# How Adversarial Assumptions Influence Re-identification Risk Measures: A COVID-19 Case Study

Xinmeng Zhang[1(✉)], Zhiyu Wan[2], Chao Yan[2], J. Thomas Brown[2], Weiyi Xia[2], Aris Gkoulalas-Divanis[3], Murat Kantarcioglu[4], and Bradley Malin[1,2,5]

[1] Department of Computer Science, Vanderbilt University, Nashville, TN, USA
Xinmeng.zhang@vanderbilt.edu
[2] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[3] IBM Watson Health, Cambridge, MA, USA
[4] Department of Computer Science, University of Texas at Dallas, Dallas, TX, USA
[5] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

**Abstract.** The COVID-19 pandemic highlights the need for broad dissemination of case surveillance data. Local and global public health agencies have initiated efforts to do so, but there remains limited data available, due in part to concerns over privacy. As a result, current COVID-19 case surveillance data sharing policies are based on strong adversarial assumptions, such as the expectation that an attacker can readily re-identify individuals based on their distinguishability in a dataset. There are various re-identification risk measures to account for adversarial capabilities; however, the current array insufficiently accounts for real world data challenges - particularly issues of missing records in resources of identifiable records that adversaries may rely upon to execute attacks (e.g., 10 50-year-old male in the de-identified dataset vs. 5 50-year-old male in the identified dataset). In this paper, we introduce several approaches to amend such risk measures and assess re-identification risk in light of how an attacker's capabilities relate to missing records. We demonstrate the potential for these measures through a record linkage attack using COVID-19 case surveillance data and voter registration records in the state of Florida. Our findings demonstrate that adversarial assumptions, as realized in a risk measure, can dramatically affect re-identification risk estimation. Notably, we show that the re-identification risk is likely to be substantially smaller than the typical risk thresholds, which suggests that more detailed data could be shared publicly than is currently the case.

**Keywords:** Data sharing · Re-identification risk · COVID-19 · Health data · Data privacy

## 1 Introduction

The Coronavirus Disease 2019 (COVID-19) outbreak caused a global pandemic that has resulted in devastating and sustained health and economic crisis [1]. As of May 2022,


J. Domingo-Ferrer and M. Laurent (Eds.): PSD 2022, LNCS 13463, pp. 361–374, 2022.
https://doi.org/10.1007/978-3-031-13945-1_25

there have been over 80 million confirmed cases (i.e., a person with laboratory confirmation of COVID-19 infection) in the United States and over 500 million worldwide. Though expected to become endemic at some point, COVID-19 continues to be a public health problem with waves of infection that are likely to reoccur for some time [2]. In this respect, it provides a clear justification for the creation of more timely case reporting strategies and surveillance efforts.

Public health departments typically rely on a case surveillance process to routinely collect information that is critical for disease control and prevention [3]. Case surveillance reports contain data on various infected individuals, including demographics, symptoms, epidemiologic characteristics (e.g., case confirmed date and location), health conditions, characteristics of hospitalizations, clinical outcomes, and exposure history. When surveillance data is made accessible at the population scale, it can enable faster responses to health emergencies and support data-driven public health research [4, 5]. Over the past several years, several resources of COVID-19 case surveillance data have been made available for public use. For instance, the World Health Organization (WHO) requests all member states to report data at a fidelity no less than national-level aggregated counts of confirmed cases, deaths, and hospitalizations within 48 h of detection [6]. In the United States, the Centers for Disease Control and Prevention (CDC) reports aggregate case and death counts, as well as person-level data that includes age, race, ethnicity, state, and county of residence of those infected [7, 8].

Despite the need to share COVID-19 case surveillance data, concerns about privacy have been raised due to the sensitive nature of the information [9–11]. There are particular concerns that the identities of the corresponding individuals could be inadvertently exposed. In public datasets, typically referred to as anonymised or de-identified data, it is obvious that direct identifiers, such as personal names, national ID numbers, and detailed residential addresses must be removed. However, it is possible that indirect or, what is often referred to as, quasi-identifiers (QIDs) [12], such as the demographic data shared in the CDC's COVID-19 datasets, can indicate small groups of patients in a de-identified dataset, which creates an opportunity for re-identification [13].

It is anticipated that attackers will rely upon QIDs to attempt to match de-identified records to accessible identified datasets through record linkage mechanisms [14, 15]. Prior studies have measured re-identification risks for QIDs by considering the degree of distinguishability within the de-identified dataset [16, 17]. The notion of $k$-anonymity [13] leads to a typical risk threshold applied in this case, whereby a de-identified dataset is considered protected if, and only if, each combination of QIDs appears at least $k$ times in the dataset. Currently, the CDC relies on this notion of privacy and releases two datasets for COVID-19 case surveillance—one for public use and the other for scientific use—at a level of 11- and 5-anonymity, respectively [7, 18].

