometimes to dramatic spikes in the PK11. This finding is informative for dynamic policy implementation, where the data sharer can protect against such unintended privacy risks by increasing the case count threshold at which the **** policy should be applied. Furthermore, we selected policies in our implementation 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 could be improved through more accurate forecasts and a model that accounts for potential case load overestimation.

Second, our approach does 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 *k*-anonymous policy in our evaluation does not fully incorporate the privacy protection mechanisms of retrospective de-identification. Without suppression, a static generalization policy is unlikely to meet the PK11 threshold. It is even less likely to do so when we increase date granularity from the original design. Though our evaluation makes assumptions that hinder the *k*-anonymous policy's privacy protections, where such assumptions are made to standardize the policies in our comparison, our evaluation still illustrates the weaknesses of a static policy applied to a dynamic dataset. A constant generalization may provide sufficient privacy during some periods (often those of higher infection rates), but not others. Thus, the generalization must be flexible or more information must be suppressed with the influx of fewer disease cases. We show how not adapting the generalization can produce periods of increased privacy risk, but future work should compare the amount of information preserved by a static generalization policy that meets the privacy threshold via suppression to that of a dynamically adjusting generalization policy without suppression.

Fourth, as we aim to generally support public data sharing, we focus on privacy risk without measuring the utility of a data generalization policy. Though we provide the data sharer with policy options, from which they can choose how to prioritize sharing quasi-identifier information, and our approach generally supports surveillance utility in terms of providing granular date information and timely updates, we do not address the more complex problem of policy planning. For instance, maximizing the granularity of one quasi-identifier early in the time series could hinder policy flexibility in the future. In the scenario where another quasi-identifier becomes important to public health research later, the data sharer may want to change the generalization of previously released data to complement the new priority. However, if the earlier policy has already consumed the available privacy risk, the policy may not be altered without potentially exposing patients' identities. Previously released data may be shared again with more detail, but not less. Future work should quantitatively measure data utility to inform data sharers in policy planning.

Fifth, the privacy risk estimation framework depends on random sampling methods that may not realistically simulate the pandemic spread of disease. We assign an equal likelihood of infection to all uninfected county residents at any given time in the simulations, and do not allow reinfections. In reality, the actual likelihood varies according to contact patterns of infectious individuals (i.e., through households or at work)[106,107], and reinfections are possible, though not likely in the case of COVID-19[90]. Still, we 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[108]. Further framework refinement should address the possibility of reinfection for diseases for which reinfection is more likely.

Sixth, 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.

Finally, while this paper focuses on de-identification through generalization, an alternative approach would rely on the principle of differential privacy. Differential privacy offers formal privacy guarantees[64]; but as has been recently noted[66], realizing this definition in practice requires injecting noise into the data, a strategy that is not appropriate for every data sharing scenario. Moreover, the CDC's COVID-19 datasets apply generalization and suppression[17]. Therefore, to be consistent with the CDC's current practice, we focused our framework's application on data generalization policies

3.5 Conclusion

Disease surveillance data is variable, between geographic areas and over time. As such, data must be consistently updated in a timely manner. To support public health research and the public's situational awareness during a pandemic, the data must also contain granular date information. The privacy risk estimation framework we propose enables a prospective approach to surveillance data de-identification. In contrast to traditional methods, prospective policy selection offers increased flexibility, with intermittent consistency, to support near-real time data dissemination. Moreover, we show that forecast-driven de-identification offers better privacy protection than the static data sharing policy application.

3.6 Availability of data and material

All data used herein are publicly available. The datasets include: the United States Census PCT12 Tables[88], the Johns Hopkins COVID-19 tracking data[89], and the CDC COVID-19 Ensemble Forecasts[100,109].

Chapter 4

Supporting COVID-19 Disparity Investigations with Dynamically Adjusting Case Reporting Policies

4.1 Introduction

The novel coronavirus disease 2019 (COVID-19) pandemic has disproportionately affected segments of society in the United States (U.S.). African American, Hispanic/Latino, and Native American communities have suffered higher risks of infection[43], hospitalization[44], and mortality[6] than other racial and ethnic groups, and the infection fatality rate has exhibited a direct correlation with age[76]. In recognition of these differential outcomes, researchers and policy makers have sought to quickly identify disparities to inform timely interventions amid an evolving pandemic. For instance, after discovering imbalanced infection and mortality rates between racial and ethnic groups, the state of Michigan increased testing resources and access to primary care physicians for minority subpopupulations[78]. Due in part to these and other policy decisions, from April to November 2020, the percentage of COVID-19 cases in Michigan corresponding to African Americans dropped from 41% to 8%[80].

Despite such efforts, the differential impact of COVID-19 remains poorly understood. Attempts to determine potential sources of disparities, including socioeconomic factors and the differential incidence of pre-existing conditions, have been hindered by limitations in access to data[6]. Notwithstanding informaticians' significant efforts to develop infrastructure and tools to monitor the spread of COVID-19[10,11], fewer resources have been allocated to publicly disseminate robust person-level information. Much of the publicly available data in the U.S. have not included racial or ethnic information, and data that do include this information are typically limited to aggregated counts at the state level[6–8]. Though several initiatives have formed patient-level COVID-19 data repositories, such as the National COVID Cohort Collaborative (N3C) of the U.S. National Institutes of Health[13] and the COVID-19 Case Surveillance datasets from the Centers for Disease Control and Prevention (CDC)[15,16], most of the repositories are not readily open to the public or include data shared in real time[10].

Patient privacy is one of the primary factors limiting person-level COVID-19 data sharing[34]. When publicly disseminating data, many organizations capturing COVID-19 data may be subject to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule[20] and related laws. Though HIPAA, and state laws such as the California Consumer Protection Act[23], permit sharing de-identified data, the process of de-identifying pandemic data is nontrivial. It has been shown that a data recipient can exploit prior knowledge to re-identify individual records from the shared quasi-identifying features (i.e., attributes such as age and race that can, in combination, uniquely represent individuals[28]). As such, HIPAA provides two alternative methods to achieve de-identification and minimize the re-identification risk. The first, Safe Harbor, specifies 18 direct (e.g., name and residential address) and quasi-identifying features (e.g., geocodes corresponding to fewer than 20,000 residents) that must be removed. However, the Safe Harbor method requires historical data to be shared with an uncertainty period of a year – achieved by generalizing date of event to year of event and imposing a delayed publication schedule – rendering it ineffective to detect disparate trends in a timely manner[20]. The second method, Expert Determination, allows data to be de-identified by the application of "generally accepted statistical and scientific principles"[20] so that "the risk is very small that the information could be used to identify an individual who is a subject of the information."[110]

Following Expert Determination, a method was recently proposed to publicly share de-identified patient-level epidemiological data in near-real time[111]. Relying on a framework to forecast the privacy risk of sharing the data at different levels of granularity, the approach adaptively generates data generalization strategies according to the influx of new records. In comparison to traditional de-identification methods, the dynamic policy approach maintains the re-identification risk below a threshold based on state and federal standards more frequently, under several adversarial scenarios, when sharing granular date information with consistent updates (e.g., daily or weekly). Though the dynamic policy approach was designed to support pandemic data sharing, its ability to support disparity investigations has yet to be systematically evaluated.

In this paper, we determine how well data shared under the dynamic policy approach enables the detection of disproportionately elevated infection rates within a specific subpopulation. Such COVID-19 disparities have fluctuated longitudinally, emerging and dissipating as subpopulation outbreaks[6,44]. As such, our evaluation applies an outbreak detection algorithm to measure the timeliness and accuracy at which disparities can be detected. We also evaluate the fairness of detection performance, in terms of enabling similar disparity detection times and accuracy between regions and subpopulations. We compare several versions of the dynamic policy to policies resembling those applied to two current, publicly available COVID-19 datasets: 1) the CDC's COVID-19 Case Surveillance Public Use Data with Geography[15] and 2) the aggregated case counts that have been used in several disparity investigations[81].

4.2 Methods

This section begins with a description of the different types of de-identification methods considered in our evaluation. Next, we describe how we simulate disparities in infectious disease surveillance data. We then provide details regarding how we detect disparities with an outbreak detection algorithm for each de-identification method. Finally, we review our experimental design and performance evaluation measures.

4.2.1 Data sharing policies and assumptions

Our analysis focuses on five different data sharing policies and the extent to which they support the timely, accurate, and fair identification of disparities within two Tennessee counties with very different demographic compositions: 1) Davidson County, a relatively large metropolitan region, and 2) Perry County, a relatively rural region. Table 1 displays the counties' population demographics according to recent estimates from U.S. Census Bureau[88].

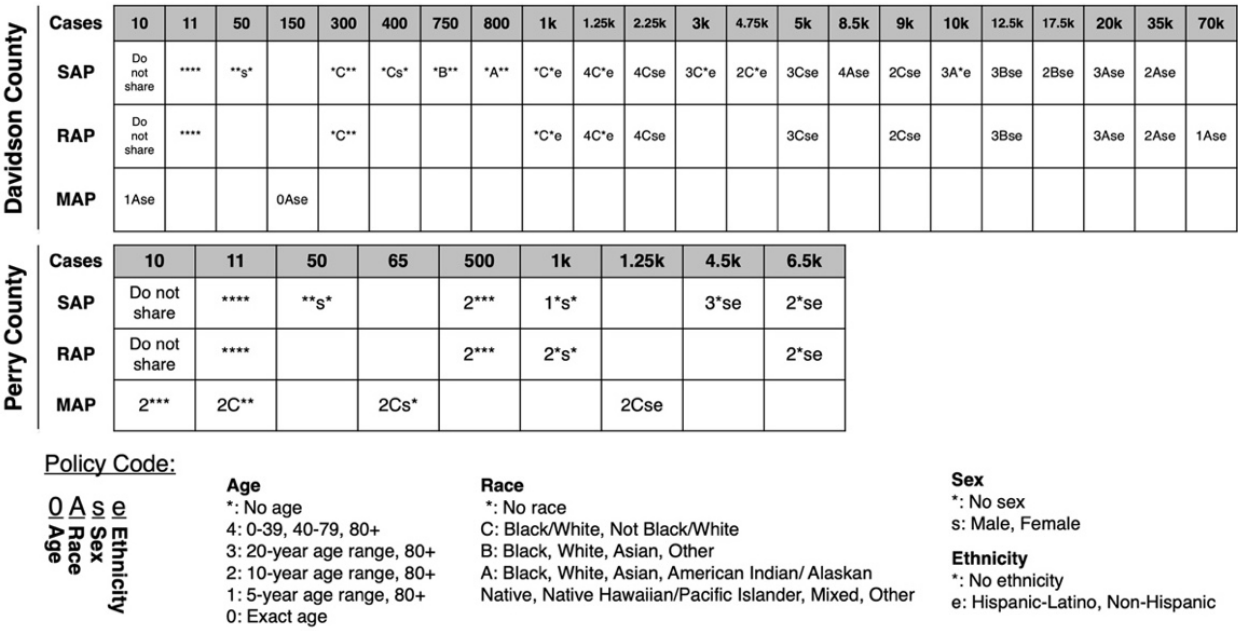
A data sharing policy describes the format of the shared dataset, including the granularity at which each quasi-identifier is transformed and the schedule by which the dataset is updated. The quasi-identifiers considered in this study are race, ethnicity, age, sex, county of residence (following the format of the CDC's surveillance datasets), and date of diagnosis. Each policy is designed to minimize the privacy risk against a potential adversary; i.e., a recipient with certain background knowledge who may attempt to re-identify individual records[28]. To mitigate re-identification risk, the quasi-identifiers can be converted into a more generalized form (e.g., converting year of birth to 5-year age intervals) to increase the number of records that correspond to each unique combination of quasi-identifier values, or equivalence class[112]. Adversarial modeling is critical for policy selection. Assuming too strong an adversary could overestimate the privacy risk and unnecessarily coarsen the data, while assuming too weak an adversary could expose patient identities.[56] Here, we consider adversaries who vary in terms of their background knowledge and their motivation to attempt re-identification. We define dynamic policies according to one of two standard privacy risk measures, each designed to measure the re-identification risk against a different type of adversary. The first measure is referred to as the PK11 risk, which is the proportion of records in the dataset that reside in an equivalence class of size less than 11[111]. The equivalence class size of 11 is typically incorporated into federal[94] and state-level[23] guidance. The PK11 risk measures the privacy risk against an adversary who knows an individual is in the dataset and a subset of the individual's quasi-identifier information. In this setting, an adversary attempts re-identification to learn additional sensitive information included in the patient dataset (e.g., comorbidities[97]) corresponding to the target individual with identity. The second measure is the marketer risk, which measures the average risk of each record in patient the dataset in the context of the underlying population[59]. A record's risk is computed as one over the size of the corresponding equivalence class in the population. The marketer risk assesses the privacy risk against an adversary who attempts to re-identify each record in the patient dataset by matching the quasi-identifiers in the shared dataset to those in a separate, identified dataset. A common example of the latter is a voter registration list[28,59].

