2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995513

Privacy-Preserving Publishing of Individual-Level Pandemic Data Based on a Game Theoretic Model

Abinitha Gourabathina[#]
Department of Operations Research & Financial Engineering
Princeton University
Princeton, USA
abinitha@princeton.edu

Zhiyu Wan[#,*]
Department of Biomedical Informatics
Vanderbilt University Medical Center
Nashville, USA
zhiyu.wan.1@vumc.org

J. Thomas Brown
Department of Biomedical Informatics
Vanderbilt University Medical Center
Nashville, USA
james.t.brown@vanderbilt.edu

Chao Yan
Department of Biomedical Informatics
Vanderbilt University Medical Center
Nashville, USA
chao.yan.1@vumc.org

Bradley A. Malin[*]
Department of Biomedical Informatics
Vanderbilt University Medical Center
Nashville, USA
b.malin@vumc.org

Abstract—**Sharing individual-level pandemic data is essential for accelerating the understanding of a disease. For example, COVID-19 data have been widely collected to support public health surveillance and research. In the United States, these data need to be de-identified before being released to the public due to privacy concerns. However, current data publishing approaches for individual-level pandemic data, such as those adopted by the U.S. Centers for Disease Control and Prevention (CDC), have not flexed over time to account for the dynamic nature of infection rates. Thus, the policies generated by these strategies may either raise privacy risks or impair the data utility (or usability). To optimize the tradeoff between privacy risk and data utility, we introduce a game theoretic model that adaptively generates policies to publish individual-level COVID-19 data according to infection dynamics. We model the data publishing process as a two-player Stackelberg game between a data publisher and a data recipient and then search for the best strategy for the publisher. In this game, we consider 1) the average accuracy of predicting future case counts for all demographic groups, and 2) the mutual information between the original data and the released data. We use COVID-19 case data from Vanderbilt University Medical Center from March 2020 to December 2021 to demonstrate our model and evaluate its effectiveness. The experimental results show that our game theoretic model outperforms all baseline approaches, including those adopted by CDC, while maintaining low privacy risk.**

Index Terms—**COVID-19, game theory, pandemic data, case prediction, data sharing, privacy-preserving data publishing**

I. Introduction

The COVID-19 pandemic has highlighted the importance of publishing infectious disease data for surveillance and trend analyses [1]. Individual-level data publication is the release of characteristics of COVID-19 patients, such as age, gender, race, and geographic area, to the public. Sharing data in a timely manner can support a wide variety of public health research endeavors, such as modeling disease transmission and understanding the biological mechanisms behind infection [2]. Such data sharing can support investigations into disease spread and methods of disease prevention [3] as well as health disparities and equity in demographic groups [4]. In recognition of such benefits, various organizations have worked to broaden access to large epidemiological datasets, such as the U.S. Centers for Disease Control and Prevention's (CDC) US COVID-19 Case Surveillance datasets [5]. While advances in surveillance have led to rapid growth in the management and treatment of the disease over the past few years, public data sharing on a wide scale remains limited [6], partly due to privacy concerns. Many organizations have only published total case counts in a given US state or county rather than the details of patient-level features. In the United States, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) governs how organizations may publish individual-level health data, also known as patient-level data [7]. 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." However, transforming data into a de-identified form requires more than just removing identifiers like a person's name or address. Numerous demonstration attacks have shown that, with the right background knowledge, a data recipient can leverage seemingly de-identified information in the data, such as age, gender, race, and/or ZIP code, to re-identify the individuals to whom the data corresponds using population data [8].

The CDC uses generalization and suppression techniques [9] to mitigate risk, when publishing individual-level COVID-19 data, where information, such as age, gender, race, and/or ZIP code is generalized (i.e., published with a coarser granularity) or removed from a record [10]. However, the CDC's data publishing approaches do not take the dynamic nature of a pandemic into account, nor does it have much flexibility

This research was funded, in part, by the following grants: NIH grants RM1HG009034, R01HG006844, T15LM007450, and NSF grant CNS-2029651.

[#] Who contributed equally to this work
[*] To whom correspondence should be addressed

as data are either generalized or suppressed. When working with a dynamic pandemic dataset that would benefit from regular updates or revision, the policy would benefit from being flexible, which requires it to be determined according to the time of the data release and data itself. In 2022, Brown et al. [11] introduced an approach that can maximize privacy and utility by dynamically adjusting case reporting policy in near-real time when reporting individual-level pandemic data; however, they did not formally measure data utility in their framework. In a subsequent study [4], Brown et al. formally evaluated the data utility in dynamic de-identification policies to make it clear if the resulting data is expected to be helpful for the anticipated end users. In both studies [4], [11], Brown et al. modelled a data recipient that receives the data and utilizes it to re-identify individuals. This attempt of re-identification is labeled an "attack." In this adversarial model, they did not consider the rationality of the data recipient, where the data recipient would only attack the data if they expect the benefit they would receive from an attack would outweigh the costs of an attack. More specifically, they only consider the worst-case scenario in which the adversary in their model always attacks. This is problematic because in real scenario, an adversary will unlikely to attack if the cost of attack is too high. Thus, it can lead to an over-estimation of the privacy risk of the released dataset, which could lead to less data being shared.

In this paper, we introduce a game theoretic model that adaptively generates policies to publicly share de-identified individual-level pandemic data in a privacy-preserving manner. We model the data sharing process as a two-player Stackelberg game between a data publisher and a data recipient and then recommend an optimal policy for the data publisher on a consistent schedule. Additionally, the payoff functions in the game theoretic model formally models and integrates data utility and data privacy measures.

II. Related Work

A. Game Theory for Sharing Health Data with Privacy

Several privacy-preserving algorithms have been developed to support the dissemination of sensitive health data. Game theoretic models have been used in a variety of privacy problems related to health data [12], [13] and have also been applied to privacy-preserving health data publishing [14]–[16] recently. For health data, there is the privacy risk that an attacker could use available population data to identify an individual based on features like age, gender, race, and/or ZIP code [17]. But releasing these attributes has utility by promoting trend analyses and better healthcare surveillance. These models are based on principles from game theory to search for the optimal publishing policy with the best tradeoff between lowering privacy risk and publishing useful data from a set of varying publishing policies in a systematic manner in order to serve the data publisher. Each policy translates into a different amount of corresponding risk and utility.