The CDC's data publication policies are based on strong adversarial assumptions. Measures of privacy that focus solely on the degree of distinguishability within the dataset to be shared (as $k$-anonymity does) assume that the recipient of the data is aware that a named individual of interest is in the sample. However, this is a worst-case scenario and weaker adversarial scenarios can be, and in many cases are, considered [19]. Specifically, distinguishability in the de-identified only creates a potential for intrusion. For a re-identification attack to be successful, the recipient of the data either needs to

know the identity of the corresponding individuals according to some prior experiences (i.e., background knowledge) or they need to demonstrate re-identification by linking the records to an external, identified dataset through QIDs [19, 20]. This is important to recognize because the estimation of risk in these situations could be quite lower than in the worst-case scenario. In recognition of this fact, alternative approaches estimate re-identification risks based on population uniqueness [21–23]. This perspective, realized in the *k*-map model [24] for instance, assumes the attacker only knows that the targeted individual in the sample was drawn from a broader population of individuals, such that uniqueness in the dataset is insufficient to claim re-identification success. This model is used when the data sharer has a reasonable expectation of the identified resources that will be leveraged for an attack.

The aforementioned risk measures assume that all individuals below a threshold are equally at risk; by contrast, the marketer risk measure assumes a record's risk is inversely proportional to the number of records it relates to [25]. This risk measure typically assumes that the de-identified dataset is a subset (or a sample) of the identified dataset. However, in reality, both the de-identified and the identified datasets are samples from a broader population, and they do not necessarily demonstrate a sub-/super-set relationship. As a consequence, and as we show in this paper, there can be combinations of QIDs in the de-identified dataset that do not exist in the identified dataset. Similarly, the number of people with a certain combination of QIDs in the de-identified data could be larger than that observed in the identified dataset. For example, imagine that there are 10 patients in the de-identified dataset who are male and 50 years old, but that there are only 5 individuals present in the identified dataset who exhibit the same combination of QIDs. This raises a question about how missing records should be handled in the risk calculation. To the best of our knowledge, current re-identification risk measures do not explicitly address such real world challenges.

Our study introduces novel re-identification risk measures to fill in the gap between the previously proposed risk estimation methods and challenges caused by missing records. Our study extends traditional risk measures to address missing record challenges and allow data sharers to evaluate re-identification risk under various assumptions of an attacker's capability. To demonstrate how different assumptions could affect the estimation of risks, we perform a re-identification risk analysis for case surveillance data of COVID-19 and voter registration records in the United States. Our findings indicate that the re-identification risks vary according to adversarial assumptions. Using an actual record linkage test, we show that the external re-identification risk is likely to be substantially smaller than 0.09, which corresponds to the CDC's intended threshold of 11-anonymity. Our findings suggest that more detailed data could be shared publicly than the current generalization level.

## 2 Methods

In this paper, the *internal* dataset refers to the de-identified patient-level data to be shared. Formally, this is represented as a set $D$ of $n$ individuals $d_1, d_2, \ldots, d_n$ defined over a set of quasi-identifying features $Y_1, Y_2, \ldots, Y_m$. The records for these individuals can be partitioned into a set of equivalence classes (i.e., the set of unique combinations of

quasi-identifying values) $q_1, q_2, \ldots, q_J$. Let $f_i$ be the number of records in $D$ for the equivalence class $q_j$ associated with record $d_i$.

In addition, we assume there exists one or more *external* datasets that potentially contains the identities of the individuals whose records in the internal dataset are at risk for re-identification. A typical example of such a dataset in the US that has been leveraged for re-identification purposes is a voter registration list [15]. Set $E$ of $N$ individuals $e_1, e_2, \ldots, e_N$ is defined over the same set of quasi-identifying features $Y_1, Y_2, \ldots, Y_m$. Let $F_i$ be the number of records in $E$ for the equivalence class $q_j$ associated with record $d_i$ in $D$. Records from the internal dataset and the external dataset are linked if they share the same set of quasi-identifying features $Y_1, Y_2, \ldots, Y_m$.

We represent $q_j$ with $f_i$ larger than $F_i$ as invalid classes, denoted as $q_{j\_invalid}$. There are $n_r$ individuals from $D$ who are in $q_{j\_invalid}$.

### 2.1 Internal Marketer Risk Measure

Based on the formulation introduced by Dankar *et al.* [24, 25], we define an ***Internal Marketer (IM) Risk*** measure:

$$IM\ \ Risk(D) = \left( \frac{\sum_{i=1}^{n} \frac{1}{f_i}}{n} \right) = \frac{J}{n} \tag{1}$$

which corresponds to the probability that a record in a de-identified dataset can be correctly linked to a targeted individual through QIDs. This measure assumes the adversary knows that a specific individual is in the de-identified dataset. As a result, it represents a worst-case scenario for the data sharer.

### 2.2 Record Linkage and External Risk Measures

As alluded to earlier, the external dataset is typically a sample from a larger population and, for some equivalence classes, patients in the internal dataset could be linked to a fewer number of identified persons. This incorrectly implies that the probability of correct re-identification is larger than 1. We introduce three new measures to correct this sampling issue under specific adversarial assumptions.