Table 4.1. 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)

To evaluate how disparity detection performance varies when protecting against adversaries of differing strength, we develop a distinct dynamic policy for each of three different adversaries. 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**), follows the PK11 policy proposed by Brown, et al[111]. This policy assumes the adversary knows a target individual's demographic information and relative COVID-19 date of diagnosis. Under this assumption, the data sharer fixes the quasi-identifier generalization strategy at the beginning of each week, according to the privacy risk estimation framework's forecasted PK11 risk, where the strategy defines the granularity of each quasi-identifier's representation. This allows quasi-identifier granularity to fluctuate with the infection rate while maintaining weekly consistency. In this paper, we 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[111], we also assume the data sharer can estimate the number of daily cases that will accrue in the coming week within ± 5 cases.



Davidson County

Cases	10	11	50	150	300	400	750	800	1k	1.25k	2.25k	3k	4.75k	5k	8.5k	9k	10k	12.5k	17.5k	20k	35k	70k
SAP	Do not share	****	**s*		*C**	*Cs*	*B**	*A**	*C*e	4C*e	4Cse	3C*e	2C*e	3Cse	4Ase	2Cse	3A*e	3Bse	2Bse	3Ase	2Ase	
RAP	Do not share	****			*C**				*C*e	4C*e	4Cse			3Cse		2Cse		3Bse		3Ase	2Ase	1Ase
MAP	1Ase			0Ase																		

Perry County

Cases	10	11	50	65	500	1k	1.25k	4.5k	6.5k
SAP	Do not share	****	**s*		2***	1*s*		3*se	2*se
RAP	Do not share	****			2***	2*s*			2*se
MAP	2***	2C**		2Cs*			2Cse		

Policy Code:

0 A s e

Age Race Sex Ethnicity

Age
*: No age
4: 0-39, 40-79, 80+
3: 20-year age range, 80+
2: 10-year age range, 80+
1: 5-year age range, 80+
0: Exact age

Race
*: No race
C: Black/White, Not Black/White
B: Black, White, Asian, Other
A: Black, White, Asian, American Indian/ Alaskan Native, Native Hawaiian/Pacific Islander, Mixed, Other

Sex
*: No sex
s: Male, Female

Ethnicity
*: No ethnicity
e: Hispanic-Latino, Non-Hispanic

Figure 4.1. 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.

The reasonable adversary policy (**RAP**) protects against an adversary who knows a target individual's demographic information, but not their diagnosis date. This is likely a more reasonable assumption due to the difficulty of ascertaining a patient's exact date of diagnosis[56,111]. In this 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. This method constrains successive generalization strategies to represent demographic quasi-identifiers with equal or greater granularity than previous strategies determine.

The marketer adversary policy (**MAP**) protects against an adversary with an identified dataset about a population that does not include diagnosis dates. In this setting, the data sharer updates the generalization strategy applied to all records on a weekly basis, similar to RAP, but they use the privacy risk estimation framework to choose policies according to a marketer risk threshold of 0.01. We estimate the marketer risk under the assumption the adversary has an identified dataset that covers every population resident - a worst-case scenario. Figure 1 displays the dynamic policy search results guiding the three dynamic policies for both Davidson and Perry counties. For Davidson county, we rely upon generalization strategies that prioritize race and ethnicity granularity. For Perry county, we rely upon strategies that prioritize age and sex granularity.

The ***k*-anonymous** policy resembles that applied to the CDC's COVID-19 Case Surveillance Public Use Data with Geography[15]. This policy shares age group (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 residency, and month of diagnosis (as date of diagnosis is considered a quasi-identifier). Due to the generalized month of diagnosis, we assume the dataset is updated on the first day of each month. For simplicity and to match the dynamic policy implementation, the *k*-anonymous policy defined for this investigation differs from the CDC's policy only in that it does not strategically suppress quasi-identifiers to ensure each equivalence class holds at least 11 records (11-anonymity[112]). Notably, the CDC's policy suppresses around 3% of each quasi-identifier to achieve 11-anonymity[99].

The **Marginal Counts** policy resembles the non-person-level data displayed in state COVID-19 dashboards[11] that have been used in several disparity investigations[81]. Though most racial data have been shared at the state level, for consistency with the other policies, we 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. We assume the dataset shared under this policy is updated on a daily basis. Table 4.2 summarizes the five de-identification policies' details.

Table 4.2. 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

4.2.2 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[113]. For our evaluation, we 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[89], 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[6,44], the baseline distribution is generated by randomly sampling individuals from the population, without replacement. To simulate a disparity, we disproportionately sample from the affected subpopulation.

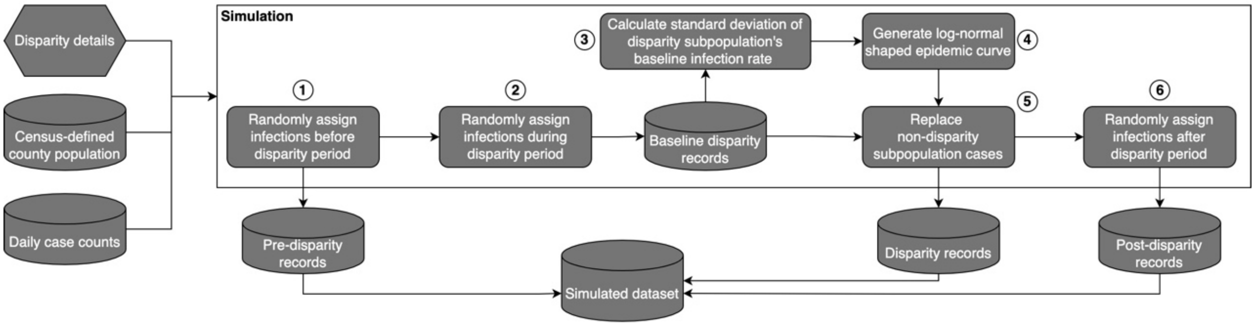

Disparity details
Census-defined county population
Daily case counts
Simulation
1
Randomly assign infections before disparity period
2
Randomly assign infections during disparity period
Baseline disparity records
3
Calculate standard deviation of disparity subpopulation's baseline infection rate
4
Generate log-normal shaped epidemic curve
5
Replace non-disparity subpopulation cases
6
Randomly assign infections after disparity period
Pre-disparity records
Disparity records
Post-disparity records
Simulated dataset

Figure 4.2. The pipeline for simulating disparity data in this study.

Figure 4.2 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 a representative county population generated from U.S. Census population count data[88]. We 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, we first calculate the standard deviation of the subpopulation's baseline infection rate during the disparity period (3). We then generate a log-normal shaped epidemic curve[114] (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. We rely upon a log-normal shaped curve, following the standard practice in the literature, to approximate real world epidemic curves[113,115]. 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. We 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, we 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[114]. 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, we 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.

4.2.3 Disparity detection

We apply the What's Strange About Recent Events (WSARE)[83] algorithm to detect disparate infection rates. We utilize WSARE because it detects and explains anomalous patterns within categorical, person-level data without requiring large amounts of historical data for hypothesis testing. Moreover, WSARE has been implemented in several real world settings, including American and Israeli outbreak detection monitoring systems[83].

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 can consist 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[116], 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[83]. In this study, we 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. We apply WSARE 2.0 to each de-identification policy. To further evaluate SAP, where the quasi-identifier generalization varies within the dataset, we additionally apply a variation of WSARE 3.0. Our variation generates a baseline by randomly sampling up to 10,000 county residents from the U.S. Census population statistics.

We 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 our comparison between policies, we convert the *k*-anonymous policy's month of diagnosis to date of diagnosis by randomly assigning a date within the month to each record. We generate assignments by randomly sampling the date with replacement, where each date within the month is equally weighted. For the Marginal Counts policy, we consider a single covariate that includes all race, ethnicity, age group, and sex values. Finally, for comparison, we apply WSARE 2.0 to the raw data.

4.2.4 Experimental design

We repeat each experiment for both Davidson and Perry county, to evaluate performance in counties of varying size and diversity. The first, which we call the **Broad** experiment, evaluates how well each of the de-identification policies enables disparity detection at different significance thresholds. We 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 our simulation method, we were 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). We 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. We then measure the proportion of the datasets in which the disparity is detected. We 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. We 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, we assign a detection time of 90 days, or twice the disparity duration. Finally, we 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, we 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.

Next, the **Fairness** experiment evaluates how the de-identification policies may bias detection between subpopulations. In this context, the smaller the difference between the proportion of disparities detected and the smaller the difference between the disparity detection times, the fairer we consider the performance. We 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. We apply WSARE to search for the best single component increase at a significance threshold of 0.05. We measure bias, or the lack of fairness, between subpopulations 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. We calculate both feature-specific standard deviations (e.g., race-specific deviations to measure racial bias) and standard deviations across all subpopulations. Additionally, we 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. We 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.

4.2.5 Code availability

All experiments are performed using Python (version 3.10). The code for our experiments can be found at https://github.com/vanderbiltheads/PandemicDataPrivacy.

4.3 Results

4.3.1 Broad Experiment

Our first experiment broadly evaluates how well the de-identification policies enable disparity detection, for both Davidson and Perry County. We first measure the proportion of the 50 experiment datasets in which the disparity is accurately detected, at each statistical significance threshold. Figures 4.3 and 4.4 present the results for Davidson and Perry County, respectively.

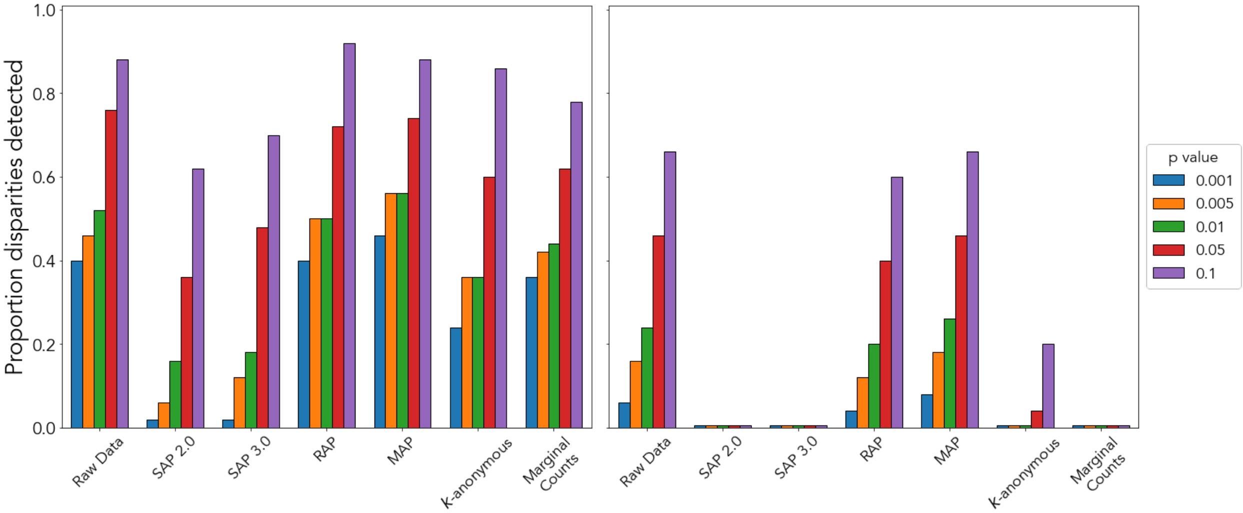


Figure 4.3. Proportion of detected disparities for Davidson County, TN, in which at least one of the simulated disparity features (left) and both features (right) are detected. The proportion is out of 50 different experiment datasets.

