Revisiting Assumptions for Membership Inference on Summary Statistics

Pascal Berrang
University of Birmingham
p.p.berrang@bham.ac.uk

Mark Ryan
University of Birmingham
m.d.ryan@bham.ac.uk

Kiera Wooldridge
University of Birmingham
kjw522@student.bham.ac.uk

Abstract

Research studies routinely publish summary statistics such as means and standard deviations to promote transparency while protecting participant privacy. Membership inference attacks (MIAs) can exploit these statistics to determine whether a specific individual contributed to a study, posing a risk especially in biomedical and health-related settings. However, existing attacks assume the adversary holds the *exact* data used in the study, an assumption that rarely holds when data evolves over time. Moreover, prior work has not quantified how much of the reported accuracy stems from true individual identification rather than from group-level traits shared within disease cohorts.

We investigate the robustness and interpretability of two standard attacks—the L_1-distance test and the log-likelihood ratio (LLR) test—under realistic conditions where the adversary has only noisy, partial, or temporally mismatched data. We derive a theoretical lower bound on inference error that cleanly separates a statistical term governed by pool size and feature dimensionality from a signal term capturing disease-driven shifts. Empirical evaluation on cross-sectional and longitudinal miRNA datasets, validated on Fitbit activity data, confirms that both attacks tolerate substantial noise and missing features, but that real-world temporal drift degrades accuracy far more steeply than synthetic perturbations predict, and that this degradation is individual-specific. We further show that attack accuracy on disease-specific cohorts exceeds that on size-matched random pools by approximately 10%, a separation that grows almost threefold when measured by true-positive rate at 1% false-positive rate. Moreover, individuals sharing disease traits but absent from the study are frequently misclassified as members, indicating that a substantial component of reported accuracy reflects shared condition rather than individual membership.

1 Introduction

Large-scale studies in health, genomics, and behavioral science routinely publish *summary statistics*—such as means and standard deviations—rather than individual-level records. This practice aims to balance scientific transparency with participant privacy and is standard in fields ranging from genome-wide association studies to wearable-sensor research. However, a line of work on membership inference attacks (MIAs) has shown that even such aggregate information can be exploited: given a target individual's data and the published statistics of a study cohort, an adversary can infer whether that individual contributed to the study [2, 20, 46]. The resulting privacy risks are significant, particularly in biomedical research where study participation may reveal sensitive health conditions or genetic predispositions, and where legal frameworks such as the GDPR and HIPAA impose strict protections on participant data.

These prior results, however, rest on assumptions that limit their practical interpretability. Existing attacks typically assume that the adversary possesses the *exact* datapoint used in the study. This assumption rarely holds for data that evolves over time, such as gene expression profiles, which fluctuate with age, health, and circadian rhythms [57], or daily activity measurements from wearable sensors. An adversary who obtains a later or independently collected sample may face substantial discrepancies with the original study data.

A second limitation concerns the *interpretation* of attack success. Prior work has noted that attacks perform better on disease-specific cohorts than on random pools [2], but the extent to which reported accuracy reflects shared group-level traits rather than true individual identification has not been quantified. Without disentangling these two sources of signal, it is difficult to assess the actual privacy risk posed by publishing summary statistics.

In this work, we investigate the robustness and interpretability of membership inference attacks on summary statistics under realistic conditions. We focus on two widely studied attacks, the L_1-distance test [20] and the log-likelihood ratio (LLR) test [2], and systematically evaluate their performance when the adversary has only noisy, incomplete, or temporally mismatched knowledge of the target individual. We complement this empirical investigation with a theoretical analysis that characterizes the fundamental limits of membership inference as a function of pool size, feature dimensionality, and disease-driven shifts in feature distributions. Our evaluation spans cross-sectional and longitudinal microRNA (miRNA) expression datasets as well as a Fitbit activity dataset, covering both molecular and behavioral data modalities.

Contributions. Our contributions are as follows:

- We evaluate membership inference attacks under **relaxed adversarial assumptions**, showing that both L_1 and LLR tests tolerate substantial noise and missing features, but with complementary weaknesses: the L_1 test resists noise that inflates variance estimates, while the LLR retains its advantage when features are missing. An adversary who can estimate their own noise regime could select the superior test at each operating point.
- We demonstrate, using longitudinal data, that **real-world temporal drift** degrades both attacks far more steeply than synthetic perturbations predict. Controlled ablations show that this degradation is individual-specific: drift vectors

Proceedings on Privacy Enhancing Technologies 2026(4), 542–560

https://doi.org/10.56553/popets-2026-0134

from even the most similar individuals, or predicted from population-level models, fail to reproduce it.
- We derive a **theoretical lower bound** on the minimum inference error for any statistical test, expressed in terms of pool size n, feature count m, and per-feature distributional shifts δ_j. The bound cleanly separates a statistical indistinguishability term from a disease-signal term, providing practitioners with a way to estimate the inherent difficulty of membership inference for their specific setting.
- We **quantify the contribution of group-level signal** to reported attack accuracy on summary statistics. Across all 19 disease groups, case-pool accuracy exceeds that of size-matched random pools by a mean of approximately 0.10 AUC. Individuals sharing disease traits but absent from the study are frequently misclassified as members.

Takeaway. Together, these results suggest that the privacy risk from publishing summary statistics, while real, is more nuanced than prior evaluations indicate: attacks are robust to imperfect adversarial knowledge, yet much of their reported accuracy reflects shared group-level traits rather than true individual identification.

Paper outline. The remainder of the paper is organized as follows. Section 2 introduces background on miRNA data and existing attacks; Section 3 formalizes the threat model; Section 4 derives the theoretical bound; Section 5 presents the empirical evaluation; Section 6 surveys related work; and Section 7 concludes.

2 Background

This section introduces the concepts underlying our analysis. We first describe microRNAs (miRNAs), which serve as our primary case study. We then discuss the role of summary statistics in research data sharing, and finally review the two membership inference attacks that our evaluation builds on.

2.1 MicroRNAs

MicroRNAs (miRNAs) are small, non-coding RNA molecules that regulate gene expression by binding to messenger RNA (mRNA) and inhibiting protein translation. They influence approximately 60% of human protein-coding genes, with over 5,000 miRNAs identified in humans [15]. Because of this broad regulatory role, miRNAs have been implicated in a wide range of diseases, including neurodegenerative disorders, cardiovascular diseases, diabetes, and cancer [14, 21, 23, 29, 42, 56].

A **miRNA expression profile** quantifies miRNA activity in a given cell or tissue at a specific point in time, typically via quantitative PCR, microarrays, or next-generation sequencing. Researchers increasingly use these profiles as biomarkers for disease diagnosis and prognosis, particularly through non-invasive blood-based assays [21, 35]. However, miRNA expression levels naturally fluctuate over time due to age, diet, circadian rhythms, and general health [57]. This temporal variability introduces noise into expression data and complicates longitudinal tracking, a challenge we formalize in our threat model (Section 3).

2.2 Summary Statistics in Research

In biomedical and health-related studies, researchers commonly publish *summary statistics*—such as means, standard deviations, or percentiles—rather than individual-level data. These aggregates provide a concise representation of the dataset and are widely reported in scientific publications and collaborative projects. Although intended to preserve privacy, summary statistics still capture meaningful group-level differences and often form the basis of statistical comparisons between study cohorts, making them precisely the information an adversary could exploit to infer study membership.

Such statistics arise across diverse data modalities. In miRNA datasets, they describe distributions of expression levels across individuals; in wearable-sensor datasets such as Fitbit, they describe activity patterns across participants. In both settings, the tension between transparency, reproducibility, and participant privacy motivates the membership inference attacks we review next.

2.3 Membership Inference against Summary Statistics

We build on prior work [2, 20] and focus on two widely used membership inference attacks against summary statistics: one based on the L_1-distance and the other on the log-likelihood ratio (LLR). By the Neyman–Pearson lemma, the LLR is the most powerful test under the Gaussian and independence assumptions, which makes it the natural benchmark for our analysis. Because both attacks draw only on per-feature pool and population statistics, any other attack restricted to the same published quantities would face comparable degradation under noise and temporal drift. Throughout this section, we follow the notation of [2].

In the standard membership inference setting, an adversary aims to determine whether a given target individual's datapoint $v \in \mathbb{R}^m$ belongs to a particular dataset (the *pool*), given access to summary statistics of both the pool and a broader *population*. The adversary compares how well the target's data aligns with each group, framing the comparison as a statistical hypothesis test.

Both attacks test the same null and alternative hypotheses:

- **Null hypothesis H_0:** The target is not part of the pool (their data is drawn from the population).
- **Alternative hypothesis H_1:** The target is part of the pool (their data is drawn from the pool distribution).

The test statistic, based either on distance or likelihood, determines whether to reject H_0 in favor of H_1.

Notation. For an individual $x \in \mathbb{R}^m$, let x_j denote their value for feature $j \in \{1, \ldots, m\}$ (e.g., the expression level of miRNA j, or an activity measurement in the Fitbit dataset). We denote by μ_j and σ_j the mean and standard deviation of feature j across the *population*, and by $\hat{\mu}_j$ and $\hat{\sigma}_j$ the corresponding values in the *pool*.

2.3.1 L_1-Distance. The L_1-based attack compares, feature by feature, how close the target v's values are to the population mean versus the pool mean [2, 20]. For each feature j, the per-feature distance difference is:

$$D(v_j) = |v_j - \mu_j| - |v_j - \hat{\mu}_j|.$$

A positive value indicates that the target is closer to the pool than to the population, suggesting membership; a negative value

suggests non-membership; and zero indicates equal distance. To reach a membership decision, we aggregate these differences into the vector $\Delta^v = (D(v_1), \ldots, D(v_m))$ and test whether its mean is significantly greater than zero.

Under the assumption that features are independent and identically distributed, the Central Limit Theorem justifies approximating the sample mean $\bar{\Delta}^v$ as normally distributed [44]. We therefore apply a one-sided one-sample Student's t-test with test statistic:

$$T_v = \frac{\bar{\Delta}^v}{s/\sqrt{m}},$$

where s is the sample standard deviation of Δ^v and m is the number of features. If T_v exceeds a chosen threshold t, the adversary rejects the null hypothesis and infers membership [2, 20].

2.3.2 Likelihood Ratio Test. The log-likelihood ratio (LLR) test is a classical approach to binary hypothesis testing. By the Neyman–Pearson lemma [37], it achieves the maximum possible power at any fixed false-positive rate, making it the strongest possible attack in this framework [2].

Under the assumption that feature values are normally distributed and independent, the LLR compares how likely the target's feature vector is under each distribution. The test statistic decomposes as a sum over features:

$$LLR(v) = \sum_{j=1}^{m} \left(\frac{(v_j - \mu_j)^2}{2\sigma_j^2} - \frac{(v_j - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2} + \log \frac{\sigma_j}{\hat{\sigma}_j} \right).$$

Each term reflects whether feature j is better explained by the population or the pool distribution. A higher LLR score indicates that the target's data is more probable under the pool, supporting the membership hypothesis.

In practice, the assumptions of independence and normality are approximate. For miRNA expression, Backes et al. [2] report that roughly 60% of features can be treated as independent, and in our cross-sectional dataset 82% of miRNAs pass a normality test. These figures make the LLR a credible benchmark for bounding attack performance, but they also imply that departures from the assumed model are inevitable, a point whose practical consequences we investigate in Section 5.

3 Threat Model

Building on the attacks described in Section 2.3, we now formalize the adversarial setting that the remainder of this paper evaluates. The key departure from prior work is that we do not require the adversary to hold the exact data used in the study, a relaxation that reflects realistic data-acquisition scenarios and substantially affects attack performance.

Setting. An adversary seeks to determine whether a target individual contributed to a study whose summary statistics have been published. Means and standard deviations are the summary statistics most commonly reported in biomedical publications [1, 10, 11, 16, 19, 24, 25, 27, 28, 32, 34, 43, 47, 59, 61], and hence most relevant to realistic disclosure scenarios.

We represent each individual as a feature vector $v \in \mathbb{R}^m$ and denote the study cohort as the *pool* Pool $\subset \mathbb{R}^m$. The pool is distinct from the general population Pop and typically represents a group sharing a condition such as a disease, though our model applies equally to arbitrary subgroups. The adversary observes two sources of information: (i) the mean (and optionally the standard deviation) of each feature in both Pool and Pop, which may be published, estimated from public data, or leaked; and (ii) a sample of the target's data of the same type. Pool membership is sensitive because studies routinely collect information beyond these accessible features (such as detailed diagnoses, treatment assignments, or longitudinal health records) enabling the adversary to link the individual to richer data elsewhere in the study [22]. Small pools are the most vulnerable in this setting: as n grows, the attacker's minimum-error bound rises and AUC falls monotonically, so a large n acts as a natural defense [2].

