

Towards Realistic Membership Inferences: The Case of Survey Data

Luke A. Bauer
lukedrebauer@ufl.edu
University of Florida

Vincent Bindschaedler
vbindsch@cise.ufl.edu
University of Florida

ABSTRACT

We consider the problem of membership inference attacks on aggregate survey data through the use of several real-world datasets and a published study as a model for the survey. We apply membership inference attacks from the literature, and discover that methodological assumptions of existing attacks produce a misleading picture of the risk. When using a more realistic methodology, experiments reveal a more nuanced picture of the risk: membership inferences do succeed, but only a small subset of individuals are highly vulnerable to them. In fact, if the adversary wishes to avoid a high false positive rate, she should perform membership inferences only when she has a reason to believe that the target participated in the survey. However, our results do not imply that publishing survey data is inherently safe. Indeed, when applying membership inference not to individuals but to hospitals, we find that highly accurate inferences are possible.

CCS CONCEPTS

• Security and privacy → Privacy protections.

KEYWORDS

Data Privacy, Membership Inference, Survey Data

ACM Reference Format:

Luke A. Bauer and Vincent Bindschaedler. 2020. Towards Realistic Membership Inferences: The Case of Survey Data. In *Annual Computer Security Applications Conference (ACSAC 2020)*, December 7–11, 2020, Austin, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3427228.3427282>

1 INTRODUCTION

Consider a survey or a hypothesis-driven academic study which is conducted by a survey curator. The curator aggregates responses to queries (i.e., questions) over a set of data records from several individuals and subsequently publishes the results. Suppose that an adversary wants to determine whether a specific target individual's data was included in the survey on the basis of the published results and perhaps some background knowledge. This is called a membership inference attack.

Membership inference attacks have been increasingly studied since the work of Homer et al. [20] who performed the first membership inference attack in the context of genome-wide association studies. However, most academic work has focused on particular settings with fairly specific types of data. For example, following Homer et al. [20], researchers have focused on genomic and related personal biological data [1, 21, 42, 49, 56]. More recently, Pyrgelis et al. [36, 37] considered the case of aggregate location data. Finally, attacking machine learning models trained on private data has received the most attention [4, 16, 29, 30, 32, 33, 41, 43, 47, 47, 53, 55].

In this paper, we apply membership inference attacks to the aggregate survey data setting, as it remains largely unaddressed by prior work on membership inference. We consider an adversary who attempts to determine if a specific target individual took part in a survey, with the goal of learning some sensitive information about the target. Inferring membership status of the target in the survey is significant for two reasons. First, because this may directly reveal sensitive information: for example, if the survey's focus is on patients that suffer from a specific medical condition, then inferring membership directly reveals that the target suffers from that medical condition. Second, because inferring membership can be used as the first step of a more complex inference attack such as an attribute inference attack.

Surprisingly, we discover two problems that make existing membership inference attacks ill-suited to our setting. The first problem is that in our setting, the queries are fixed. The second problem is that existing membership inference attacks make unrealistic (sometimes implicit) methodological assumptions that produce a distorted picture of the privacy risk. For example, prior work often takes the adversary's auxiliary data directly from the same source as the input dataset, which is ideal for the adversary. Moreover, prior work's attacks are evaluated in the 50%-inclusion frequency setting: a target individual is included in the input data with probability 0.5; whereas realistic values of inclusion frequency are arguably closer to 1% (or even 0.1%).

To address the first problem, we propose a novel *query filtering* methodology for the adversary to identify an optimal subset of queries to use for the attack (and ignores the rest). To address the second problem, we systematically revisit and relax unrealistic assumptions of prior work on membership inference. Further, to obtain a more realistic picture of the risk, we simulate inference attacks on the recent study by Rizzo et al. titled "Racial Disparities in PAD-Related Amputation Rates among Native Americans and non-Hispanic Whites: an HCUP Analysis" [40]. We use the underlying data and the query results reported in the study for our attacks.

Our experiments reveal a more nuanced picture of the risk compared to prior work: membership inference can be an effective

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACSAC 2020, December 7–11, 2020, Austin, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8858-0/20/12...\$15.00

<https://doi.org/10.1145/3427228.3427282>

attack, but only a small subset of individuals are highly vulnerable. Furthermore, from the point of view of an adversary wishing to avoid false positives, performing membership inferences only make sense when the adversary has some reason to believe that the target participated in the survey. However, this is not to say that publishing aggregate survey data is without risk. When using the same data and our inference attacks not on individuals, but on hospitals (i.e., groups of patients) we find that highly accurate inferences are possible. In particular, our results suggest that, from published studies like Rizzo et al. [40], it may be possible to infer sensitive properties of a hospital such as whether the hospital performs amputation of PAD patients at a higher rate compared to that of other hospitals.

Contributions.

- We perform the first membership inference attacks against aggregate survey data with queries taken directly from a published article.
- We propose a query filtering technique and demonstrate its importance for the success of membership inference on survey data
- We demonstrate that inclusion frequency, a critical factor often overlooked in prior research on membership inference, has a significant effect on attack success that should not be ignored going forward.
- We demonstrate the first membership inference attack that targets not individuals but organizations. Specifically, we perform membership inference with a hospital as the target and show that successful inferences have the potential to damage a hospital’s reputation.

Structure. The paper is structured as follows. In Section 2, we provide a brief overview of related work. Section 3 provides background on membership inference attacks and defenses. In Section 4, we discuss challenges of inference attacks on survey data and describe our attack methodology. The datasets we use in our experiments are described in Section 5. Experimental results on factors affecting membership inference can be found in Section 6. Our case study on PAD-related amputations is described in Section 7. In Section 8, we discuss experimental findings in the broader context of inference attacks. Finally, we conclude in Section 9.

2 RELATED WORK

Membership inference attacks have been studied extensively since the seminal paper of Homer et al. [20], which showed how to identify a target participating in a genome-wide association study by comparing allele and genotype counts to public averages. Further work on membership inference has significantly expanded our understanding of these attacks [1, 42, 49, 56]. In particular, Backes et al. [1] demonstrated a similar attack in the context of Micro-RNA data. In a different vein, Pyrgelis et al. [36, 37] explored membership inference attacks on aggregate location data.

A different line of work, focused on machine learning, started with Shokri et al. [44] who demonstrated the first membership inference attack on machine learning models. Yeom et al. [53] draws a theoretical connection between privacy and overfitting. Long et al. [29] demonstrate that overfitting is not a requirement for

membership inference. Salem et al. [41] propose simpler attacks that perform almost as well as the one proposed by Shokri et al. [44]. Membership inference attacks have been further performed in many different settings using various types of models [4, 16, 33, 43, 47, 52, 55].

There are several defenses to protect aggregate statistics from inference attacks. These defenses fall into two categories: (1) principled defenses such as differential privacy [8–10] or relaxations [3, 11, 15, 27, 31]; and (2) heuristics such as omitting to report *small cell counts*, which is a common recommendation advocated as a way of reducing the risk of re-identification [7, 38].

3 BACKGROUND & PRELIMINARIES

In this section, we provide background on membership inference.

3.1 Inference Attacks

Consider inference attacks on the basis of a published survey or scientific study. We formalize this below.

Problem statement. The survey curator collects information about n individuals into a *sample* database — we call n the *sample size*. Formally, we let $\mathbf{D}$ be the sample database of $n \geq 2$ records $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n$ where each data record $\mathbf{d}_i$ lies in some universe U . More precisely, we model the sample $\mathbf{D}$ as a collection of records taken i.i.d. from some data distribution $\mathcal{D}$.

The curator has a list of queries q_i for $i = 1, 2, \dots, m$ and he wants to publish *aggregate results* $\mathbf{s} = \{s_1, s_2, \dots, s_m\}$ where each $s_i \in \mathbb{R}$ such that each result is the output of the corresponding query: $s_i = q_i(\mathbf{D})$. For conciseness, we often write: $\mathbf{s} = q(\mathbf{D})$ to denote the survey results.

The adversary seeks to infer sensitive information about some *target* individual t (or targets individuals $\mathbf{t}$) on the basis of the (published) survey results $\mathbf{s}$. We assume that the adversary knows the list of queries q_i and (possibly) the data distribution $\mathcal{D}$. She also has access to some *background information* or *prior knowledge* π . However, she does *not* have access to the sample database $\mathbf{D}$ itself, which is the input data for the queries. For example, the prior knowledge π may consist of an auxiliary dataset $\mathbf{D}^{\text{aux}}$ that reflects statistical trends of $\mathcal{D}$.

Membership inference. Given the queries output over an unknown sample database $\mathbf{D}$, the goal of the adversary is to infer whether the data record t of a specific *target* individual is included in the sample database. More formally, the adversary must distinguish between the following hypotheses.