***Conservative External Marketer (CEM) Risk*:** In this scenario, we assume that, if a person exists in the internal dataset, he should also be included in the external dataset. However, for some $q_{j\_invalid}$, the $f_i$ may be larger than the corresponding $F_i$ in the external dataset. Thus, we add dummy records to the external dataset so that the equivalence class is of the same size as that observed in the internal dataset. We leave the external dataset unchanged for all $q_j$, where $f_i$ is no larger than the corresponding $F_i$.

Figure 1 depicts a situation in which a de-identified patient dataset is linked to an identified voter registration list. In this figure, there are three male patients who were born in 1959, but there are only two voters in the same equivalence class. Thus, to account for the "missing" patient, we add one voter record ("Imputed for Male 1959" in the upper section of Fig. 1) to the identified dataset. We assume that the attacker has

the same prior knowledge about individuals in $q_{j\_invalid}$ and in $q_j$. This yields an upper bound for re-identification risk, which is calculated as follows:

$$CEM\ \ Risk(D, E) = \frac{\sum_{i=1}^{n} \frac{1}{\max(F_i, f_i)}}{n} \tag{2}$$

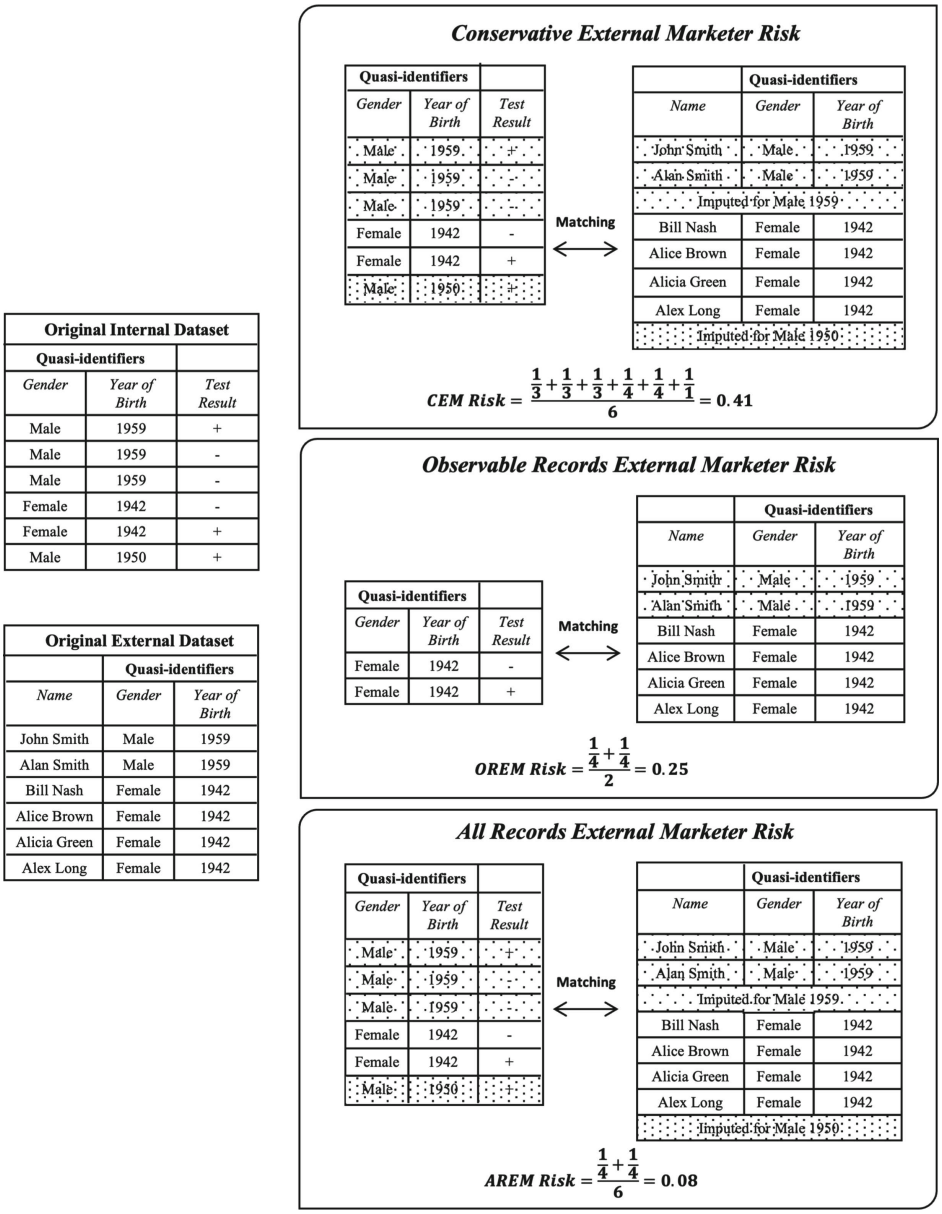


**Fig. 1.** An illustration of record linkage and risk computation for *CEM* (upper), *OREM* (middle), and *AREM* (lower).