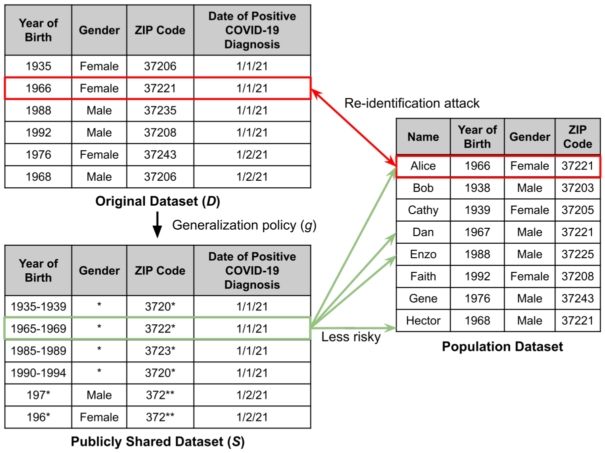
Similar to our model, these models [14]–[16] represent the data publishing process as a Stackelberg game, a game in which one player, the leader, makes an action to which the second player, the follower, makes another action as a result of the leader's action. Essentially, the publisher of the data is the leader and chooses a policy to publish data. Then, the recipient may decide whether to attack or not. Notably, not all recipients will attempt to try and compromise the privacy of people. The previous models assumed the existence of a "malicious" data recipient who would attempt to attack for expected benefit. We made the same assumption. In addition, we try to measure how "benign" data recipients may use the data to benefit society. Thus, instead of measuring data utility based solely on information entropy [14], we measure it using mutual information and the accuracy of predicting future case counts, representing one type of downstream uses of the data.

B. Publishing Individual-Level Pandemic Data with Privacy

There are several prior studies that provide guidance into privacy-preserving methods for publishing data from electronic health records [18], but there is limited research on privacy-preserving algorithms for publishing dynamic and time-critical datasets [19]. COVID-19 data ideally would be published on a consistent schedule, as a pandemic is on-going and the data is dynamic. Moreover, COVID-19 data should be published as early as possible to ensure rapid usage.

In 2021, Lee et al. described the CDC's approach to publish two versions of individual-level COVID-19 data [10]. Their approach is based on the k-anonymity privacy model [17]. k-anonymity is a property of dataset that requires each record match other $k-1$ records in the dataset on a set of quasi-identifying (QID) features such as age, gender, and race [17]. The approach works as follows: when there are less than k individuals who all have the same features, one of the features is generalized or suppressed [17] so that there are no longer less than k individuals with the same features. The publicly-accessible version includes fewer features and was generalized or suppressed to satisfy 11-anonymity . The restricted version includes more features and satisfies 5-anonymity. Our game theoretic model also uses generalization and/or suppression, but our model does not guarantee that the released dataset satisfies k-anonymity.

In 2022, Brown et al. introduced a framework to dynamically adapt de-identification polices for near-real time sharing of patient-level surveillance data via simulation [11]. It was shown that dynamically changing case-reporting policy leads to more data dissemination while maintaining low privacy risk. In a subsequent study, Brown et al. evaluated the data utility in dynamic de-identification policies in terms of the effectiveness and fairness of outbreak detection among demographic groups [4]. Similarly, in this work, our model can generate dynamic policies to maximize data utility and privacy. By contrast, we measured the data utility quantitatively, considered a rational adversarial model, and used a real-world individual-level COVID-19 dataset in evaluation experiments. When evaluating the data utility, we considered both the general data quality based on mutual information and the accuracy of the prediction of case counts among demographic groups.



Year of Birth	Gender	ZIP Code	Date of Positive COVID-19 Diagnosis
1935	Female	37206	1/1/21
1966	Female	37221	1/1/21
1988	Male	37235	1/1/21
1992	Male	37208	1/1/21
1976	Female	37243	1/2/21
1968	Male	37206	1/2/21

Original Dataset (*D*)

Year of Birth	Gender	ZIP Code	Date of Positive COVID-19 Diagnosis
1935-1939	*	3720*	1/1/21
1965-1969	*	3722*	1/1/21
1985-1989	*	3723*	1/1/21
1990-1994	*	3720*	1/1/21
197*	Male	372**	1/2/21
196*	Female	372**	1/2/21

Publicly Shared Dataset (*S*)

Name	Year of Birth	Gender	ZIP Code
Alice	1966	Female	37221
Bob	1938	Male	37203
Cathy	1939	Female	37205
Dan	1967	Male	37221
Enzo	1988	Male	37225
Faith	1992	Female	37208
Gene	1976	Male	37243
Hector	1968	Male	37221

Population Dataset

Fig. 1. **An example illustrating data structures in the publishing and the attacking scenarios.** The original COVID-19 case dataset is a list of records that each contains the quasi-identifier values (year of birth, gender, and ZIP code) and the date of positive diagnosis. The published dataset on 1/1/21 contains generalized values with 5-year bins, gender values removed, and exact ZIP codes. Because the values are generalized, there is less risk of re-identification compared to the original dataset. This example shows how different generalization policies can be applied to different days.

III. Methods

When deciding how to publish COVID-19 case data, it is important to consider how this data can be utilized. While publishing individual-level COVID-19 case data has several benefits to most data recipients, such as understanding the nature of the pandemic and the future of the ongoing virus, there may also be malevolent data recipients that seek to leverage external datasets to re-identify de-identified COVID-19 case data. Fig. 1 shows how published de-identified data can be linked to an exact identified individual in an external population dataset based on shared feature values. In this example, we can see how if the original de-identified dataset D were to be published, a data recipient could re-identify the female born in 1966 in ZIP code 37221 as Alice in the external population dataset. However, by modifying the publicly shared dataset S by binning the feature values into broader ranges, there is less risk of accurately linking the record to a specific individual; there are more individuals with matching features, thereby lowering the chance of an attacker successfully re-identifying any individual. This method of coarsening the data into broader concepts is known as generalization [9]. A policy that publishes data via generalization is called a generalization policy. In this work, a generalization policy, g, applies rules of binning the data to a set of data records.

