Robust De-anonymization of a prototypical Electronic Health Record System

Karnav Popat*, Swati Waghdhare¶, Sandeep Budhiraja‖, Anurag Agrawal†§ and Subhashis Banerjee*‡§

*Computer Science, Ashoka University
†Trivedi School of Biosciences, Ashoka University
‡Centre for Digitalisation, AI and Society
§Koita Centre for Digital Health
¶Department of Endocrinology, Max Hospital, Saket, New Delhi
‖Institute of Internal Medicine, Max Hospital, Saket, New Delhi

***Abstract*—The rapid adoption of electronic health records (EHRs) by both the public and private sectors of Indian healthcare has presented significant opportunities for healthcare innovation. At the same time, concerns have been raised about the privacy and security of personal healthcare records. This paper investigates the vulnerability of a prototypical EHR used by Max Healthcare, a leading Indian healthcare provider, to de-anonymization attacks. Using a model inspired by Narayanan & Shmatikov [20], we demonstrate that purportedly "anonymized" EHRs are susceptible to robust re-identification with minimal auxiliary information. We find that as few as four data points on an individual are enough to de-anonymize them in a dataset. We outline a novel neighbourhood attack that exploits familial relationships in EHRs to enable the de-anonymization of entire family networks even when individual family-members' data is completely secret. We show that the integration of EHRs into the budding Indian digital health ecosystem without other robust privacy measures exposes participating patients to complete medical and non-medical de-anonymization.**

1. Introduction

Electronic health records (EHRs) have become integral to modern healthcare systems, offering significant benefits for patient care, research, and policy-making. India's healthcare ecosystem, one of the largest in the world, is at a critical juncture of evolving prototypes and standards for digitalization. However, while the rapid digitalization of health records has ushered in opportunities for innovation and efficiency, it has also heightened concerns about data privacy and security [21].

The Digital Information Security in Healthcare Act (DISHA) and the Digital Personal Data Protection (DPDP) Act represent two pivotal legislative efforts aimed at addressing these concerns. DISHA focuses specifically on healthcare, mandating stringent protocols for secure storage, sharing, and processing of health data. It provides a broad framework for Indian Digital Health Data (DHD) to be turned into EHRs and shared with different hospitals and healthcare centres. It requires fiduciaries to provide guarantees of purpose limitation and anonymity. Complementing this, the DPDP Act provides a broader framework for personal data protection, including health data, and introduces key principles such as purpose limitation, data minimization, and the rights of data principals. The Indian medical industry and government machinery have become exceedingly optimistic on EHRs as the pathway to the digitalization of the healthcare system [25] [22], with pioneers from both the private and public sectors. [15] [24]

A data sharing model that can protect user privacy is essential for such digitalization endeavours. Data sharing is crucial not only for research, but also for effective intelligence gathering and epidemiology, data analytics and training of machine learning models, and topic discovery using health data. However, privacy preserving data sharing requires effective operational models for purpose limitation – ensuring that no function creep is possible – and neither of the Acts define the standards tightly enough. Data anonymization is by far the most common approach for purpose limitation of data sharing, though it is well known in computer science that it is rarely effective.

Despite efforts to anonymize sensitive health data, significant research has shown that anonymization methods are severely vulnerable to de-anonymization attacks, particularly in the context of high-dimensional and sparse datasets [20] [3] [4]. Medical data, by the nature of the information required to be stored, is often exceedingly sparse. Sparse data is especially easy to de-anonymize [17]. Specifically, this sparsity means that when adversaries possess auxiliary datasets or background knowledge, the risk of privacy breaches is further amplified. In the medical context, background knowledge about a medical subject can be easily obtained digitally or through day-to-day interactions. Existing research has shown that medical datasets can be extremely compromising to patient privacy. [10] [11] [13]

An absolute notion of privacy might be that no information about an individual should be learnable with access to a statistical database that may not already be learnt without any such access. Indeed, this was the notion that was originally introduced by Dalenius [2] and later termed as *inferential privacy* by Ghosh and Kleinberg [14]. In her celebrated result, Dwork [6] not only proved that absolute inferential privacy is impossible to achieve, but also observed that if the adversary has access to arbitrary auxiliary information, an individual's inferential privacy would be violated even when she doesn't participate in the database, because information about her could be leaked by correlated information of other individuals. This led to the development of the notion of *differential privacy* [6] [8] [9], which measures the difference in the information gained by the adversary when the individual's data is collected vs. when it is not collected; thus it measures the *additional* privacy risk an individual incurs by participating in a database. This framework provides a mathematically rigorous approach to privacy, leveraging calibrated noise to protect against a wide range of adversarial attacks, including those leveraging auxiliary information.