- $H^{\text{in}}: t \in \mathbf{D}$
- $H^{\text{out}}: t \notin \mathbf{D}$

Existing membership inference attacks fall into two categories: likelihood-ratio attacks and machine learning-based attacks.

Likelihood ratio attack. Inspired by statistical hypothesis testing, this attack uses likelihood ratios to distinguish between H^{in} and H^{out} . The attack was originally proposed by Homer et al. [20] and subsequently refined by Sankararaman et al. [42]. The idea is to compute the likelihood of H^{in} and of H^{out} and decide based on their ratio. If the ratio is larger than some threshold (e.g., 1), then the attack outputs IN, otherwise OUT.

The likelihood ratio is computed as:

$$\text{llr} = \frac{p^{\text{in}}}{p^{\text{out}}}, \quad (1)$$

where p^{in} (resp. p^{out}) denotes the probability of the observed output under H^{in} (resp. H^{out}). The calculation of p^{in} and p^{out} depend on how the likelihood term is modeled, which (in part) depends on the specific setting considered.

The adversary can trade-off the probability of false positive and false negatives by changing the threshold compared to the ratio in Eq. (1). For example, the adversary can set a threshold $0 \leq \tau \leq 1$ and then predict IN if $\text{llr}/(1 + \text{llr}) \geq \tau$ (and OUT otherwise).

Machine learning. Alternatively, the problem of predicting membership can be recast as a binary classification task. In other words, one can train a classification model to predict membership (i.e., IN or OUT) using features derived from the survey results. Prior work has used various classification models for this such as logistic regression, random forest, and neural networks. As features, one can use various functions of the output. For example, in our setting we could simply take the i^{th} feature to be the output of the i^{th} query.

After training the classification model, membership status for a target t can be predicted by querying the model on the list of features derived from the observed output. The model returns a confidence value or probability p that reflects the plausibility of H^{in} . Depending on how conservative the adversary wants to be, she sets a threshold $0 < \tau < 1$ and predicts IN if $p \geq \tau$.

Metrics. There are several ways to quantify the success of membership inference attacks. Prior work often uses traditional machine learning metrics for classification: accuracy (proportion of correct predictions); precision (proportion of correct predictions among IN predictions); and recall (proportion of included targets who were predicted as IN). In the case where the evaluation dataset is such that exactly half of the targets are included in the sample database (and half are not), an alternative is to use the adversary’s advantage. The *advantage* is defined by Yeom et al. [53] as the difference between the true positive rate and false positive rate.

Beyond membership. Other inference attacks can be performed on the basis of membership inference such as attribute inference. In addition, in Section 4.4, we propose two novel *group membership* inference attacks. These attacks infer the membership of an entire group of targets (e.g., a family or the patients of a specific hospital) in the sample database.

3.2 Defenses

Differential Privacy. Differential privacy [9, 10] is a theoretically-sound defense against membership inference. Informally, differential privacy ensures that the influence of any single record on the query output is bounded. This translates into an upper-bound on the success of *any* membership inference attack. In particular, as argued by Yeom et al. [53], if the defense satisfies ϵ -differential privacy, then the advantage of any membership inference attack is at most $e^\epsilon - 1$.

We note that there exist some differential privacy mechanisms customized for the GWAS settings [13, 24, 54]. Unfortunately, in

our setting, differential privacy mechanisms are difficult to apply without incurring catastrophic utility loss (Section 7.3).

Suppressing Small Counts. A heuristic defense, often advocated for in practice, is to remove or suppress (i.e., not report on) count queries with very small outputs (i.e., involving only a small number of individuals). For example, consider the following query “how many male patients over 95 years old have coronary atherosclerosis?”¹ If this number is below a small threshold (e.g., 10) then the survey curator may decide not to report this query output (or always report 0 or < 10). This defense strategy is often referred to as *suppressing small counts* and there exists several variants including the suppression of the output of queries with count less than 16 [7] or even prohibiting any studies with data size less than 100 [38].

4 ATTACKING SURVEY DATA

In this section, we discuss the motivation for considering survey data as well as the unique challenges of performing inference attacks on aggregate statistics from survey data: queries are fixed and limited. This is in contrast to other settings such as the machine learning model setting where the adversary can query the model arbitrarily. We also discuss how the methodology of prior work can be made more realistic in our setting. Finally, we describe our attack methodology and propose two novel group membership inference attacks.

4.1 Why Survey Data?

Surveys are ubiquitous nowadays and survey results are disseminated in various ways such as online, academic papers, popular magazines, etc. And these surveys often contain various kinds of sensitive information about individuals such as their finances, medical history, relationships, religious or political opinions, etc. In particular, there are a number of academic articles that provide survey results for healthcare patients with specific medical conditions [26, 28, 48]. In this context, the threat of successful inferences is clear even when the survey results are aggregated over all participants. For example, if an individual’s participation in one such a survey is discovered, then it becomes known that they suffer from the medical condition considered. This represents a significant privacy violation.

To the extent that survey results are regularly published without privacy safeguards beyond anonymizing participants answers and aggregating them, it is essential that we understand the potential privacy risk. Especially, given that membership inference (and inference attacks in general) have so far largely focused on other domains such as machine learning, genomic data, or location data. In contrast to these domains, the survey data setting is not only understudied but also particularly challenging.

4.2 Challenge: Fixed Queries

In the machine learning setting [29, 41, 44, 53], the adversary is able to query the model on chosen data points. In particular, given a target t , membership inference attacks typically query the model on t because the model’s output on t is often significantly different

¹Coronary atherosclerosis refers to a disease or damage of the major blood vessels in the heart (e.g., as a result of the build up of plaque.)

when the target included in the model’s training data compared to when it is excluded.

This is in stark contrast with the aggregate survey setting considered in this paper. In the aggregate survey setting, the survey curator decides (ahead of time) on a set of queries $q_1, q_2, \dots, q_m$ for the survey. Consequently, the adversary only observes the output of the queries on the unknown sample database $\mathbf{D}$ (i.e., she observes $q_1(\mathbf{D}), q_2(\mathbf{D}), \dots, q_m(\mathbf{D})$). As a result, the adversary has no direct way of customizing the attack to a specific target t , except for possibly ignoring queries which may mislead her. In fact, filtering out misleading queries is essential (Section 6.2).

Queries may mislead the adversary for two reasons. First, a query may be such that the target t has no influence on it. This happens whenever the query output is the same regardless of whether $t \in \mathbf{D}$. For example, if the target is a 40 year old individual, the outcome of the query “number of individuals under 25 with a diabetes” is useless for the purpose of inferring membership of t in the database. Second, given the adversary’s prior knowledge π , her interpretation of the outcome of some query q_j may mislead her (e.g., $q_j(\mathbf{D})$ may suggest that H^{in} is true even if in fact $t \notin \mathbf{D}$) if the sample data does not reflect the distribution estimated from the auxiliary data. As a concrete example, consider the HCUP NIS databases (Section 5), which we use for some of our experiments. The percentage of all diagnoses in 1995 that are HIV is 6.4% compared to 2.1% in 2006 and .4% in 2012.

For these reasons, it is crucial that the adversary performs the attack on a subset of the available queries containing only *informative* queries. We propose a novel methodology to select such a subset of queries in Section 4.4.

4.3 Towards A More Realistic Methodology

To provide an accurate assessment of the risk of inferences attacks in our setting, we need to simulate attacks with a methodology that accurately reflects a real-world adversary. In this paper, we take steps towards a more realistic methodology. Specifically: (1) we select realistic queries by extracting them directly from published studies in the literature, whenever possible; (2) we use “less-than-ideal” (e.g., stale or outdated) data as the adversary’s auxiliary data; and (3) we consider the effect of varying inclusion frequency on the attack (in contrast to prior work which consistently sets the inclusion frequency to 50%).

Selecting realistic queries. To accurately reflect the inference attacks that exists in the real-world, we need to use realistic queries (e.g., queries with results likely to be published). To ensure this, we pull our queries directly from academic studies [34] that analyze datasets used in our evaluation. These queries correspond to statistics that are of interest to researchers and thus are likely reflective of queries used in actual surveys. We provide further details in Section 5.3. For the evaluation of our attacks in Section 7, we use exactly the queries reported in the recent paper by Rizzo et al. [40] titled: “Racial Disparities in PAD-Related Amputation Rates among Native Americans and non-Hispanic Whites: an HCUP Analysis.”