We assume the data publisher only chooses amongst generalization policies for simplicity of comparison between the different policies and approaches. In this work, a record means a row of feature values that corresponds to a patient. Each record has a set of quasi-identifiers (QID), which are identifiers that are not explicitly identifiable like a patient's name or address but can be combined to identify an individual. Each policy has a unique level of generalization for each QID features and is represented by a vector according to the level of generalization in the order of QID features. For instance, if the three QID features are (1) year of birth, (2) gender, and (3) ZIP code, whose generalization hierachies are shown in Fig. 2, then the generalization policy $\langle 1,1,0 \rangle$ means 5-year age bins, no specific gender, and original ZIP codes. For each day of data release, we select the policy based with the highest payoff for the data publisher among the total set of generalization policies. To calculate the payoff, we first calculate the risk and utility of the data published according to the generalization policy.

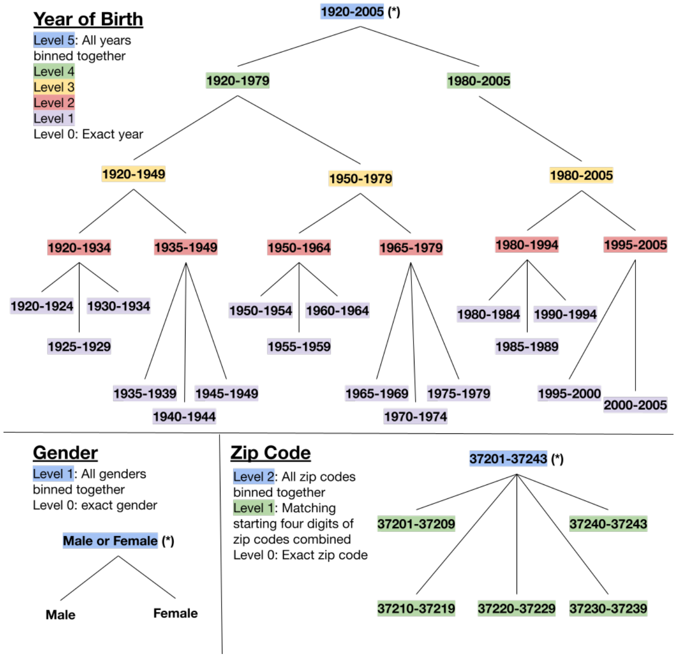


Fig. 2. **The generalization hierarchies for three features used in this paper: (1) year of birth, (2) gender, and (3) ZIP code.** This hierarchical representation shows the specific bin sizes for each of the features and how the data can be more generalized as the bins become larger. A generalization policy is represented as a unique level of generalization for each feature.

We use a measure of risk called marketer risk [20] that has been used in various privacy polices [11], [14], [16]. The marketer risk r for policy g is calculated as the probability of correctly linking each record in the dataset D to the corresponding identified record in the population and taking the average across all records. The i^{th} record of the dataset D is represented as d_i. The total number of records in the dataset D is n. Thus, the privacy risk of a dataset, given a generalization policy, can be represented as follows:

$$r(D,g) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\phi(d_i,g)} \tag{1}$$

where $\phi(d_i, g)$ is the number of individuals in the population that match record d_i when record d_i is generalized according to generalization policy g.

The utility u for a policy will be calculated is the weighted sum of two different measures. The first is mutual information, which is a measure of the similarity between the original dataset and the generalized dataset. This is similar to an established measure based on the loss of information entropy used by Wan et al. in their model [14]. Intuitively, as the original dataset and the generalized dataset become more dissimilar, the generalized dataset will be less useful. To compute the mutual information, we generate a list of numbers from each of the two datasets. The original dataset has the number of cases for each combination of feature values. The generalized dataset also has the number of cases for each combination of feature values, which can be generalized values. Then, we compare the values of counts between the original dataset and the generalized dataset to compute the mutual information:

$$u(D, g) = I(d, S(D, g)) \tag{2}$$

where u is the utility function, I is the mutual information function, and $S(D, g)$ is the publicly shared dataset according to generalization policy g given the initial dataset D.

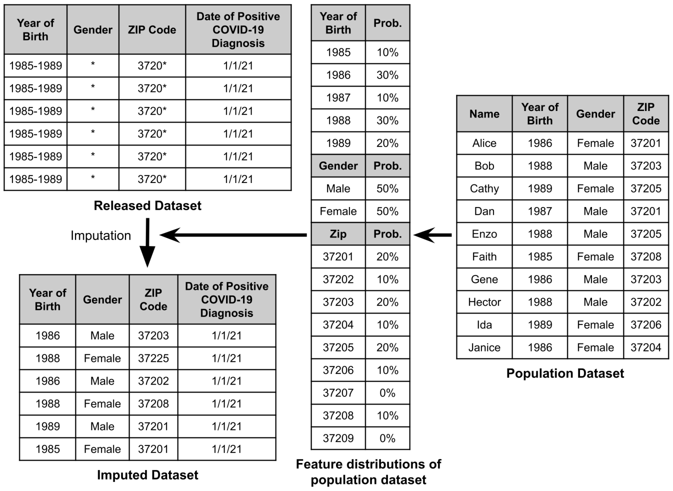
$$I(D_0, D_1) = H(D_0) + H(D_1) - H(D_0, D_1) \tag{3}$$

where H is the information entropy function.

For our second metric for data utility, we consider common uses of COVID-19 data in the public health sphere and the motivations of publishing COVID-19 data. Researchers have found that significant life-saving response efforts were taken due to epidemiological modeling based on open data, underscoring the importance of understanding trends in data and patient-level information to predict future outbreaks [21] [22]. Overall, the use of data-sharing of outbreak data is primarily analyzing current trends among the patient population and utilizing this information to forecast future trends. Woolhouse et al. state that the most critical component of outbreak management is the dissemination of data in a manner that allows for modeling future outbreak behavior [3].