Relaxation: noisy and partial adversarial knowledge. Prior attacks assume that the adversary possesses the *exact* datapoint used in the study [2, 20]. This assumption is unrealistic for data that evolves over time or varies across measurement conditions. MiRNA expression profiles, for instance, fluctuate with age, health, and circadian rhythms [57]; daily activity records from wearable sensors shift with seasonal and behavioral changes. An adversary who obtains a later or independently collected sample will therefore face discrepancies with the original study data.

We accordingly do not require the adversary's sample to match the exact datapoint used in the study; instead, the sample in (ii) may be *noisy* (due to natural variation or measurement error) or *incomplete* (with missing features), reflecting a broad range of realistic acquisition scenarios. This relaxation subsumes the exact-data setting of prior work as a special case.

4 Theoretical Analysis

Before turning to our empirical evaluation, we establish a theoretical lower bound on the error of any membership inference test based on summary statistics. This bound serves two purposes. First, it characterizes how pool size n and feature dimensionality m influence the statistical difficulty of membership inference, providing context for interpreting the robustness experiments in Section 5. Second, it cleanly separates this statistical term from a disease-signal term driven by per-feature distributional shifts, revealing when reported attack accuracy may reflect group-level traits rather than individual identification.

We model both the pool and the target as draws from multivariate Gaussian distributions and allow a subset of features to differ between the pool and the population due to a condition-related mean shift. This modeling choice is motivated by the empirical properties of our datasets discussed in Section 2.3 and by the tractability of Gaussian distributions for deriving closed-form bounds. While real data may deviate from strict normality, the model captures the key structural effects (mean shifts, variance scaling, and feature independence) that govern attack feasibility.

Our analysis focuses on inference from *means* alone, motivated by the original L_1-based test. The results extend to the LLR test when per-feature variances are equal between the population and the pool (i.e., $\hat{\sigma}_j^2 = \sigma_j^2$), since the LLR test statistic then depends only on mean proximity.

The hypotheses are:

- H_0: The individual $v \notin S$; the pool mean $\hat{\mu}$ is independent of v.
- H_1: The individual $v \in S$; the pool mean includes v, introducing a dependence.

Theorem 4.1 (Membership Inference Bound). *Let P and Q denote the joint distributions of $(v, \hat{\mu})$ under H_1 and H_0 respectively, where $v \in \mathbb{R}^m$ has components $v_j \sim \mathcal{N}(\mu_j + \delta_j, \sigma_j^2)$ under H_1 and $v_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ under H_0, and the pool mean $\hat{\mu}$ is computed over n individuals with each $x_j \sim \mathcal{N}(\mu_j + \delta_j, \sigma_j^2)$.*

Then, the minimum achievable error for any test distinguishing H_1 from H_0 satisfies:

$$\text{min error} \geq \frac{1}{2}\left(1 - \sqrt{\frac{m}{4}\log\left(\frac{n}{n-1}\right) + \frac{1}{4}\sum_j \frac{\delta_j^2}{\sigma_j^2}}\right).$$

The bound is independent of the population distribution itself; it depends only on the joint distribution of $(v, \hat{\mu})$. The full proof appears in Appendix A.

Assumptions and Tightness of Theorem 4.1

Gaussian features. Each feature j of each individual is drawn from a univariate Gaussian: population members from $\mathcal{N}(\mu_j, \sigma_j^2)$, pool members from $\mathcal{N}(\mu_j + \delta_j, \sigma_j^2)$, where $\delta_j = 0$ for unaffected features.
Feature independence. Features $j = 1, \ldots, m$ are mutually independent.
Equal variances. The variance σ_j^2 is the same in the pool and the population for each feature j.
Equal priors. H_0 and H_1 are assumed equally likely. In practice, membership probability is very small, which only increases the adversary's Bayes-optimal error. The equal-prior setting therefore represents a best case for the adversary and yields a conservative (privacy-pessimistic) bound.

Tightness. The proof applies Pinsker's inequality as its sole relaxation, which is tight up to a constant factor when $D_{\mathrm{KL}}(P\|Q)$ is moderate [6]. For large divergences ($D_{\mathrm{KL}} > 2$), which arise when the disease signal is strong or the pool very small, Pinsker's upper bound on total variation exceeds 1 and becomes vacuous; in this regime the Bretagnolle–Huber inequality $\mathrm{TV}(P, Q) \leq \sqrt{1 - e^{-D_{\mathrm{KL}}(P\|Q)}}$ provides a tighter bound [4]. This distinction matters little for our conclusions: the large-divergence regime corresponds to settings where the minimum error is near zero regardless of the inequality used.

Relating the bound to empirical metrics. The minimum error is the lowest misclassification rate achievable by any test under equal priors on H_0 and H_1. Our empirical evaluation, by contrast, reports the area under the ROC curve (AUC), which is equivalent to the probability that a randomly chosen member receives a higher test score than a randomly chosen non-member [13]. The two metrics are related monotonically: a minimum error of $\frac{1}{2}$ implies AUC = 0.5 (random guessing), while a minimum error of 0 is consistent with AUC = 1.0 (perfect separation). The bound therefore constrains empirical performance: when the minimum error is close to $\frac{1}{2}$, no attack can achieve AUC substantially above 0.5; when it is close to 0, high AUC is possible in principle but not guaranteed for any particular attack. We use this correspondence throughout Section 5 to compare observed AUC values against the theoretical limits.

4.1 Special Cases

We now instantiate the bound for two scenarios that correspond to different relationships between the pool and the population.

No distributional shift ($\delta_j = 0$ for all j). When the pool is statistically indistinguishable from the general population, reflecting a baseline in which disease has no measurable impact on the features of interest, the bound simplifies to:

$$\text{min error} \geq \frac{1}{2}\left(1 - \sqrt{\frac{m}{4}\log\left(\frac{n}{n-1}\right)}\right).$$

In this setting, any successful inference must exploit statistical noise and small-sample effects alone: as the pool size n grows, the contribution of any single individual to the pool mean diminishes, and the minimum error approaches $\frac{1}{2}$ (random guessing). Because the bound depends only on the joint distribution of $(v, \hat{\mu})$ and not on the population, this case also covers the scenario in which both v and the pool are drawn from the same *diseased* distribution; the relevant question is whether the pool distribution differs from the population, not whether either is "healthy."

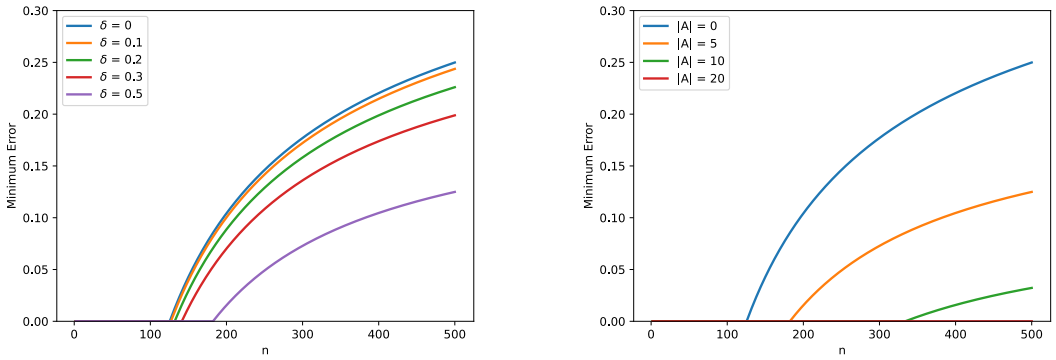
To place this result in context, Theorem 2 of Backes et al. [2] relates the power β of a one-sided t-test to the false-positive rate α as $z_\alpha + z_{1-\beta} \approx \sqrt{2m}/n$. Their result characterizes a *specific* test statistic under idealized conditions; Theorem 4.1, by contrast, bounds the performance of *any* test, regardless of its form.

Condition-driven shift in a small feature subset. The more realistic scenario arises when the pool differs from the population in a small number of features due to disease-related shifts. Let $A \subseteq \{1, \ldots, m\}$ with $|A| \ll m$ denote the affected features, so that $\delta_j \neq 0$ for $j \in A$ and $\delta_j = 0$ otherwise. Such small $|A|/m$ ratios are characteristic of biomedical data. In our cross-sectional dataset ($m = 466$ features after preprocessing), we identified disease-associated miRNAs from the literature for 19 conditions [1, 8, 10, 11, 16, 18, 19, 24, 25, 27, 28, 30, 32–34, 36, 38, 43, 47, 49, 52–54, 58, 59, 61] The number of affected features $|A|$ ranges from 1 (e.g., prostate cancer, sarcoidosis) to 47 (melanoma), yielding $|A|/m$ ratios between 0.002 and 0.10. The per-disease counts are tabulated in Appendix B, Table 4.

The full bound, including the disease-signal term $\frac{1}{4}\sum_{j\in A} \delta_j^2/\sigma_j^2$, now applies. Three factors increase the minimum error, and thus make inference harder: larger per-feature variance σ_j^2, smaller shifts δ_j, and fewer affected features $|A|$. When all three are combined, the bound approaches that of the no-shift case, and the attack offers little advantage over random guessing.

4.2 Interpretation

Figure 1a shows how the minimum error varies with pool size n and shift magnitude δ, for $m = 500$, $|A| = 5$, and $\sigma^2 = 1$. Even minimal disease signals (e.g., $\delta = 0.1$) substantially reduce the minimum error for small pools: below $n = 100$, the bound equals zero, indicating that accurate inference is feasible in principle. As n



(a) $m = 500, |A| = 5, \sigma^2 = 1$ **(b)** $m = 500, \delta = 0.5, \sigma^2 = 1$

Figure 1: Minimum error rates based on Theorem 4.1 for varying parameters.

increases, however, each individual's contribution to the pool mean is diluted, and the minimum error rises toward $\frac{1}{2}$.

Figure 1b varies the number of affected features $|A|$ instead, confirming that the disease-signal term can easily dominate the statistical term. Even when only a handful of features differ from the population, the minimum error remains low for small pools.

These results yield two insights for interpreting our empirical evaluation. First, the statistical term $\frac{m}{4}\log\left(\frac{n}{n-1}\right)$ establishes a *fundamental baseline*: for large pools, even high-dimensional feature spaces are insufficient for reliable detection, regardless of how much auxiliary information the adversary holds. Observed degradation under noise or missing features (Section 5) can be compared against this limit. Second, the disease-signal term $\frac{1}{4}\sum_j \delta_j^2/\sigma_j^2$ reveals that even modest condition-driven shifts in a handful of features can dominate the bound, suggesting that high attack accuracy on disease-specific cohorts may largely reflect shared group-level traits rather than individual membership.

An important caveat applies: because the bound is a *lower* bound on error, a near-zero value does not rule out accurate inference *in principle*. It also does not guarantee that any specific attack achieves this performance; Section 5 tests whether the L_1 and LLR attacks approach these limits in practice.

Theoretical Findings

Our bound on minimum inference error separates two terms: a *statistical* term (pool size and feature dimensionality) and a *signal* term (condition-related shifts in a small subset of features). Even small shifts in a few features can drive the bound to near-zero for small pools, meaning high attack accuracy need not imply individual identification.

5 Evaluation

We now turn from theoretical limits to empirical performance. Our experiments proceed in two stages. We first stress-test the L_1 and LLR attacks under degraded adversarial knowledge (noise, missing features and temporal drift) to establish how robust they are in practice (Section 5.3). We then use the patterns revealed by these robustness experiments, together with the two-term decomposition from Theorem 4.1, to determine how much of the observed accuracy stems from pool-size effects versus shared disease signal (Section 5.4). We begin by describing the datasets and experimental setup common to both stages.

Across both stages we run eleven experiments, summarised in Table 1. In every experiment the adversary observes the pool's and the population's per-feature means and standard deviations; the experiments differ only in what the adversary additionally knows about the target individual.

5.1 Datasets

Our evaluation requires datasets that (i) support comparison between disease-specific and randomly composed pools, to isolate group-level signal, and (ii) contain repeated measurements over time, to evaluate robustness under temporal drift. We therefore require uniform longitudinal coverage, so that missing data does not introduce noise beyond the temporal drift under study.

We use three publicly available datasets spanning both molecular and behavioral data modalities. The two miRNA datasets are available from the Gene Expression Omnibus (GEO) repository; the Fitbit dataset is available on Kaggle. We collected no additional data. For the miRNA datasets, the local ethics committees approved the original studies and patients gave informed consent before sample collection. The Fitbit dataset is licensed CC0 and was collected from users who consented to share their data.