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.

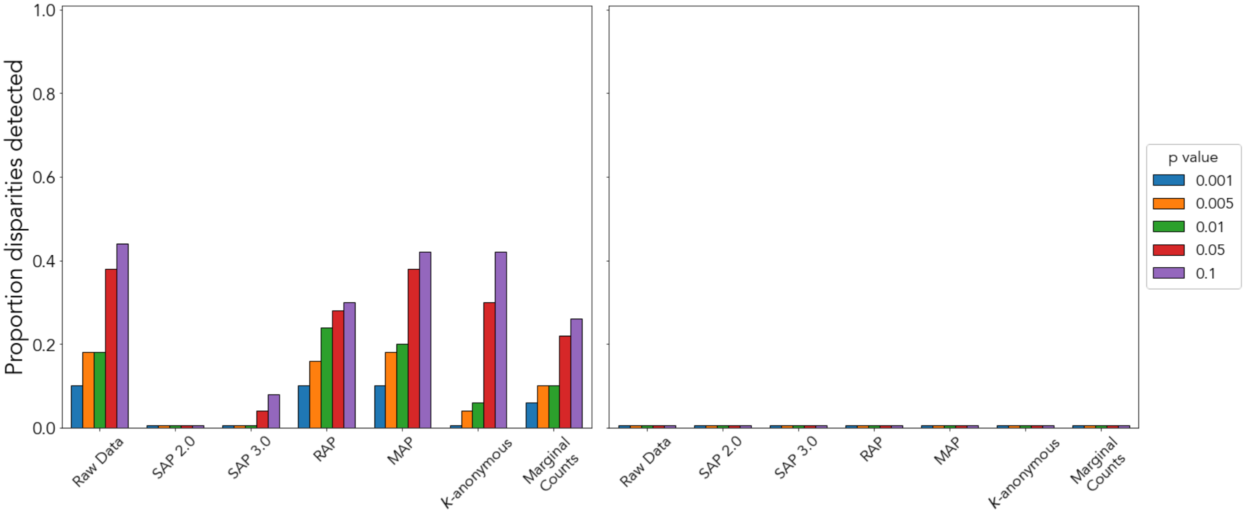


Figure 4.4. Proportion of detected disparities for Perry County, TN, in which at least one of the simulated disparity features (left) and both features (right) are detected. The proportion is out of 50 different experiment datasets.

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.

We next consider the detection times and false positives generated by each data sharing policy. We 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 4.5 and 4.6 present the AMOC curves for Davidson and Perry County, respectively.

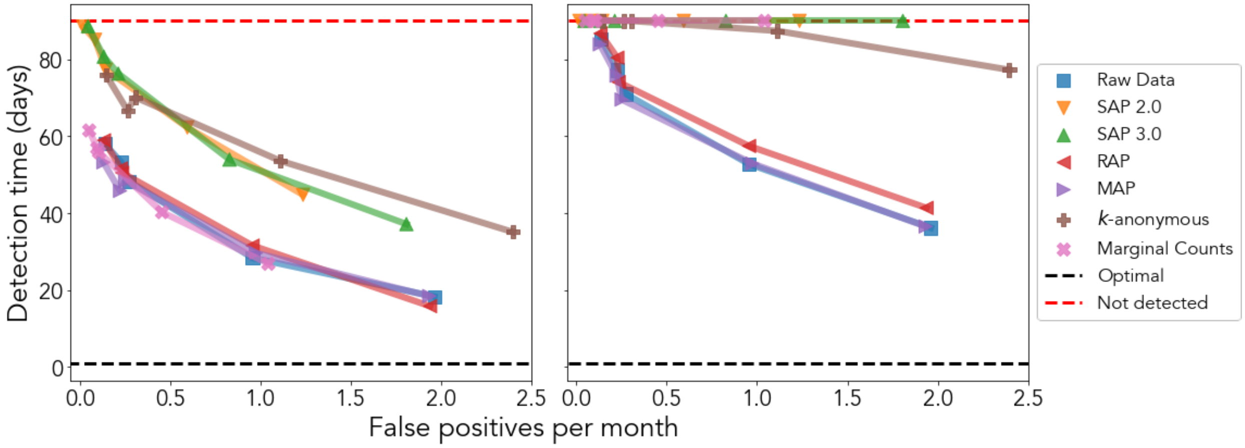


Figure 4.5. AMOC curves for Davidson County, TN, for detecting at least one of the simulated disparity features (left) and both features (right). Each point is the average of 50 different experiment datasets.

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.

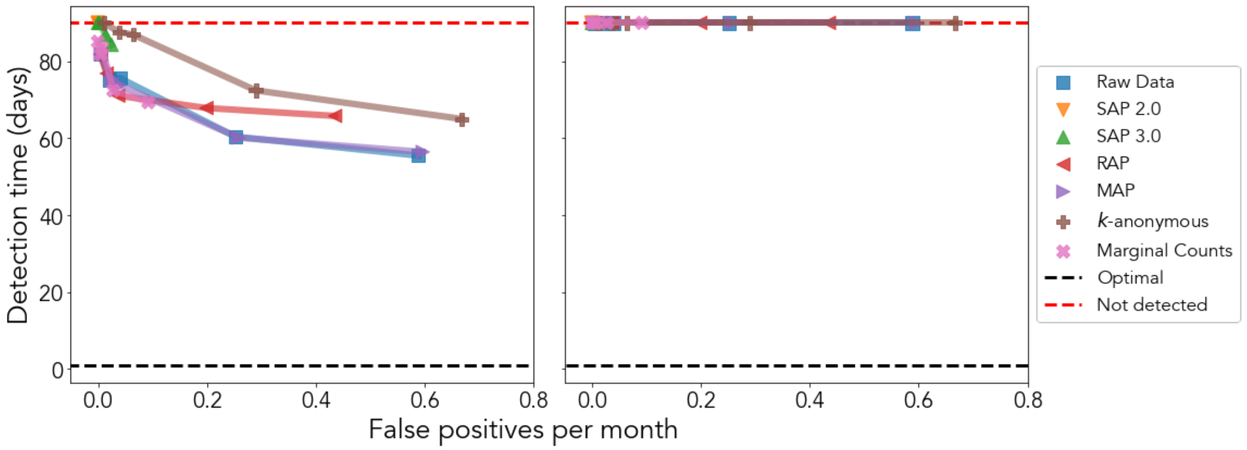


Figure 4.6. AMOC curves for Perry County, TN, for detecting at least one of the simulated disparity features (left) and both features (right). Each point is the average of 50 different experiment datasets.

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.

4.3.2 Fairness Experiment

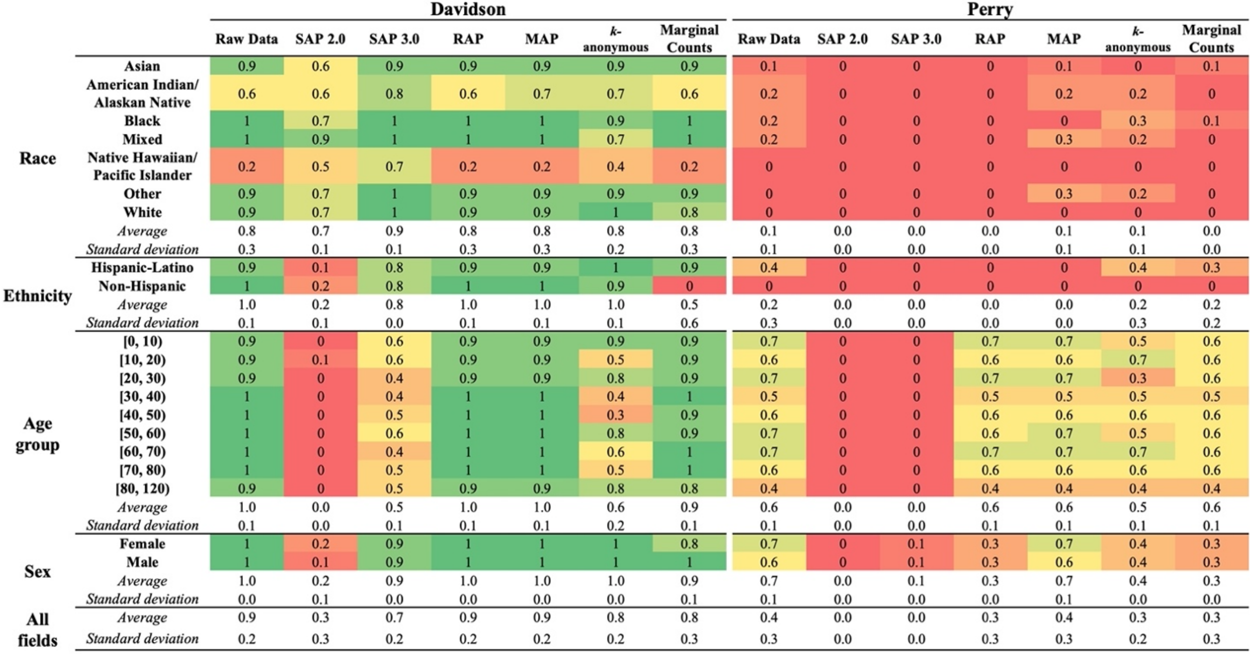
To ensure a de-identification policy supports the detection of disparities in different subpopulations to a similar degree, we evaluate each de-identification policy's ability to support the detection of disparities in different subpopulations. We first test for any statistically significant differences between the de-identification policies and the raw data. Then, we measure the variation in performance across groups, where a larger variation suggests less fair performance. The hypothesis test results are presented in Tables 4.3 and 4.5. The fairness results are presented in Tables 4.4 and 4.6.

Table 4.3. 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.

Table 4.4. 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

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.

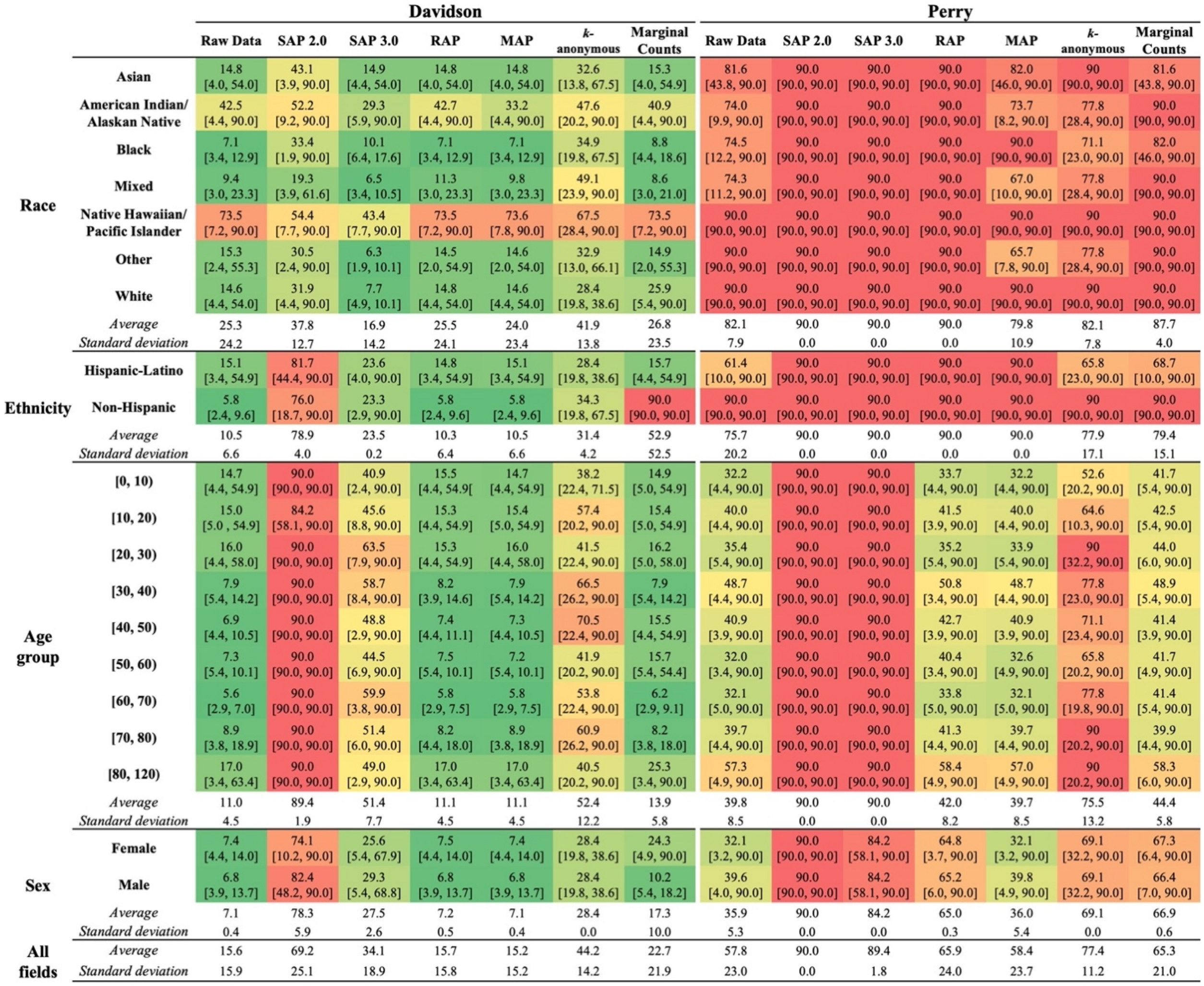
In Perry County, only the MAP and *k*-anonymous policies produced p-values greater than 0.05 in the McNemar tests. The SAP implementations produced the most significant p-values. In fact, the SAP policy does not support disparity detection for almost any group. This is because SAP does not share many 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 4.5. 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 4.6. 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]

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. It should be noted that RAP and MAP are relatively fair across groups, 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.