While differential privacy has shown significant promise in protecting aggregate query results, its application to high-dimensional and sparse datasets, such as electronic health records (EHRs), remains a challenge. This is mainly because making people indistinguishable in sparse high-dimensional dimensional datasets requires adding so much noise, that the utility of the data becomes questionable. Research building on Dwork's foundational work has explored constructions like the Laplace mechanism [8] [5] and extensions to interactive and non-interactive settings [18], yet these approaches often struggle to balance privacy guarantees with data utility in sparse domains.

The privacy risks of releasing and using "anonymized" micro-data (such as EHRs) have been widely documented [26]. Landmark studies include the re-identification of a Massachusetts hospital discharge database [23], which demonstrated how voter registration records could be used to identify patients, and privacy breaches involving anonymized AOL search data [1]. Narayanan & Shmatikov [20] extended these findings by demonstrating robust de-anonymization techniques on large, sparse datasets, highlighting that even a small amount of auxiliary information can compromise privacy. Their work underscored the fundamental challenges in anonymizing high-dimensional datasets while preserving their utility for research and analysis, which has only seen success in limited scopes [19] [7] [12].

In this paper, we use the model suggested by Narayanan & Shmatikov [20] to demonstrate that EHRs from a major Indian hospital that are purportedly "anonymized", are vulnerable to robust de-anonymization with even a small amount of background information. We specify both the formal model of privacy breach as well as the practical scenarios of EHR de-anonymization by demonstrating a de-anonymization attack on "de-identified" individual-level data from Max Healthcare using minimal auxiliary information. We outline a novel attack specific to sparse medical data, and validate it on a synthetically derived dataset. We highlight the privacy risks associated with making anonymization guarantees to released sensitive data which is statistically de-anonymizable. Finally, we apply the framework of differential privacy to examine the problem and outline mitigation strategies for more effective anonymization of such sparse EHRs.

The findings from this study underscore the critical need for alternatives and deeper analysis of data sharing models.

2. Model

Define a database D consisting of n records and m features. We are interested in Electronic Health Records in the context of personal health data, i.e. information about individual patients and treatments. We thus take each $i \in \{1, 2, \ldots, n\}$ to represent the index of one patient and each $j \in \{1, 2, \ldots, m\}$ to represent the index of one medical attribute or feature. For simplicity, we will primarily refer to a "record" r_i, one patient's row of data consisting of $x_{i0}...x_{im}$ for patient i.

Each observation x_{ij} can be of many types, but we can simplistically represent each observation as one of four types:

- A date, representing personal information or the date of a medical procedure
- A numeric value, representing the quantification of some personal or medical information
- A boolean value
- A string observation containing rich information about the patient

Even more simplistically, we can reduce the majority of observations in a medical database to a simple binary. For example, the string $gender$ feature can be reduced to is_female holding $True$ for a female patient or $False$ otherwise. Similarly, a numeric medical information can be reduced to $is\_abnormal$ representing whether the value is outside of the normal range or not. This helps us model a weaker form of attack where an adversary might not directly know information in the database but has background knowledge that informs general assumptions.

To compare two records, we need a similarity function that compares each attribute of the two records and returns a similarity score between 1 (identical) and 0 (completely distinct).

Following the convention of Narayan & Shmatikov, we refer to the set of non-null attributes of a record as the *support* of the record $supp(r)$. The support of two records

is $supp(r) \cup supp(r')$.