Auxiliary data quality. In the machine learning model setting, the auxiliary dataset is often taken to be a subset of the data that is disjoint from the model’s training data. This is ideal for the adversary. In this paper, we relax this assumption. Specifically, we

assume that the sample database $\mathbf{D}$ is taken from some distribution $\mathcal{D}$ over the data universe U . And, we assume that the auxiliary dataset is sampled from some distribution $\mathcal{D}'$ over the data universe U . In the ideal case, we may have $\mathcal{D} = \mathcal{D}'$ but we consider the general case where the two distributions can differ arbitrarily. In experiments, we take a different year or version for the auxiliary data distributions than for the distribution from which the sample database is taken. We also consider *cross-dataset* attacks in Section 6.4, where we use different (and unrelated) data sources for the two distributions.

Inclusion frequency. The typical methodology to quantify the success of membership inference is to take the inclusion frequency of the target t in $\mathbf{D}$ to be exactly 50%. In other words: a (target) participant is included (or not) with *equal probability*. Although this assumption is used pervasively in related work, it is unrealistic and could produce an inaccurate picture of the risk. To understand this, imagine a study conducted and published by a small town hospital with 50 local participants (i.e., patients). Suppose the town has 5000 inhabitants and Alice (the adversary) wants to infer whether her neighbor Bob was one of the study participants. Alice would be justified (given no other information) to think that the chance of Bob being a study participant is about 1%. On the one hand, membership inference experiments that simulate a 50% inclusion frequency could provide a misleading picture of the risk. On the other hand, one might reason that a correct inference is a correct inference and thus why should it matter what the inclusion frequency is? We settle this question experimentally in Section 6.3 and Section 7, finding that inclusion frequency is a critical methodological factor that further research should not ignore! We note that concurrent work by Jayaraman et al. [23] studies similar concerns in the machine learning models setting.

4.4 Attacks & Methodology

In this section, we describe our attack methodology including our novel query filtering technique, and our proposed group membership inference attacks.

Likelihood ratio. We model the likelihood term (Eq. (1)) by treating each query independently and assuming that the query outputs samples from a Gaussian distribution with means and standard deviations calibrated to the respective hypotheses (i.e., H^{in} and H^{out}). The rationale for this assumption is the principle of maximum entropy.² It is known that the Gaussian distribution has the largest entropy among all distributions with a given mean and variance [5]. Concretely, we compute the ratio as:

$$\text{llr} = \prod_{i=1}^m \frac{p_i^{\text{in}}(s_i)}{p_i^{\text{out}}(s_i)},$$

where $s_i = q_i(\mathbf{D})$ that is the output of the i^{th} query and $p_i^{\text{in}}(s_i)$ (resp. $p_i^{\text{out}}(s_i)$) is the probability of observing s_i if H^{in} holds (resp. if H^{out} holds).

The independent Gaussian assumption means that $p_i^{\text{in}}(s_i)$ is the density of $\mathcal{N}(\mu_i^{\text{in}}, (\sigma_i^{\text{in}})^2)$ at s_i , where $\mu_i^{\text{in}}, \sigma_i^{\text{in}}$ are the means and

² Informally, the principle says that, after taking into account what is known about a distribution, we should prefer the distribution that (among all possible candidates) has the largest entropy.

standard deviations for query i (under H^{in}) estimated from the auxiliary information. The density $p_i^{\text{out}}(s_i)$ alongside with the corresponding mean μ_i^{out} and standard deviation, σ_i^{out} are defined analogously.

Machine learning. We use a logistic regression model as classifier. We train the model on features $v_1, v_2, \dots, v_m$ with $v_i = \frac{s_i - \mu_i}{\sigma_i}$, where μ_i and σ_i are estimated from the auxiliary information. To produce each training example, we sample n records from the auxiliary data $\mathbf{D}^{\text{aux}}$, either including the target t or not. We then compute the queries on this sample data and label it IN or OUT depending on whether the target t was included. This technique can be used to train a model for a specific target t , or it can be used with a large set of targets such that statistical trends in feature distribution for IN vs. OUT examples emerge.

Parameter estimation. Given target t , we estimate the parameters (i.e., $\mu_i^{\text{in}}, \mu_i^{\text{out}}, \sigma_i^{\text{in}}, \sigma_i^{\text{out}}$ for $i = 1, 2, \dots, m$) for the likelihood ratio attack and machine learning by repeating the following procedure k times for each query i .

- (1) Sample $\mathbf{D}^{\text{out}} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n\}$ so each $\mathbf{d}_j \sim \mathbf{D}^{\text{aux}}$.
- (2) Set $\mathbf{D}^{\text{in}} = \{t, \mathbf{d}_2, \dots, \mathbf{d}_n\}$.
- (3) Compute $s_i^{\text{in}} = q_i(\mathbf{D}^{\text{in}})$ and $s_i^{\text{out}} = q_i(\mathbf{D}^{\text{out}})$.
- (4) Output $(s_i^{\text{in}}, s_i^{\text{out}})$.

From this, we obtain k sample pairs $(s_{i,l}^{\text{in}}, s_{i,l}^{\text{out}})$ for $l = 1, 2, \dots, k$. The means and standard deviation (for query i) are estimated using maximum likelihood estimation. For example, $\mu_i^{\text{in}} = k^{-1} \sum_{l=1}^k s_{i,l}^{\text{in}}$.

The above parameter estimation procedure is justified by the Glivenko-Cantelli Theorem [50] which, in this context, says that the empirical CDF of the s_i^{in} (and s_i^{out}) converges (uniformly) to the CDF of q_i on $\mathbf{D}^{\text{aux}}$ as k increases towards ∞ . Observe that, if auxiliary data $\mathbf{D}^{\text{aux}}$ comes from the same distribution as sample database $\mathbf{D}$, then this estimates the true CDF of queries on the survey samples.

Query filtering. Given the queries, the adversary should select the subset of queries expected to maximize attack accuracy. Our proposed strategy for this looks at each query q_i independently and aims to determine if q_i is *informative*. Informative queries are ones that are expected to help the adversary. Other queries are filtered out, that is the adversary simply ignores the results. The challenge in determining if a given query q_i is informative is that the adversary does not know the sample data $\mathbf{D}$.

We propose to determine the informativeness of a query q_i as follows. First, we estimate the parameters $\mu_i^{\text{in}}, \sigma_i^{\text{in}}, \mu_i^{\text{out}}, \sigma_i^{\text{out}}$ given the target t . Then, a query is deemed *informative* if its output s_i is a plausible sample from either IN or OUT estimated distributions. Intuitively, if s_i is an unlikely sample for *both* the estimated IN and OUT distributions, then it is unlikely that the estimated distribution (from the auxiliary data) is statistically reflective of the sample database $\mathbf{D}$. More formally, we define query q_i to be *informative* iff:

$$p_q \leq \Phi_i^{\text{in}}(s_i) \leq 1 - p_q \quad \text{or} \quad p_q \leq \Phi_i^{\text{out}}(s_i) \leq 1 - p_q, \quad (2)$$

where $\Phi_i^{\text{in}}(\cdot)$ and $\Phi_i^{\text{out}}(\cdot)$ are the CDFs of the i^{th} query under H^{in} and H^{out} , respectively; and $0 < p_q < 1/2$ which is a threshold parameter. In addition, if the output of the query shows absolutely

no change when the target is IN vs OUT, then the query is deemed uninformative and is discarded.

Attribute inference. If the adversary knows that a given target t is included in the sample database (e.g., from a prior successful membership inference or due to some external information) then she may attempt to infer the value of some sensitive attribute about the target that she does not know. For example, the adversary may know that the target was treated at a specific hospital and know the total amount of the medical bill but not know the exact medical diagnosis. To infer this, the adversary can perform a membership inference attack on $t_1, t_2, \dots, t_k$ where t_i denotes the target's data extended with the i^{th} possible value for the sensitive attribute. If the membership inference attack on t_j returns the highest probability for IN, then the adversary infers that the value of the sensitive attribute is the j^{th} possible value.

Group membership. In a group membership inference attack, the adversary wants to determine whether target set $\mathbf{t}$ is included in the sample, i.e.: $\mathbf{t} \subset \mathbf{D}$. (The alternative hypothesis is that $\mathbf{t} \cap \mathbf{D} = \emptyset$.) In other words, we assume that either all records in the target set are included in the sample, or none of them are.

We propose two ways to perform group membership inference attacks. The first, which we call *individual influence*, runs one membership inference attack for each target in the group and aggregates the results. The second, which we call *group-as-a-record*, considers the target set as though it were a single record and then runs membership inference.

To build on generic membership inference attacks, we model the underlying attack as a membership inference *oracle* O_π^{MI} . Given a target t , survey results $\mathbf{s}$ and queries q , the oracle outputs a probability p where p represents the plausibility of H^{in} (and $1 - p$ is the plausibility of H^{out}).

Let $\mathbf{t} = \{t_1, t_2, \dots, t_l\}$.