***Observable Records External Marketer* (OREM) *Risk*:** In this setting, we assume that patients in the internal dataset with no corresponding records in the external dataset (as defined by their QID) are protected by their lack of presence. As a result, in this

measure, we assume they are not at risk of re-identification. Thus, these patients are removed from the computation. As shown in the middle section of Fig. 1, three male patients who were born in 1959 and one male patient who was born in 1950 are removed from the internal dataset in the linkage process. This measure yields a risk that is no greater than the upper bound and is calculated as follows:

$$OREM\ \ Risk(D, E) = \frac{\sum_{i=1}^{n-n_r} \frac{1}{F_i}}{n - n_r} \tag{3}$$

***All Records External Marketer (AREM) Risk*:** In this setting, we assume that the attacker has no knowledge about individuals in $q_{j\_invalid}$ (i.e., in the equivalence classes that do not have enough corresponding records in the external dataset). In the examples shown in the bottom section of Fig. 1, we add dummy records to the external dataset in the same manner as *CEM* risk. As a result, this risk is calculated as follows:

$$AREM\ \ Risk(D, E) = \frac{\sum_{i=1}^{n-n_r} \frac{1}{F_i}}{n} \tag{4}$$

It should be recognized that *AREM* risk is a combination of the other two risk measures. The numerator is the same as that in the *OREM*, while the denominator is the same as that in *CEM*.

## 3 Experiments

We use two real datasets to demonstrate how risk is influenced by adversarial assumptions. For the internal dataset, we use case line data for COVID-19 confirmed cases in the state of Florida (FL) as of June 3, 2021 [26]. This dataset is updated daily and includes the following information about infected individuals: 1) residential county, 2) age, 3) gender, 4) FL residency status, and 5) record date. For the external dataset, we use the FL's voter registration list as of June 8, 2020, the latest dataset accessible at the time of this study. The voter registration list includes an individual's 1) full name, 2) gender, 3) date of birth, 4) race, 5) residential address, 6) ZIP code, 7) county and 8) contact information (such as email address). For the purpose of this study, we use county, year of birth (YOB), and gender as quasi-identifiers. From the internal dataset, we remove 5% of records 1) whose patient ID, county, gender, and diagnosis date are unknown, 2) have an age below 0, or 3) those are not FL residents. Table 1 provides a summary of the datasets used in the experiments.

The COVID-19 case-line data covers January 5, 2020, to June 1, 2021. Since rapid growth in cases is a characteristic of pandemic patient-level data, we evaluate risk at each three-month interval till June 1, 2021, yielding six time points: April 1, July 1, and October 1, 2020, and January 1, April 1, and June 1, 2021.

To investigate how different policies affect the risk across demographic groups, we designed 12 alternative case-reporting policies, as shown in Table 2. These are defined by the QID generalization levels, such that policies P1 through P11 vary in their generalization of age and sex. The policies include six potential generalizations of age and two potential generalizations of sex. Here, *suppressed* indicates that the corresponding

**Table 1. Summary of the dataset used in this study.** *a b c* represents the first quartile, median, and third quartile. $d \pm e$ represents the mean and one standard deviation. *f g%* indicates that the percentage of *f* patients (in a given category) is *g%* among all patients.

| Characteristic | Distribution | | | |
|---|---|---|---|---|
| | COVID-19 Case line | | Voter Registration | |
| Age | 23 38 55 | 39.5±20.6 | NA | |
| Date of Birth | NA | | 1912-12-12 to 2020-05-31 | |
| Race/Ethnicity | NA | | | |
| Non-Hispanic White | NA | | 9,009,488 | 65.7% |
| Hispanic | NA | | 2,420,628 | 17.6% |
| Non-Hispanic Black | NA | | 1,950,476 | 14.2% |
| Other Races/Ethnicities | NA | | 328,628 | 2.3% |
| Gender | | | | |
| Male | 441,413 | 46.4% | 6,346,193 | 46.2% |
| Female | 508,316 | 53.5% | 7,363,027 | 53.7% |
| Number of counties | 67 | | 67 | |
| Event date | 2020-01-05 to 2021-06-01 | | NA | |

QID is reported as a null value for all individuals and, thus, the corresponding QID is not used in the linkage experiments. The *current* policy corresponds to the generalization level for the actual COVID-19 case line data from the FL Department of Health.

The re-identification experiments are composed of four steps: 1) apply the policy to the COVID-19 data, 2) harmonize the patients' demographic characteristics in the COVID-19 database with the FL voter database, 3) match the de-identified patient database with the identified voter database, and 4) compute the re-identification risk measures.

In these experiments, we link patients by their county, YOB, and gender. YOB is not directly available in the COVID-19 database and could be inferred from the age of the patient at the COVID-19 positive test event date, but there is an ambiguity in the transformation. For example, imagine that we observe a patient who is 30 years old, for whom the event date is March 1$^{st}$, 2021. This patient's date of birth could be as early as March 1, 1990, but as late as March 1, 1991. To address this issue, we create a YOB range for each patient with (*event date* $-$ *age* $-$ 1, *event date* $-$ *age*). In the situation where age is generalized to a range in different case-reporting policies, such as a 30-year range, the aforementioned example's age is generalized to 30–59. The YOB lower bound is 1961 and the upper bound is 1991. In general, the *YOB lower bound* is (*event date* $-$ *age range upper bound* $-$ 1 and the *YOB upper bound* is (*event date* $-$ *age lower bound*). We compare the YOB of voter records to *YOB lower bound* and *YOB upper bound* of patient records for the linkage.