The second utility measure corresponds to the accuracy of the COVID-19 case count predictions for the next day of data release. Given that one of the primary uses of COVID-19 data, or any outbreak data, is to forecast and predict case count distributions, we use accuracy of predicting case counts as a measure of data utility in our model. Accuracy of case count prediction is a metric that will allow for data-dependent selection of generalization policies that accounts for the dynamic nature of COVID-19 data publication. First, we generalize the data according to a generalization policy g for a given day. Then, we de-generalize the values in a sense by imputing values according to the distributions of values for each QID in the population dataset (see Fig.3). After imputing values for each record and calculating the counts for each demographic group for the given day, we use these counts along with the previous t_p days' counts to develop a sequence of time series for each demographic group and then use the Random Forest Regressor in Scikit-learn, a machine learning package in Python, to predict the counts for the next day. We utilize the Random Forest Regressor due to the sparsity of features, as in the low number of QID, when predicting future case counts. To calculate the utility, we find the average relative error of the case count predictions for the next day based on the selection of the generalization policy g across all demographic groups M in the dataset D. For the i^{th} demographic group m_i in M, we compare the predicted count $n_p(i)$ and actual count $n_a(i)$. We further denote the number of total demographic groups as n_M. Thus, the utility of a dataset, given a generalization policy, can be represented as follows:

$$u(D, g) = 1 - \frac{1}{n_M} \sum_{i=1}^{n_M} \frac{n_a - n_p}{n_p} \tag{4}$$



Year of Birth	Gender	ZIP Code	Date of Positive COVID-19 Diagnosis
1985-1989	*	3720*	1/1/21
1985-1989	*	3720*	1/1/21
1985-1989	*	3720*	1/1/21
1985-1989	*	3720*	1/1/21
1985-1989	*	3720*	1/1/21
1985-1989	*	3720*	1/1/21

Year of Birth	Gender	ZIP Code	Date of Positive COVID-19 Diagnosis
1986	Male	37203	1/1/21
1988	Female	37225	1/1/21
1986	Male	37202	1/1/21
1988	Female	37208	1/1/21
1989	Male	37201	1/1/21
1985	Female	37201	1/1/21

Year of Birth	Prob.
1985	10%
1986	30%
1987	10%
1988	30%
1989	20%
Gender	**Prob.**
Male	50%
Female	50%
Zip	**Prob.**
37201	20%
37202	10%
37203	20%
37204	10%
37205	20%
37206	10%
37207	0%
37208	10%
37209	0%

Name	Year of Birth	Gender	ZIP Code
Alice	1986	Female	37201
Bob	1988	Male	37203
Cathy	1989	Female	37205
Dan	1987	Male	37201
Enzo	1988	Male	37205
Faith	1985	Female	37208
Gene	1986	Male	37203
Hector	1988	Male	37202
Ida	1989	Female	37206
Janice	1986	Female	37204

Fig. 3. **An example illustrating the imputation process.** To predict counts for each demographic group from the shared dataset, we must work with generalized values. To do so, we impute values from the generalized records by imputing according to the proportion of each QID value in the external population dataset.

We then simulate a Stackelberg game. First, the publisher, who is the leader, publishes the data according to a generalization policy and then the data recipient, the follower, decides whether to attack the data (i.e., re-identify the published records). Here, we assume that the publisher knows the adversary has access to a population dataset with identifying features (e.g., name) and a set of QID features (e.g., age, gender and ZIP code). To facilitate the publisher to update a pandemic dataset on a daily basis, the Stackelberg game need to be modelled and solved for each day of data release.

We use the payoff measure to assess each of the data publisher's strategy, which is a generalization policy. The publisher's payoff is a function of both players' strategy sets and the dataset to be generalized. It is dependent upon several key models including the adversarial model, the privacy model and the utility model. In this case, the publisher's payoff P is the publisher's expected gain G from publishing the dataset D in the manner dictated by the generalization policy g minus the expected loss L due to any successful privacy attacks.

$$P(D, g) = E[G(D, g)] - E[L(D, g)] \tag{5}$$

The expected gain G is calculated by multiplying the utility u of publishing the data by the constant parameter β that represents the benefit that the publisher receives by sharing the record in its original form. Clearly, the expected gain of publishing data increases when the data has higher utility.

$$E[G(D,g)] = \beta * u(D,g) \tag{6}$$

The expected loss L is calculated by the decision of whether the attacker decides to attack $\mathbb{1}_{attack}$ and the product of the risk r by the constant parameter λ which represents the publisher's loss for one record due to a successful re-identification. We model the data recipient as a rational attacker, such that they only attempt re-identification if his or her associated benefits exceed the costs. The benefits for the attacker are posited to be equal to the loss that the publisher would face (e.g., any fines levied for the privacy breach by federal regulators).

$$E[L(D,g)] = \mathbb{1}_{attack}(D,g) * \lambda * r(D,g) \tag{7}$$

where

$$\mathbb{1}_{attack}(D,g) = \begin{cases} 1, & E[L] > C \\ 0, & E[L] \leq C \end{cases} \tag{8}$$

given that C is a parameter that represents the adversary's cost to launch a re-identification attack towards one record.

Thus, the total payoff is calculated as

$$P(D,g) = \beta * u(D,g) - \mathbb{1}_{attack}(D,g) * \lambda * r(D,g) \tag{9}$$

IV. Results

A. Experimental Settings

1) Datasets

We use two datasets in our experiments. The Vanderbilt University Medical Center (VUMC) COVID-19 dataset is derived from a VUMC dataset contains all patient cases that were treated at VUMC for COVID-19 from March 11, 2020 to December 19, 2021. It is composed of 9,632 records with the following four features: year of birth, gender, 5-digit ZIP code, and the date of positive diagnosis. The Nashville Voter Registration dataset is derived from the latest Davidson County Voter Registration list [23] in July 2021. It contains 337,681 records with the following four features: name, year of birth, gender, and 5-digit ZIP code. The Nashville Voter Registration dataset serves as a population dataset and is used to calculate the marketer risk for patients in the VUMC dataset. For this investigation, we focus on records with a ZIP code in the city of Nashville for both datasets.