We then specify our similarity function as:

$$Similar(r, r') = \frac{\sum Sim(r_i, r'_i)}{|supp(r) \cup supp(r')|}$$

where $Sim(r_i, r'_i)$ is a comparison function that checks the distance of the values of r_i and r'_i in the case of numeric values, or checks if they are within some tolerance of each other otherwise. We therefore call two records "0.5 similar" if half their attributes satisfy the similarity condition of $Sim(r_i, r'_i)$, and so on. $\sum Sim(r_i, r'_i)$ is therefore the total similarity of two records on each of their common attributes. We outline the variations of the similarity measure that we experiment with in Section 4.2.2.

2.1. Dataset Sparsity

A pre-requisite for our auxiliary information deanonymization attack is that the dataset must be *sparse*. A *sparse* dataset is one where:

- A majority of records have non-null values for only a small proportion of features
- A majority of features have non-null values for only a small proportion of records

More formally, we can say that a dataset is sparse if the probability of two unique records having the same non-null values is low (below some threshold δ) [20]. We can therefore call a dataset sparse if

$$Pr[Similar(r, r') > \epsilon \quad \forall \quad r' \neq r] < \delta$$

We know that personal EHR data is sparse because the vast majority of patients only suffer from, and are tested for, specific ailments. As a result, any database of electronic health records of patients will necessarily be sparse. We show this for our dataset in Section 4.1 and Fig. 3.

2.2. De-anonymization algorithm

To demonstrate the lack of privacy in the dataset, we demonstrate a background-information attack on a prototypical EHR modelled as our database D.

We model an adversary that obtains a small amount of background information $aux(r)$ about a person representing record r in D. This $aux(r)$ is a subset of the information about the person available in record r, with some error tolerance and/or limitations. We outline the different variations of $aux(r)$ that we will experiment with in Section 4.2.1.

We hold that a successful de-anonymization attack constituting a privacy breach could be conducted if the adversary can successfully use $aux(r)$ to reidentify the patient, i.e, extract a record's information from the database D with some degree of certainty. We can quantify the certainty requirement as ϵ, the minimum similarity score between the auxiliary information and the record above which we consider it re-identified.

A de-anonymization attack is thus successful if

$$A[D, aux(r)] \rightarrow r' \in D \mid Similar(aux(r), r') > \epsilon$$

We can define the simplest adversarial algorithm A as:

- Calculate the similarity $Scores$ of r using $Similar(r, r') = \sum Sim(r_i, r'_i) \quad \forall\ r_i \in aux(r)$ for every $r' \in D$
- If $max(Scores) \geq \epsilon$ and $len(max(Scores)) == 1$, return the record with the maximum score.
- Otherwise, return the k records in $max(Scores)$ (we consider this a failure to identify the record)

This algorithm simply matches the auxiliary information against every record in D and calculates the similarity score. If the record r' with the highest similarity is unique, it returns r'. If the best match is not unique, it returns all the records which have the highest similarity score.

2.3. Auxiliary information selection $aux(r)$

For each record r that we attempt to de-anonymize, we have to select the background information $aux(r)$ which will be available to the adversary. We implement three variations in the way this auxiliary information is selected.

2.3.1. k most informative features. To establish a lower-bound on the number of features required to de-anonymize a record, we first grant the adversary the k most informative features that are present for that record. This represents the worst-case scenario where the adversary happens to have the information which would be most useful in de-anonymizing the victim.

For this purpose, We measure informativeness by the rarity of the feature. Specifically, we select the k features in $supp(r, r')$ with the lowest $supp(c)$, indicating that those features least often have non-negative values for any patient record.

2.3.2. Randomly selected features. A more realistic example is an adversary who has access to some arbitrary information set about a person and wants to use that external background knowledge to identify the victim in D. We attempt to de-anonymize each record using k features selected randomly from the non-null features of the target record r. In cases where we must select k features but $supp(r) < k$, we calculate the similarity score on $len(supp(r))$ features instead.

To reduce the variance of results, we take the average of the similarity scores from ten instances of the random selection of features for each record. Note that this has no impact for records where $supp(r) < k$ or $supp(r) \approx k$.

2.3.3. Adversary-accessible features. However, neither of these approaches truly represent any realistic scenario. Some attributes, by virtue of being less common, are much more informative for de-anonymization than others; consequently, random selection can be misleading. It is not very likely that an adversary would have access to specific information such as the patient's sodium serum or squamous epithelial cell blood test values. However, there is a pool of more "public" features that we can guess an adversary would be more likely to observe or gain access to: height, allergies, clubbing, major medical history, etc.