Cross-sectional dataset (GSE61741). [1] This dataset contains miRNA profiles from 1,049 individuals across 19 disease groups and a control group, each represented by 848 miRNA expression values obtained from blood samples at a *single timepoint*. Backes et al. [2] used this dataset for membership inference, enabling direct comparison with prior results. Following their preprocessing approach, we filter out miRNAs with median expression below 50, yielding 466 features. After preprocessing, expression values range from 1 to 39,186 with per-feature means between 54.4 and 36,364.4,

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61741

Table 1: Overview of empirical experiments. The adversary observes the pool's and the population's per-feature means and standard deviations; the second column describes additional information about the target.

Experiment	Adversary's view of the target	Section	Figure/Table
Noise	Target sample with adaptive Gaussian noise of scale $\alpha\sigma_j$ per feature	§5.3.1	Fig. 2, Tab. 2
Missing features	Target sample restricted to a subset of features	§5.3.1	Fig. 3
Combined noise + missing features	Both perturbations applied together	§5.3.1	App. Fig. 10–11
Real temporal drift	Target sample drawn from a later timepoint t_i ($i \geq 0$) while the pool's statistics are computed at t_0	§5.3.2	Tab. 3
Independent Gaussian surrogate	Target's drift replaced by feature-wise Gaussian noise matched to real per-feature variance	§5.3.2	Tab. 3
NN drift swap	Target's drift vector replaced by the nearest neighbor's	§5.3.2	Tab. 3, App. Fig. 14
LOO conditional mean	Target's drift predicted by a leave-one-out linear model of drift on baseline expression	§5.3.2	Tab. 3
Pool-size sweep	Exact target sample; pool size n varied within a single disease	§5.4.1	Fig. 5
19-disease validation	Exact target sample; case-pool versus random-pool comparison repeated across all 19 diseases	§5.4.2	Fig. 6
Disease-feature stratification	Exact target sample restricted to a subset of features to (or away from) literature-identified disease miRNAs	§5.4.2	Fig. 7
Transferred-individual test	Exact target sample drawn from non-members who share the pool's disease	§5.4.2	App. Fig. 8a–8b

and per-feature standard deviations between 39.6 and 9,365.4. Alternative normalization techniques exist [51] but we retain the original pipeline for comparability.

Longitudinal dataset (GSE68951). [2] To evaluate attacks under real-world temporal drift, we use a longitudinal dataset of repeated plasma miRNA measurements from 26 patients with lung cancer and 12 non-cancerous control samples. Each patient was sampled at up to eight timepoints (starting with a two-week interval, followed by three-month gaps) yielding 1,205 miRNAs per sample. We exclude the non-cancerous lung disease samples as they lack longitudinal information. In our experiments, we only use the first 6 timepoints to ensure even longitudinal coverage for all 26 patients. Expression values are $\log_2$-transformed and range from -5.4 to 15.7, with per-feature means between -1.6 and 13.3 and per-feature standard deviations between 0.2 and 1.6. Across consecutive timepoints the median per-feature absolute change is 0.30 $\log_2$ units ($\approx$1.2-fold) and the mean is 0.43 $\log_2$ units ($\approx$1.3-fold), comparable to one within-feature standard deviation; the 90th percentile reaches 0.95 $\log_2$ units ($\approx$1.9-fold).[3] The repeated measurements allow us to test whether the attacks withstand this natural variation.

Longitudinal dataset (Fitbit). [4] To test whether our findings extend beyond molecular data, we include the Fitbit Fitness Tracker Data dataset from Kaggle, comprising daily activity submissions from 35 users (step count, distance, active minutes, and calories, among others). We omit two near-constant features (*SedentaryActiveDistance* and *LoggedActivitiesDistance*, zero in over 91% of entries) and retain 11 features. Submission counts range from 8 to 32 per user; we cap at 8 to align every user at the minimum available coverage. As with the longitudinal miRNA dataset, this structure lets us evaluate inference across submissions from the same individual.

5.2 Experimental Design

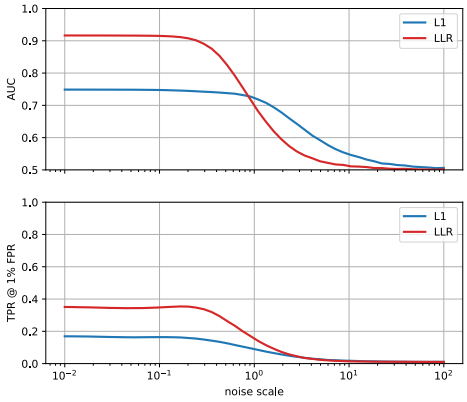
The three datasets described above differ in structure, dimensionality, and whether they support disease-specific comparisons. We now describe the experimental configurations that map each dataset to our research questions, together with a shared evaluation protocol.

Pool–population split. Each experiment compares a pool against a reference population. All patients not used in the pool are included in the population. We exclude pool members from the population statistics, unlike prior work [2] which treated the pool as a subset of the population. This choice is more favorable to the adversary and avoids biasing reference statistics in small datasets.

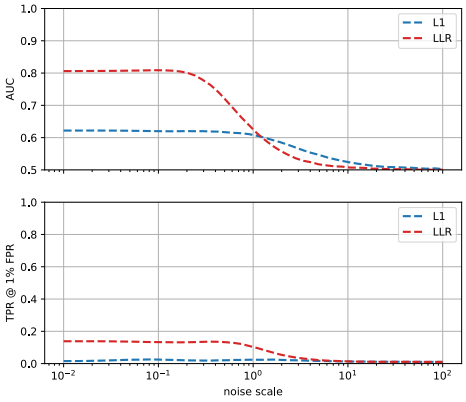
Pool composition. On the cross-sectional dataset, we evaluate two pool types. *Case pools* consist of all individuals sharing a disease condition (e.g., the 65 prostate cancer patients). *Random pools* are sampled uniformly at matching size. Comparing the two isolates the contribution of shared disease traits to attack accuracy. The longitudinal datasets (miRNA and Fitbit) contain patients who all share a single condition or have no identifying unique feature traits (the Fitbit dataset has no labeled disease status, so we treat its 35 users as a homogeneous population), precluding case-versus-random comparisons. We instead sample random pools of $n = 8$ individuals and use the remainder as the population.

Attack implementation. Given a pool–population split, we compute per-feature means and standard deviations for both groups and apply the L_1-distance and LLR tests (Section 2.3). Each attack scores every individual as a potential target v and classifies them as a member when $T_v > t$. We report AUC and TPR @ 1% FPR and repeat all experiments at least 2,000 times for stable estimates. Reported values are means over these iterations; standard errors of

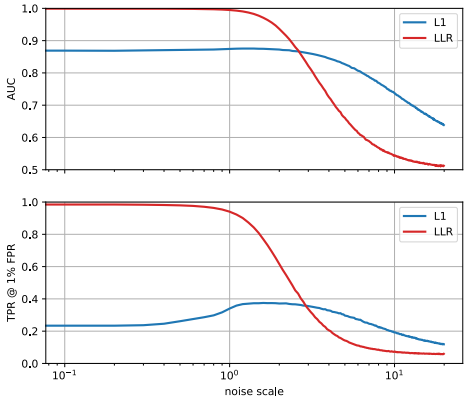
[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68951
[3] A distributional plot can be found in the Appendix, Figure 12a.
[4] https://www.kaggle.com/datasets/arashnic/fitbit



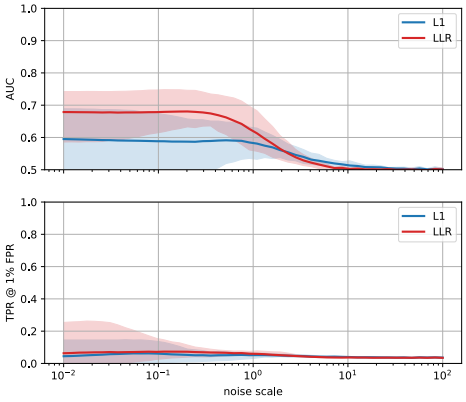
(a) Prostate cancer case pool of the cross-sectional dataset ($n = 65$).



(b) Random pool of the cross-sectional dataset ($n = 65$).



(c) Average, minimum and maximum values of all time-points of the longitudinal miRNA dataset ($n = 8$).



(d) Average, minimum and maximum values of first 8 submissions of the longitudinal Fitbit dataset ($n = 8$).

Figure 2: For each panel, top: AUC for the L_1 and LLR tests under increasing adaptive Gaussian noise $\alpha \cdot \sigma_j$ per feature; bottom: corresponding true-positive rate at 1% false-positive rate.

the mean are below 0.002 for AUC and below 0.005 for TPR @ 1% FPR across all configurations.

Perturbation protocol. We degrade the adversary's knowledge along two dimensions. In *noise experiments*, we add independent Gaussian noise with standard deviation $\alpha \cdot \sigma_j$ to each feature j of the target, where σ_j is the feature's population standard deviation; α thus measures perturbation in units of natural variation. In *partial-data experiments*, we progressively reduce the feature set by removing at most two features at a time from target, pool, and population simultaneously. Section 5.3.1 evaluates each dimension independently and in combination.

5.3 Attack Performance under Realistic Conditions

The theoretical analysis in Section 4 characterizes the fundamental limits of membership inference; we now test how the L_1 and LLR attacks behave in practice when the adversary's knowledge is degraded. The central finding is that the two attacks have *complementary weaknesses*: the L_1 test resists noise but loses power as features are removed, while the LLR degrades sharply under noise yet retains its advantage under partial data. Real-world temporal drift proves more damaging than either synthetic perturbation alone, and its effect is individual-specific rather than attributable to shared population-level structure.

5.3.1 Noise and Missing Features. We first evaluate each perturbation dimension independently, then examine their interaction.

Table 2: Summary of noise and crossover thresholds across configurations. $\alpha_{0.6}$: noise scale at which AUC drops below 0.6. α^*: crossover noise scale above which L_1 outperforms LLR.

Configuration	n	m	$\alpha_{0.6}$ LLR	$\alpha_{0.6}$ L_1	α^*
Prostate cancer (case)	65	466	1.9	4.9	0.9
Random pool	65	466	1.3	1.3	1.3
Renal cancer (case)	20	466	3.4	8.7	0.8
Random pool	20	466	2.8	6.0	1.3
Longitudinal miRNA	8	1,205	6.6	$>\alpha_{\max}$	2.7
Fitbit	8	11	1.3	<0.01	2.3

Noise. Figure 2 presents AUC under increasing adaptive noise across all three datasets (renal cancer replications appear in Appendix B, Figure 9). Both attacks tolerate substantial perturbation: on the prostate cancer case pool ($n = 65$), for instance, the LLR remains above AUC 0.6 until $\alpha \approx 1.9$, while the L_1 test resists until $\alpha \approx 4.9$. Table 2 reports the corresponding thresholds for all configurations and reveals two structural patterns. First, smaller pools tolerate more noise, consistent with the theoretical bound's prediction that smaller n amplifies individual contributions to the pool mean. Second, higher feature dimensionality m extends noise tolerance, directly illustrating the role of m in the statistical term of Theorem 4.1. Both patterns are visible across case and random pools alike.

Figure 2 also reports the corresponding TPR @ 1% FPR, the operating point used in prior work to assess whether membership inference reliably identifies *specific* individuals [7]. Both metrics decay monotonically with α and the L_1/LLR crossover persists, but it shifts to substantially higher noise for TPR than for AUC: on the prostate cancer case pool, AUC crossover occurs at $\alpha \approx 0.9$ while TPR crossover only occurs at $\alpha \approx 3.4$. The LLR therefore retains its advantage at the operating point most relevant for identifying specific individuals across a much broader noise range than the AUC crossover alone would suggest. At zero noise the LLR TPR reaches 0.35 on the prostate cancer case pool and 0.14 on the size-matched random pool, so under exact-data assumptions the attack identifies a meaningful fraction of individuals; as α rises both curves collapse toward the 1% floor.

The mechanism behind the L_1–LLR difference is straightforward. The L_1 test depends only on per-feature means; because the added noise has zero mean, it averages out over many features, preserving the signal that the L_1 statistic exploits. The LLR also depends on variance estimates, which noise systematically inflates.[5] At low noise, however, the picture reverses: the LLR exploits richer distributional information and achieves higher AUC when that information is not yet corrupted. A consistent crossover pattern emerges across all configurations (Table 2): below some noise scale α^*, the LLR outperforms L_1; above it, L_1 is superior. We use 0.6 as the AUC threshold for failed performance [9]. This crossover appears in random pools as well, confirming that it reflects a fundamental trade-off rather than an artifact of the disease signal.

[5]Under independence, $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$; when features are dependent, the inflation is compounded: $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\mathrm{Cov}(X, Y)$ [44].