**Table 2.** Case-reporting generalization policy rules.

| Policy | Age generalization level | Sex generalization level |
|---|---|---|
| P1 | Suppressed | Suppressed |
| P2 | 60 year range: 0–59, 60 + | |
| P3 | 30 year range: 0–29, 30–59, 60–89, 90 + | |
| P4 | 15 year range: 0–14, 15–29, 30–44, … | |
| P5 | 5 year range: 0–4, 509, 10–14, 15–19, … | |
| P6 | 1 year range | |
| P7 | Suppressed | Male/Female |
| P8 | 60 year range: 0–59, 60+ | |
| P9 | 30 year range: 0–29, 30–59, 60–89, 90+ | |
| P10 | 15 year range: 0–14, 15–29, 30–44, … | |
| P11 | 5 year range: 0–4, 509, 10–14, 15–19, … | |
| Current policy | 1 year range | |

# 4 Results

## 4.1 Internal Risk Evaluation

We evaluate the *IM Risk* at the end of each time period starting on the date of the first confirmed case. As time proceeds, both the number of patients and the number of unique QIDs groups grow, but at different rates. Figure 2 shows the risks for each policy. Following the U.S. Institute of Medicine report [27] and European Medicines Agency guidelines [28], we set a risk threshold of 0.09 (which corresponds to 11-anonymity for a public dataset) as an acceptable level of risk for our following analysis. It can be seen that in April 2020, all of the policies exhibited risks higher than the threshold. In July 2020, October 2020, and January 2021, policies P1-P3, P1-P4, and P1-P5 satisfy the requirement, respectively. After April 2021, all of the policies were under the threshold.

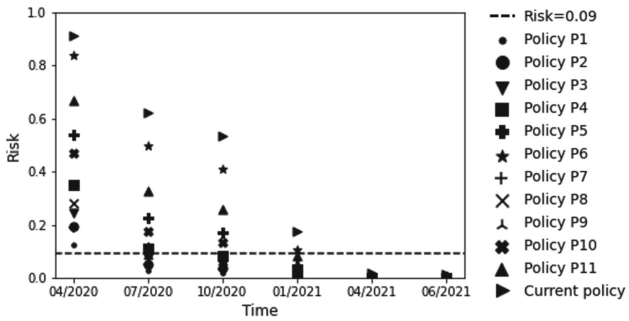


**Fig. 2.** *IM Risk* evaluated as a function of time.

### 4.2 External Risk Evaluation

We compare the *CEM*, *OREM*, and *AREM* risks by linking the FL COVID-19 case-line data to FL's voter registration list. Risk is evaluated at each of the six time points, the results of which are summarized in Fig. 3. Recall that the *CEM Risk* is an upper bound of the external marketer risk. In this case, some patients may not be matched to the corresponding voters in the voter registration list, but we added dummy records in the external dataset to acknowledge the existence of the missing records. It can be seen in Fig. 3A that only policies P6 and P11, as well as the current policy, achieve risks that are higher than the 0.09 threshold in April 2020. Still, these risks are lower than the *IM Risk*. By July 2020, policy P6 and the current policy's risks are higher than 0.09. By October 2020, only the current policy has a risk higher than 0.09. After January 2021