2) Parameter Settings

We compare our *Game-theoretic* approach with three other data publishing approaches as baselines. The first is the *No-protection* approach, where each record is published with its original values at the end the corresponding date of diagnosis.

The second baseline is the *CDC-based* approach based on a k-anonymity privacy model [10] which has been adopted by CDC to release individual-level COVID-19 case data. CDC updates data on a monthly basis [10]. Both the CDC-based approach in our experiments and one of the actual CDC approaches achieve 5-anonymity. However, the ways they achieve 5-anonymity are different. We directely used the k-anonymization algorithm from a off-the-shelf anonymization software ARX [24], [25] instead of implementing the k-anonymization algorithm used by Lee et al. [10].

The third baseline is the *Dynamic* approach, which is based on the approach used by Brown et al. [11] to dynamically adjust data generalization policy considering the marketer risk. In their approach, all records in the released dataset have the same generalization policy at each releasing time (daily or weekly). We changed this setting by allowing each record to have its own generalization policy. This change should improve the performance of the *Dynamic* approach in terms of the average payoff of patients. We set the acceptable threshold for marketer risk as 0.01 as it was set in Brown et al.'s work [11]. Note that, for all approaches, the generalization policy for a record will not change once the record is released.

For the game theoretic model, we set β to \$10, λ to \$500 and C to \$1. These parameters are set according to a practical scenario as described in the following. The CDC received 27.77 billion dollars in total to prevent, prepare, and respond to COVID-19 by 2021. We assume 5% of the budget is assigned to decision makers in CDC for sharing the individual-level data to the public or third-parties. The population of the US is 308,745,538 in 2020 according to the US census [26]. On May 2, 2022, the highest positive rate of COVID-19 tests in a state among all states in the US was 46.3%. We assume that 46.3% of the US population will have the experience to be tested positive at a certain point of time when the CDC stops updating the dataset. We set β as the estimate of the average budget for each record that will be included in the dataset.

$$\beta = \$27.77 \times 10^9 \times 0.05/308,745,538/0.463 \approx \$10 \tag{10}$$

We assume the loss of the data publisher for each re-identified record, λ, is proportional to the fine paid to a federal regulator for a data breach as reported on the Office for Civil Rights' Wall of Shame [14], [27]. As such, we set $\lambda = \$500$ according to the average fine per record. We set the cost of the adversary, C, for accessing each identified record in the population dataset to \$1 (based on the discounted price for a report from www.intelius.com).

We used an off-the-shelf implementation of the Random Forrest algorithm to simulate the downstream prediction task to calculate the predictive utility. In addition, we set 21 days as the minimal number of historical data points that are required as the input of the prediction algorithm.

In our experiments, based on the available QID features in the datasets, year of birth, gender, and ZIP code, we consider a fixed set of 36 generalization policies that are represented as vectors according to generalization hierarchies as shown in Fig. 2, where we can see those three features have 6, 2, and 3 levels, respectively.

3) Evaluation Measures

In our experiments, we use the average of the daily average payoff over 21 months as the main effectiveness measure (see Eq. 9). In addition, we also calculate the average of the daily average utility (or privacy) of records over 21 months (see

Eqs. 2 and 4). The privacy of a released record is defined as 1 minus the marketer risk [20] of that record (see Eq. 1). The utility of all records released in one day is defined as the number of records times a utility function based on predictive utility and/or mutual information (see Eqs. 2 and 4).

B. Experiments with Mutual Information Only

In this set of experiments, we set the utility function as the mutual information between original data and released data.

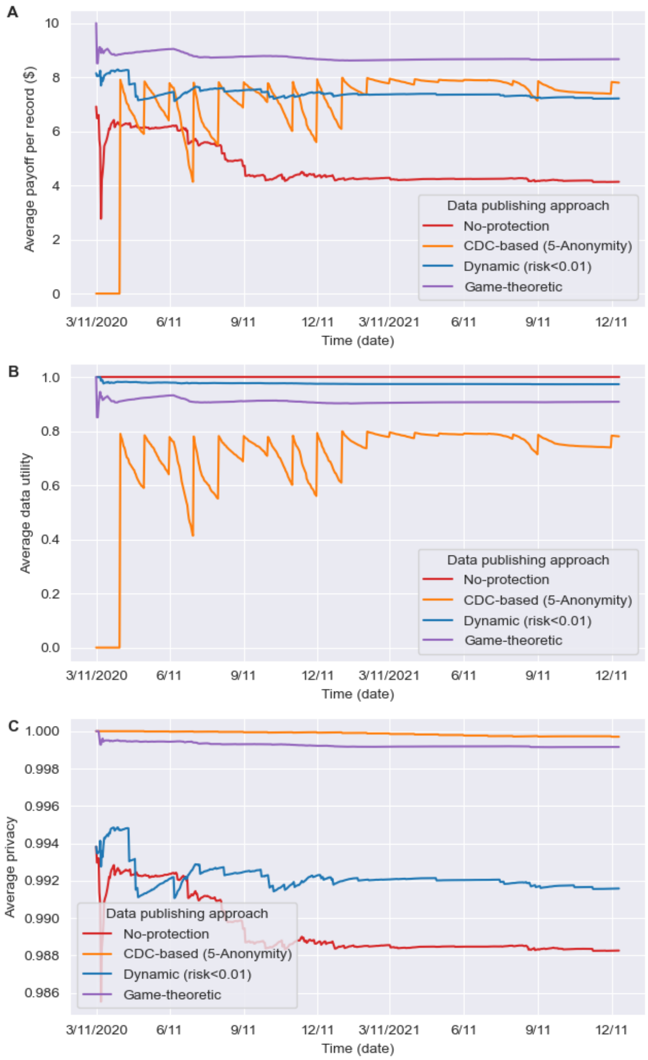
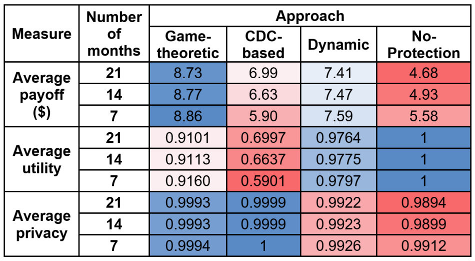


Fig. 4. **The effectiveness measures of policies on each day considering only mutual information in the utility measure.** (A) Average payoff of the data publisher. (B) Average data utility. (C) Average privacy.