- (1) *Individual influence* strategy:

- Let $p_i = O_\pi^{\text{MI}}(t_i, q, \mathbf{s})$ for $i = 1, 2, \dots, l$.
- If $\prod_{i=1}^l \frac{p_i}{1-p_i} \geq 1$, then output IN.
- Otherwise output OUT.

- (2) *Group-as-a-record* strategy: estimate the parameters as previously described while considering the entire target set $\mathbf{t}$ as though it were a single record.

Concretely, in step (2) of the parameter estimation procedure, we replace the first $l = |\mathbf{t}|$ records by $\mathbf{t}$ to obtain $\mathbf{D}^{\text{in}}$. We then run either the likelihood ratio attack or the machine learning attack and output whatever it outputs.

Group membership inferences open the door for adversarial inferences that go beyond compromising an individual's privacy. In addition to exploiting the correlated behaviors of related individuals (e.g., a family whose members either all participate in a survey, or none do), these attacks can potentially allow to infer information dependent on the inclusion of any semantic or otherwise meaningful grouping of individuals. In Section 7.4, we apply group membership inference attacks to hospitals.

5 DATASETS

To experimentally validate our attacks, we use two real-world datasets. The datasets are chosen because they contain information likely to be published in academic studies or surveys. In addition, multiple years are available, which enables us to use older years as reasonable representatives for the auxiliary data that a realistic adversary may have.

5.1 HCUP Nationwide Inpatient Sample

The Healthcare Cost and Utilization Project (HCUP), is a collection of healthcare databases collected by a Federal-State-Industry partnership. The Nationwide Inpatient Sample (NIS) database [18] is the largest publicly available collection of inpatient data from hospitals all across the U.S., and contains data from over 7 million hospital stays each year. We used data from years 2008, 2012, and 2013. Each year contains fine-grained (anonymized) medical records for each patient hospital visit, including diagnoses, procedures, severity of condition, as well as general information about each hospital. We process the data by first selecting attributes according to the queries and query templates we consider (Section 4.3).

To deal with missing values (usually indicated by a negative number) we pre-processed the data by discarding records with missing values, except for the RACE attribute because many hospitals did not report this attribute. Additionally, we binned TOTAL_CHARGES into 10 equal-sized bins spanning this attribute’s range when it was not used in a mean or median query. Age was binned into intervals of 5 years, to be consistent with how this attribute is provided in the released Texas data. Finally, we linked together records in the main table with the severity and hospital information, which are included in separate files.

5.2 Texas Health Care Information Collection

The Texas Health Care Information Collection (THCIC) Inpatient Public Use Data File (PUDF) [46] contains information on discharges from Texas hospitals by quarter since 1999.

The files provide information on each discharge including diagnosis, cause of visit, procedures, and basic information about the patient and encounter. We used a similar method to pre-process the data as for the HCUP-NIS database. The selected attributes are shown in Table 1.

5.3 Queries & Query Templates

Unfortunately, it is not always possible to find published academic studies for all data sets of interest that list a concrete set of queries. Thus, we expand our queries by deriving *query templates* from the available literature. Query templates generalize a specific query into a set of related *plausible* queries. For example, the query “proportion of individuals (i.e., records) who are Asian and over 70 years old” can be generalized as: “proportion of individuals who are [race] and over [age] years old.” Here [race] and [age] are placeholders which can be instantiated to other values of the respective attributes to produce a larger set of queries from the given template.

Concretely, we created query templates using queries found in many different survey papers [12, 19, 35, 39, 45]. We list our query templates in Table 2. Most of the time these are proportion queries as they are easy to generate, while also being the most commonly

Table 1: Selected Texas [46] Attributes.

Name	Description
PROVIDER_NAME	Hospital Name
TYPE_OF_ADMISSION	Type of Admission
SOURCE_OF_ADMISSION	Source of Admission
SPEC_UNIT_1	Most Used Special Unit
PAT_STATE	Patient’s Home State
PAT_STATUS	Status Upon Leaving
SEX_CODE	Patient’s Sex
RACE	Patient’s Race
ETHNICITY	Hispanic (Yes or No)
ADMIT_WEEKDAY	Weekday Admitted
LENGTH_OF_STAY	Length of Stay
PAT_AGE	Patient’s Age, Binned
FIRST_PAYMENT_SRC	Patient’s Primary Payer
TOTAL_CHARGES	Hospital’s Total Charges
ADMITTING_DIAGNOSIS	Admitting Diagnosis
PRINC_ICD9_CODE	Primary Procedure

Table 2: General Query Template Samples

Data Set	Query
Texas	Proportion of patients that are [AGE]
Texas	Proportion of patients with [TYPE_OF_ADMISSION]
Texas	Proportion of patients with [ADMITTING_DIAGNOSIS]
Texas	Proportion of patients with [FIRST_PAYMENT_SRC] and [LENGTH_OF_STAY]
Texas	Proportion of patients [Race] and [AGE] at [PROVIDER_NAME]

reported queries. For the attack on Rizzo et al. [40], we use queries pulled directly from the tables in the paper.

6 EVALUATION: FACTORS AFFECTING MEMBERSHIP INFERENCE

In this section, we systematically explore the factors that impact the success of membership inference attacks. Unless otherwise specified, we take our sample database D from the Texas 2012q1 data, and use the Texas 2011q1 dataset as our auxiliary data. We use 200 proportion queries generated from templates (Section 5.3), a sample size of 100 entries, a query filtering threshold $p_q = 0.1$, and 100 randomly selected targets.

6.1 Number of Queries & Sample Size

How does the number of queries and the sample size influence attack success? We vary the number of queries from 5 to 200, and the sample size from 50 to 2000. We perform the likelihood ratio and machine learning attacks.

Results. Figs. 1a and 1b show the results. As expected, the attack success increases with the number of queries. Interestingly, with only a few queries (e.g., <25) the adversary’s advantage is already significant, although there are diminishing returns after about 50.

Also, as expected, the sample size has a significant effect on attack success. For small samples, the advantage is very high, reaching an average advantage of 0.83 for samples of 30 records. However, the advantage decreases starkly as the sample size increases. But, even

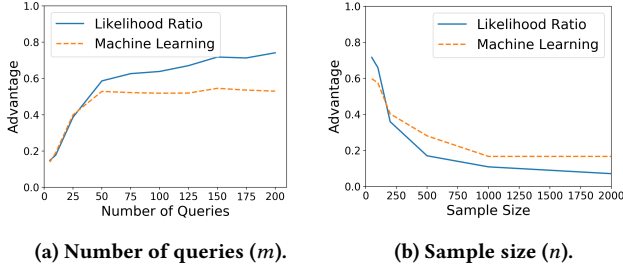


Figure 1: Adversary's advantage for varying number of queries (1a; sample size 100) and sample size (1b; 200 queries) for the likelihood ratio and machine learning attacks.

for large sample sizes (e.g., 2000), it remains possible to detect the influence of particularly vulnerable targets.

Does the likelihood ratio attack perform better than the machine learning attack? From Fig. 1a, it is tempting to answer in the affirmative as we observe that the machine learning attack does not appear to benefit from a large number of queries (i.e., anything larger than $m = 50$). Moreover, for small samples (i.e., $n \leq 200$) the likelihood ratio attack has significantly better performance (Fig. 1b). But for larger sample size (i.e., $n > 200$), this effect is reversed and the machine learning attack has a larger advantage. (The effect is particularly visible for $n \geq 1000$.) This surprising finding suggests that the machine learning attack is more robust to a large noise-to-influence ratio.

6.2 Query Filtering

As explained in Section 4.4, query filtering can significantly increase the success of membership inference attacks. We explore its effect by quantifying the attack success while varying the query filtering threshold p_q . We vary p_q from 0 (no filtering) to 0.2 (stringent filtering).

Table 3: Precision and recall for both membership inference attacks for $p_q = 0.2$ and for no filter. In addition, it shows the average number of queries that are deemed useful for both filter thresholds. Note that in the no filter case, we still filter out queries that the target cannot influence.

		Filtering ($p_q = 0.2$)	No Filter
	Number of queries	78	119
LLR	Precision	0.902	0.796
	Recall	0.806	0.88
ML	Precision	0.698	0.649
	Recall	0.957	0.946

Importance of filtering. It can be seen in Table 3, that for the likelihood ratio attack, filtering decreases recall but significantly improves precision, leading to an overall increase in advantage. For the machine learning attack, filtering increases both recall and precision, but has a smaller effect than for the likelihood ratio attack. While both attacks still achieve relatively high advantage without filtering, filtering significantly increases the advantage (i.e., by 0.15

in our experiments). In addition, we find that filtering has an even stronger impact when the auxiliary data is stale or of lower quality. Specifically, filtering increases the advantage by 0.3 when using Texas 2006 as the auxiliary dataset for attacking Texas 2012.