4.4 Discussion and Conclusions

To support COVID-19 disparity investigations, we evaluate how accurately, timely, and fairly disparate subpopulation outbreaks can be detected from data shared under three different dynamic policies and two policies derived from current public datasets. The results suggest that in larger, more heterogenous populations like Davidson County, TN, the RAP and MAP enable better disparity detection performance, for single and double-feature disparities, than the other policies. The *k*-anonymous policy's generalization of date of diagnosis hinders the ability to detect more specific, multi-feature disparities whilst generating more false positives. Though the policy can support accurate detection of large single-feature disparities, its monthly data publication schedule significantly delays time to detection. The Marginal Counts policy's lack of joint statistics prevents the detection of more than one demographic feature defining the disparity. In smaller, more homogenous populations like Perry County, TN, disparity detection is generally more challenging. Though RAP and MAP enable higher disparity detection rates and faster detection times than the other policies, they do not enable simultaneous detection of multiple disparity features.

The fairness in detection performance supported by each de-identification policy is more nuanced. In terms of producing similar detection times between racial groups in Davidson County, the SAP implementations outperform the other policies, including the raw data. Note, SAP's detection times are not shorter than those of other policies; they are more similar between racial groups. In fact, RAP and MAP support the detection of a greater proportion of disparities than SAP, and at earlier times, across all racial groups except for disparities occurring in the NH/PI subpopulation. This result, and the variation in time to detection between racial groups using the raw data, highlight the difficulty of detecting disparities in super-minority populations. Yet, fairness in terms of racial disparity detection is only part of the picture. SAP produces the fairest disparity detection in terms of race in Davidson County, but not in terms of age group. SAP 2.0 fails to detect nearly all age group disparities, where SAP 3.0 detects such disparities with a greater standard deviation in detection time than RAP and MAP. This is because SAP, after prioritizing racial granularity over age granularity, shares less granular information overall to mitigate the privacy risk of sharing data with a stronger adversary. Finally, when evaluating fairness with respect to all demographic subpopulations, the *k*-anonymous policy supports the fairest detection times in both Davidson and Perry counties. Nevertheless, the average detection time in Davidson County is nearly 3 times that of RAP and MAP, and MAP enables the detection of more disparities overall in both counties.

We would like to note that, in some cases, the de-identification policies outperform the raw data in detecting disparities. This finding seems counter-intuitive, but it follows similar findings in previous studies in which generalization dampened the noise in the data to improve downstream application performance. For instance, Deleger et al. measured medication extraction performance on clinical notes that had been de-identified (PHI removed) through various NLP and manual review methods. Compared to the raw data, several de-identification methods enabled marginally improved sensitivity and precision in extracting medication information[117].

In this study, we evaluate several dynamic policies, each designed to meet a privacy risk threshold against adversaries with different types of background knowledge. We do not, however, advocate for which policy should be implemented. Rather, our 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[56].

Despite the merits of this investigation, we wish to highlight several limitations that can guide future extensions of our work. First, our evaluation measures the ability to detect a disparity without quantifying how accurately the disparity is represented by the data sharing policy. Though data representation may be sufficient for accurate detection, it is likely the data sharing policies distort disparity features (e.g., severity or duration). Moreover, our simulated data does 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.

Second, our experiments using simulated data do not consider the effect of suppressing values (to achieve *k*-anonymity privacy guarantees[112]) and missing data on disparity detection. We illustrate their potential impact in the Case Studies, but do not quantify the results. Future work should quantify the robustness the policies' performance under suppression and varying levels of missingness.

Third, our evaluation relies on a single outbreak detection algorithm. It is possible that other outbreak detection algorithms improved disparity detection performance and fairness. Notably, however, 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[86,118]. 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[85]. In fact, to the extent of the authors' 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)[119]. Alternatively, different statistical methods, such as regression[6], 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, we focus our 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 are unfair in terms of providing similar disparity detection performance between counties. Future work should extend our fairness to analyze performance differences between all counties in a state or country.

In conclusion, we show that when protecting against a potential adversary of reasonable strength – an adversary who, at most, knows a target individual's 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.

Chapter 5

Summary

5.1 Discussion

In this thesis, I introduce 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 case loads while incorporating the uncertainty of who will be infected in the future. In Chapter 3 and Appendix 2, I demonstrate how dynamically changing the policy per the framework's recommendations maintains 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 Chapter 4, I show 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 enhance surveillance utility. 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 show how these last two features are crucial for effective disease monitoring[38,39], as they reduce 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 demonstrate 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 (referred to as SAP in Chapter 4), I fix county of residence and date of diagnosis granularity while increasing or decreasing the demographic granularity with the influx of new disease case records. I make 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 show 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 increases the data sharer's ability to share more granular information over time.

The disparity detection evaluation's results suggest that both in large, urban populations and small, rural populations, RAP and MAP enable 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 Chapter 4) generalization of date of diagnosis induces 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 increases detection times. The Marginal Counts policy can detect disparities in a timely manner, but its removal of joint distributions prevents the complete characterization of multi-feature disparities. Though SAP 3.0 outperforms SAP 2.0, it still provides suboptimal detection performance for both counties.

The fairness in detection performance supported by each de-identification policy is more nuanced. In terms of producing similar detection times between racial groups in Davidson County, the SAP implementations outperform the other policies, including the raw data. Note, SAP's detection times are not shorter than those of other policies; they are more similar between racial groups. In fact, RAP and MAP support the detection of a greater proportion of disparities than SAP, and at earlier times, across all racial groups except for disparities occurring in the NH/PI subpopulation. This result, and the variation in time to detection between racial groups using the raw data, highlight the difficulty of detecting disparities in super-minority populations. Yet, fairness in terms of racial disparity detection is only part of the picture. SAP produces the fairest disparity detection in terms of race in Davidson County, but not in terms of age group. SAP 2.0 fails to detect nearly all age group disparities, where SAP 3.0 detects such disparities with a greater standard deviation in detection time than RAP and MAP. This is because SAP, after prioritizing racial granularity over age granularity, shares less granular information overall to mitigate the privacy risk of sharing data with a stronger adversary. Finally, when evaluating fairness with respect to all demographic subpopulations, the *k*-anonymous policy supports the fairest detection times in both Davidson and Perry counties. Nevertheless, the average detection time in Davidson County is nearly 3 times that of RAP and MAP, and MAP enables the detection of more disparities overall in both counties.

In this thesis, I evaluate 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. This investigation shows 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[55,56].

5.2 Limitations and future directions

Despite the merits of this work, 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 remain relatively robust. Policies chosen from forecasted counts are typically similar or close to those chosen from actual case counts. And when overestimating the number of cases, the privacy risk does 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 exceeds 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 does 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 assign an equal likelihood of infection to all uninfected county residents at any given time in the simulations, and do not allow reinfections. In reality, the actual likelihood varies according to contact patterns of infectious individuals (i.e., through households or at work)[106,107], and reinfections are possible, though not likely in the case of COVID-19[90]. 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[108]. 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 evaluation in Chapter 4 measures the ability to detect a disparity without quantifying how accurately the disparity is 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 does 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.

Sixth, my experiments using simulated data do not consider the effect of suppressing values and missing data on disparity detection. As *k*-anonymity is often achieved in practice through suppression[112] and real-world data is rarely complete, future work should quantify the robustness the policies' performance under suppression and varying levels of missingness.

Seventh, the data utility evaluation in Chapter 4 relies on a single outbreak detection algorithm. It is possible that other outbreak detection algorithms improve performance and fairness. Notably, however, as discussed in Section 2.5, 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[86,118]. 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[85]. 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)[119]. Alternatively, different statistical methods, such as regression[6], 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 focus 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 are 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.

5.3 Conclusion

Disease surveillance data is variable, between geographic areas and over time. As such, data must be regularly updated in a timely manner. 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 introduce 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 show that forecast-driven de-identification offers better privacy protection than the static data sharing policy application. Moreover, I show 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.

5.4 Acknowledgements

This research was sponsored in part by grants from the National Science Foundation (CNS2029651 and CNS2029661) and National Institutes of Health (T15LM007450).

Appendix 1

My role in manuscript development

In the first manuscript, comprising Chapter 3, my role included designing the framework and privacy model, writing the computer code, performing the experiments, analyzing the results, and preparing the manuscript. Regarding the second manuscript, comprising Chapter 4, I designed the analysis, including the de-identification policies, the synthetic data generation method, and the experimental approach; wrote the computer code; executed the experiments; analyzed the results, and prepared the manuscript.

Appendix 2

Supplementary Information for Chapter 3

A2.1. Framework algorithm inputs

The Monte Carlo privacy risk estimation framework's 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 set of quasi-identifier values (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. We obtain the county distributions for the quasi-identifiers listed in Table 1 from the U.S. Census PCT12 tables[88].

Each PCT12 table contains joint statistics on age, sex, and county for a given Census-defined race[88]. 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. We 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. We then subtract these statistics from the PCT12H table. The new remainder is the number of non-white, Hispanic-Latino residents. We 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, we 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[120]. 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 data sharing policy's generalization 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. We use the Johns Hopkins COVID-19 tracking data for COVID-19 disease case times series[89]. 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. Different measures consider different types of re-identification attacks. We show the PK risk in Chapter 3, which is evaluated on a window of disease case records throughout the time series. Here, we show the marketer risk measure, which is evaluated on the cumulative dataset at each time point.