To demonstrate vulnerability to specific adversaries, we demonstrate three different realistic adversaries who acquire background information through plausible means and are able to successfully de-anonymize the patient.

Malicious colleague/acquaintance: In our day-to-day activities, we interact with a large number of acquaintances of uncertain intent. Any of these individuals would possess some knowledge of our personal information and lifestyle (travel habits, allergies, etc.) or could acquire it if determined. For example, a malicious colleague would know the answers to:

- Do you have any skin allergies?
- How old were you when you started consuming alcohol regularly?
- How often do you Travel or go on Vacation?
- Have you travelled domestically in last one Year?

Note that this is not just information that a colleague would know; it is information that many people, including public-facing individuals, publish on the internet for anyone to see. Instagram and Twitter accounts reveal detailed social activity including data points that reveal or are highly correlated to relevant medical information. Many professions, which can be easily identified through legitimate means, are similarly highly correlated with medically relevant information. Many of these datapoints can also be inferred from one other because of strong correlation, such as between the consumption of alcohol and nicotine. The bar for obtaining this information, as a result, is quite low.

Malicious medical insider: EHRs are designed to be used and shared by medical institutions and professionals to make treatment simpler. Insider attacks or improprieties in the handling of this data can cause severe vulnerabilities. The minimum information that such a vulnerability would expose would be the patient's current medical state and the medical history required to diagnose them. This exposure includes features like:

- Diabetes Mellitus (Y/N)
- Year of Start
- Year of End
- Duration
- Surgery Name (Y/N)
- Year of Surgery
- Place of Surgery

Once again, note that this doesn't just include the example case of a regular general practitioner or doctor who attended to the patient personally. While that is certainly one vector to obtain the information, it is also possible for this information to be provided to the adversary through badly purpose-limited systems such as ABDM.

Smart medical app: There are several health-related apps and services which collect medical data from consumers who voluntarily offer it to the operators for utility benefits. The subset of features collected by smartwatches and such services is not nearly enough (since they would only have access to vitals and lifestyle/travel habits). However, there are several smart home blood test services and similar services which collect medical information, sometimes with vulnerable security and obscure terms and conditions. This includes:

- Blood pressure
- Temperature
- Absolute monocyte count
- Drug name (Y/N)
- Strength
- Frequency
- Duration

By attempting de-anonymization on the database D using auxiliary information drawn from different subsets of features representing realistic adversaries, we can show that the database is vulnerable to real-world malicious actors under more realistic assumptions.

2.4. Similarity scoring function $Sim(r_i, r'_i)$

To simulate different ranges of inaccuracy in the adversary's knowledge, we can add error tolerances to our comparison function, i.e., reduce the precision with which the adversary knows a feature.

2.4.1. Perfect information: . The most simple model of an attack is an adversary that directly obtains k features of auxiliary information about a particular record r from the database D. We assume that this information is perfect, i.e. there is no noise added to the database and no imperfections in the adversary's information.

However, especially in the medical context, only a limited range of adversaries would have perfect information from the database. Given that features include numeric values like blood test results, it's more likely that an adversary such as an insider or an eavesdropper would have less specific information about certain attributes.

2.4.2. 20% error tolerance: . In the event of the adversary "guessing" or extrapolating features such as medical measurements, or deriving them from a different source, there would be errors in their information. To investigate this scenario, We allow an error tolerance of 20% in all numeric features. If $r_i \in aux(r)$ is within 20% of $r'_i \in r'$, $Sim()$ returns a 1, otherwise 0.

Note that, as Narayanan & Shmatikov established, there is no formal difference between an error rate in the adversary's information and statistical noise added to D for security. As a result, this scenario models both the cases where either the adversary has made slightly incorrect "guesses" about the information or the EHR has statistical noise added to it to protect against adversarial attack.

2.4.3. Presence of non-null attributes: . A result of the severe sparsity of the database is that the mere presence of a non-null attribute for a record r is revealing, regardless of its value. Thus, we now ignore the values altogether and only check whether or not the attribute is non-null. If both r_i and r'_i contain a non-negative value (for eg. $is\_abnormal$ is $True$ for both), $Sim()$ returns a 1, otherwise 0. This allows us to model a scenario with much lower requirements on the attacker, who need only know general information such as whether a patient has any allergy at all, or has ever been tested for a particular disease.