Attack parameter tuning. The adversary controls two parameters: (1) the query filtering threshold p_q , and (2) the IN prediction threshold τ . Since both parameters influence the success of the attack and the trade-off between precision and recall, the adversary should tune the parameters based on her objectives, which could be to maximize precision, maximize recall, or some combination of the two. To explore this experimentally, we perform a grid search on these parameters. Specifically, we take $p_q = 0.0, 0.1, 0.2$ and $\tau = 0.5, 0.9, 0.99, 0.999, 0.9999$.

We find that to maximize precision, the adversary should set $p_q = 0.2$ and $\tau = 0.9999$ (or higher) for the likelihood ratio attack, and $p_q = 0.1$ and $\tau = 0.99$ for the ML attack. In contrast, to maximize recall, the adversary should set $p_q = 0$ (no filter) and $\tau = 0.5$ for the likelihood ratio attack, and $p_q = 0.1$ and $\tau = 0.5$ for ML attack.

While the optimal values for p_q and τ that maximize precision and recall may vary slightly between datasets and attacks, these experiments show how the adversary can tune these parameters to find more targets, or to get more precise results.

6.3 Inclusion Frequency

Consider the assumed inclusion frequency of 50% used in previous membership inference attacks. This assumption may be greatly overestimating attack capabilities. To test this hypothesis, we perform membership inference attacks at inclusion frequencies of 50%, 10%, and 1%. In addition, we examine how varying the ratio threshold τ , for which the likelihood ratio outputs IN (vs. OUT), is able to offset some of the effect of low inclusion rates. When the inclusion frequency is 50%, τ is set to 0.5, but when it is more likely that the target is OUT than IN, the values of τ can be increased arbitrarily to require more evidence before declaring IN. Fig. 2 shows the precision and recall trade-off that can be achieved when varying τ .

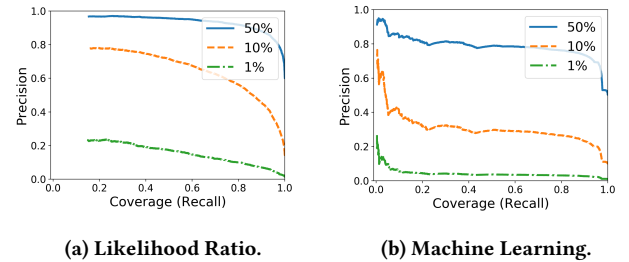


Figure 2: Precision versus recall for the likelihood ratio attack (Fig. 2a) and the machine learning attack (Fig. 2b) for inclusion frequencies 50%, 10%, and 1%.

Results. As shown in Fig. 2, when the inclusion frequency (f) is very low (i.e., 1%), it is difficult, even for an otherwise powerful attack, to correctly infer membership (IN). The IN prediction threshold (τ) can be tuned to compensate for this effect to some degree. However, we find that precision never reaches the same values as

for the 50% inclusion frequency case (even if the adversary completely sacrifices recall). Tuning the IN prediction threshold for the machine learning attack has less of an effect than for the likelihood ratio attack, except for extremely low recall cases. These results suggest that overlooking low inclusion frequencies scenarios can result in overestimating the membership inference risk.

Inclusion frequency and query filtering. We find that query filtering greatly increases the effectiveness of tuning the IN prediction threshold τ , especially in low inclusion frequency scenarios. Without query filtering, increasing τ to 0.999 in the likelihood ratio attack with a 1% inclusion frequency only increases the precision from 0.06 to 0.08. In contrast, setting the query filtering threshold to $p_q = 0.1$, the precision goes from 0.06 to 0.28. Our results suggest that query filtering allows mitigating (to some degree) the effects of a low inclusion frequency, which otherwise makes attacks much less likely to succeed. Although we note that filtering has a greatly reduced effect for the machine learning attack (compared to the likelihood ratio attack), where it only increases precision from 0.02 to 0.04. This is consistent with observations in Section 7.2, that machine learning is unable to significantly mitigate a low inclusion frequency.

6.4 Auxiliary Data Quality

Does the membership inference attack still perform well if the auxiliary data is of poor quality? More generally, we are interested in understanding how auxiliary data quality impacts the attack’s success. For this, we perform an experiment in two parts using the likelihood ratio attack.

First, we purposefully pick old (stale) auxiliary data. Specifically, we use the Texas dataset from 2006q1 as auxiliary data while keeping the 2012q1 Texas dataset to extract the sample D . This means that there is a 6-year gap between the auxiliary data and the sample data distributions. Second, we perform a *cross-dataset* attack using HCUP-NIS data as auxiliary data. Specifically, we use the 2008 HCUP-NIS dataset to attack the Texas 2012q1 dataset. We also reverse this attack by using the Texas dataset as the auxiliary and the HCUP-NIS dataset as the target. It should be noted that for a cross-dataset attack the possible queried attributes are severely reduced, due to different collection methods or one dataset not containing some attributes.

To obtain an objective measure of the distance between two distributions, we use the two-samples Kolmogorov-Smirnov (KS2) test [6] and the Kullback-Leibler divergence [25]. The KS2-test is a nonparametric method to determine whether two distributions (given by their ECDF estimated from the respective samples) differ. We use the KS2-test statistic itself (computed using SciPy’s `ks_2samp` function) as the distortion metric; the range of this metric is $[0, 1]$. The KL-divergence is computed in the usual way, but we add a very small smoothing factor of $\xi = 0.0001$ to all the probabilities to avoid numerical instability. Since the KL-divergence function returns different results depending on which dataset is given as sample data and which is given as auxiliary data, we compute the KL-divergence for both cases and average the two results. (This is sometimes called the Jensen-Shannon divergence.) In both cases, we compute the KS2-test statistic and KL-divergence over 100 samples from the distributions of query outputs (and then average

over all queries). For reference the corresponding distance between Texas 2006 and Texas 2012 in query output distributions yields a KS2-test statistic of 0.44 and a KL divergence of 3.37.

Results. We observe in Table 4 that for the Texas data, even with a 6-year gap, the advantage only decreases by 0.06. This shows that while year gaps do decrease advantage, they do *not* thwart membership inferences. It should be noted, that this small decrease in advantage is partially due to query filtering eliminating queries with too much of a difference between years. Without query filtering the 6-year difference results in a 0.21 decrease in advantage.

Table 4: Advantage of the likelihood ratio attack for varying quality of auxiliary data. To vary the quality, we use different datasets and different years for the sample and auxiliary data. ‘T’ denotes the Texas dataset and ‘H’ denotes HCUP-NIS. The two rightmost columns show *cross-dataset* attacks where the Texas data is used as sample but HCUP-NIS as the auxiliary (or vice-versa).

Sample	T ’12	T ’12	T ’12	T ’12	T ’12	H ’08
Auxiliary	T ’12	T ’11	T ’10	T ’06	H ’08	T ’12
Advantage	0.74	0.74	0.73	0.68	0.32	0.26
KS2-test	0.08	0.15	0.27	0.44	0.49	0.49
KL-divergence	0.10	0.50	0.74	3.37	3.72	3.72

6.5 Vulnerability

Are all targets (similarly) vulnerable to membership inference? (We note that this question has seen some study in related work, e.g., [2, 52].) Intuitively, targets that are “outliers” in the sample, such as individuals with uncommon attribute values (e.g. a rare disease) should be significantly more vulnerable. To explore this, we perform our likelihood ratio and machine learning attacks on 100 randomly chosen targets. We can then explore the vulnerability of targets by considering the empirical CDF of the advantage (over the targets). We also use bootstrap resampling [51] to compute approximate confidence intervals at the 95%-level.

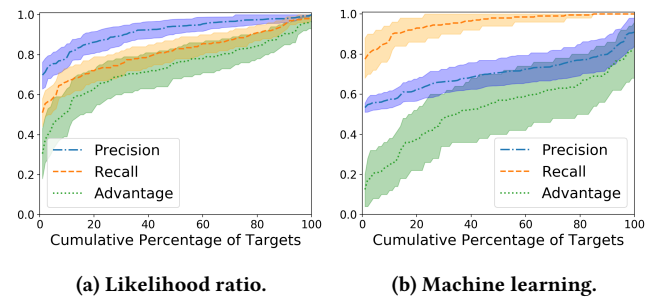


Figure 3: Empirical CDF of attack performance for 100 randomly selected targets. Each data point shows the statistic value with the surrounding shaded region showing the 95%-level confidence interval. There is a significant difference between the most vulnerable target and the least vulnerable target.