The payoff of the data publisher averaged across all records at the end of each day is shown in Fig. 4(A). It is evident that the *Game-theoretic* approach achieves the best average payoff for the data publisher on each day. The *CDC-based* approach outperforms the other two baselines on most days; however, the average payoff the *CDC-based* approach brings to the data publisher is unstable due to the low frequency of data publishing. The *Dynamic* approach outperforms the *No-protection* approach on each day and outperforms the *CDC-based* approach in limited days. For the entire published dataset, the *Game-theoretic* approach brings 11.2% (109.7%) more average payoff to the data publisher in comparison to the *CDC-based* (*No-protection*) approach.

The data utility averaged across all existing records at the end of each day is shown in Fig. 4(B). It is evident that the *Game-theoretic* approach outperforms the *CDC-based* approach on each day in terms of the average data utility per record. The *Dynamic* approach outperforms the *Game-theoretic* approach on each day. For the entire published dataset, the *Game-theoretic* approach leads to 16.4% more (9.1% less) average data utility in comparison to the *CDC-based* (*No-protection*) approach.

The privacy averaged across all existing records at the end of each day is shown in Fig. 4(C). It is evident that the *Game-theoretic* approach outperforms the *Dynamic* and *No-protection* approaches on each day in terms of the average data privacy per record. The *CDC-based* approach outperforms the *Game-theoretic* approach on each day. For the entire published dataset, the *Game-theoretic* approach leads to 0.1% less (1.1% more) average data privacy in comparison to the *CDC-based* (*No-protection*) approach.



Measure	Number of months	Approach			
		Game-theoretic	CDC-based	Dynamic	No-Protection
Average payoff ($)	21	8.73	6.99	7.41	4.68
	14	8.77	6.63	7.47	4.93
	7	8.86	5.90	7.59	5.58
Average utility	21	0.9101	0.6997	0.9764	1
	14	0.9113	0.6637	0.9775	1
	7	0.9160	0.5901	0.9797	1
Average privacy	21	0.9993	0.9999	0.9922	0.9894
	14	0.9993	0.9999	0.9923	0.9899
	7	0.9994	1	0.9926	0.9912

Fig. 5. **The average effectiveness measures of policies over three long periods of time considering only mutual information in the utility measure.**

The average effectiveness measures of policies averaged across all days during a certain period of time (namely, the first 7, 14, or 21 months) considering only mutual information in the utility measure are shown in Fig. 5. It is evident that the *Game-theoretic* approach outperforms all other approaches in terms of the data publisher's average daily payoff. In addition, the *Game-theoretic* approach outperforms the *Dynamic* and *No-protection* approaches in terms of the average daily privacy and outperforms the *CDC-based* approach in terms of the average daily utility. The *Dynamic* approach outperforms the *CDC-based* approach over these three periods of time in terms of the data publisher's average daily payoff as well as the average daily data utility. As the period of time in comparison increases from 7 months to 21 months, the data publisher's average daily payoff decreases by using the *Game-theoretic* approach or the *Dynamic* approach. Over the period of the 21 months presented in our dataset, the *Game-theoretic* approach brings 24.9%, 17.8%, and 86.5% more average daily payoff to the data publisher in comparison to the *CDC-based* approach, *Dynamic* approach, *No-protection* approach, respectively.

C. Experiments with Predictive Utility & Mutual Information

In this set of experiments, we set the utility function to the average of 1) the mutual information between the original data and the released data, and 2) the predictive utility which is defined as the accuracy of the COVID-19 case count predictions, given existing data.

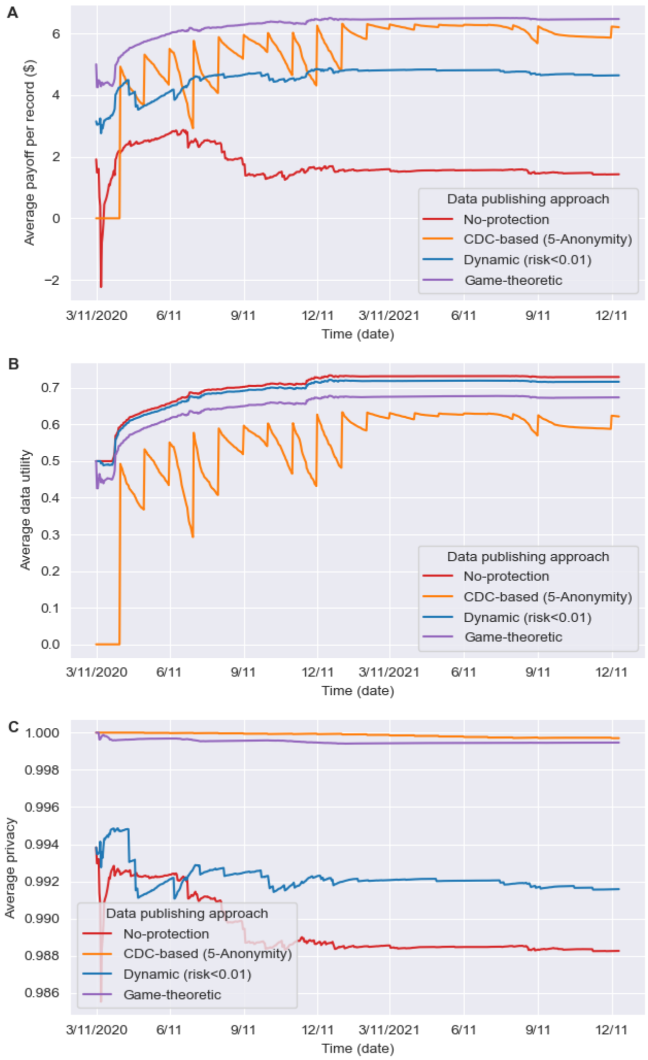
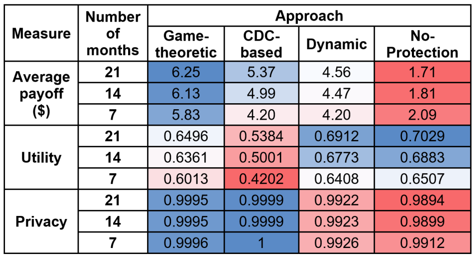