To implement this comparison, we select every feature m in the Max Health database where $count(mode(m)) > \delta\% * count(m)$. For $\delta = 50$, this is true for 89% of the 629 features in the database (558 features). For $\delta = 66$, this is true for 83% of the database (521 features). For each of these, we replace the modal value with $False$, and any other value with $True$.

This approach has the consequence of showing that our attack is robust to protection schemes such as bucketization [16] which prevent raw numeric values from being released.

2.5. Neighbourhood attack algorithm

We can further outline a novel attack made possible by the nature of personal EHRs: the neighbourhood re-identification attack. Literature on privacy in social network data has found that "neighbourhood attacks" can reveal information about a victim even if the victim themselves is not present in the database. Even more easily, neighbourhood attacks can be used to take a single de-anonymized record and identify multiple records in the database without any incremental auxiliary information.

This attack is possible because personal EHRs such as the Max Health dataset require significant amounts of information on a person's familial history to help identify hereditary and environmental disorders. Our dataset includes 231 features relating to relatives' medical history, as illustrated in Figure 1. In a dataset that is sufficiently sparse, it is likely that two records with corresponding medical histories of at least two relatives, will be from the same family tree.

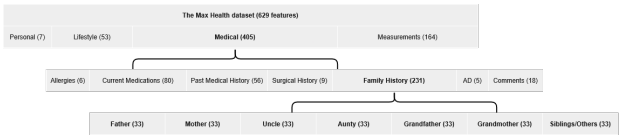

Figure 1. The hierarchy of family-related features in the Max Health EHR

We model the attack using the same database D as before. We add the condition that each $r \in D$ must contain some n attributes which contain information about the individual's relatives. We denote these attributes as r_{mi} through $r_{(m+n)i}$, where m denotes the index of the attribute in r, and i is the index of the relationship to the individual represented by r. We take an adversary that uses background information $aux(r)$ to successfully identify the record $r' = r \in D$ using the algorithm A from the previous section.

We denote this de-anonymized record as $r*$, and the records of the relatives of patient $r*$ as $r*_i$. We know that there are some attributes $r*_{mi}$ through $r*_{(m+n)i}$ which hold medical information on the relative i of the patient represented by $r*$.

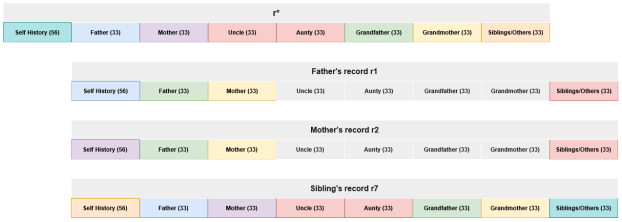

Figure 2. A family of records in D. Corresponding colours indicate fields that can be used as $aux(r*_i)$ once $r*$ is identified.

By treating this information as the auxiliary information $aux(r*_i)$ available to the adversary to de-identify the relative, we can construct an adversarial algorithm similar to A which would allow the adversary to de-identify every relative of the patient whose information is present in $r*$.

This neighbourhood attack is thus successful if

$$A'|D, r*| \rightarrow r*_i \in D \mid Similar(r*_{mi},\ r*_i) > \epsilon$$

We can define the neighbourhood attack's adversarial algorithm A' as:

- Use algorithm A with $aux(r)$ to successfully de-anonymize a record $r*$
- For any i where $r*_{mi}$ is non-null, assign $aux(r*_i) = r*_{mi}$

- Use this auxiliary information with A and D to find the record $r*_i$ representing the relative of $r*$
- Recursively repeat steps 2 and 3 for all $r*_{imj}$ in $r*_i$, representing the relatives of the relatives.
- Repeat steps 2, 3 and 4 for all values of i in $r*$

The consequences of this attack are severe; by identifying one patient's record through the acquisition of a small, single-digit number of information points, it is possible to set off a chain reaction which de-anonymizes every member of their immediate and extended family who have a medical history. Since many families in India consult the same doctor or visit the same hospital when needed, this represents a significant attack vector at the household level.