A2.2 Framework algorithm – PK risk

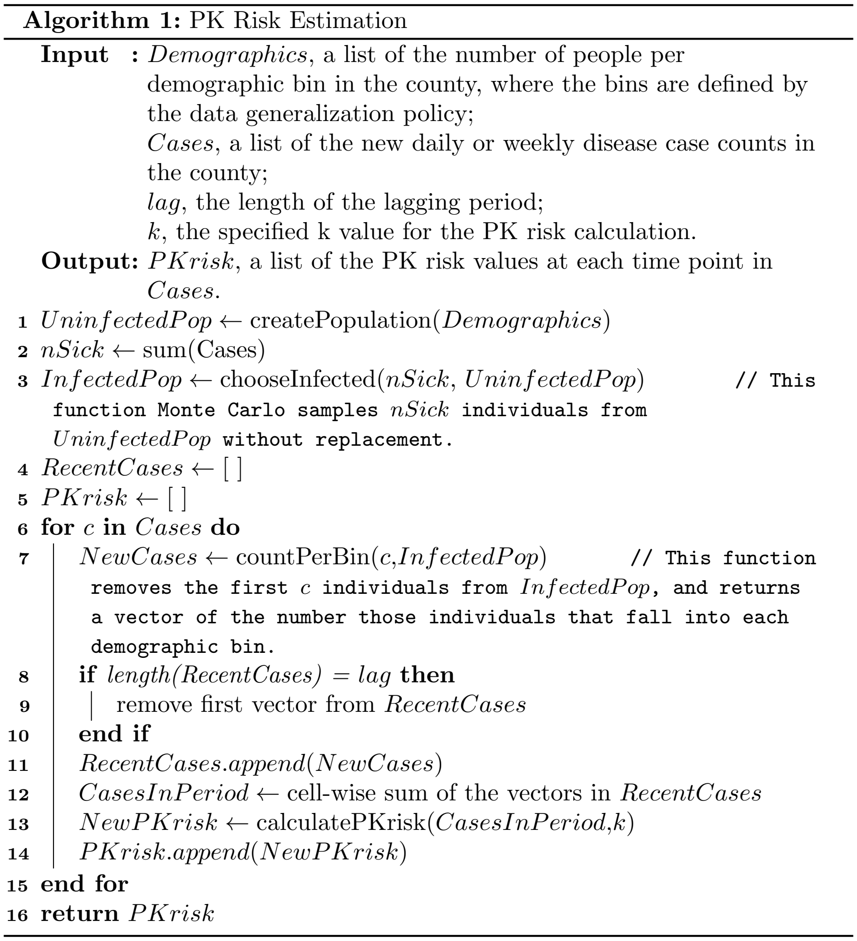
The algorithm follows the process described in Figure E1 to evaluate the privacy risk. The algorithm first creates the uninfected population from the input demographic distribution, where each county resident is uniquely represented by their demographic group (step 1). It then sums each value in *Cases* 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 combination of quasi-identifier values 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[93]. 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

$$\frac{\sum_{k=1}^{K-1} \sum_{j=1}^{J} I\left(f_j = k\right) \cdot k}{n} \qquad \text{(E1)}$$

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  $UninfectedPop \leftarrow$ createPopulation($Demographics$)
2  $nSick \leftarrow$ sum(Cases)
3  $InfectedPop \leftarrow$ chooseInfected($nSick$, $UninfectedPop$)        // This
     function Monte Carlo samples $nSick$ individuals from
     $UninfectedPop$ without replacement.
4  $RecentCases \leftarrow$ [ ]
5  $PKrisk \leftarrow$ [ ]
6  for $c$ in $Cases$ do
7      $NewCases \leftarrow$ 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     $CasesInPeriod \leftarrow$ cell-wise sum of the vectors in $RecentCases$
13     $NewPKrisk \leftarrow$ calculatePKrisk($CasesInPeriod$,$k$)
14     $PKrisk.append(NewPKrisk)$
15 end for
16 return $PKrisk$
```

Figure E1. PK risk estimation algorithm.

A2.3. PK risk algorithm complexity

We 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. E1, 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.

A2.4. Policy search

The policy search, whose results are summarized in Figure 3.3, applies the framework to calculate the PK11 estimates for combinations of data generalization policy, case count number, and county in a brute force manner. The feature generalization options for the policies follow the generalization hierarchy presented in Figure 3.2. The PK11 risk is calculated on 1,000 simulations, for each policy, case count, and county combination. We then compare the upper bound of the 95% quantile range of the simulations to the PK11 threshold. We choose to represent the PK11 distribution by the empirical confidence interval's upper bound to increase the likelihood the outcome remains below the threshold in practice. If the upper bound is less than or equal to 0.01 for every county in the size category, the policy is marked as an acceptable policy. When a policy does not meet the threshold, the policy and its parent policies are removed from the list of potential policies for the remainder of counties in the county size category, at the case count value. We acknowledge that previously developed policy search strategies[53,54] could be applied to the framework's output, but this type of policy search is not a focus of this paper nor necessary to search the limited number of policies presented here.

Sampling without replacement requires the total county population to equal or exceed the case number . Therefore, policy search combinations are restricted to the county sets meeting this requirement. For example, the results for the (0, 1k) county size category at 300 cases were generated from the US counties with a total population in the interval [300, 1000).

We choose to summarize the policy results into tables to aid in readability and facilitate downstream applications. Summarization first groups results by population size to broadly incorporate its effect on the underlying entropy in each county's demographic distribution, which influences the privacy risk[121]. Summarization also involves limiting the policies listed in each cell to two or less. Policies listed in each cell are those in the search results that avoid suppressing quasi-identifier fields, prioritize age granularity, and are not child policies of any that meet the threshold.

A2.5. Estimating the risk of the dynamic policy

The data generalization policy can be chosen according to the expected number of new cases based on the results of the policy search. We use the CDC ensemble prediction model's one week estimates for policy selection in the evaluation (Table 3.2) and case studies (Figures 3.4 and 3.5). We specifically used the model's point estimates, calculated as the median of the point estimates of the various prediction models.

We calculate each county's PK11 estimates under the dynamic policy selection in the following manner. For a given county, we simulate the first policy listed in each cell of the summarized policy search results (Figure 3.3) for the corresponding county size. For example, if the county had less than 1,000 residents per the U.S. Census, we run 1,000 simulations for each ****, **s*, 4***, and 4*s*.

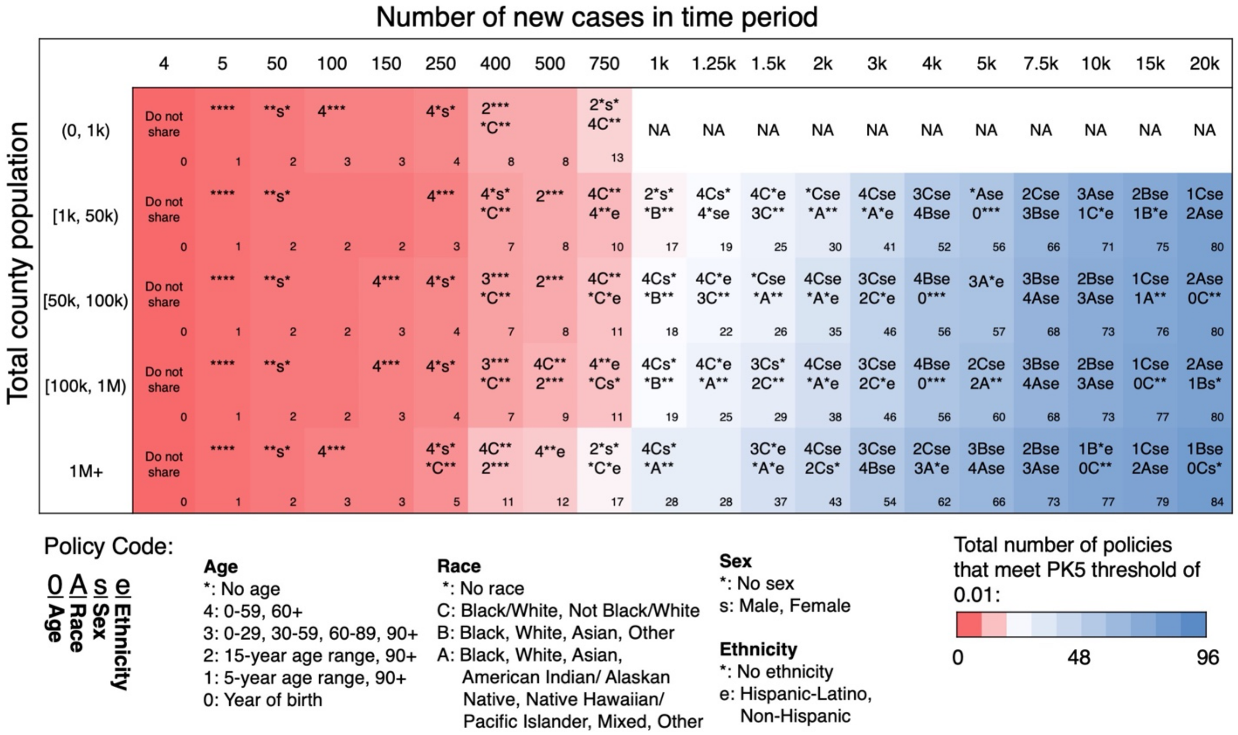
Each simulation uses the Johns Hopkins surveillance data to define the disease case time series from January 23, 2020 through October 23, 2021. This simulates sharing the actual number of disease records throughout the COVID-19 pandemic.

After calculating the county's privacy risk estimates under each policy, we use the forecasted counts to select which of the policies to apply at the beginning of each week. For the daily release schedule, we select the policy according to the minimum forecasted rolling sum within the upcoming week. For the weekly release schedule, we use the forecasted weekly increase. We then concatenate the PK11 estimates from each policy's simulations by weekly increments, following the sequence of selected policies. The concatenated list of privacy risk estimates represents the county's PK11 risk of sharing the actual number of disease records per day or per week when dynamically selecting the data generalization policy.

We estimate the privacy risk of the dynamically adapting policy in this manner for is all counties in the evaluation in Table 3.2, under two release schedules, and for the case studies. The case studies include an additional variation, where each week's policy is selected from the actual case numbers (from the Johns Hopkins data) instead of the forecasted case quantities. We add this variation to evaluate how well the dynamic policy meets the PK11 threshold when perfectly predicting future case loads.

A2.6. Additional PK risk policy search results

We repeat our policy search for two additional *k* values: *k* = 5 and *k* = 20, while maintaining a threshold of 0.01 in both cases. We summarize the results in Supplemental Figures E2 and E3, which are formatted in the same manner as Figure 3.3.



Number of new cases in time period

Total county population

	4	5	50	100	150	250	400	500	750	1k	1.25k	1.5k	2k	3k	4k	5k	7.5k	10k	15k	20k
(0, 1k)	Do not share 0	**** 1	**s* 2	4*** 3	3	4*s* 4	2*** *C** 8	8	2*s* 4C** 13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[1k, 50k)	Do not share 0	**** 1	**s* 2	2	2	4*** 3	4*s* *C** 7	2*** 8	4C** 4**e 10	2*s* *B** 17	4Cs* 4*se 19	4C*e 3C** 25	*Cse *A** 30	4Cse *A*e 41	3Cse 4Bse 52	*Ase 0*** 56	2Cse 3Bse 66	3Ase 1C*e 71	2Bse 1B*e 75	1Cse 2Ase 80
[50k, 100k)	Do not share 0	**** 1	**s* 2	2	4*** 3	4*s* 4	3*** *C** 7	2*** 8	4C** *C*e 11	4Cs* *B** 18	4C*e 3C** 22	*Cse *A** 26	4Cse *A*e 35	3Cse 2C*e 46	4Bse 0*** 56	3A*e 57	3Bse 4Ase 68	2Bse 3Ase 73	1Cse 1A** 76	2Ase 0C** 80
[100k, 1M)	Do not share 0	**** 1	**s* 2	2	4*** 3	4*s* 4	3*** *C** 7	4C** 2*** 9	4**e *Cs* 11	4Cs* *B** 19	4C*e *A** 25	3Cs* 2C** 29	4Cse *A*e 38	3Cse 2C*e 46	4Bse 0*** 56	2Cse 2A** 60	3Bse 4Ase 68	2Bse 3Ase 73	1Cse 0C** 77	2Ase 1Bs* 80
1M+	Do not share 0	**** 1	**s* 2	4*** 3	3	4*s* *C** 5	4C** 2*** 11	4**e 12	2*s* *C*e 17	4Cs* *A** 28	28	3C*e *A*e 37	4Cse 2Cs* 43	3Cse 4Bse 54	2Cse 3A*e 62	3Bse 4Ase 66	2Bse 3Ase 73	1B*e 0C** 77	1Cse 2Ase 79	1Bse 0Cs* 84

Policy Code:

Age
*: No age
4: 0-59, 60+
3: 0-29, 30-59, 60-89, 90+
2: 15-year age range, 90+
1: 5-year age range, 90+
0: Year of birth

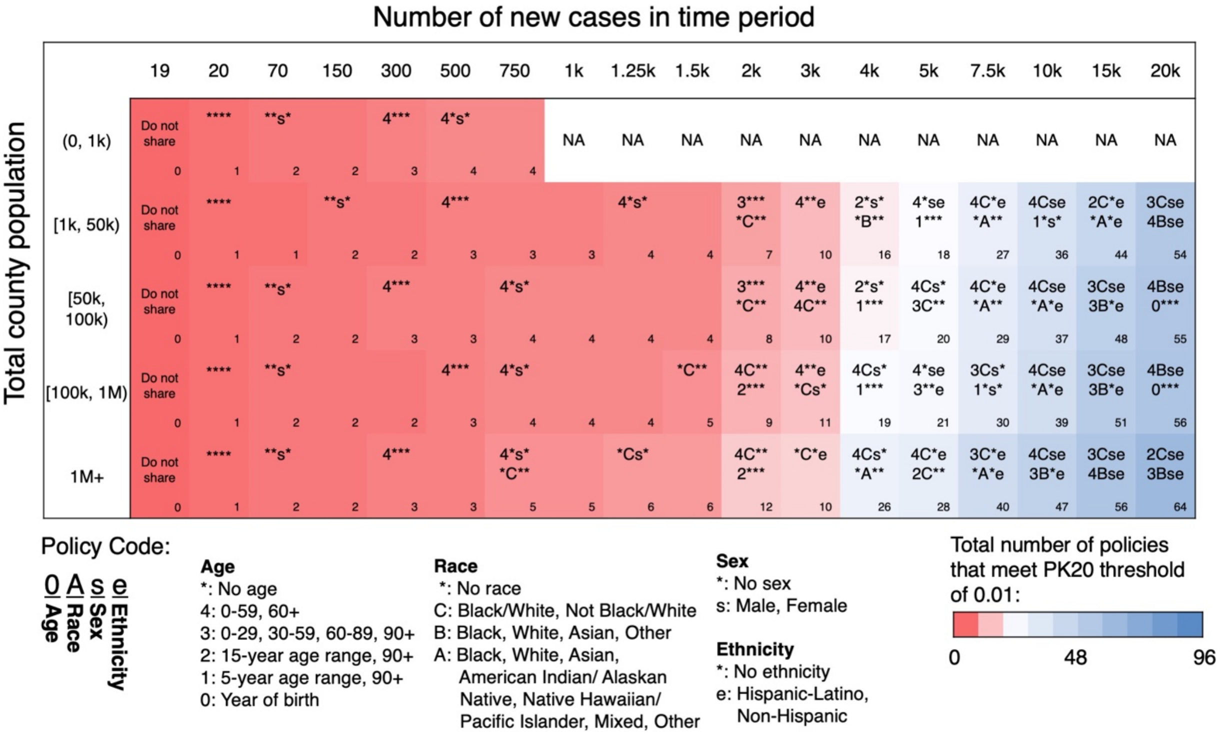
Race
*: No race
C: Black/White, Not Black/White
B: Black, White, Asian, Other
A: Black, White, Asian, American Indian/ Alaskan Native, Native Hawaiian/ Pacific Islander, Mixed, Other

Sex
*: No sex
s: Male, Female

Ethnicity
*: No ethnicity
e: Hispanic-Latino, Non-Hispanic

Total number of policies that meet PK5 threshold of 0.01:

Figure E2. 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.



Number of new cases in time period

Total county population	19	20	70	150	300	500	750	1k	1.25k	1.5k	2k	3k	4k	5k	7.5k	10k	15k	20k
(0, 1k)	Do not share 0	**** 1	**s* 2	2	4*** 3	4*s* 4	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
[1k, 50k)	Do not share 0	**** 1	1	**s* 2	2	4*** 3	3	3	4*s* 4	4	3*** *C** 7	4**e 10	2*s* *B** 16	4*se 1*** 18	4C*e *A** 27	4Cse 1*s* 36	2C*e *A*e 44	3Cse 4Bse 54
[50k, 100k)	Do not share 0	**** 1	**s* 2	2	4*** 3	3	4*s* 4	4	4	4	3*** *C** 8	4**e 4C** 10	2*s* 1*** 17	4Cs* 3C** 20	4C*e *A** 29	4Cse *A*e 37	3Cse 3B*e 48	4Bse 0*** 55
[100k, 1M)	Do not share 0	**** 1	**s* 2	2	2	4*** 3	4*s* 4	4	4	*C** 5	4C** 2*** 9	4**e *Cs* 11	4Cs* 1*** 19	4*se 3**e 21	3Cs* 1*s* 30	4Cse *A*e 39	3Cse 3B*e 51	4Bse 0*** 56
1M+	Do not share 0	**** 1	**s* 2	2	4*** 3	3	4*s* *C** 5	5	*Cs* 6	6	4C** 2*** 12	*C*e 10	4Cs* *A** 26	4C*e 2C** 28	3C*e *A*e 40	4Cse 3B*e 47	3Cse 4Bse 56	2Cse 3Bse 64

Policy Code:
0 A s e
Age Race Sex Ethnicity
Age
*: No age
4: 0-59, 60+
3: 0-29, 30-59, 60-89, 90+
2: 15-year age range, 90+
1: 5-year age range, 90+
0: Year of birth
Race
*: No race
C: Black/White, Not Black/White
B: Black, White, Asian, Other
A: Black, White, Asian, American Indian/ Alaskan Native, Native Hawaiian/ Pacific Islander, Mixed, Other
Sex
*: No sex
s: Male, Female
Ethnicity
*: No ethnicity
e: Hispanic-Latino, Non-Hispanic
Total number of policies that meet PK20 threshold of 0.01:
0
48
96

Figure E3. 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.

A2.7. Framework algorithm – marketer risk

A variety of privacy risk measures have been developed to account for different portions of the privacy risk distribution and different types of re-identification scenarios[37]. In the main body of the manuscript, we apply the Monte Carlo framework to estimate the PK risk, an upper bound of the re-identification risk. Here, we use the framework to estimate the amortized re-identification risk, also known as the marketer risk[59]. These two measures are not necessarily mutually exclusive. The PK risk considers the most unique records in the dataset, while the marketer risk measures the average uniqueness of each record in the context of the surrounding population. We do not suggest which measure dictates the best privacy protection; rather, we provide an illustration of how to apply the framework under another privacy perspective. We leave the decision of how to use the measures to the data sharer.

The marketer risk considers a different attack scenario, 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[28,30]. Not every county resident registers to vote, but for simplicity, we assume in our analysis the data recipient possesses an identified dataset containing every county resident. This assumption models the worst-case scenario when considering the marketer risk. We further assume the dataset contains all demographic information listed in Table 1, except for the date of diagnosis. 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. First, the marketer risk is evaluated on the cumulative dataset at each time point as date of diagnosis is no longer considered a quasi-identifier, 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. Supplemental Figure E4 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 1: 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 $Cases$.
 1 $UninfectedPop \leftarrow$ createPopulation($Demographics$)
 2 $nSick \leftarrow$ sum(Cases)
 3 $InfectedPop \leftarrow$ chooseInfected($nSick$, $UninfectedPop$)        // This
     function Monte Carlo samples $nSick$ individuals from
     $UninfectedPop$ without replacement.
 4 $AllCases \leftarrow$ zero vector of the same dimension as $Demographics$
 5 $MarketerRisk \leftarrow [\,]$
 6 for $c$ in $Cases$ do
 7     $NewCases \leftarrow$ 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     $AllCases \leftarrow AllCases + NewCases$
 9     $NewMarketerRisk \leftarrow$ calculateMarketerRisk($AllCases$)
10     $MarketerRisk.append(NewMarketerRisk)$
11 end for
12 return $MarketerRisk$
```