Fig. 6. **The effectiveness measures of policies on each day considering both predictive utility and mutual information in utility measure.** (A) Average payoff of data publisher. (B) Average data utility. (C) Average privacy.

The average effectiveness measures of policies averaged across all existing records at the end of each day with this new utility function is shown in Fig. 6. The main observations remain the same with this utility function. For the entire published dataset, the *Game-theoretic* approach brings 4.4% (352.9%) more average payoff to the data publisher in comparison to the *CDC-based* (*No-protection*) approach, leads to 8.4% more (7.7% less) average data utility in comparison to the *CDC-based* (*No-protection*) approach, and leads to 0.02% less (1.1% more) average data privacy in comparison to the *CDC-based* (*No-protection*) approach.



Measure	Number of months	Approach			
		Game-theoretic	CDC-based	Dynamic	No-Protection
Average payoff ($)	21	6.25	5.37	4.56	1.71
	14	6.13	4.99	4.47	1.81
	7	5.83	4.20	4.20	2.09
Utility	21	0.6496	0.5384	0.6912	0.7029
	14	0.6361	0.5001	0.6773	0.6883
	7	0.6013	0.4202	0.6408	0.6507
Privacy	21	0.9995	0.9999	0.9922	0.9894
	14	0.9995	0.9999	0.9923	0.9899
	7	0.9996	1	0.9926	0.9912

Fig. 7. **The average effectiveness measures of policies over three long periods of time considering both predictive utility and mutual information in the utility measure.**

The average effectiveness measures of policies averaged across all days during a certain period of time (namely, the first 7, 14, or 21 months) considering both predictive utility and mutual information in the utility measure are shown in Fig. 7. It is evident that the *Game-theoretic* approach outperforms all other approaches in terms of the data publisher's average daily payoff. In addition, the *Game-theoretic* approach outperforms the *Dynamic* and *No-protection* approaches in terms of the average daily privacy and outperforms the *CDC-based* approach in terms of the average daily utility. The *CDC-based* approach outperforms the *Dynamic* approach in terms of the data publisher's average daily payoff and vice versa in terms of the average daily data utility. As the period of time in comparison increase from 7 months to 21 months, the data publisher's average daily payoff increases by using all approaches except the *No-protection* approach. Over the period of the 21 months presented in our dataset, the *Game-theoretic* approach brings 16.4%, 37.1%, and 265.5% more average daily payoff to the data publisher in comparison to the *CDC-based* approach, *Dynamic* approach, *No-protection* approach, respectively.

V. Discussion and Conclusion

Our game theoretic model presents several notable takeaways. First, it is effective to use a game theoretic model to optimize the tradeoff between privacy and utility when publishing a pandemic dataset that needs to be updated on a regular basis. Second, the policies recommended by our game theoretic model tend to share much more data than those from the CDC-based approach while maintaining low privacy risk. The CDC-based approach, by contrast, is a more conservative approach in our adversarial setting, where the adversary has limited knowledge and makes rational attacking decisions. Third, by considering an essential downstream analysis task using published data (i.e., case count prediction using a machine learning model), we demonstrate how a game theoretic model can be tailored to a dynamic privacy-preserving data sharing problem. Fourth, although health equity has not been explicitly measured in our model, our model explicitly considers the data utility across subpopulations such that the optimal policy out of modeling can be regarded as ethical and fair.

Limitations exist in our model, which provide directions for future improvement. First, this study focuses on an adversarial setting including one adversary and one type of privacy attack. A natural extension of the work is to consider the possibility of multiple adversaries and multiple types of adversarial models. Second, our model requires that all records released on the same day have the same generalization policy, reducing the flexibility of choosing generalization policies. Removing this constraint and adopting more efficient algorithms (such as genetic algorithms) to explore a larger strategy space can further improve the effectiveness of the game theoretic model. Third, the prediction model for the case counts is based on the Random Forest algorithm, which might be suboptimal in prediction performance. Fourth, we remove demographic features such as race and ethnicity from the set of quasi-identifying features in our datasets because the Davidson county voter registration list has a high missing rate for values of these two features. This setting makes CDC-based approach appears conservative in protecting privacy. In other words, the scenario we considered might be less risky than the real-world scenario. This can be resolved by testing on more datasets with more available features and no or fewer missing values. Finally, we obtained the population dataset from only one source, which may not contain all the records in the patient dataset. In the future, we should seek population data from other sources (e.g., US census) as well for better evaluation.

VI. Acknowledgment

Z.W. and B.A.M. conceived and designed the study. A.G., Z.W., and B.A.M. created models and developed the methodology. C.Y. and Z.W. prepared data. A.G. and Z.W. performed the experiments. J.T.B. provisioned materials/analysis tools. A.G., Z.W., and B.A.M. analyzed and presented the results. A.G. and Z.W. wrote the original draft. All authors wrote the final manuscript. The first two authors contributed equally.

All source code and all relevant results can be found at https://github.com/zhywan/covid19-privacy-game.