There are two limitations to this attack:

- In cases where the medical history of the relative is not present in the original de-anonymized record $r*$, either because of the relative has no medical history or because of incomplete self-reporting, it is not possible to de-anonymize the relative. It is still possible, however, to accomplish the lesser de-anonymization of revealing whether the relative's record is in the database D or not.
- The attack can only be propagated upwards. Due to the nature of heredity, in our dataset and in the majority of EHRs described in literature, medical information on the patients' descendants is never included. As a result, while the attack can be used to de-anonymize the patient's family tree, it cannot be used iteratively to de-anonymize the entire dataset.

However, these limitations do not take away from the primary advantage of this attack; it severely reduces the background information necessary to de-anonymize multiple people in a dataset by making use of networks within the records.

3. Data & Methods

3.1. The Max Health Dataset

To test our de-anonymization model on real-world EHRs, we use an anonymized (de-identified) dataset provided by Max Health.

Table 1 contains a summary of the features in the provided data. The dataset is long-format and cross-sectional, containing 2,692 patient records with 629 possible features for each patient. 442 of the 629 features consist of survey responses, relying on self-reporting by the patient or relatives, along with information from past medical records. 187 features consist of measurements and observations including the results of medical tests and information such as height, weight, pallor, etc.

Type	Category	Feature	Count
Survey	Personal	Personal	7
Survey	Lifestyle	Sleep	1
Survey	Lifestyle	Physical Activity	15
Survey	Lifestyle	Food habits	5
Survey	Lifestyle	Alcohol History	12
Survey	Lifestyle	Travel History	5
Survey	Lifestyle	Social & Personal History	15
Survey	Medical	Allergies	6
Survey	Medical	Current Medications	80
Survey	Medical	Past Medical History	56
Survey	Medical	Past Surgical History	9
Survey	Medical	Family History	231
Observation	Medical	Anthropometric Data	5
Observation	Medical	Investigation Comments	18
Observation	Measurement	Vitals	4
Observation	Measurement	Systemic Examination	8
Observation	Measurement	General Examination	13
Observation	Measurement	Blood Test Report	136
Observation	Measurement	Radiographs/Spirometry	3
		Total	**629**

TABLE 1. MAX HEALTH DATASET FEATURES

The features in the dataset can also be categorized based on the information they contain. There are 7 features holding personal data such as gender, date of birth, marital status, etc. There are 53 features pertaining to the patient's responses on their lifestyle, such as travel habits, alcohol consumption, etc. There are 382 features containing medical information, such as allergies, patient's history with specific diseases (whether they have had it and when,), and patients' relatives' history with the diseases. Finally, there are 164 features representing various tests and measurements, from pulse and temperature to platelet count and various other blood test components.

This dataset, as expected of an individual-level set of personal health records, is sparse. The majority of records have fewer than 200 out of 692 features with non-negative values, and the majority of features have fewer than 100 records with non-negative values. A full description of the dataset with detailed features and examples is included in Appendix A.

Per our formal specification of sparsity, we can show that the Max Health dataset is sparse by evaluating the $Similarity(r, r')$ of each pair of rows in the dataset. If the dataset is sparse, we would expect the majority of records r to have no record r' with a high degree of similarity.

Fig. 3 demonstrates that this is true for the dataset. Fig. 4 similarly shows that this is equally true when considering the sparsity of the features themselves. As we can see, most of the features are only non-negative for a single-digit number of patients; this is why we can be certain that having even a small amount of information about a person's medical history, habits, etc. can be strongly indicative.