Results. The results are shown in Fig. 3, where we observe a large difference in attack advantage between different targets. For the

likelihood ratio attack the difference in advantage between the best and worst target is about 0.67, whereas it is 0.61 for machine learning. Note that given the width of the 95%-confidence intervals, this cannot be explained by inherent randomness in the experiment: *some individuals are truly much more vulnerable to membership inference (than others)*. We explore this further in Section 7.2.

7 CASE STUDY: PAD-RELATED AMPUTATIONS

In this section, we evaluate our attacks on a recently published medical article by Rizzo et al. [40] titled: “Racial Disparities in PAD-Related Amputation Rates among Native Americans and non-Hispanic Whites: an HCUP Analysis.” This article studies the use of amputation in treating peripheral artery disease (PAD) across America. The study finds that (for an unknown reason) hospitals in the western U.S. use amputation on Native Americans at a much higher rate than on non-Hispanic Whites. This presents serious concerns because amputation is a last-resort to deal with PAD-related conditions. Revascularization of the limb is generally preferred since it does not result in the loss of a limb. Difference in the rate of amputation could indicate discrimination. While PAD and its treatment is not necessarily sensitive, if membership is inferred, a patient’s medical history is exposed without their consent, which is a serious breach of privacy. As we explore in Section 7.4, a successful membership inference can also be damaging to the physician that performed the procedure or the hospital where it occurred, since inclusion in this study could indicate discrimination.

7.1 Data Setup and Challenges

Rizzo et al. [40] is a large study that uses multiple years of HCUP NIS data. It examines the relationships between many different attributes and the race of patients. For both non-Hispanic Whites and Native American patients, the study explores patient attributes such as age and gender, severity information such as condition severity and comorbidities, hospital information such as region and number of beds, and the procedures performed. To simulate an attack using data extracted from this paper, we attempt to reproduce the study and in the process reconstruct all reported statistics as queries.

The study uses HCUP NIS data from years 2006 to 2013. Only the 2008, 2012, and 2013 years of the HCUP NIS data were available to us, so we use those to reproduce the study. Unfortunately, the authors do not specify the exact ICD-9-CM diagnoses codes used to filter out the subsets of HCUP NIS patients suffering from PAD-related conditions that were used for the study. They also do not explain what procedure (if any) was used to handle changes in the HCUP NIS sampling methodology during the 8 years of data considered. As a result, we could not reproduce fully the findings reported in Rizzo et al. [40], but we recreated the data analysis of the paper on the available HCUP NIS data using online resources [14] to establish the list of ICD-9-CM codes for PAD-related conditions as well as standard HCUP policies [17]. For reference, we end up with 22,313 patients matching the study’s criteria in the HCUP NIS 2008 data compared to Rizzo et al. [40] who report 19,773 for the same year.

In total, the study contains 132 individual queries, 18 of which we do not use due to attributes not transferring between years and not knowing how the authors processed them. Out of the 114 remaining queries, exactly half (57) target Native American patients whereas the other half target non-Hispanic White patients. The population of patients that match the study criteria is large in contrast to the number of queries and the precision of the reported figures in Rizzo et al. [40]. As a result, we do not expect most targets to be vulnerable (we confirmed this through experiments). However, Native American patients are in a minority (their population varies from 100 to 1500 depending on the number of years considered) and thus may be vulnerable. Focusing on this population, an adversary may use up to 57 queries in her inference attack.

7.2 Membership Inference Attacks

We used both the likelihood ratio attack and the machine learning attack while tweaking various parameters and factors such as inclusion frequency, the query threshold, and the auxiliary dataset.

Experimental Setup. Since we did not have access to all the years used in the paper, we used HCUP NIS 2008 as the auxiliary dataset to attack HCUP NIS 2012 and a sample size of 100, which corresponds with the population of Native American patients in several years.

In addition, we used a query threshold p_q of 0.1, and an inclusion frequency of 50%. In select experiments, we additionally varied the parameters and factors as follow: query threshold $p_q = 0.0, 0.1$, and 0.2 ; inclusion frequency $f = 50\%, 10\%$, and 1% , and we used the 2008, 2012, or 2013 versions of the HCUP NIS data as the auxiliary dataset.

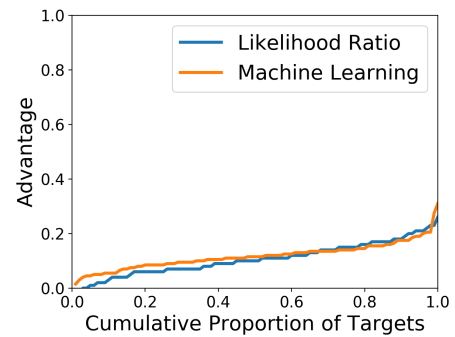


Figure 4: Empirical CDF of attack performance for likelihood ratio and machine learning attacks on a published medical article [40].

Results. The results of the membership inference attacks on Native American targets are shown in Fig. 4. The average advantage of both attacks is 0.11. However, we find that for approximately 7% of targets, the attack achieves an advantage above 0.20. Moreover, we find that this group of highly vulnerable targets are consistently vulnerable (i.e., across different samples of the data). That said, we were unable to discern any specific feature or pattern that would explain why these individuals are more vulnerable than the typical patient matching the study criteria (other than being Native American).

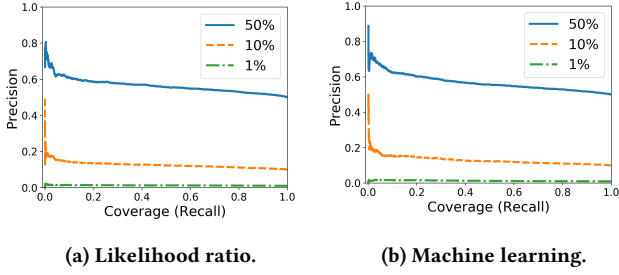


Figure 5: Precision versus recall for the likelihood ratio (Fig. 5a) and machine learning (Fig. 5b) attacks at varying inclusion frequencies.

Factors affecting attack’s success. Upon closer examination of the data, we find that the reasons for the average advantage being low are: (1) the small number of useful queries, and (2) the large difference between the auxiliary and target dataset. While there are 57 potential queries for the attack to use, we find that on average only about 11 of them are actually informative. This is because many of the queries are examining different values of the same attribute, for which the target can only influence one, in addition to query value differences between the target and auxiliary datasets.

We also find that the auxiliary dataset has a significant effect on advantage. While using data from years 2008 and 2013 both lead to attacks with advantages of about 0.11, when two separate subsets of 2012 HCUP NIS data are used as the auxiliary dataset and sample database, the advantage increases to 0.17.

Query filtering and inclusion frequency. Query filtering is essential for the attack. Without it the advantage is 0.07. Figs. 5a and 5b show recall versus precision for varying inclusion frequencies for the likelihood ratio and machine learning attacks. We observe that the inclusion frequency is a significant factor in this attack scenario (even more than it was for attacks in Section 6) due to the overall reduction in attack confidence. This further emphasizes the need to consider inclusion frequency when evaluating membership inference attacks in practice. Our results show that more powerful attacks can mitigate to some extent the effect of low inclusion frequency by tuning the prediction threshold. However, for less powerful attacks, the prediction threshold can often only be tuned a little bit before the attack’s precision goes to zero. Due to this, a low inclusion frequency can make successful inferences (almost) impossible.

7.3 Defenses

We apply two common defenses: differential privacy and small-count suppression.

Differential privacy. To apply differential privacy in this setting we apply noise sampled from $\text{Lap}(\frac{\Delta}{\epsilon})$ independently to each query result. Here Δ is the global sensitivity of the query. (For example, for a query computing a proportion over a dataset of n records, $\Delta = \frac{1}{n}$.) We then perform the likelihood ratio attack. Since we have multiple queries we must calculate the *overall effective privacy budget* ϵ' using composition. Specifically, if m queries are used then

we calculate:

$$\epsilon' = \min \left\{ \epsilon \sqrt{2m \ln \frac{1}{\delta}} + m\epsilon(e^\epsilon - 1), m\epsilon \right\}. \quad (3)$$

Eq. (3) combines advanced composition [10, Theorem 3.20] with sequential composition, where we set $\delta = 10^{-5}$. We vary the *per-query* budget ϵ from 0.05 to 2.0. Table 5 shows the advantage of the likelihood ratio attack for varying values of the *per-query* budget ϵ , and the overall privacy budget ϵ' using Eq. (3). For reference, we also calculate the $e^\epsilon - 1$ upper-bound on the advantage [53]. As a reference for the amount of change that is made to each query result we report the Mean Absolute Percentage Error (MAPE), which shows the percent change with reference to true query value. Specifically, if a query’s output value without the defense is v , but the defense outputs v' instead, then the MAPE is $|\frac{v-v'}{v}|$ scaled as a percentage. (For example, a defense that induces no distortion (i.e., $v' = v$) produces a MAPE of 0%, whereas one that doubles the value (i.e., $v' = 2v$) produces a MAPE of 100%.)