Missing features. Missing features present a qualitatively different challenge: they reduce the number of terms in each test's statistic without distorting the surviving estimates. Figure 3 shows that both attacks tolerate substantial feature reduction: from 466 to 100 features, AUC decreases by less than 5 percentage points on the prostate cancer case pool. This resilience aligns with the theoretical bound: for $n = 65$, the minimum error approaches 0 at $m = 400$, rises to approximately 0.19 at $m = 100$, and reaches 0.40 at $m = 10$. TPR @ 1% FPR in Figure 3 decays more sharply than AUC as features are removed: on the prostate cancer case pool the LLR TPR falls from 0.35 at $m = 466$ to 0.26 at $m = 100$, while AUC barely moves (from 0.92 to 0.88). Removing features therefore erodes the LLR's ability to identify specific individuals faster than AUC alone suggests.

Crucially, partial data *reverses* the relative ranking of the two attacks. Across the full range of feature counts, the LLR maintains a lead of roughly 15 percentage points over the L_1 test (Figure 3a). Removing features reduces the LLR's summands without inflating its surviving variance estimates. The L_1 test, by contrast, loses the averaging effect that underlies its noise robustness: with fewer features, the sample mean becomes noisier, and the t-statistic loses power. The two attacks thus have complementary weaknesses (noise for the LLR, sparse features for L_1) and which matters more depends on the constraint the adversary faces.

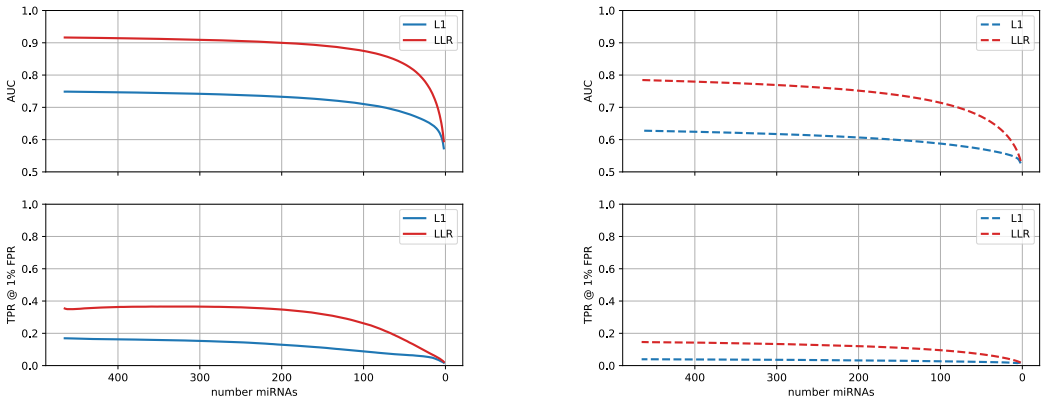
When both perturbations are applied jointly, the interaction is sub-additive (Appendix B, Figures 10–11): once feature reduction has degraded the signal, less residual signal remains for noise to destroy. An adversary who can estimate their own noise level could therefore select the superior test at each operating point, achieving the upper envelope of both curves; prior work evaluated each attack in isolation and did not consider this adaptive strategy.

5.3.2 Real-World Temporal Drift. The synthetic experiments above perturb each feature independently. Real temporal drift, by contrast, degrades attack accuracy far more steeply, and the source of this degradation turns out to be individual-specific. In other words, each person's own variation between timepoints is large enough to defeat membership inference. We demonstrate this through a series of controlled experiments that progressively add structure to the perturbation model; none of these surrogates reproduces the collapse observed under real drift.

Concretely, in each temporal-drift experiment the adversary observes the pool's per-feature means and standard deviations (computed from the pool members' samples at the baseline timepoint t_0) together with a sample of the target individual drawn from a later timepoint t_i with $i \geq 0$. The attack is unchanged from the non-temporal case; the adversary simply substitutes this later target sample for the exact baseline sample assumed in prior work.

Synthetic noise underestimates real drift. On the longitudinal miRNA dataset, both attacks collapse to near-random levels after the first timepoint[6]: the LLR falls to AUC 0.557 and the L_1 test to 0.571 (Table 3). The baseline AUC at t_0 is itself moderate, consistent with prior reports that plasma-based miRNA assays have lower signal-to-noise than tissue-based assays [2]; our analysis therefore focuses on the additional degradation from t_0 onward. Because synthetic noise

[6]As can be seen in Appendix B, Figure 12b.



(a) Prostate cancer case pool ($n = 65$). **(b) Random pool ($n = 65$).**

Figure 3: For each panel, top: AUC versus feature count for the L_1 and LLR tests on the cross-sectional dataset; bottom: corresponding true-positive rate at 1% false-positive rate.

matched to the same per-feature variance degrades both attacks far more gradually (Figure 2c), marginal feature-level variance alone cannot explain the real-drift collapse. The remaining experiments in this section identify the responsible structure.

The degradation is individual-specific. To determine what drives the gap between synthetic and real performance, we test whether the responsible structure is individual-specific or shared across individuals. Concretely, we swap each individual's temporal change with that of a similar person and check whether the attack recovers. We define a *drift vector* as the per-feature difference between an individual's later measurement t_i and their *baseline* (first timepoint t_0), and assign each individual the drift vector of their nearest neighbor in baseline feature space (Euclidean distance). The nearest neighbor sits at only the 17th percentile of all pairwise distances, reflecting the approximate equidistance typical of high-dimensional spaces; yet this single swap raises LLR AUC from 0.557 to 0.804 (Table 3). Swapping drift vectors, no matter how similar, breaks the structure that degrades the attack, indicating that the effect is tied to each person's own temporal evolution rather than to a shared population-level pattern. This conclusion is robust to the choice of distance metric and the number of neighbors k: Euclidean, cosine, and correlation distances yield nearly identical results across $k \in \{1, \ldots, 25\}$ (Appendix B, Figure 14). At $k = 20$ the chosen neighbor lies at median pairwise distance (effectively a random draw); AUC plateaus there near 0.90 and TPR @ 1% FPR near 0.58, both well above the real-drift level. The structure that degrades the attack therefore resides in each target's own drift; no other individual's drift, similar or not, reproduces it.

Population-level structure does not explain the collapse. If drift were a predictable function of baseline expression, a population-level model could capture individual-specific effects without per-person data. We test this with a leave-one-out (LOO) conditional mean: for each individual, we fit a linear regression of drift on baseline from all other individuals and predict that person's drift deterministically. This model captures whatever linear baseline–drift coupling exists in the population. The LOO conditional mean yields LLR AUC of 0.821, comparable to the nearest-neighbor swap (0.804) and far above the real-drift level (0.557). The population-level linear relationship between baseline and drift does not account for the observed degradation.

Summary of drift experiments. Table 3 collects all four models. The key finding is that only each person's own temporal evolution degrades the attack to near-random levels; the tested surrogate models fail to reproduce the collapse. The table reveals a noteworthy asymmetry between the real-drift and nearest-neighbor settings: substituting the closest individual's drift in baseline space recovers attack performance to AUC 0.80, while the target's own drift degrades it to AUC 0.55. If individual drift varied randomly with respect to identifying structure, the two settings would degrade the attack by similar amounts. The observed gap therefore indicates that each person's drift carries structure specific to that person. Structure that no other individual's drift reproduces, including that of their nearest neighbor, and that our leave-one-out regression (limited to linear, population-wide predictors of drift from baseline features) does not capture. Understanding what generates this structure (e.g., external factors influencing miRNA expression), and whether it can be exploited by a stronger attacker or used as a basis for defense, is an important open question for future work. TPR at 1% FPR tells the same story: only real drift collapses TPR to near the 1% floor, while every surrogate model preserves a substantial identification signal.

The Fitbit dataset (Figure 4) exhibits a qualitatively different pattern: AUC begins lower (0.63 for L_1 and 0.75 for LLR at the first submission) and declines gradually rather than collapsing. This smoother degradation likely reflects both the lower feature dimensionality ($m = 11$) and the less volatile nature of behavioral drift compared to molecular fluctuation. Because the Fitbit cohort has no labeled disease, it acts as a control for the steep miRNA collapse:

Table 3: AUC and TPR @ 1% FPR under four drift models on the longitudinal miRNA dataset. Real drift uses the individual's own later timepoint; Gaussian noise matches per-feature variance; LOO predicts drift via linear regression; NN assigns the nearest neighbor's real drift vector.

	AUC		TPR @ 1% FPR	
Drift model	LLR	L_1	LLR	L_1
Real drift	0.557	0.571	0.079	0.059
Independent Gaussian	0.998	0.907	0.981	0.558
LOO conditional mean	0.821	0.588	0.444	0.175
NN (k=1, Euclidean)	0.804	0.621	0.410	0.134

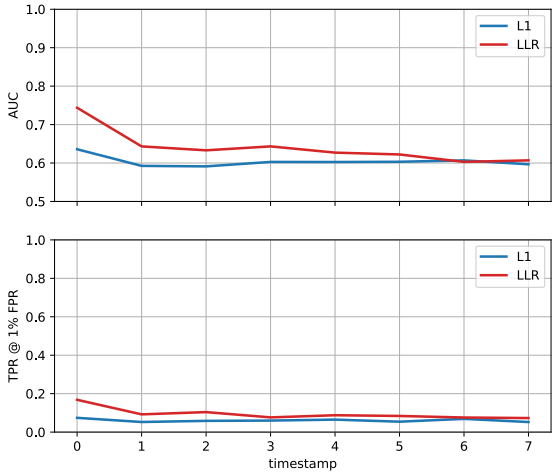


Figure 4: AUC (top) and true-positive rate at 1% false-positive rate (bottom) for the L_1 and LLR tests across all 8 daily submissions of the longitudinal Fitbit dataset ($n = 8$).

the absence of an analogous collapse here suggests that the longitudinal miRNA degradation reflects the volatility of plasma miRNA expression rather than any property unique to cancer patients.

Empirical Findings: Attack Robustness

Both attacks tolerate substantial noise and missing features with complementary weaknesses: the L_1 test resists noise (mean preservation) while the LLR resists feature reduction (variances remain intact). Real-world temporal drift, by contrast, degrades both attacks far more steeply than any synthetic model predicts, and the collapse is individual-specific—no surrogate (independent noise, nearest-neighbor drift, or population-level predictor) reproduces it.

5.4 What Drives Attack Success

The preceding experiments show that both attacks tolerate substantial noise and missing features, but they do not reveal *what* the attacks are detecting. Our theoretical bound (Theorem 4.1) decomposes the minimum inference error into two independent components: a *statistical term* governed by pool size n and feature dimensionality m, and a *signal term* driven by condition-related shifts δ_j. Because the bound is a lower limit on error, it does not predict the performance of any particular attack; it does, however, provide a framework for designing controlled experiments that isolate each component. We now use such experiments to determine how much of the observed accuracy stems from each source.

5.4.1 The Role of Pool Size. The statistical term $\frac{m}{4}\log\left(\frac{n}{n-1}\right)$ predicts that larger pools dilute each individual's contribution to the pool mean, making inference harder. We test this prediction by varying n within a single disease, which holds the disease composition approximately constant.

We subsample case pools and size-matched random pools at sizes ranging from $n = 5$ to $n_{\max}$ for three diseases: prostate cancer ($n_{\max} = 65$), Wilms tumor ($n_{\max} = 124$), and ovarian cancer ($n_{\max} = 24$). Figure 5 shows the results.

Both case-pool and random-pool AUC decrease monotonically with n for all three diseases, confirming the statistical term's prediction. At the smallest pool size ($n = 5$), the LLR test approaches perfect discrimination (AUC $\approx$ 1.0) regardless of pool type, because a handful of individuals leaves a strong imprint on the pool mean; the L_1 test starts lower (AUC $\approx$ 0.89–0.95), consistent with its weaker use of distributional information. As n grows, random-pool AUC drops substantially (for Wilms tumor (LLR), from 1.0 at $n = 5$ to 0.72 at $n = 124$) while case-pool AUC declines more gently (from 1.0 to 0.90). A persistent gap between case and random curves remains at every pool size, indicating that pool-size effects alone do not account for the full attack accuracy on disease-specific cohorts.

The TPR panels of Figure 5 show the same pattern at the operating point relevant for identifying specific individuals: case-pool TPR@1%FPR sits substantially above its random-pool counterpart at every n, and the gap widens with signal strength. At each disease's largest pool size, the LLR case-vs-random TPR gap is 0.193 for prostate cancer ($n = 65$: 0.354 vs. 0.161), 0.081 for Wilms tumor ($n = 124$: 0.202 vs. 0.121), and 0.426 for ovarian cancer ($n = 24$: 0.750 vs. 0.324). Both TPR curves decay with n, mirroring AUC.