Fig. 5 shows that our dataset is sparse by considering

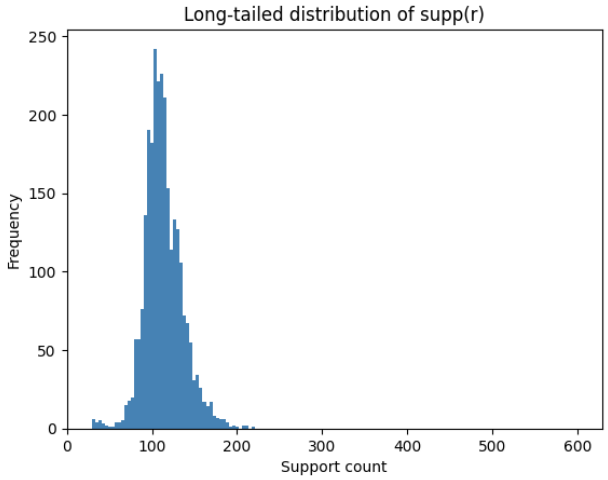

Figure 3. Long-tailed distribution of number of non-negative columns per person

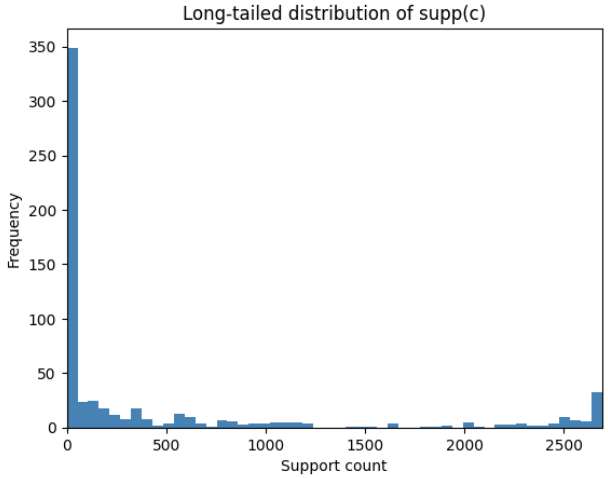

Figure 4. Long-tailed distribution of number of non-negative records per attribute

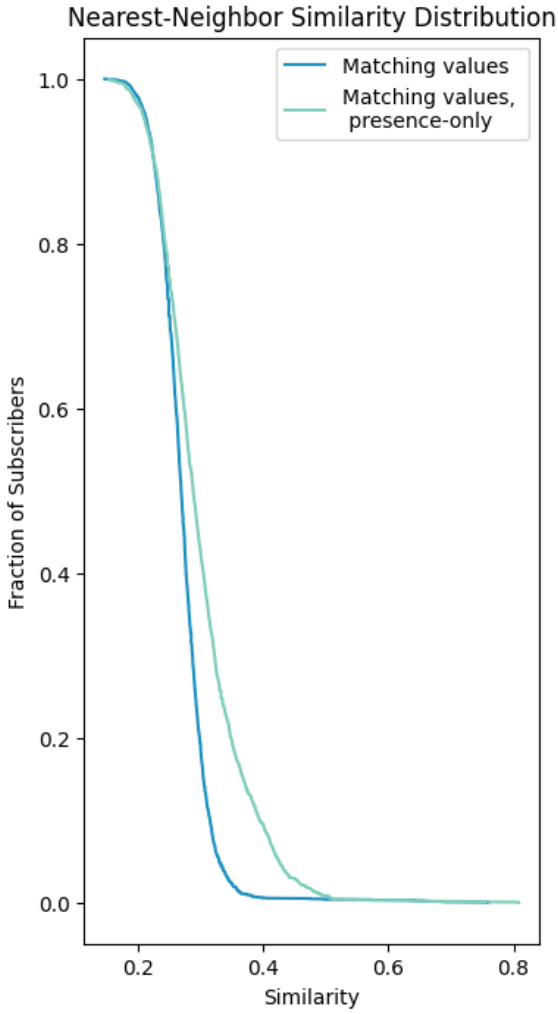

Figure 5. The majority of records in the Max Health dataset have no other record with $Similarity > 0.4$

each record's most similar record in the dataset. By comparing every feature in every record against every other record, we can show that for the vast majority of records, practically all records are at least 60% dissimilar from every other record; even when considering only the presence of a non-negative value, practically all records are at least 50% dissimilar from every other record.

3.2. Experiments

3.2.1. De-anonymization attack. We conducted a series of thirty experiments, with different combinations of the auxiliary information selection function and similarity function.

With the auxiliary information $aux(r)$ available to the adversary, we use the adversarial algorithm $A(D, aux(r))$ to attempt to find the equivalent unique record in the database. To compare $aux(r)$ against the records in D, we calculate the $Similarity\ Score$ using $Sim(r_i, r'_i)$.