We observe a large gap between the advantage of the attack and the bound. This is consistent with Jayaraman and Evans [22] who observe that the bound is often quite loose in practice. Low ϵ values (i.e., very high privacy) thwart the attack as expected. While the attack accuracy is reduced even at $\epsilon = 1$, it is not until $\epsilon = 0.1$ that the attack starts to fail even for vulnerable targets. At this point, the MAPE is around 87%, which means that on average a deviation of magnitude equal to 87% of the true value has occurred. So it would be difficult (if not impossible) to extract meaningful information from the results.

Table 5: Advantage and utility loss (MAPE) due to achieving differential privacy. We also show the upper-bound on the advantage from Yeom et al. [53]. Due to the number of queries, the effective privacy budget (ϵ') is significant even for a per-query budget (ϵ) of 0.1.

Per-query (ϵ)	0.05	0.1	0.5	1.0	2.0
Overall (ϵ')	0.16	1.1	5.5	11	22
Bound ($e^{\epsilon'} - 1$)	0.174	2.0	243	59873	$\approx 10^9$
Advantage	0	0.01	0.06	0.08	0.11
MAPE	174.68	87.38	17.46	8.75	4.28

Small count suppression. To apply small count suppression we simply omit (i.e., do not use for the attack) queries with values less than a threshold amount. We evaluate results for threshold counts from 1 to 40, although 20 is the largest threshold we have observed being used in practice. Table 6 shows that the effect of small count suppression is relatively small until all queries with values less than 40 are suppressed, at which point most interesting information has been removed.

7.4 Hospital Inference Attacks

In addition to membership inference attacks on individual patients we also performed attacks on the hospitals used to generate the dataset. These attacks represent a serious risk for hospitals involved in the study. The study of Rizzo et al. [40] discovered that hospitals

Table 6: Effects of suppressing small counts. Any query reporting numbers less than the threshold is removed from the attack. The advantage and number of usable queries for each threshold are also shown.

Suppression Threshold	1	5	10	20	40
Advantage	0.11	0.09	0.09	0.07	0.02
Number of Queries	11	10	8	7	3

in the western region of the U.S. performed more amputations on Native Americans than non-Hispanic Whites, compared to hospitals in other regions which performed the procedure at roughly the same rate for both groups. While the study cannot state the exact reason for this disparity, suggested possible reasons include a failure of the Indian Healthcare Service (IHS), differences between the western populations of Native Americans, or discrimination by physicians [40].

Regardless of the reason, this motivates us to consider a possible threat to enterprise privacy. Specifically, if one were to publish a survey of hospitals, could an adversary successfully perform an inference attack to determine membership of a hospital in the survey? Or worse yet, infer whether a given hospital had a higher than normal rate of PAD-related amputation for Native Americans (compared to for non-Hispanic Whites)?

Group membership inference. To perform hospital attacks we first restructured the database so that it consists of hospitals instead of individual records. To do this, we grouped patients by hospital. We then formed test sets of these groups by selecting some number of hospitals and concatenating all the patients in each group together into one set, which we could then query. We performed the two group attacks as described in Section 4.4.

We performed the attacks with 20 hospitals per set, which produced individual patient set sizes of 100–300, which is in the same range as the yearly Native American populations. Aside from this, we performed the attacks using a similar setup as the previous experiments. We used the HCUP NIS 2008 data as the auxiliary dataset to attack a sample database taken from HCUP NIS 2012 data. We used a query threshold p_q of 0.1, and assumed an inclusion frequency of 50%. Note that an inclusion frequency of 50% here is not necessarily unrealistic because a survey of hospitals in a given region could very well include a large proportion of all hospitals in that region.

As Fig. 6a shows, while the average advantage of 0.13, is not much higher than the single membership attacks, a greater portion of targets are much more vulnerable. Over 20% of targets have an advantage greater than 0.20, with some even reaching 0.50, compared to about 10% of target reaching above 0.20 advantage in the previous attacks. These attacks are especially threatening for larger hospitals with a greater number of patients, since they have a much higher influence.

Attribute inference. We performed an attribute inference attack on the hospital’s rate of performing amputations on PAD patients. Specifically, the goal of the attack is to infer if the rate for a given hospital is greater than or less than 10%. This attack could be even more damaging than the group membership inference, as it could

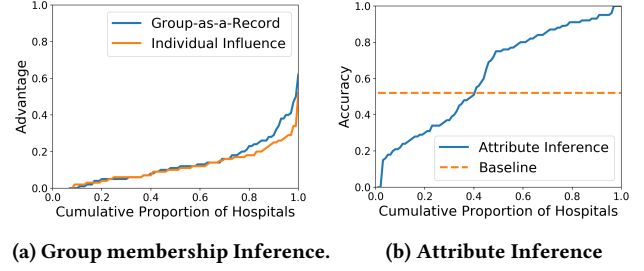


Figure 6: Empirical CDF of hospital inference attack performance for both group attacks, and for the attribute inference attack. Group-as-a-Record is slightly more powerful than the individual inference attack, but both group attacks represent a privacy risk. The attribute inference attack performs significantly better than the baseline accuracy

directly point to some hospital as having a worse overall standard of care, or suggest that discrimination may be occurring. This attack could not be performed on the initial set of queries because only 2 queries actually showed this rate. To offset this we added 17 queries that showed the relationship between hospital amputation rate and other attributes, such as age, sex, insurance, income quartile, hospital size, and hospital location/teaching status. We believe that augmenting the set of queries this way is not unrealistic as all of these relationships and more were stated to be explored by the authors, but the corresponding results were not reported by Rizzo et al. [40]. As baseline for this attribute inference attack, we assume a naive adversary that always guesses the most likely value.

The results are shown in Fig. 6b. The attribute inference attack, with an average accuracy of 0.64, performs significantly better than the baseline (accuracy 0.52). In addition, many targets (hospitals) are extremely vulnerable with over 20% of targets having an accuracy greater than 90%.

8 DISCUSSION

In this section, we discuss a few keys findings in the broader context of inference attacks.

Query filtering is critical. When queries are fixed (and not chosen by the adversary), selecting an optimal set of queries is essential. Our experimental findings suggest that query filtering significantly improves attack accuracy in almost all cases and is particularly critical whenever the quality of the auxiliary data is less than ideal, as well as in any low inclusion frequency scenarios (e.g., $f \ll 50\%$).

Inclusion frequency should be varied and reported. All of our experimental results show that, even with proper parameter tuning, membership inferences are less successful in low inclusion frequency scenarios. Prior work on membership inference has overlooked inclusion frequency as a factor and almost exclusively evaluated the case $f = 50\%$. To obtain an accurate assessment of attacks’ success, future research must consider evaluation using more realistic inclusion frequencies.

Attacks with low quality auxiliary data can succeed. Inference attack success depends on the quality of the adversary’s auxiliary data. That said, we found that attacks can still be largely successful with stale auxiliary data or even auxiliary data from an unrelated dataset. We hope that future research on membership inference will consider scenarios with less than ideal auxiliary data.

Defenses do work, but at a significant cost. We find that both principled defenses such as differential privacy and heuristic defenses such as suppressing small count queries are effective at mitigating membership inferences when tuned adequately. However, this comes at a significant cost in terms of utility. In particular, to completely eliminate our membership inference attack on Rizzo et al. [40], the overall differential privacy budget needs to be calibrated to $\epsilon = 1$ or less, which leads to a mean absolute percentage error of over 80% on query results.

Some individuals are more vulnerable than others. Our findings show that in a realistic setting, membership inference attacks are much less likely to succeed than what may be inferred from prior work’s results. However, this does not mean they never succeed. Indeed, we discover a subset of targets whose members are consistently and unavoidably vulnerable to inference attacks. In one case, we find that these vulnerable individuals are members of minority groups. This is consistent with findings from related work by Yaghini et al. [52] in the machine learning models setting.

Publishing aggregate surveys data is *not* always safe. Even in scenarios where the risk to individuals’ privacy is low, publishing aggregate statistics without mitigation poses a serious risk. We propose and apply group membership inference attacks targeting hospitals that show that inference attacks can pose a significant threat to enterprise privacy.

9 CONCLUSIONS

We study the problem of adversarial inferences in the aggregate survey data setting. By applying existing membership inference attacks, we uncover challenges of this often overlooked setting and we discover unrealistic methodological assumptions at the root of related work. Through extensive experiments on real-world data and targeting an actual published academic study, we systematically relax these assumptions and quantify the success of inference attacks. We believe that our findings can provide a baseline measure of the privacy risk that can be used to help guide future research as well as practitioners who publish aggregate survey data.