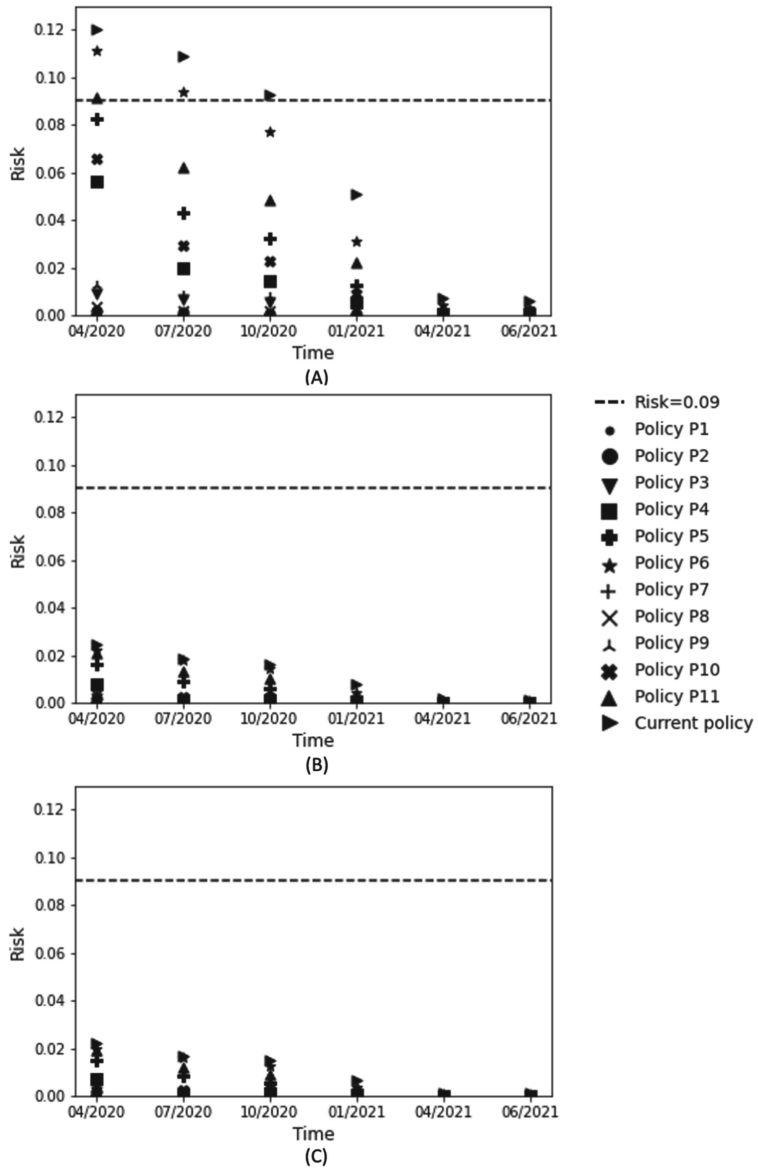


**Fig. 3.** External risk evaluation with the A) *CEM*, B) *OREM*, and C) *AREM* risks over time.

(one year after the first case-line data was released), all policies exhibit *CEM* risk that is smaller than 0.09.

Next, we analyze the *OREM* (i.e., evaluated with all valid records) and *AREM* (i.e., evaluated with all the records) risks. As shown in Figs. 3B and 3C, none of the policies have risks higher than 0.09 after April 2020.

### 4.3 Internal vs. External Risks

As anticipated, a comparison of *IM Risk* and external risks shows that risks decrease for all policies once a real identified dataset is factored into the risk assessment. However, the rate of change in risk is not constant across all policies. To illustrate this fact, we defined a risk reduction rate from the *IM Risk* to the external risk as follows:

$$\begin{aligned} &Risk\ \ reduction\ \ \ rate \\ &= \frac{Internal\ \ \ marketer\ \ \ risk\ \ - \ \ External\ \ \ marketer\ \ \ risk}{Internal\ \ \ marketer\ \ \ risk} \times\ 100\% \end{aligned} \tag{5}$$

Figure 4 shows the risk reduction rate for each policy. In this figure, the arrows indicate a hierarchical structure, where moving up the hierarchy means the data becomes more specific. The lattice graphs illustrate the partially ordered generalization levels between policies. Here, the current policy corresponds to the most specific policy. The arrow between two policies indicates a decrease in the generalization level of one of the QID variables. Specifically, an orange arrow indicates a change in age generalization level, whereas a blue arrow indicates a change in sex generalization level but remains the same level of age generalization.

The results show that policies P1 and P7 exhibit the largest reduction rates with respect to the three external marketer risk measures. When compared to policy P1, change in the age generalization level from completely suppressed to the 30-year-old range (i.e., P3) result in risk reduction rate decreases by 20%. When the age generalization level is less strict (e.g., 15-year-old range, 5-year-old range, and 1-year-old range), the effect on the risk reduction rate is almost constant.

### 4.4 Risk Reduction Rate

Figure 5 depicts the risk reduction rates for different measures. It was observed that the average risk reduction rate for the current policy evaluated with the *CEM Risk* is 24% larger than the *AREM Risk*. As the age generalization level becomes more specific, the average risk reduction rate differences between the *CEM Risk* and the *AREM Risk* grow (i.e., from P1 to P6, and P7 to the current policy). However, for all policies, there is no difference between the average risk reduction rate for the *OREM* and *AREM* risks.

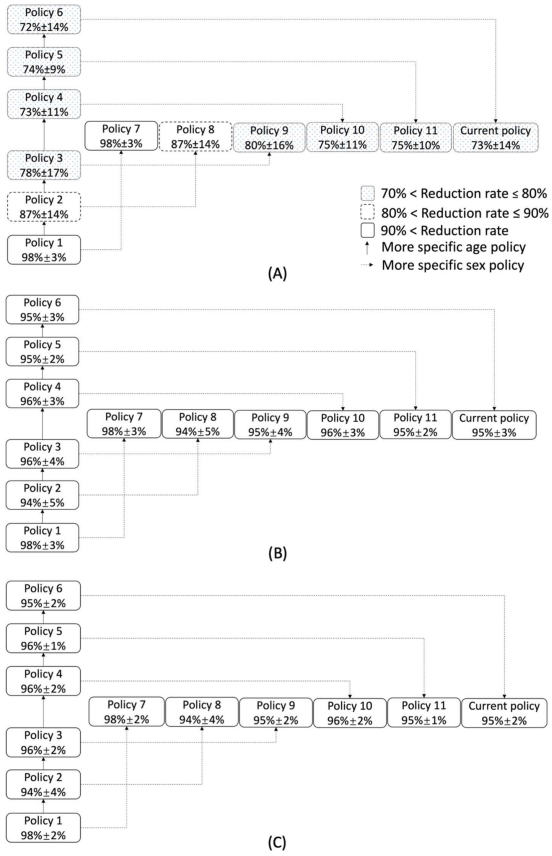


**Fig. 4.** Risk reduction rates (average of the six time points ± 1 standard derivation) from the *IM* risk to the external risks for policies organized in generalization hierarchy: A) *CEM*, B) *OREM*, and C) *AREM* risks.