Figure E4. Marketer risk estimation algorithm.

The marketer risk is calculated following the formulation from Dankar and El Emam[59]. In Eqn. E2, 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.

$$\frac{\sum_{j=1}^{J} \frac{f_j}{F_j}}{n} \qquad \text{(E2)}$$

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 (13).

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.

A2.8. Marketer risk-based policy search

We apply the framework to search the same policy space as before (described in Figure 3.2) and identify data sharing policies that likely meet a marketer risk threshold 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 we calculate the marketer risk on 1,000 independent simulations. From the 1,000 simulations, we 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. We summarize the results in Supplemental Figure E5.

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 Supplemental Figure E5, we select a single path for each county population category and generate a corresponding results table.

Supplemental Figure E5 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. There 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.

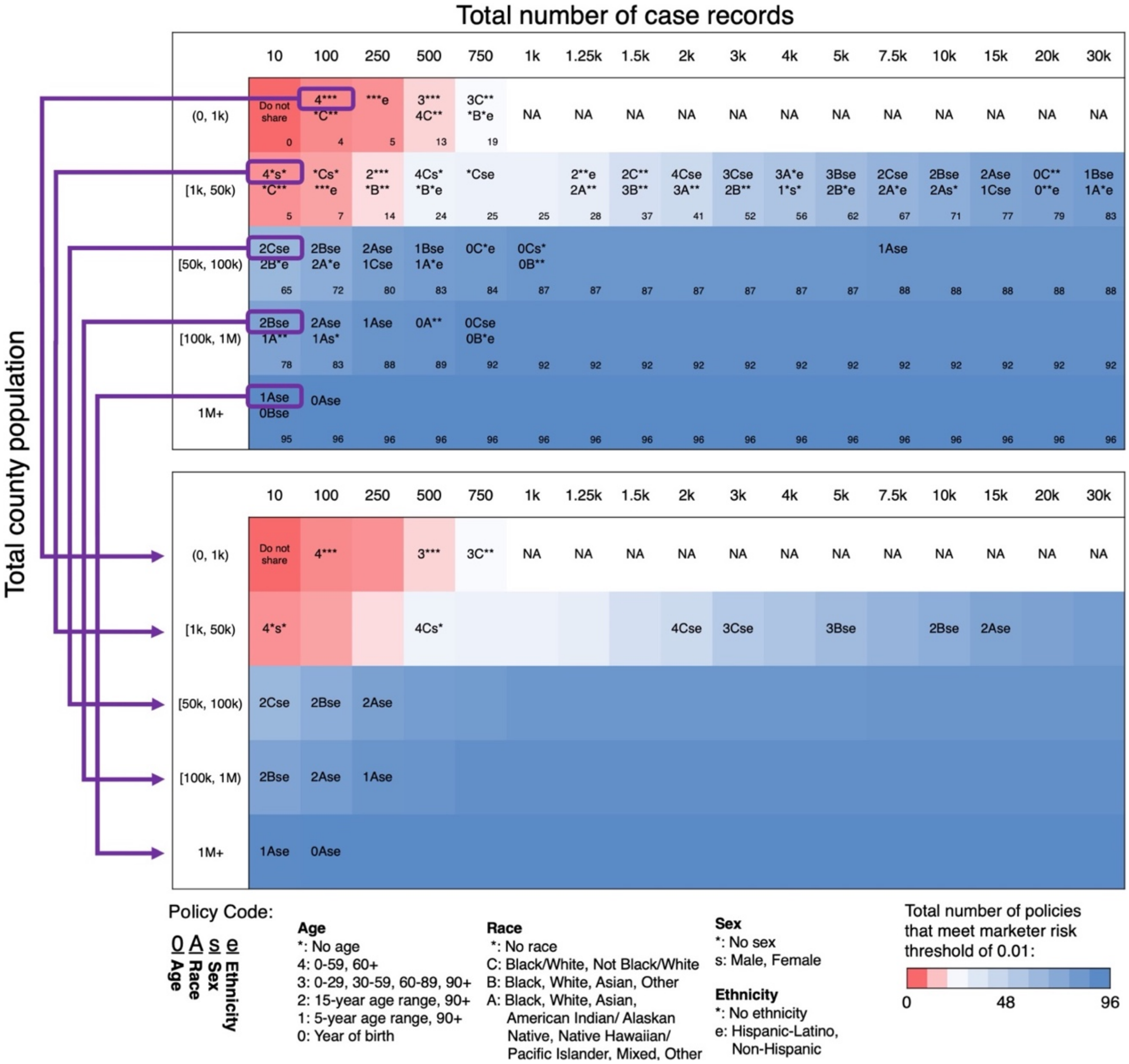


Figure E5. (Top) Policies with a marketer risk upper bound (calculated as the upper bound of the 95% quantile range of 1,000 framework simulations) less than or equal to 0.01 at varying disease case volume thresholds. A four-character alphanumeric code indicates the policy's generalization levels. All 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.

To further illustrate the relationship between the marketer risk and the size of the dataset, we 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 Supplemental Figure E5) is applied to a daily release schedule and allows for 532 potential demographic groups. Supplemental Figure E6 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. We 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.