5.4.2 The Role of Disease Signal. The persistent case–random gap in the pool-size sweep (Figure 5) reflects the signal term $\frac{1}{4}\sum_j \delta_j^2/\sigma_j^2$ in Theorem 4.1, which captures the contribution of condition-driven mean shifts. If the attacks exploit shared disease traits rather than individual membership, two predictions follow: (i) disease-specific case pools should be systematically easier to attack than size-matched random pools, and (ii) disease-associated features should carry disproportionate discriminative power. We test both predictions through three complementary analyses.

Case pools versus random pools. We computed case-pool and random-pool AUC (LLR) for all 19 disease groups in the cross-sectional dataset. Figure 6A shows that random-pool AUC is influenced almost entirely by pool size, ranging from 0.97 for the smallest pool (Stomach tumor, $n = 13$) to 0.72 for the largest (Wilms

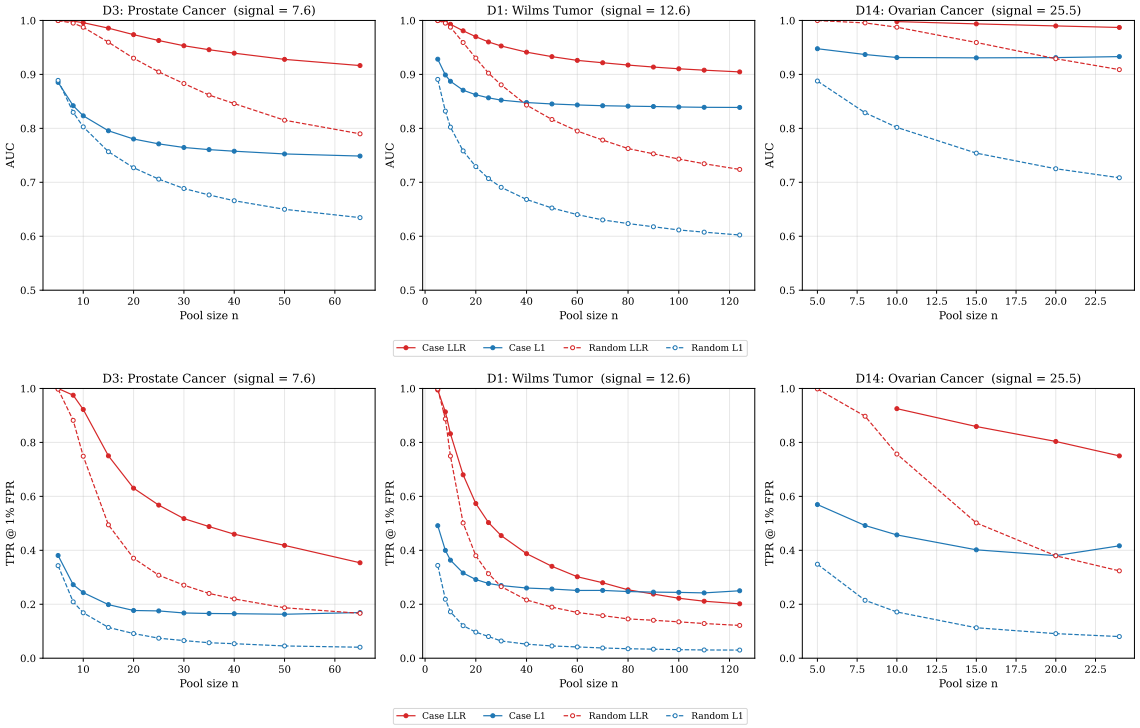


Figure 5: AUC versus pool size n and TPR @ 1% FPR versus pool size n for three diseases. Solid lines: case (disease-specific) pools; dashed lines: size-matched random pools.

tumor, $n = 124$). Case-pool AUC is consistently higher but does not track n as cleanly, because each disease contributes a different signal strength. To quantify this effect, we define the empirical signal strength as $S = \frac{1}{4}\sum_j \delta_j^2/\sigma_j^2$. The case–random gap in the pool-size sweep (Figure 5) widens with S: it is narrowest for prostate cancer ($S = 7.6$; case–random LLR gap of 0.13 at $n_{\max}$) and widest for ovarian cancer ($S = 25.5$; case LLR remains at 0.99 at $n = 24$ while the random LLR falls to 0.91). Figure 6B confirms that this pattern is universal: every disease lies above the identity line, with a mean case–random gap of 0.096 and a range from 0.018 (Renal cancer) to 0.180 (Wilms tumor).

The bottom row of Figure 6 extends the comparison to TPR at 1% FPR, the stricter operating point used in prior work to assess whether membership inference identifies *specific* individuals [7]. The case–random separation persists in this metric and becomes substantially larger: every one of the 19 diseases again lies above the identity line, with a mean LLR TPR gap of 0.278 between case and random pools (L_1: 0.302), roughly three times the corresponding mean AUC gap of 0.096. Gaps range from 0.041 (Ductal adenocarcinoma) to 0.709 (Sarcoidosis) and are largest for small, high-signal pools: Sarcoidosis ($n = 45$: 0.911 vs. 0.203), Ovarian cancer ($n = 24$: 0.750 vs. 0.324), and Multiple sclerosis ($n = 23$: 0.913 vs. 0.336).

This general pattern is also visible in the feature-reduction experiments of Section 5.3.1: on the prostate cancer case pool, AUC exceeds the size-matched random pool by approximately 0.15 (LLR) across the full range of feature counts (Figures 3a–3b), with all other attack parameters held constant. TPR @ 1% FPR shows the same separation.

Isolating disease-associated features. The case–random comparison shows that disease signal inflates accuracy but does not reveal which features carry that signal. To address this question, we identified disease-specific miRNAs from the medical literature [18, 19, 24, 43, 47] for several conditions and repeated the attacks using two feature subsets of equal size: the known disease-associated miRNAs, and a random sample drawn from the remaining features.

Figure 7 shows the results for three diseases: Wilms tumor ($n = 124$, $|A| = 15$), renal cancer ($n = 20$, $|A| = 16$), and ovarian cancer ($n = 24$, $|A| = 24$). In each case, disease-associated features yield higher AUC than the random subset at comparable feature counts ($\approx$15 miRNAs), with gaps of approximately 0.05–0.10 (LLR) and 0.10–0.12 (L_1). The corresponding TPR @ 1% FPR panels in Figure 7 mirror this pattern.

For ovarian cancer ($|A| = 24$), the gap is negligible at full feature count but emerges as features are reduced. This pattern suggests that non-disease miRNAs accumulate enough correlated signal to match the disease subset when many features are available, but this advantage vanishes as the feature set shrinks.[7]

Transferred-individual test. The analysis above demonstrates that disease signal inflates attack accuracy, but does not directly test

[7] An analogous stratification on the longitudinal dataset appears in Appendix B, Figure 13.

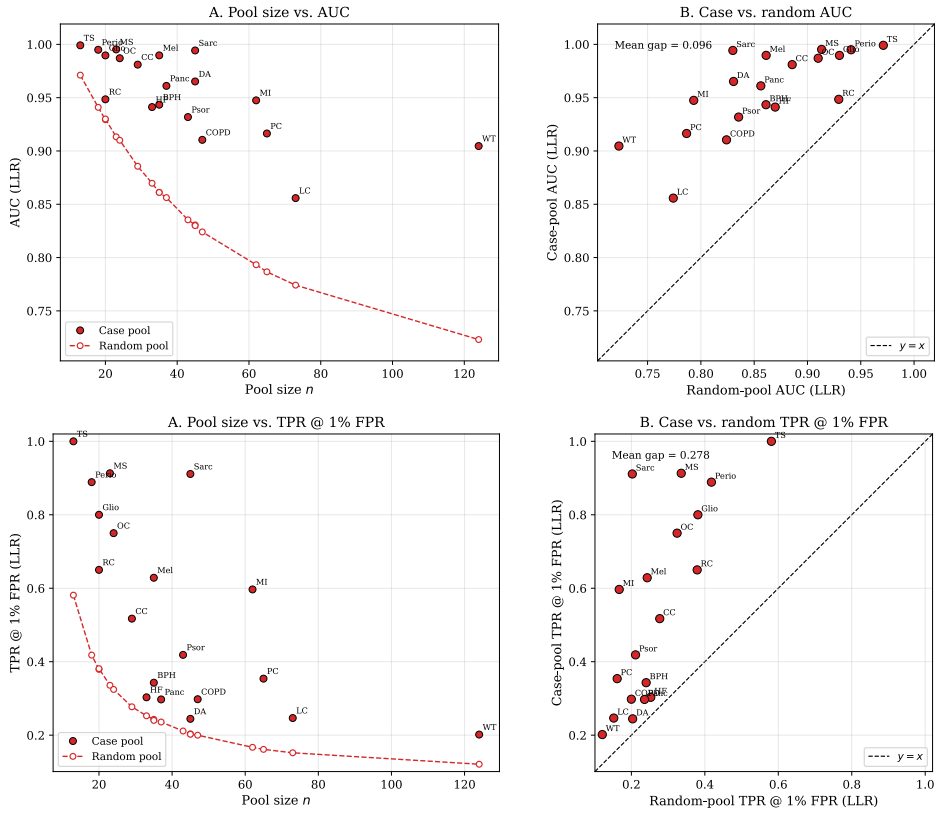
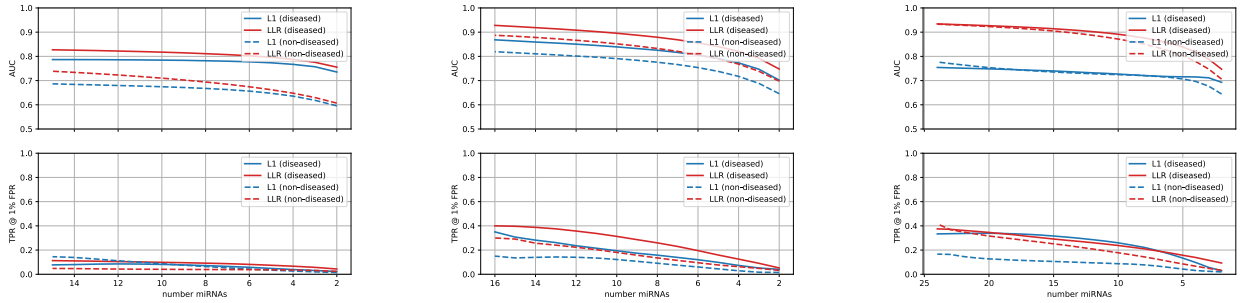


Figure 6: Membership inference accuracy (LLR test) across all 19 disease groups ($m = 466$) and size-matched random pools. (A) AUC versus pool size n. (B) Case-pool versus random-pool AUC. Bottom row: as (A) and (B), but for TPR @ 1% FPR.



(a) Wilms Tumor ($n = 124$). **(b) Renal cancer ($n = 20$).** **(c) Ovarian cancer ($n = 24$).**

Figure 7: For each subfigure, top: AUC; bottom: TPR @ 1% FPR. Disease-associated versus non-associated feature subsets on the cross-sectional dataset.

whether the attacks can distinguish a true pool member from a non-member who shares the same condition.

We probe this distinction by removing one quarter of the prostate cancer case pool ($n = 65 - 16 = 49$) and also excluding those 16 individuals from the population; we then evaluate the attacks on three groups: true pool members, population members, and the "transferred" individuals (who share the disease but were never in the pool). Their test statistics cluster with the true pool members rather than with the population (Figures 8a and 8b, Appendix B), so the attacks cannot distinguish a contributor to the dataset from a non-member who merely shares the same condition.

Together, these analyses show that the signal persisting through noise is in substantial part the shared group-level disease signal. The case–random separation is visible in both metrics but is markedly

larger in TPR @ 1% FPR: the mean LLR gap grows from 0.096 in AUC to 0.278 in TPR (L_1: 0.302), and exceeds 0.4 for several small, high-signal cohorts (e.g. Sarcoidosis, Multiple sclerosis, Ovarian cancer). Reported accuracy on summary statistics therefore reflects a mixture of individual and group-level evidence [22] and should be interpreted accordingly.

Empirical Findings: Sources of Attack Success

Attack success splits into a pool-size term and a disease-signal term: random-pool AUC and TPR track pool size, while case pools are universally easier to attack (mean LLR gap ≈ 0.10 in AUC across 19 diseases, growing roughly threefold to ≈ 0.28 in TPR @ 1% FPR). Non-members sharing the disease are frequently misclassified, so a substantial component of reported accuracy reflects shared condition rather than individual membership.