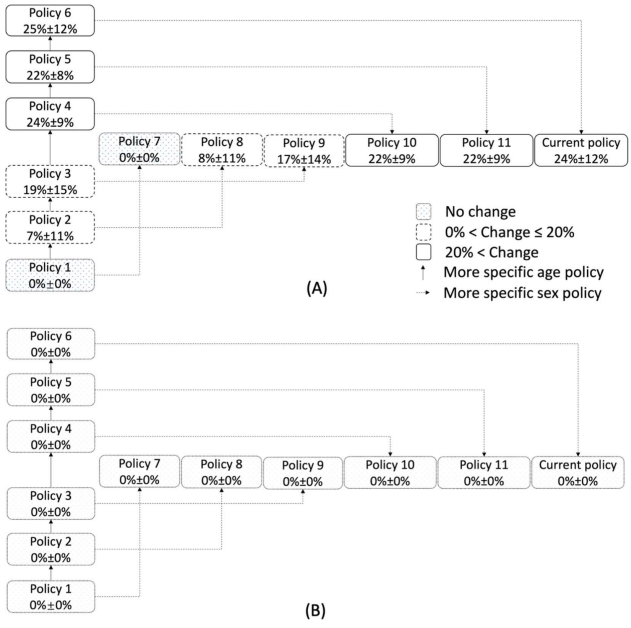


**Fig. 5.** Change in risk reduction rates (average value across six time points ± 1 standard derivation) between risks evaluated with: A) *CEM* and *AREM* risks and B) *OREM* and *AREM* risks.

## 5 Discussion and Conclusion

As this work shows, re-identification risk measures have insufficiently addressed real world data challenges, particularly the missing records in an identified resource that an attacker is expected to leverage. The external marketer risk measures we introduced show that missing records can contribute to risk in different ways depending on adversarial assumptions. Our experiments with FL COVID-19 case line data show that such assumptions non-trivially affect re-identification risk estimation. In particular, our results reveal that the risks under all 12 policies are below the typical risk threshold of 0.09 as of April 2021 for the internal marketer risks. The *CEM*, *OREM*, and *AREM* risks are below the threshold as of January 2021, April 2020, and April 2020, respectively. It suggests that more detailed data could be shared publicly than the current generalization policy.

Further, in our comparison of risk reduction rates, we observed that there is no difference in the risk reduction rates between the *OREM* and the *AREM Risk*. This suggests that the data sharer could use either risk measure considering an attacker's decision to target the invalid groups does not affect external marketer risk estimation. We also observed that the risk reduction rates between the *IM Risk* and external risks are relatively stable for all policies evaluated with the *OREM* and *AREM* risks. This suggests that data sharers could use the *IM Risk* as a proxy to estimate the external risks for a data sharing policy based on the reduction rate and use that estimation for other policies. Finally, the reduction rate between the *IM Risk* and the *CEM Risk* is the smallest. Thus, the *CEM Risk* is a more reliable measure compared to the other two measures. The risk reduction rate could be utilized as an approximation for the *CEM Risk* from the *IM Risk* when the external dataset is not accessible.

There are also several limitations we wish to acknowledge as opportunities for future investigations and improvements. First, residency does not always imply the place where an individual currently lives. Our study assumes residency per the voter registration list is equal to the residency in the internal dataset. Second, we only evaluate the risks with COVID-19 case surveillance data from Florida. Yet different states may adopt different approaches to collecting, generalizing, and releasing medical data. Third, the FL's voter registration list is updated on a monthly basis, such that an analysis of the recency of the voter data for the COVID data should be assessed to determine the influence on risk. One particularly notable question is how to maximize the re-identification risk-utility trade-off when data is released with dynamically updated policies [29]. Fourth, our study evaluates re-identification risks without considering the benefits inherent in sharing data and the attacker's gain from the re-identification attack. Future studies could use economic arguments (e.g., based on game theory) to analyze the balance between privacy and utility [30].

**Funding.** This study was supported by the funding sources: grants CNS-2029651 and CNS-2029661 from the National Science Foundation and training grant T15LM007450 from the National Library of Medicine.