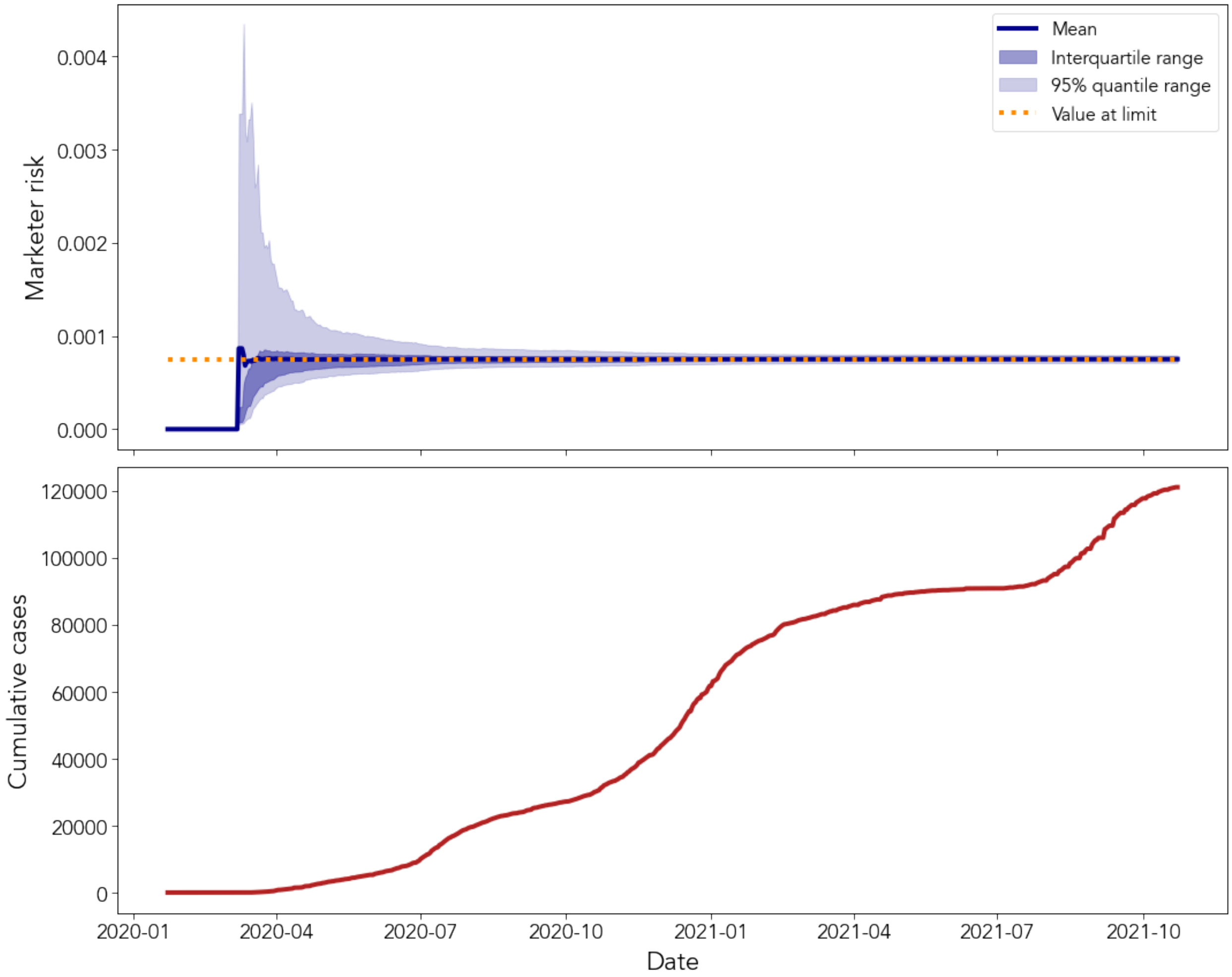


Figure E6. Marketer risk estimation of 1Ase policy applied to daily releases of COVID-19 disease case surveillance data in Davidson County, TN. The expectation and quantile ranges were calculated from 1,000 independent simulations. The marketer risk is evaluated each day (Top) on the cumulative number 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. E4.

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. E2) moves toward its limit, as shown in Eqn. E3:

$$\frac{\sum_{j=1}^{J} 1}{N} = \frac{J}{N} \quad \text{(E3)}$$

where N is the size of the identified dataset/total population. Eqn. E3 approximates the expected marketer risk estimated by the framework's algorithm. The orange dotted line in Supplemental Figure E6 was calculated using Eqn. E4:

$$\frac{\sum_{j=1}^{\tilde{J}} 1}{N} = \frac{\tilde{J}}{N} \quad \text{(E4)}$$

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.

A2.9. Dynamic policy evaluation – marketer risk

We repeat the evaluation from the main text of the paper, this time for the marketer risk-based policy search results. The succession of policies applied follows the generalization paths displayed in the bottom table in Supplemental Figure E5. Again, we consider a daily and a weekly release schedule in the context of the COVID-19 pandemic. Since date of diagnosis is not a quasi-identifier in this scenario, no date generalization is specified. We again compare the results of the framework-informed dynamic policy selection to statically applying the *k*-anonymous policy described in the main text.

In the marketer risk scenario, we 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) we 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). The privacy risk of sharing the actual number of case records is again estimated by inputting the Johns Hopkins COVID-19 tracking data into the framework. We restrict the results to the same time period as before for consistency. Supplementary Table E1 presents the results.

Table E1. 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]

Dynamic policy selection, guided by the framework's results, 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. 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.

A2.10. Marketer risk case studies

In this section, we demonstrate how to apply the marketer risk-based guidance, revisiting Davidson and Perry counties. The case studies select the data sharing policy on a weekly basis for a daily release schedule, in the same manner as the evaluation above. The actual marketer risk is estimated from the Monte Carlo framework, using the Johns Hopkins tracking data as input. We restrict the results to the same time interval as the case studies in the main paper. The results for Davidson County are presented in Supplemental Figure E7, and the results for Perry County are presented in Supplemental Figure E8.

As the generalization path in Supplemental Figure E5 instructs, the 1Ase policy is applied to each data release, as the size of the dataset remains above 250 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 data sharing 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.

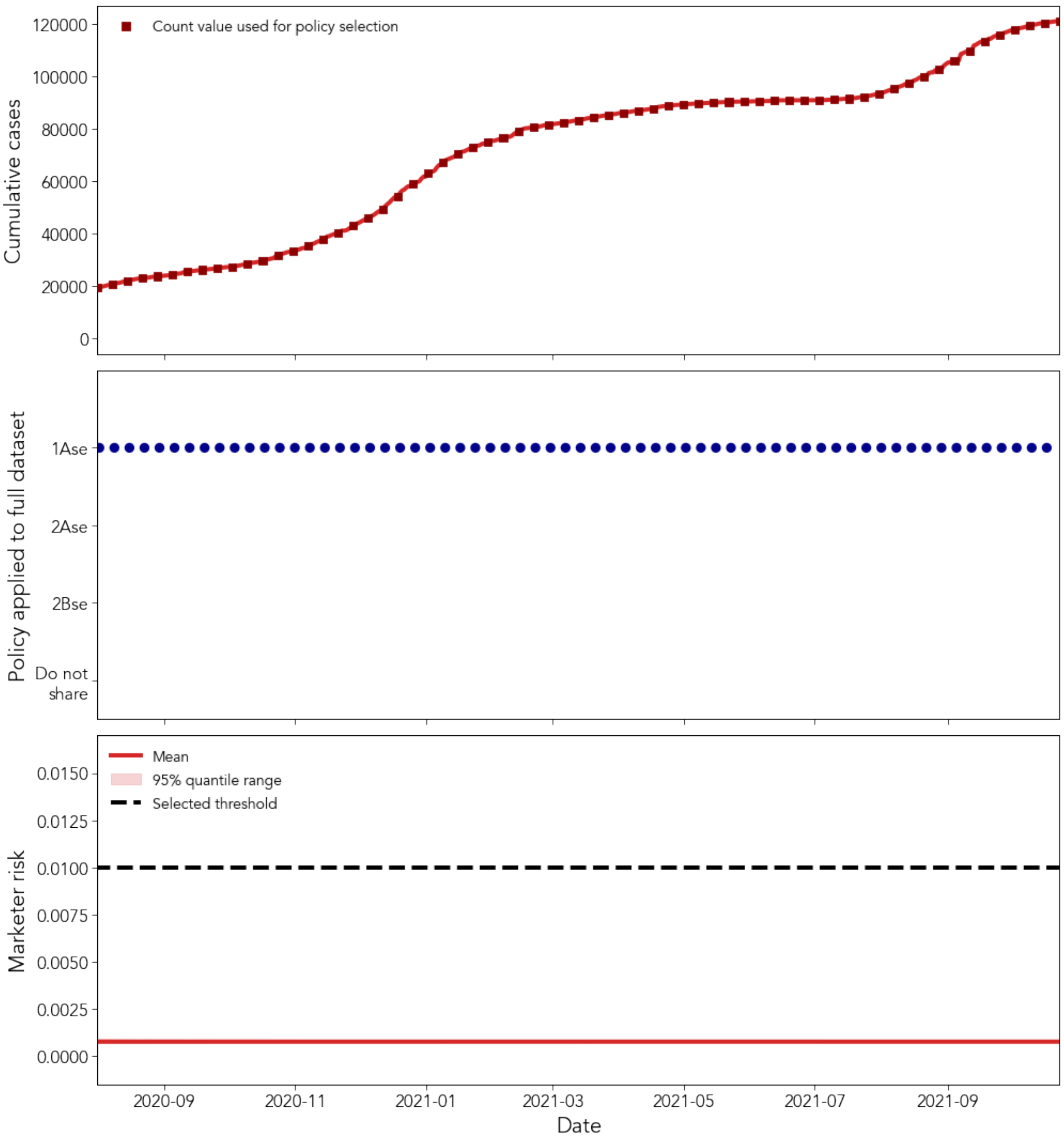


Figure E7. Dynamic policy selection applied to Davidson County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The cumulative sum of the case counts reported in Davidson County, according to the Johns Hopkins COVID-19 tracking data. The red squares represent the case record number 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 Supplementary Figure E5. (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.