6 Related Work

Our work relates to three lines of research: membership inference attacks on summary statistics, their extension to non-genomic domains, and the broader study of membership inference against machine learning models. We discuss each in turn, highlighting how our contribution differs.

MIAs on summary statistics. Membership inference against published statistics originated in genome-wide association studies (GWAS). Homer et al. [20] introduced the L_1-distance attack using allele frequencies; Wang et al. [55] later improved it by incorporating genomic correlations. Sankararaman et al. [46] showed empirically that the likelihood-ratio test outperforms the L_1 approach, and Zhou et al. [62] provided the first theoretical analysis of attack complexity. More recently, Bu et al. [5] proposed a haplotype-based attack exploiting variant presence/absence data. All of these attacks assume that the adversary holds the *exact* target datapoint. Our work relaxes this assumption and evaluates attack robustness when the adversary has only noisy, partial, or temporally mismatched data, a setting none of these prior works addresses.

Extension to other domains. Backes et al. [2] were the first to apply summary-statistic MIAs to microRNA expression data, demonstrating that the L_1 and LLR attacks transfer beyond genomics. Hagestedt et al. [17] extended MIAs to DNA methylation data and proposed a privacy-preserving alternative to Beacon-style sharing platforms. Outside the biomedical domain, Pyrgelis et al. [41] studied MIAs on aggregate location time-series and found that attack success depends on user mobility patterns, while Zhang et al. [60] introduced LocMIA, an attack on aggregated location data that notably does not require exact prior knowledge of the target's location—the closest prior threat model to ours, though on a different modality and without the group-signal question that our work also addresses.

Broader privacy context. The observation that aggregate records may not uniquely identify individuals has a long history in the privacy literature. Machanavajjhala et al. [31] introduced ℓ-diversity as a refinement of k-anonymity [50], formalizing the risk that group-level homogeneity undermines privacy even when individual records are suppressed. Our finding that MIAs on summary statistics partially detect group-level traits resonates with this insight, though we arrive at it from the adversary's perspective rather than the data-release mechanism's.

MIAs against machine learning models. A large body of work studies membership inference against trained models, where the adversary queries a model to determine whether a specific record appeared in its training set [7, 45, 48]. These attacks differ from ours in threat model: the adversary observes model outputs (predictions, confidence scores, or gradients) rather than published statistics, and the signal arises from overfitting rather than from the arithmetic relationship between an individual's data and the cohort mean.

7 Conclusion

We revisited membership inference on summary statistics under realistic adversarial conditions. Both the L_1 and LLR tests tolerate substantial noise and missing features with complementary weaknesses (L_1 resists variance inflation under noise; LLR retains advantage under partial data), so an adversary who can estimate their own noise regime could select the better test at each operating point. Real-world temporal drift, by contrast, degrades both attacks far more steeply than synthetic experiments predict, and controlled ablations show this collapse is individual-specific.

Our theoretical bound separates this accuracy into a statistical term (pool size n, feature dimensionality m) and a signal term (condition-related shifts δ_j); experiments across 19 disease groups confirm the decomposition, with case pools systematically easier to attack (mean gap ≈ 0.10) and transferred non-members frequently misclassified as members. Detecting that an individual belongs to a disease cohort is itself a genuine privacy concern, but current attacks cannot reliably distinguish true pool members from non-members who share the same condition.

Limitations and future work. Our evaluation focuses on health-related datasets; whether the same group-signal dominance and complementary attack weaknesses arise in other domains remains open. Our experiments span three datasets, including two longitudinal datasets in which consecutive submissions show only weak temporal correlation. Confirming our findings across a wider range of modalities (particularly longitudinal data with strong within-individual stability) is a natural direction for future work. A further open question concerns the temporal-drift collapse itself (Section 5.3.2): the structured, person-specific component of individual drift that degrades the attack is reproduced neither by any other individual's drift nor by a population-level model. Understanding what generates it and whether it can be exploited by a stronger attacker or harnessed as a defense is an important direction for future work. The theoretical bound rests on strong assumptions (Gaussian, independent features with equal variances) and our longitudinal experiments show that each can be violated in practice. Relaxing these assumptions, extending the bound to handle feature correlations, and quantifying defenses such as differentially private summary releases are natural next steps.

Acknowledgments

The code and data underlying this work are publicly available at https://github.com/privacy-UoB/pmia-summary-stats. The authors used generative AI-based tools to revise the text, improve flow, and correct any typos, grammatical errors, and awkward phrasing. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Christian Ascoli, Yue Huang, Cody Schott, Benjamin A Turturice, Ahmed Metwally, David L Perkins, and Patricia W Finn. 2018. A circulating microRNA signature serves as a diagnostic and prognostic indicator in sarcoidosis. *American journal of respiratory cell and molecular biology* 58, 1 (2018), 40–54.

[2] 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.

[3] P Bickel, P Diggle, S Fienberg, U Gather, I Olkin, and S Zeger. 2009. Springer series in statistics. *Principles and Theory for Data Mining and Machine Learning. Cham, Switzerland: Springer* (2009).

[4] Jean Bretagnolle and Catherine Huber. 2006. Estimation des densités: risque minimax. In *Séminaire de Probabilités XII: Université de Strasbourg 1976/77*. Springer, 342–363.

[5] Diyue Bu, Xiaofeng Wang, and Haixu Tang. 2021. Haplotype-based membership inference from summary genomic data. *Bioinformatics* 37, Supplement_1 (2021), i161–i168.

[6] Clément L Canonne. 2022. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198* (2022).

[7] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*. IEEE, 1897–1914.

[8] Chao Cheng, Qiang Wang, Wenjie You, Manhua Chen, and Jiahong Xia. 2014. MiRNAs as biomarkers of myocardial infarction: a meta-analysis. *PloS one* 9, 2 (2014), e88566.

[9] Şeref Kerem Çorbacıoğlu and Gökhan Aksel. 2023. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish journal of emergency medicine* 23, 4 (2023), 195–198.

[10] Xiaomin Dang, Xiaoyan Qu, Weijia Wang, Chongbing Liao, Ying Li, Xiaojin Zhang, Dan Xu, Carolyn J Baglole, Dong Shang, and Ying Chang. 2017. Bioinformatic analysis of microRNA and mRNA Regulation in peripheral blood mononuclear cells of patients with chronic obstructive pulmonary disease. *Respiratory Research* 18, 1 (2017), 4.

[11] Rhonda Daniel, Qianni Wu, Vernell Williams, Gene Clark, Georgi Guruli, and Zendra Zehner. 2017. A panel of MicroRNAs as diagnostic biomarkers for the identification of prostate cancer. *International journal of molecular sciences* 18, 6 (2017), 1281.

[12] John Duchi. 2007. Derivations for linear algebra and optimization. *Berkeley, California* 3, 1 (2007), 2325–5870.

[13] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[14] Andrew P Feinberg and M Daniele Fallin. 2015. Epigenetics at the crossroads of genes and the environment. *Jama* 314, 11 (2015), 1129–1130.

[15] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* 19, 1 (2009), 92–105.

[16] Shiwei Guo, Hao Qin, Ke Liu, Huan Wang, Sijia Bai, Shiyi Liu, Zhuo Shao, Yanan Zhang, Bin Song, Xiaoya Xu, et al. 2021. Blood small extracellular vesicles derived miRNAs to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis. *Clinical and Translational Medicine* 11, 9 (2021), e520.

[17] Inken Hagestedt, Mathias Humbert, Pascal Berrang, Irina Lehmann, Roland Eils, Michael Backes, and Yang Zhang. 2020. Membership Inference Against DNA Methylation Databases. In *Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE.

[18] SFM Häusler, A Keller, PA Chandran, K Ziegler, K Zipp, S Heuer, M Krockenberger, JB Engel, A Hönig, M Scheffler, et al. 2010. Whole blood-derived miRNA profiles as potential new tools for ovarian cancer screening. *British journal of cancer* 103, 5 (2010), 693–700.

[19] Qingfang He, Yirong Fang, Feng Lu, Jin Pan, Lixin Wang, Weiwei Gong, Fangrong Fei, Jingjie Cui, Jieming Zhong, Ruying Hu, et al. 2019. Analysis of differential expression profile of miRNA in peripheral blood of patients with lung cancer. *Journal of clinical laboratory analysis* 33, 9 (2019), e23003.

[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.

[21] Weili Huang. 2017. MicroRNAs: biomarkers, diagnostics, and therapeutics. *Bioinformatics in MicroRNA research* (2017), 57–67.

[22] Jun Ishii, Hiroyuki Maeomichi, and Ikuo Yoda. 2012. Privacy-Preserving Statistical Analysis Method for Real-World Data. In *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 807–812.

[23] Peter A Jones and Stephen B Baylin. 2007. The epigenomics of cancer. *Cell* 128, 4 (2007), 683–692.

[24] Petra Leidinger, Valentina Galata, Christina Backes, Cord Stähler, Stefanie Rheinheimer, Hanno Huwer, Eckart Meese, and Andreas Keller. 2015. Longitudinal study on circulating miRNAs in patients after lung cancer resection. *Oncotarget* 6, 18 (2015), 16674.

[25] Petra Leidinger, Andreas Keller, Anne Borries, Jörg Reichrath, Knuth Rass, Sven U Jager, Hans-Peter Lenhof, and Eckart Meese. 2010. High-throughput miRNA profiling of human melanoma blood samples. *BMC cancer* 10, 1 (2010), 262.

[26] David A Levin and Yuval Peres. 2017. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.

[27] Chunjian Li, Zhijuan Fang, Ting Jiang, Qiu Zhang, Chao Liu, Chenyu Zhang, and Yang Xiang. 2013. Serum microRNAs profile from genome-wide serves as a fingerprint for diagnosis of acute myocardial infarction and angina pectoris. *BMC medical genomics* 6, 1 (2013), 16.

[28] Xize Li, Qingyu Wang, Zhouhan Xu, Xuesi Yang, Ruiqi Zhao, Haocheng Wang, Xueling Li, and Jiping Zeng. 2025. Screening and evaluation of specific blood MiRNAs as potential biomarkers in diagnostics of gastric Cancer. *Scientific Reports* 15, 1 (2025), 22974.

[29] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, et al. 2005. MicroRNA expression profiles classify human cancers. *nature* 435, 7043 (2005), 834–838.

[30] Xinting Ma, Juhua Zhou, Yin Zhong, Linlin Jiang, Ping Mu, Yanmin Li, Narendra Singh, Mitzi Nagarkatti, and Prakash Nagarkatti. 2014. Expression, regulation and function of microRNAs in multiple sclerosis. *International journal of medical sciences* 11, 8 (2014), 810.

[31] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)* 1, 1 (2007), 3–es.

[32] Anveshika Manoj, Mohammad K Ahmad, Gautam Prasad, Durgesh Kumar, Abbas A Mahdi, and Manoj Kumar. 2023. Screening and validation of novel serum panel of microRNA in stratification of prostate cancer. *Prostate International* 11, 3 (2023), 150–158.

[33] Manuel Mayr, Anna Zampetaki, and Stefan Kiechl. 2013. MicroRNA biomarkers for failing hearts? , 2782–2783 pages.

[34] Pablo Micó-Martínez, Pedro J Almiñana-Pastor, Francisco Alpiste-Illueca, and Andrés López-Roldán. 2021. MicroRNAs and periodontal disease: A qualitative systematic review of human studies. *Journal of periodontal & implant science* 51, 6 (2021), 386.

[35] Patrick S Mitchell, Rachael K Parkin, Evan M Kroh, Brian R Fritz, Stacia K Wyman, Era L Pogosova-Agadjanyan, Amelia Peterson, Jennifer Noteboom, Kathy C O'Briant, April Allen, et al. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the national academy of sciences* 105, 30 (2008), 10513–10518.

[36] Ellie TY Mok, Jessica L Chitty, and Thomas R Cox. 2024. miRNAs in pancreatic cancer progression and metastasis. *Clinical & Experimental Metastasis* 41, 3 (2024), 163–186.

[37] Jerzy Neyman and Egon Sharpe Pearson. 1933. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231, 694-706 (1933), 289–337.

[38] Ana Peterlin, Karolina Počivavšek, Danijel Petrovič, and Borut Peterlin. 2020. The role of microRNAs in heart failure: a systematic review. *Frontiers in Cardiovascular Medicine* 7 (2020), 161.

[39] Mark S Pinsker. 1964. Information and information stability of random variables and processes. *Holden-Day* (1964).

[40] Yury Polyanskiy and Yihong Wu. 2019. Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616* (2019).

[41] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2020. Measuring membership privacy on aggregate location time-series. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4, 2 (2020), 1–28.

[42] Irfan A Qureshi and Mark F Mehler. 2011. Advances in epigenetics and epigenomics for neurodegenerative diseases. *Current neurology and neuroscience reports* 11 (2011), 464–473.