References

[1] N. K. Ibrahim, “Epidemiologic surveillance for controlling covid-19 pandemic: types, challenges and implications,” *J. Infect. Public Health*, vol. 13, no. 11, pp. 1630–1638, 2020.

[2] S. Bansal, G. Chowell, L. Simonsen, A. Vespignani, and C. Viboud, “Big Data for Infectious Disease Surveillance and Modeling,” *J. Infect. Dis.*, vol. 214, no. suppl_4, pp. S375–S379, 2016.

[3] M. E. J. Woolhouse, A. Rambaut, and P. Kellam, “Lessons from ebola: Improving infectious disease surveillance to inform outbreak management,” *Sci. Transl. Med.*, vol. 7, no. 307, pp. 307rv5–307rv5, 2015.

[4] T. Brown, Z. Wan, A. Gkoulalas-Divanis, M. Kantarcioglu, and B. Malin, “Supporting covid-19 disparity investigations with dynamically adjusting case reporting policies,” in *Proceedings of the 2022 American Medical Informatics Association Annual Fall Symposium (AMIA)*, 2022, p. in press.

[5] CDC, “Covid-19 case surveillance public use data with geography.” [Online]. Available: https://data.cdc.gov/CaseSurveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w

[6] L. Gardner, J. Ratcliff, E. Dong, and A. Katz, “A need for open public data standards and sharing in light of covid-19,” *Lancet Infect. Dis.*, vol. 21, no. 4, p. e80, 2021.

[7] Health and Human Services Department, “Standards for privacy of individually identifiable health information,” 2000. [Online]. Available: https://www.federalregister.gov/documents/2000/12/28/00-32678/standards-forprivacy-of-individually-identifiable-health-information

[8] P. Golle, “Revisiting the uniqueness of simple demographics in the us population,” in *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, ser. WPES ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 77–80.

[9] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 571–588, 2002.

[10] B. Lee, B. Dupervil, N. P. Deputy, W. Duck, S. Soroka, L. Bottichio, B. Silk, J. Price, P. Sweeney, J. Fuld, J. T. Weber, and D. Pollock, “Protecting privacy and transforming covid-19 case surveillance datasets for public use,” *Public Health Reports*, vol. 136, no. 5, pp. 554–561, 2021, pMID: 34139910.

[11] J. T. Brown, C. Yan, W. Xia, Z. Yin, Z. Wan, A. Gkoulalas-Divanis, M. Kantarcioglu, and B. A. Malin, “Dynamically adjusting case reporting policy to maximize privacy and public health utility in the face of a pandemic,” *J. Am. Med. Inform. Assoc.*, vol. 29, no. 5, pp. 853–863, 2022.

[12] M. Li, D. Carrell, J. Aberdeen, L. Hirschman, J. Kirby, B. Li, Y. Vorobeychik, and B. A. Malin, “Optimizing annotation resources for natural language de-identification via a game theoretic framework,” *J. Biomed. Inform.*, vol. 61, pp. 97–109, 2016.

[13] C. Yan, B. Li, Y. Vorobeychik, A. Laszka, D. Fabbri, and B. Malin, “Get your workload in order: Game theoretic prioritization of database auditing,” in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 1304–1307.

[14] Z. Wan, Y. Vorobeychik, W. Xia, E. W. Clayton, M. Kantarcioglu, R. Ganta, R. Heatherly, and B. A. Malin, “A game theoretic framework for analyzing re-identification risk,” *PLOS ONE*, vol. 10, no. 3, pp. 1–24, 2015.

[15] Z. Wan, Y. Vorobeychik, W. Xia, E. W. Clayton, M. Kantarcioglu, and B. Malin, “Expanding access to large-scale genomic data while promoting privacy: a game theoretic approach,” *Am. J. Hum. Genet.*, vol. 100, no. 2, pp. 316–322, 2017.

[16] Z. Wan, Y. Vorobeychik, W. Xia, Y. Liu, M. Wooders, J. Guo, Z. Yin, E. W. Clayton, M. Kantarcioglu, and B. A. Malin, “Using game theory to thwart multistage privacy intrusions when sharing data,” *Sci. Adv.*, vol. 7, no. 50, p. eabe9986, 2021.

[17] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[18] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, “Publishing data from electronic health records while preserving privacy: A survey of algorithms,” *J. Biomed. Inform.*, vol. 50, pp. 4–19, 2014, special Issue on Informatics Methods in Medical Privacy.

[19] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, “Anonymization of longitudinal electronic medical records,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 413–423, 2012.

[20] F. K. Dankar and K. El Emam, “A method for evaluating marketer re-identification risk,” in *Proceedings of the 2010 EDBT/ICDT Workshops*, ser. EDBT ’10. New York, NY, USA: Association for Computing Machinery, 2010.

[21] J.-P. Chretien, C. M. Rivers, and M. A. Johansson, “Make data sharing routine to prepare for public health emergencies,” *PLoS Med.*, vol. 13, no. 8, p. e1002109, 2016.

[22] C. Rivers, J.-P. Chretien, S. Riley, J. A. Pavlin, A. Woodward, D. Brett-Major, I. Maljkovic Berry, L. Morton, R. G. Jarman, M. Biggerstaff *et al.*, “Using “outbreak science” to strengthen the use of models during epidemics,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–3, 2019.

[23] Nashville.org, “Order a voter list.” [Online]. Available: https://www.nashville.gov/departments/elections/candidates/order-voter-list

[24] F. Prasser, J. Gaupp, Z. Wan, W. Xia, Y. Vorobeychik, M. Kantarcioglu, K. Kuhn, and B. Malin, “An open source tool for game theoretic health data de-identification,” in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 1430.

[25] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, “Flexible data anonymization using arx—current status and challenges ahead,” *Software: Practice and Experience*, vol. 50, no. 7, pp. 1277–1304, 2020.

[26] US Census Bureau, “Decennial census of population and housing datasets.” [Online]. Available: https://www.census.gov/programs-surveys/decennial-census/data/datasets.2020.List_327707051.html

[27] Office for Civil Rights, “Breach notification rule.” [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html