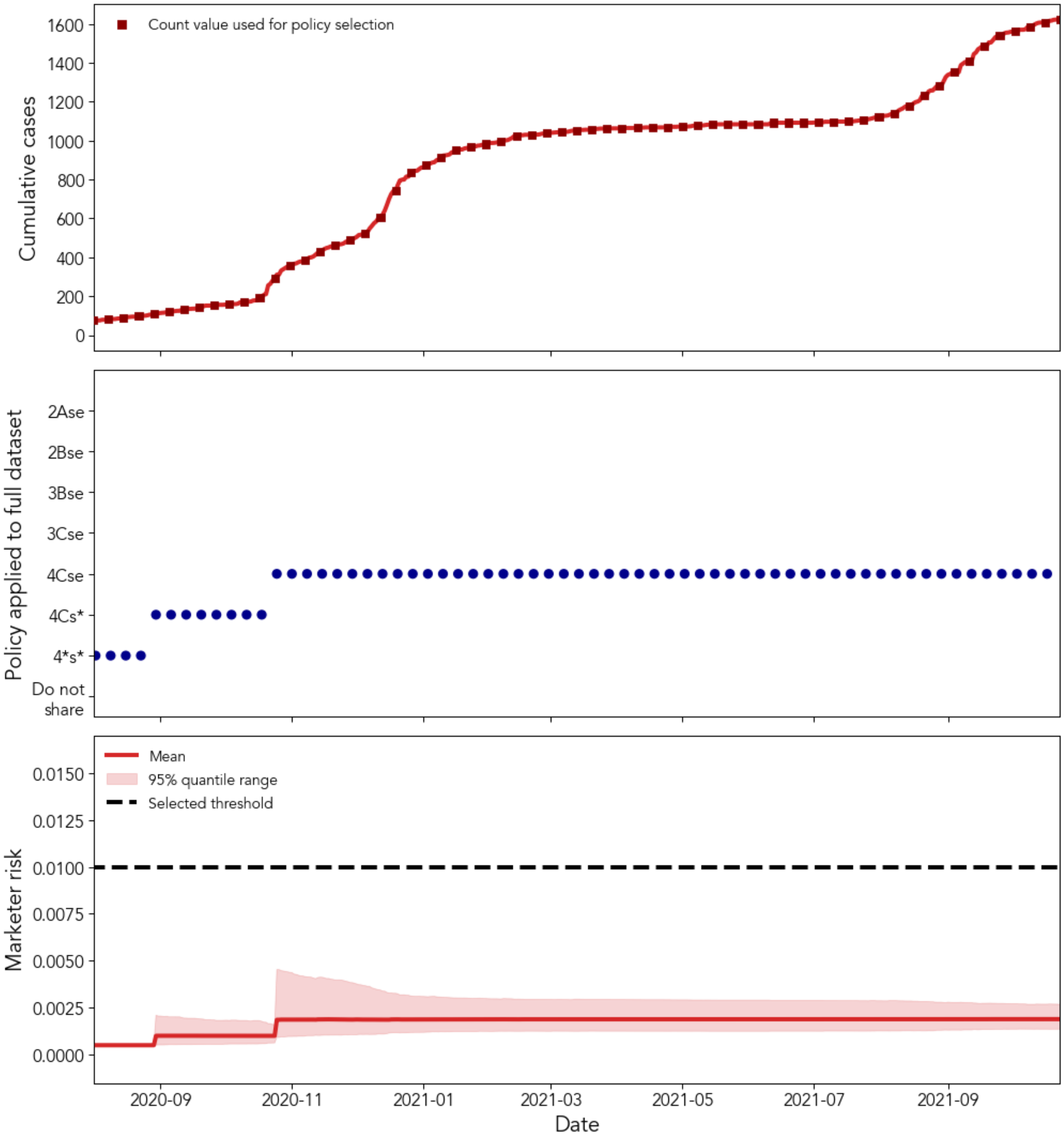


Figure E8. Dynamic policy selection applied to Perry County, TN in the COVID-19 pandemic (August 2, 2020 to October 23, 2021). (Top) The cumulative sum of the case counts reported in Davidson County, according to the Johns Hopkins COVID-19 tracking data. The red squares represent the case record number 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 Supplementary Figure E5. (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.

References

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

2. 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.

3. 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.

4. 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.

5. 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 [Internet]. 2020 Mar 16; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7228381/

6. 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.

7. 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.

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

9. 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/

10. 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.

11. 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. Journal of the American Medical Informatics Association. 2021 Jul 1;28(7):1363–73.

12. 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.

13. 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. Journal of the American Medical Informatics Association [Internet]. 2020 Aug 17;(ocaa196). Available from: https://doi.org/10.1093/jamia/ocaa196

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

15. COVID-19 Case Surveillance Public Use Data with Geography | Data | Centers for Disease Control and Prevention [Internet]. Available from: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

16. COVID-19 Case Surveillance Restricted Access Detailed Data | Data | Centers for Disease Control and Prevention [Internet]. Available from: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t

17. 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.

18. Maxmen A. Massive Google-funded COVID database will track variants and immunity. Nature [Internet]. 2021 Feb 24; Available from: https://www.nature.com/articles/d41586-021-00490-5

19. Ness RB, Joint Policy Committee S of E for the. Influence of the HIPAA Privacy Rule on Health Research. JAMA. 2007 Nov 14;298(18):2164–70.

20. Rights (OCR) O for C. Summary of the HIPAA Privacy Rule [Internet]. HHS.gov. 2008 [cited 2020 Dec 22]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

21. 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.

22. 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

23. 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

24. 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/

25. 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

26. The Utah Consumer Privacy Act: Utah Becomes Fourth US State with Comprehensive Privacy Law [Internet]. JD Supra. [cited 2022 Apr 12]. Available from: https://www.jdsupra.com/legalnews/the-utah-consumer-privacy-act-utah-2977882/

27. Golle P. Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of the 5th ACM workshop on Privacy in electronic society [Internet]. New York, NY, USA: Association for Computing Machinery; 2006. p. 77–80. (WPES '06). Available from: http://doi.org/10.1145/1179601.1179615

28. Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. 2000;34.

29. 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.

30. 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.

31. El Emam K, Dankar FK. Protecting Privacy Using k-Anonymity. J Am Med Inform Assoc. 2008;15(5):627–37.

32. Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA L Rev. 2009 2010;57(6):1701–78.

33. Piller C. Data secrecy may cripple U.S. attempts to slow pandemic. Science. 2020 Jul 24;369(6502):356–8.

34. Maxmen A. Why the United States is having a coronavirus data crisis. Nature [Internet]. 2020 Aug 25; Available from: https://www.nature.com/articles/d41586-020-02478-z

35. Office for Civil Rights (OCR). Methods for De-identification of PHI [Internet]. HHS.gov. 2012. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

36. Cassa CA, Grannis SJ, Overhage JM, Mandl KD. A Context-sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection. Journal of the American Medical Informatics Association. 2006 Mar 1;13(2):160–5.

37. 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.

38. Thacker SB, Qualters JR, Lee LM. Public Health Surveillance in the United States: Evolution and Challenges* [Internet]. [cited 2021 Jan 20]. Available from: https://www.cdc.gov/MMWR/preview/mmwrhtml/su6103a2.htm

39. 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.

40. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. Lancet Digit Health. 2020 Apr;2(4):e201–8.

41. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform. 2004 Jun;37(3):179–92.

42. Samreth D, Arnavielhe S, Ingenrieth F, Bedbrook A, Onorato GL, Murray R, et al. Geolocation with respect to personal privacy for the Allergy Diary app - a MASK study. World Allergy Organization Journal. 2018 Jul 16;11(1):15.

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

44. 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.

45. Hauser C. Is Your Vaccine Card Selfie a Gift for Scammers? Maybe. The New York Times [Internet]. 2021 Feb 6; Available from: https://www.nytimes.com/2021/02/06/health/covid-vaccination-card.html

46. 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.

47. Rights (OCR) O for C. 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 [cited 2022 Apr 11]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

48. Office for Civil Rights, HHS. Standards for privacy of individually identifiable health information. Final rule. Fed Regist. 2002 Aug 14;67(157):53181–273.

49. Utah Consumer Privacy Act Newest State Privacy Act Signed into Law [Internet]. The National Law Review. [cited 2022 Apr 11]. Available from: https://www.natlawreview.com/article/utah-consumer-privacy-act-newest-state-privacy-act-signed-law

50. Rights (OCR) O for C. Disclosures for Emergency Preparedness - A Decision Tool: Limited Data Set (LDS) [Internet]. HHS.gov. 2007 [cited 2022 Apr 11]. Available from: https://www.hhs.gov/hipaa/for-professionals/special-topics/emergency-preparedness/limited-data-set/index.html

51. Sweeney L, Yoo JS, Perovich L, Boronow KE, Brown P, Brody JG. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Technol Sci. 2017;2017:2017082801.

52. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Journal of the American Medical Informatics Association. 2011 Jan 1;18(1):3–10.

53. 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 [Internet]. 2015 Mar 25;10(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4373733/

54. 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.

55. 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 [Internet]. Rochester, NY: Social Science Research Network; 2012 Jul. Report No.: ID 2076397. Available from: https://papers.ssrn.com/abstract=2076397

56. 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

57. 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.

58. Sweeney L. ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. Int J Unc Fuzz Knowl Based Syst. 2002 Oct;10(05):571–88.

59. Dankar FK, El Emam K. A method for evaluating marketer re-identification risk. In: Proceedings of the 1st International Workshop on Data Semantics - DataSem ’10 [Internet]. Lausanne, Switzerland: ACM Press; 2010. p. 1. Available from: http://portal.acm.org/citation.cfm?doid=1754239.1754271

60. 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.

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

62. CMS Cell Size Suppression Policy | ResDAC [Internet]. [cited 2022 Jan 10]. Available from: https://resdac.org/articles/cms-cell-size-suppression-policy

63. 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

64. Dwork C. Differential privacy. In: Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II [Internet]. Berlin, Heidelberg: Springer-Verlag; 2006. p. 1–12. (ICALP'06). Available from: https://doi.org/10.1007/11787006_1

65. 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 [Internet]. 2018 [cited 2022 Apr 11]; Available from: https://www.ssrn.com/abstract=3338027

66. 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

67. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, et al. Approximation Algorithms for k-Anonymity. In: Journal of Privacy Technology (JOPT) [Internet]. Edinburgh, UK,; 2005 [cited 2022 Apr 13]. Available from: http://ilpubs.stanford.edu:8090/645/

68. Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly System. In: Proceedings of the AMIA Annual Fall Symposium [Internet]. 1997 [cited 2022 Apr 13]. p. 51–5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233452/

69. 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/

70. LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian Multidimensional K-Anonymity. In: 22nd International Conference on Data Engineering (ICDE'06) [Internet]. Atlanta, GA, USA: IEEE; 2006 [cited 2022 Apr 13]. p. 25–25. Available from: http://ieeexplore.ieee.org/document/1617393/

71. Wang K, Fung BCM. Anonymizing sequential releases. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06 [Internet]. Philadelphia, PA, USA: ACM Press; 2006 [cited 2022 Apr 13]. p. 414. Available from: http://portal.acm.org/citation.cfm?doid=1150402.1150449

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

73. 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.

74. Jeffery C, Ozonoff A, White LF, Nuño M, Pagano M. Power to Detect Spatial Disturbances under Different Levels of Geographic Aggregation. J Am Med Inform Assoc. 2009;16(6):847–54.

75. 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.

76. 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.

77. 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.

78. 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

79. 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.

80. Keating D, Cha AE, Florit G. ‘I just pray God will help me’: Racial, ethnic minorities reel from higher covid-19 death rates [Internet]. Washington Post. [cited 2022 Feb 4]. Available from: https://www.washingtonpost.com/graphics/2020/health/covid-race-mortality-rate/

81. Gross CP, Essien UR, Pasha S, Gross JR, Wang S yi, Nunez-Smith M. Racial and Ethnic Disparities in Population Level Covid-19 Mortality [Internet]. medRxiv; 2020 [cited 2022 Mar 2]. p. 2020.05.07.20094250. Available from: https://www.medrxiv.org/content/10.1101/2020.05.07.20094250v1

82. 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.

83. Wong WK, Moore A, Cooper G, Wagner M. What’s Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Learning Research. 2005;6(66):1961–98.

84. Wong WK. Data mining for early disease outbreak detection [Internet] [Ph.D.]. [United States -- Pennsylvania]: Carnegie Mellon University; [cited 2022 Apr 8]. Available from: https://www.proquest.com/docview/305203527/abstract/3CAFA28A3D1A4E0BPQ/1

85. 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.

86. 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.

87. 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 [Internet]. 2020 Aug 10;0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30635-6/abstract

88. Population Census Tables [Internet]. The United States Census Bureau. 2016. Available from: https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

89. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020 May 1;20(5):533–4.

90. 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.

91. Walters CE, Meslé MMI, Hall IM. Modelling the global spread of diseases: A review of current practice and capability. Epidemics. 2018 Dec;25:1–8.

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

93. 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.

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

95. 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

96. 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

97. 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.

98. 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.

99. 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

100. 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.

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

102. 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.

103. 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

104. 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/

105. 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/

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

107. 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.

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

109. Center for Disease Control and Prevention. Coronavirus Disease 2019 (COVID-19) [Internet]. Centers for Disease Control and Prevention. 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html

110. Rights (OCR) O for C. The HIPAA Privacy Rule [Internet]. HHS.gov. 2008 [cited 2022 Feb 9]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

111. 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. Journal of the American Medical Informatics Association. 2022 Feb 19;ocac011.

112. Sweeney L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. Int J Unc Fuzz Knowl Based Syst. 2002 Oct 1;10(05):557–70.

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

114. Lotze T, Shmueli G, Yahav I. Simulating Multivariate Syndromic Time Series and Outbreak Signatures. SSRN Journal [Internet]. 2007 [cited 2021 Nov 15]; Available from: http://www.ssrn.com/abstract=990020

115. Sartwell PE. The Distribution of Incubation Periods of Infectious Disease. American Journal of Epidemiology. 1995 Mar 1;141(5):386–94.

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

117. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2013;20(1):84–94.

118. 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.

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

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

121. Airoldi EM, Bai X, Malin BA. An Entropy Approach to Disclosure Risk Assessment: Lessons from Real Applications and Simulated Domains. Decis Support Syst. 2011 Apr 1;51(1):10–20.