[43] Martina Redova, Alexandr Poprach, Jana Nekvindova, Robert Iliev, Lenka Radova, Radek Lakomy, Marek Svoboda, Rostislav Vyzula, and Ondrej Slaby. 2012. Circulating miR-378 and miR-451 in serum are potential biomarkers for renal cell carcinoma. *Journal of translational medicine* 10, 1 (2012), 55.

[44] John A Rice and John A Rice. 2007. *Mathematical statistics and data analysis*. Vol. 371. Thomson/Brooks/Cole Belmont, CA.

[45] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).

[46] 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.

[47] Jana Schmitt, Christina Backes, Nasenien Nourkami-Tutdibi, Petra Leidinger, Stephanie Deutscher, Markus Beier, Manfred Gessler, Norbert Graf, Hans-Peter Lenhof, Andreas Keller, et al. 2012. Treatment-independent miRNA signature in blood of Wilms tumor patients. *BMC genomics* 13, 1 (2012), 379.

[48] 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.

[49] Albert Sufianov, Sema Begliarzade, Tatiana Ilyasova, Yanchao Liang, and Ozal Beylerli. 2022. MicroRNAs as prognostic markers and therapeutic targets in gliomas. *Non-coding RNA Research* 7, 3 (2022), 171–177.

[50] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* 10, 05 (2002), 557–570.

[51] Shirley Tam, Ming-Sound Tsao, and John D McPherson. 2015. Optimization of miRNA-seq data preprocessing. *Briefings in bioinformatics* 16, 6 (2015), 950–963.

[52] Teodora Larisa Timis and Remus Ioan Orasan. 2018. Understanding psoriasis: Role of miRNAs. *Biomedical reports* 9, 5 (2018), 367–374.

[53] Warapen Treekitkarnmongkol, Jianliang Dai, Suyu Liu, Deivendran Sankaran, Tristian Nguyen, Seetharaman Balasenthil, Mark W Hurd, Meng Chen, Hiroshi Katayama, Sinchita Roy-Chowdhuri, et al. 2024. Blood-based microrna biomarker signature of early-stage pancreatic ductal adenocarcinoma with lead-time trajectory in prediagnostic samples. *Gastro Hep Advances* 3, 8 (2024), 1098–1115.

[54] Britta Vogel, Andreas Keller, Karen S Frese, Petra Leidinger, Farbod Sedaghat-Hamedani, Elham Kayvanpour, Wanda Kloos, Christina Backe, Ann Thanaraj, Thomas Brefort, et al. 2013. Multivariate miRNA signatures as biomarkers for non-ischaemic systolic heart failure. *European heart journal* 34, 36 (2013), 2812–2823.

[55] 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*. 534–544.

[56] Laura D Wood, D Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjoblom, Rebecca J Leary, Dong Shen, Simina M Boca, Thomas Barber, Janine Ptak, et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318, 5853 (2007), 1108–1113.

[57] Wei Wu. 2018. MicroRNA, noise, and gene expression regulation. *MicroRNA and Cancer: Methods and Protocols* (2018), 91–96.

[58] Lei Xin, Jun Gao, Dan Wang, Jin-Huan Lin, Zhuan Liao, Jun-Tao Ji, Ting-Ting Du, Fei Jiang, Liang-Hao Hu, and Zhao-Shen Li. 2017. Novel blood-based microRNA biomarker panel for early diagnosis of chronic pancreatitis. *Scientific reports* 7, 1 (2017), 40019.

[59] Toshiki Yoneda, Takaaki Tomofuji, Daisuke Ekuni, Tetsuji Azuma, Takayuki Maruyama, Kohei Fujimori, Yoshio Sugiura, and Manabu Morita. 2019. Serum microRNAs and chronic periodontitis: A case-control study. *Archives of Oral Biology* 101 (2019), 57–63.

[60] Guanglin Zhang, Anqi Zhang, and Ping Zhao. 2020. Locmia: Membership inference attacks against aggregated location data. *IEEE Internet of Things Journal* 7, 12 (2020), 11778–11788.

[61] Yajie Zhang, Min Li, Yijiang Ding, Zhimin Fan, Jinchun Zhang, Hongying Zhang, Bin Jiang, and Yong Zhu. 2017. Serum MicroRNA profile in patients with colon adenomas or cancer. *BMC medical genomics* 10, 1 (2017), 23.

[62] 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: 16th European Symposium on Research in Computer Security, Leuven, Belgium, September 12-14, 2011. Proceedings 16*. Springer, 607–627.

A Proofs

Proof of Theorem 4.1. We aim to bound the minimum error of any statistical test that distinguishes between the two given hypotheses.

For both hypotheses, we have $v, x \sim \mathcal{N}(\mu + \delta, \sigma^2)$ for the target v and members $x \in S$ of the pool.

We define the hypotheses as follows:

- H_0: $v \notin S$, so $v \sim \mathcal{N}(\mu, \sigma^2)$, and $\hat{\mu} \sim \mathcal{N}(\mu + \delta, \sigma^2/n)$ as it is independent of v and models the mean of the sum of n Gaussian random variables.
- H_1: $v \in S$, so $v, x \sim \mathcal{N}(\mu + \delta, \sigma^2)$, and $\hat{\mu} = \frac{1}{n}(v + \sum_{x \in S \setminus \{v\}} x)$ which is dependent on v.

Let P and Q denote the joint distributions $(v, \hat{\mu})$ under H_1 and H_0, respectively. The optimal test between H_0 and H_1 achieves a minimum error rate of

$$\text{min error} = \frac{1}{2}(1 - \text{TV}(P, Q)),$$

where TV denotes the total variation distance, i.e., the largest difference between the probabilities that two distributions assign to the same event [26, 40]. This expression assumes equal prior probabilities for the two hypotheses, which simplifies the analysis and yields the minimax error rate. In realistic scenarios, the prior probability of membership is typically very small. While this can reduce the overall Bayes error for an adversary, it also means that reliably identifying members becomes more difficult. As such, the equal-prior setting provides a meaningful lower bound on per-member inference success and serves as a conservative, best-case analysis for privacy.

To upper bound $\text{TV}(P, Q)$, we apply Pinsker's inequality [3, 6, 39]:

$$\text{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P \| Q)}.$$

Since we assume m independent features, each normally distributed, we can compute the KL divergence per feature and sum over m features.

Distribution under H_0. For a feature j, the joint distribution of $(v_j, \hat{\mu}_j)$ is a bivariate Gaussian with:

$$\mu_{Q_j} = \begin{bmatrix} \mu_j \\ \mu_j + \delta_j \end{bmatrix}, \quad \Sigma_{Q_j} = \begin{bmatrix} \sigma_j^2 & 0 \\ 0 & \sigma_j^2/n \end{bmatrix}.$$

Distribution under H_1. Since v contributes to $\hat{\mu}$, we compute:

$$\mathbb{V}[\hat{\mu}_j] = \frac{1}{n^2}\left(\sigma_j^2 + (n-1)\sigma_j^2\right) = \frac{\sigma_j^2}{n}, \quad \text{Cov}(v_j, \hat{\mu}_j) = \frac{1}{n}\sigma_j^2.$$

Thus,

$$\mu_{P_j} = \begin{bmatrix} \mu_j + \delta_j \\ \mu_j + \delta_j \end{bmatrix}, \quad \Sigma_{P_j} = \begin{bmatrix} \sigma_j^2 & \sigma_j^2/n \\ \sigma_j^2/n & \sigma_j^2/n \end{bmatrix}.$$

KL divergence for a single feature. The KL divergence between two multivariate Gaussians with m dimensions is given by [12]:

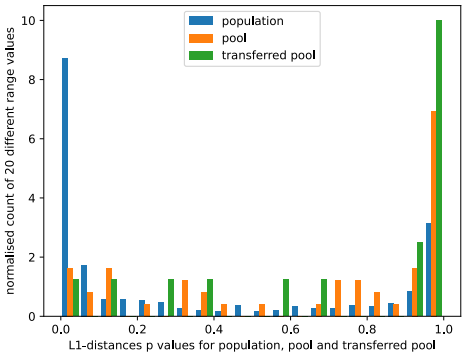
$$D_{\text{KL}}(P \| Q) =$$
$$\frac{1}{2}\left[\log \frac{\det \Sigma_Q}{\det \Sigma_P} - m + tr(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^T \Sigma_Q^{-1}(\mu_Q - \mu_P)\right]$$

We compute:

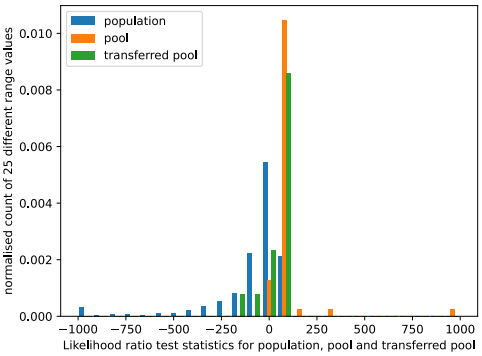
$$\det \Sigma_{Q_j} = \frac{\sigma_j^4}{n}, \quad \det \Sigma_{P_j} = \frac{(n-1)\sigma_j^4}{n^2}, \quad \frac{\det \Sigma_{Q_j}}{\det \Sigma_{P_j}} = \frac{n}{n-1}.$$

$$\Sigma_{Q_j}^{-1} = \begin{bmatrix} \frac{1}{\sigma_j^2} & 0 \\ 0 & \frac{n}{\sigma_j^2} \end{bmatrix}, \quad \Sigma_{Q_j}^{-1}\Sigma_{P_j} = \begin{bmatrix} 1 & 1/n \\ 1 & 1 \end{bmatrix}, \quad \text{tr}(\Sigma_{Q_j}^{-1}\Sigma_{P_j}) = 2.$$

$$\mu_{Q_j} - \mu_{P_j} = \begin{bmatrix} -\delta_j \\ 0 \end{bmatrix}, \quad (\mu_{Q_j} - \mu_{P_j})^T \Sigma_{Q_j}^{-1}(\mu_{Q_j} - \mu_{P_j}) = \frac{\delta_j^2}{\sigma_j^2}$$



(a) P values for the L_1-distances.



(b) Test statistics for the likelihood ratio test.

Figure 8: Histograms of test statistics when removing $\frac{1}{4}$ patients from the cross-sectional prostate cancer case pool ($n = 65-16 = 49$) while also excluding those 16 from the population.

Putting this together:

$$D_{\mathrm{KL}}(P_j \| Q_j) = \frac{1}{2}\left(\log\left(\frac{n}{n-1}\right) + \frac{\delta_j^2}{\sigma_j^2}\right).$$

Total divergence. Since features are independent:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{j=1}^{m} D_{\mathrm{KL}}(P_j\|Q_j) = \frac{m}{2}\log\left(\frac{n}{n-1}\right) + \frac{1}{2}\sum_j \frac{\delta_j^2}{\sigma_j^2}$$

Final bound on the minimum error. Applying Pinsker's inequality:

$$\mathrm{TV}(P,Q) \le \sqrt{\frac{m}{4}\log\left(\frac{n}{n-1}\right) + \frac{1}{4}\sum_j \frac{\delta_j^2}{\sigma_j^2}}.$$

Thus,

$$\text{min error} \ge \frac{1}{2}\left(1 - \sqrt{\frac{m}{4}\log\left(\frac{n}{n-1}\right) + \frac{1}{4}\sum_j \frac{\delta_j^2}{\sigma_j^2}}\right).$$

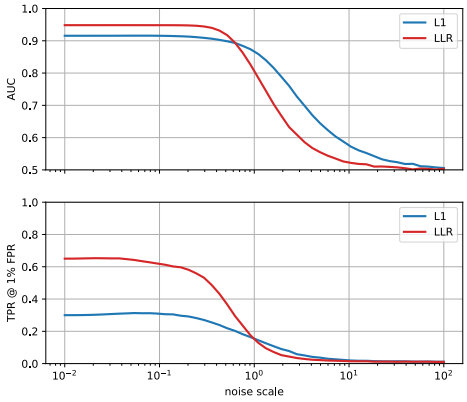
□

B Additional Plots

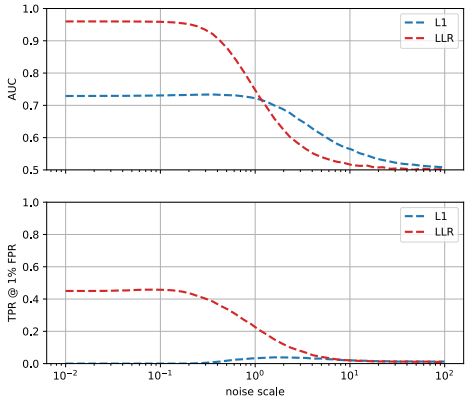
Filtering dependent features. To further quantify the role of feature dependencies within a single timepoint, we filtered the longitudinal dataset to retain only approximately independent features (Pearson $|r| < 0.9$) and repeated the attacks under synthetic Gaussian noise. As Figure 12b shows, AUC on this filtered dataset drops substantially compared to the full feature set, with the LLR test affected more severely than the L_1 test; TPR @ 1% FPR drops in step. This result directly tests the feature-independence assumption underlying Theorem 4.1: correlated features inflate apparent attack power beyond what independent features would yield, meaning that the theoretical bound is conservative in the presence of correlations. In practical terms, reported attack accuracy on correlated datasets, including our cross-sectional results and those of prior work [2], likely overstates the privacy risk, because part of the apparent signal stems from inter-feature dependencies rather than from individual membership.