## References

1. Sohrabi, C., et al.: World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). Int. J. Surg. **76**, 71–76 (2020)

2. Rodriguez-Morales, A.J., et al.: Clinical, laboratory and imaging features of COVID-19: a systematic review and meta-analysis. Travel Med. Infect. Dis. **34**, 101623 (2020)
3. CDC national surveillance. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance. html#:~:text=CDC%20uses%20national%20case%20surveillance,identify%20groups%20most%20at%20ris. Accessed 20 May 2022
4. Kostkova, P.: Disease surveillance data sharing for public health: the next ethical frontiers. Life Sci. Soc. Policy **14**(1), 1–5 (2018). https://doi.org/10.1186/s40504-018-0078-x
5. Ienca, M., Vayena, E.: On the responsible use of digital data to tackle the COVID-19 pandemic. Nat. Med. **26**, 463–464 (2020)
6. World Health Organization: Global Surveillance for COVID-19 Caused by Human Infection with COVID-19 Virus: Interim Guidance. World Health Organization, Geneva (2020)
7. Lee, B., et al.: Protecting privacy and transforming COVID-19 case surveillance datasets for public use. Public Health Methodol. **136**(5), 554–561 (2021)
8. COVID-19 Case Surveillance Public Use Data. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf. Accessed 20 May 2022
9. French, M., Monahan, T.: Disease surveillance: how might surveillance studies address Covid-19? Surveill. Soc. **18**(1), 1–11 (2020)
10. Ioannou, A., Tussyadiah, I.: Privacy and surveillance attitudes during health crises: acceptance of surveillance and privacy protection behaviours. Technol. Soc. **67**(101774) (2021)
11. Allam, Z., Jones, D.S.: On the Coronavirus (COVID-19) Outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (AI) to benefit urban health monitoring and management. Healthcare **8**(1), 46 (2020)
12. Dalenius, T.: Finding a needle in a haystack – or identifying anonymous census record. J. Official Stat. **2**(3), 329–336 (1986)
13. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncert. Fuzz. Knowl. Based Syst. **10**(5), 557–570 (2002)
14. Durham, E., Xue, Y., Kantarcioglu, M., Malin, B: Private medical record linkage with approximate matching. In: AMIA Annual Symposium Proceedings 2010, pp. 182–186 (2010)
15. Benitez, K., Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule. J. Am. Med. Inf. Assoc. JAMIA **17**(2), 169–177 (2010)
16. Skinner, C., Holmes, D.: Estimating the re-identification risk per record in microdata. J. Official Stat. **14**(4), 361–372 (1998)
17. Sweeney, L.: Simple demographics often identify people uniquely. Technical Report LIDAP-WP3, Carnegie Mellon University (2000). https://dataprivacylab.org/projects/identifiability/paper1.pdf. Accessed 21 May 2022
18. Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry. Policy on public health research and non-research data management and Access. https://www.cdc.gov/maso/policy/policy385.pdf. Accessed 20 May 2022
19. Xia, W., et al.: Enabling realistic health data re-identification risk assessment through adversarial modeling. J. Am. Med. Inform. Assoc. **28**(4), 744–752 (2021)
20. Xia, W., Kantarcioglu, M., Wan, Z., Heatherly, R., Vorobeychik, Y., Malin, BA.: Process-driven data privacy. In: 24th ACM International on Conference on Information and Knowledge Management (CIKM 2015) Proceedings, pp. 1021–1030. Association for Computing Machinery, New York, NY, USA (2015)
21. Koot, M.R., Noordende, G. van 't, de Laat C.: A study on the re-identifiability of Dutch citizens. In: 3rd Hot Topics in Privacy Enhancing Technologies (HotPETs 2010) Proceedings, pp. 35–49. Berlin, Germany (2010)
22. Golle, P.: Revisiting the uniqueness of simple demographics in the US population. In: 5th ACM Workshop on Privacy in Electronic Society Proceedings, pp. 77–80. New York, NY, USA (2006)

23. Emam, K.E., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., Verma, A.: The re-identification risk of Canadians from longitudinal demographics. BMC Med. Inf. Dec. Mak. **11**(1), 46 (2011)
24. Emam, K.E., Dankar, F.K.: Protecting privacy using k-anonymity. J. Am. Med. Inform. Assoc. **15**(5), 627–637 (2008)
25. Dankar, FK., Emam, KE.: A method for evaluating marketer re-identification risk. In: 2010 EDBT/ ICDT Workshops Proceeding Article 28, pp. 1–10. Association for Computing Machinery, New York, NY, USA (2010)
26. Florida COVID-19 Case Line Data. https://open-fdoh.hub.arcgis.com/datasets/florida-covid19-case-line-data/about. Accessed 20 May 2022
27. Institute of Medicine (IOM): Sharing clinical trial data: Maximizing benefits, minimizing risk. The National Academies Press, Washington, DC (2015)
28. European Medicines Agency: External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use, Revision 4. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_001799.jsp&mid=WC0b01ac0580b2f6ba. Accessed 20 May 2022
29. Brown, J.T., et al.: Dynamically adjusting case reporting policy to maximize privacy and public health utility in the face of a pandemic. J. Am. Med. Inform. Assoc. **29**(5), 853–863 (2022)
30. Wan, Z., et al.: A game theoretic framework for analyzing re-identification risk. PLoS ONE **10**(3), e0120592 (2015)