REFERENCES

- [1] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. 2016. Membership privacy in MicroRNA-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 319–330.
- [2] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*. 15453–15462.
- [3] Raef Bassily, Alex Groce, Justin Katz, and Adam Smith. 2013. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 439–448.
- [4] S. Chang and C. Li. 2018. Privacy in Neural Network Learning: Threats and Countermeasures. *IEEE Network* 32, 4 (July 2018), 61–67. <https://doi.org/10.1109/MNET.2018.1700447>
- [5] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.
- [6] Donald A Darling. 1957. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics* 28, 4 (1957), 823–838.
- [7] CDC Division of Cancer Prevention and Control. 2019. United States Cancer Statistics (USCS). https://www.cdc.gov/cancer/uscs/technical_notations/stat_methods/suppression.htm.
- [8] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*.
- [9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. 265–284.
- [10] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- [11] Cynthia Dwork and Guy N. Rothblum. 2016. Concentrated Differential Privacy. *CoRR abs/1603.01887* (2016). [arXiv:1603.01887](https://arxiv.org/abs/1603.01887) <http://arxiv.org/abs/1603.01887>
- [12] Peter J Fagenholz, Carlos Ferna, N Stuart Harris, Andrea J Pelletier, Carlos A Camargo Jr, et al. 2007. Direct medical costs of acute pancreatitis hospitalizations in the United States. *Pancreas* 35, 4 (2007), 302–307.
- [13] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. 2011. Privacy preserving GWAS data sharing. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 628–635.
- [14] For the Record. [n.d.]. Coding for Peripheral Artery Disease. <https://www.fortherecordmag.com/archives/101110p28.shtml>.
- [15] Johannes Gehrke, Michael Hay, Edward Lui, and Rafael Pass. 2012. Crowdbinding privacy. In *Advances in Cryptology—CRYPTO 2012*. Springer, 479–496.
- [16] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies* 2019, 1 (2019), 133–152.
- [17] HCUP. [n.d.]. Appendix A - Clinical Classification Software-DIAGNOSES. <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt>.
- [18] Healthcare Cost and Utilization Project (HCUP). 2019. HCUP Nationwide Inpatient Sample (NIS). <https://www.hcup-us.ahrq.gov/nisoverview.jsp>.
- [19] Sheleika L Hervey, Harriett R Purves, Ulrich Guller, Alison P Toth, Thomas P Vail, and Ricardo Pietrobon. 2003. Provider volume of total knee arthroplasties and patient outcomes in the HCUP-Nationwide Inpatient Sample. *JBS* 85, 9 (2003), 1775–1783.
- [20] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. 2008. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genetics* 4, 8 (2008), e1000167. <https://doi.org/10.1371/journal.pgen.1000167>
- [21] Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2013. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 1141–1152.
- [22] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1895–1912.
- [23] Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. 2020. Revisiting Membership Inference Under Realistic Assumptions. *arXiv preprint arXiv:2005.10881* (2020).
- [24] Aaron Johnson and Vitaly Shmatikov. 2013. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1079–1087.
- [25] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
- [26] Joyce C LaMori, Omar Shoheiber, Kellie Dudash, Concetta Crivera, and Samir H Mody. 2014. The economic impact of acute coronary syndrome on length of stay: an analysis using the Healthcare Cost and Utilization Project (HCUP) databases. *Journal of medical economics* 17, 3 (2014), 191–197.
- [27] Jaewoo Lee and Chris Clifton. 2012. Differential identifiability. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1041–1049.
- [28] Kristin M Lefebvre and Julia Chevan. 2015. The persistence of gender and racial disparities in vascular lower extremity amputation: an examination of HCUP-NIS data (2002–2011). *Vascular Medicine* 20, 1 (2015), 51–59.
- [29] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyu Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. 2018. Understanding Membership Inferences on Well-Generalized Learning Models. *CoRR abs/1802.04889* (2018). [arXiv:1802.04889](https://arxiv.org/abs/1802.04889) <http://arxiv.org/abs/1802.04889>
- [30] Yunhui Long, Lei Wang, Diyu Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2020. A Pragmatic Approach to Membership Inferences on Machine Learning Models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE.
- [31] Edward Lui and Rafael Pass. 2015. Outlier Privacy. In *Theory of Cryptography*. Springer, 277–305.
- [32] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 634–646.

- [33] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 739–753.
- [34] Henry S. Park, Frank C. Detterbeck, Daniel J. Boffa, and Anthony W. Kim. 2012. Impact of Hospital Volume of Thoracoscopic Lobectomy on Primary Lung Cancer Outcomes. *The Annals of Thoracic Surgery* 93, 2 (2012), 372 – 379. <https://doi.org/10.1016/j.athoracsur.2011.06.054>
- [35] Adriana Perez, H SHELTON BROWN III, and Blanca I Restrepo. 2006. Association between tuberculosis and diabetes in the Mexican border and non-border regions of Texas. *The American journal of tropical medicine and hygiene* 74, 4 (2006), 604–611.
- [36] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2018. Knock knock, who's there? Membership inference on aggregate location data. *NDSS* (2018).
- [37] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2019. Under the Hood of Membership Inference Attacks on Aggregate Location Time-Series. *arXiv preprint arXiv:1902.07456* (2019).
- [38] Rhode Island Department of Health. 2016. Rhode Island Small Numbers Reporting Policy. <https://health.ri.gov/publications/policies/SmallNumbersReporting.pdf>.
- [39] Taylor S Riall, Deepthi M Reddy, William H Nealon, and James S Goodwin. 2008. The effect of age on short-term outcomes after pancreatic resection: a population-based study. *Annals of surgery* 248, 3 (2008), 459.
- [40] John A Rizzo, Jie Chen, Chad Laurich, Angelo Santos, Brad J Martinsen, Michael P Ryan, Harry Kotlarz, and Candace Gunnarsson. 2018. Racial disparities in PAD-related amputation rates among Native Americans and non-Hispanic whites: An HCUP analysis. *Journal of health care for the poor and underserved* 29, 2 (2018), 782–800.
- [41] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed Systems Security Symposium 2019*. Internet Society.
- [42] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. 2009. Genomic privacy and limits of individual detection in a pool. *Nature Genetics* 41, 9 (2009), 965–967. <https://doi.org/10.1038/ng.436>
- [43] Reza Shokri, Martin Strobel, and Yair Zick. 2020. Exploiting Transparency Measures for Membership Inference: a Cautionary Tale. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)*. AAAI, Vol. 13.
- [44] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.
- [45] Benjamin C Sun, Jennifer A Emond, and Carlos A Camargo Jr. 2005. Direct medical costs of syncope-related hospitalizations in the United States. *The American journal of cardiology* 95, 5 (2005), 668–671.
- [46] Texas Health Care Information Collection (THCIC). 2020. Texas Inpatient Public Use Data File (PUDF). <https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm>.
- [47] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2018. Towards Demystifying Membership Inference Attacks. *CoRR* abs/1807.09173 (2018). [arXiv:1807.09173](https://arxiv.org/abs/1807.09173) <http://arxiv.org/abs/1807.09173>
- [48] Bo Wang, Zhenyu Liu, Jiaojiao Liu, Zhenchao Tang, Hongjun Li, and Jie Tian. 2016. Gray and white matter alterations in early HIV-infected patients: Combined voxel-based morphometry and tract-based spatial statistics. *Journal of Magnetic Resonance Imaging* 43, 6 (2016), 1474–1483.
- [49] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 534–544.
- [50] Jacob Wolfowitz. 1954. Generalization of the theorem of Glivenko-Cantelli. *The Annals of Mathematical Statistics* 25, 1 (1954), 131–138.
- [51] Michael Wood. 2005. Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods* 8, 4 (2005), 454–470.
- [52] Mohammad Yaghini, Bogdan Kulynych, and Carmela Troncoso. 2019. Disparate Vulnerability: on the Unfairness of Privacy Attacks Against Machine Learning. *arXiv preprint arXiv:1906.00389* (2019).
- [53] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 268–282.
- [54] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. 2014. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of biomedical informatics* 50 (2014), 133–141.
- [55] Bo Zhang, Ruotong Yu, Haipei Sun, Yanying Li, Jun Xu, and Hui Wang. 2020. Privacy for All: Demystify Vulnerability Disparity of Differential Privacy against Membership Inference Attack. *arXiv preprint arXiv:2001.08855* (2020).
- [56] Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng Wang. 2011. To Release or Not to Release: Evaluating Information Leaks in Aggregate Human-Genome Data. In *Computer Security – ESORICS 2011*, Vijay Atluri and Claudia Diaz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 607–627.