Table 4: Disease-associated miRNA counts ($|A|$) identified from the literature for each condition in the cross-sectional dataset ($m = 466$). Pool size n is the number of individuals in each disease group.

Disease	n	$\lvert A\rvert$
Wilms tumor	124	15
Lung cancer	73	2
Prostate cancer	65	1
Myocardial infarction	62	2
COPD	47	3
Sarcoidosis	45	1
Ductal adenocarcinoma	45	13
Psoriasis	43	3
Pancreatitis	37	3
Benign prostate hyperplasia	35	1
Melanoma	35	47
Non-ischaemic systolic heart failure	33	5
Colon cancer	29	10
Ovarian cancer	24	24
Multiple sclerosis	23	12
Glioma	20	6
Renal cancer	20	16
Periodontitis	18	1
Stomach tumor	13	1

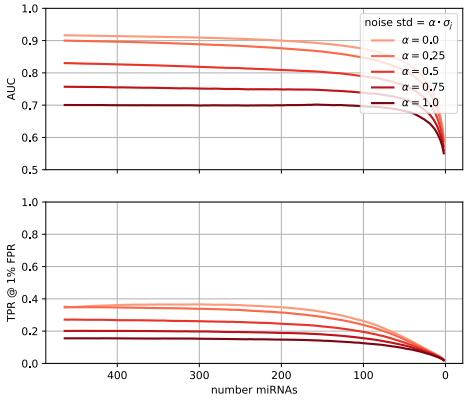


(a) Renal cancer case pool of the cross-sectional dataset ($n = 20$).

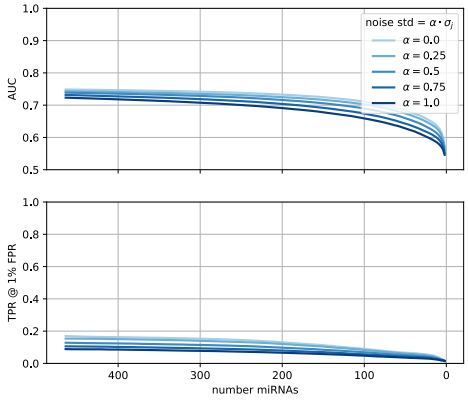


(b) Random pool of the cross-sectional dataset ($n = 20$).

Figure 9: For each subfigure, top: AUC; bottom: TPR @ 1% FPR. L_1 and LLR tests under increasing adaptive Gaussian noise ($\alpha \cdot \sigma_j$ per feature) on the renal cancer dataset. These panels complement Figure 2 in the main text.

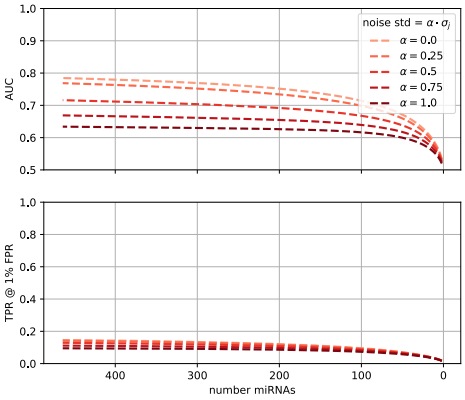


(a) LLR test on the prostate cancer case pool ($n = 65$).

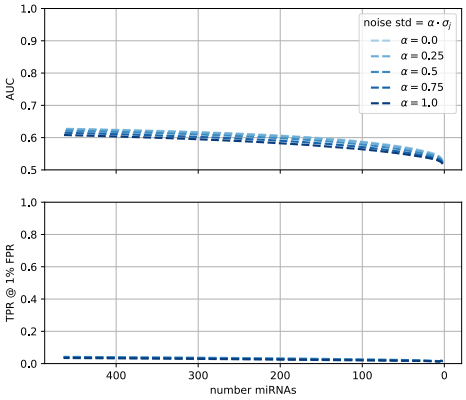


(b) L_1-distance test on the prostate cancer case pool ($n = 65$).

Figure 10: For each subfigure, top: AUC; bottom: TPR @ 1% FPR. Joint noise and feature reduction on the prostate cancer case pool. Five noise scales ($\alpha = 0, 0.25, 0.5, 0.75, 1$) are shown. This figure complements Section 5.3.1.

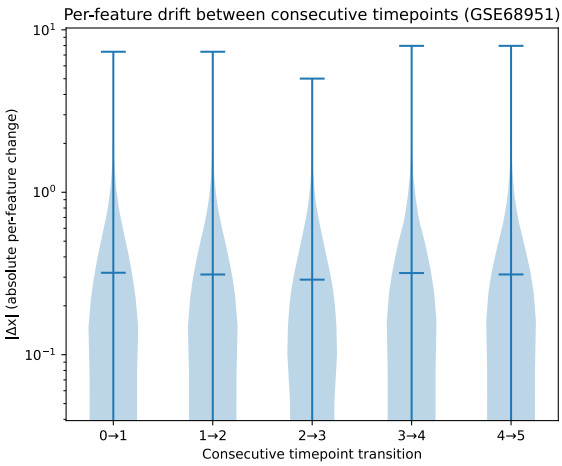


(a) LLR test on a random pool ($n = 65$).

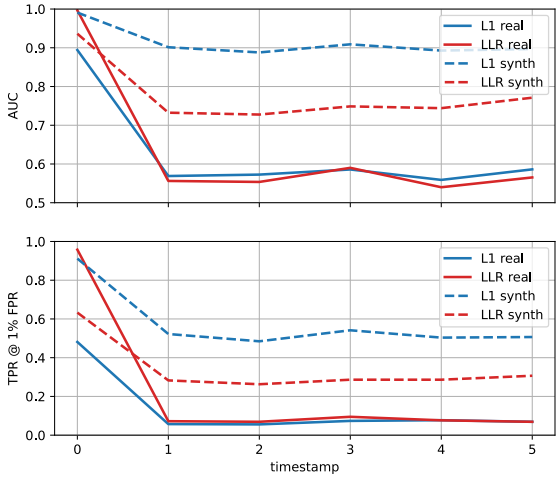


(b) L_1-distances test on a random pool ($n = 65$).

Figure 11: For each subfigure, top: AUC; bottom: TPR @ 1% FPR. Joint noise and feature reduction on the random pool ($n = 65$). Five noise scales ($\alpha = 0, 0.25, 0.5, 0.75, 1$). The five noise curves collapse as feature count decreases, reflecting the sub-additive interaction described in Section 5.3.1.

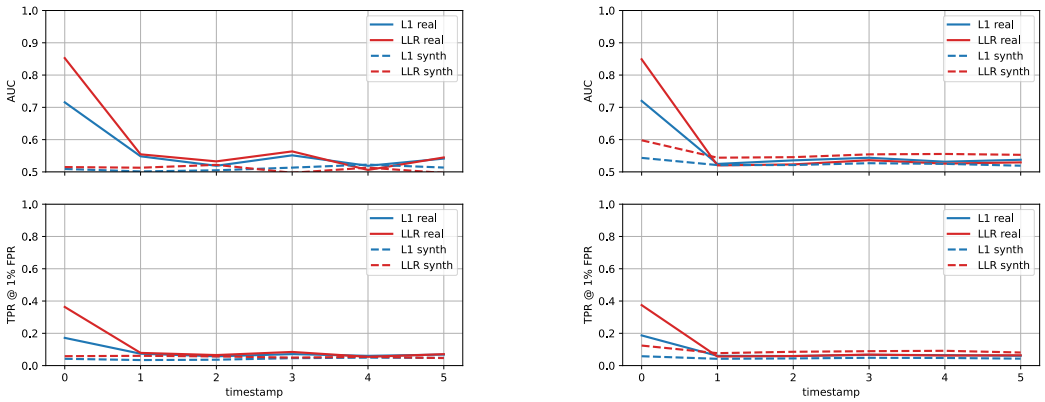


(a) Distribution of absolute per-feature changes between consecutive timepoints in the longitudinal miRNA dataset (GSE68951). Horizontal bars indicate the median; whiskers extend to the minimum and maximum. The median change is approximately 0.30 $\log_2$ units (≈1.2-fold), comparable to one within-feature standard deviation.



(b) Top: AUC; bottom: TPR @ 1% FPR. Synthetic Gaussian noise with maximum 0.9 independency between miRNAs on the longitudinal dataset. Filtering to approximately independent features reduces performance, confirming that inter-feature correlations inflate apparent attack power.

Figure 12: Temporal drift and feature-independence analysis on the longitudinal miRNA dataset.



(a) Using only over/under-regulated miRNAs relevant to lung cancer ($n = 8$).

(b) Using only the miRNAs not over/under-regulated due to lung cancer ($n = 8$).

Figure 13: For each subfigure, top: AUC; bottom: TPR @ 1% FPR. Longitudinal dataset stratified by disease-associated miRNAs.

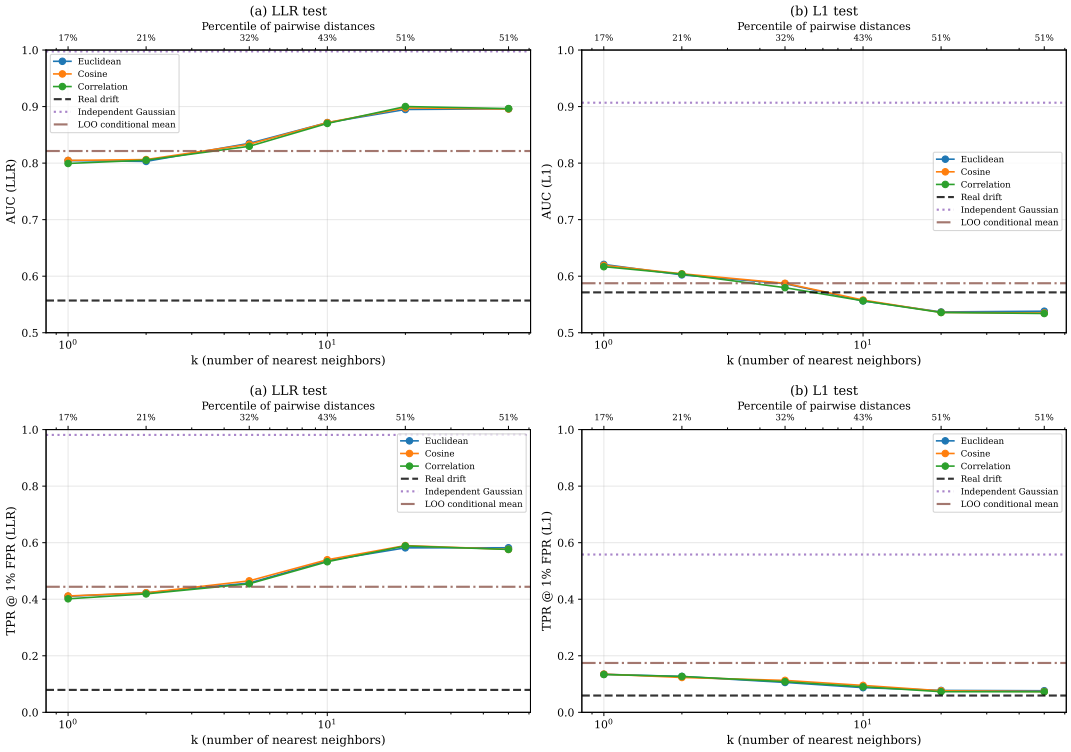


Figure 14: AUC (top) and TPR @ 1% FPR (bottom) under nearest-neighbor drift-vector swaps as a function of k and distance metric (Euclidean, cosine, correlation) on the longitudinal miRNA dataset ($n = 8$). Horizontal lines show values under real drift, independent Gaussian noise, and the LOO conditional mean predictor. (a) LLR test. (b) L_1 test. Top axis: percentile of all pairwise distances corresponding to the kth neighbor.