To evaluate the de-anonymization potential, for each record $r \in D$, we extract $aux(r)$ based on the variant of the auxiliary information function we've chosen. We then find $A|D, aux(r)| \rightarrow r'$. If $r == r'$, i.e., the indices of the source and target record match, we have successfully de-anonymized the record r.

Note that this requires a modification of our adversarial algorithm; instead of setting a minimum similarity ϵ for us to consider the record identified, we set ϵ' as the maximum similarity of r with any record except itself. Thus, if A returns a set of n records with similarity $> \epsilon$, we can consider it de-anonymized if r is the best match for itself and the second-best match has similarity $< \epsilon'$.

Our method of evaluating the security of the database is exceeding simple: given our various assumptions, what is the smallest number of features k required to be in $aux(r)$ for us to successfully de-anonymize r with no other match $> \epsilon'$? The smaller the required value of k (i.e, the adversary's background information), the less secure the database. We therefore calculate the similarity scores for each $r \in D$ for each value of k from 1 to 16 with $\epsilon = 1$ and $\epsilon' = 1$.

We carried out the experiments summarized in Table 2:

Figure	aux(r)aux(r)	Sim(r,r')Sim(r,r')
6	Most Informative	Perfect
6	Most Informative	20% Tolerance
6	Most Informative	Presence-only
7	Random	Perfect
7	Random	20% Tolerance
7	Random	Presence-only
8	Acquaintance	Perfect
8	Acquaintance	20% Tolerance
8	Acquaintance	Presence-only
9	Insider	Perfect
9	Insider	20% Tolerance
9	Insider	Presence-only
10	App	Perfect
10	App	20% Tolerance
10	App	Presence-only

TABLE 2. DE-ANONYMIZATION EXPERIMENTS

4. Results

4.1. Adversary has most informative $aux(r)$

First, we establish the worst-case scenario by providing the adversary with the k most informative features for each record measured by rarity of non-null values. Although this is not a realistic attack scenario, it provides a useful baseline for comparison against different allowed error rates.

4.1.1. Perfect information. Under the perfect information assumption, we find that only 4 features are sufficient to identify $> 90\%$ records in the Max Health database with high certainty (no other record with > 0.8 similarity). With an additional feature (5 in total), this is true of more than 97% of records.

4.1.2. 20% error tolerance. Introducing a $+-20\%$ error tolerance in matching the adversary's information against the database (equivalent to adding upto 20% of noise in the database) has next to no effect on the de-anonymization power, still requiring 4 features to identify records with high certainty.

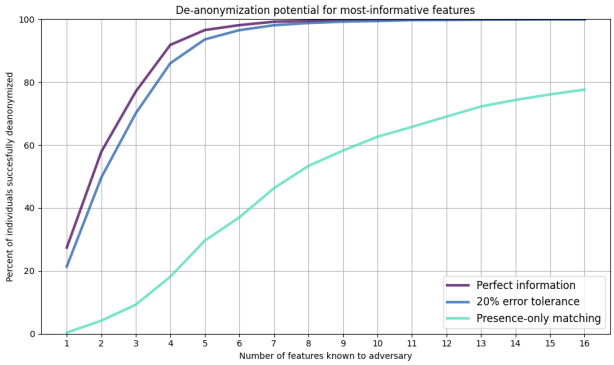

Figure 6. De-anonymization with k most informative features

4.1.3. Presence-only information. Finally, we test without matching any values, checking only to see whether the record has a non-null value for each attribute. This simulates the most robust condition where an adversary might have limited or no information about the details of a patient's medical condition, only that the patient has some history with a disease or lifestyle or whether they underwent the test in question.

We find that this reduces the de-anonymization power significantly, requiring 10 features to reach $> 60\%$ success rate for most informative features, while it fails to reach higher de-anonymization power entirely. This shows that only having access to presence/absence information is enough to find the relevant record occasionally, but mostly fails to distinguish the record from peers. This doesn't take away from the power of the attack as, if an adversary can limit the pool of candidates to a small number using this heavily-constrained information, it makes further de-anonymization much more feasible.

4.2. Randomly selected features

Our attack remains feasible for the randomly selected set of features, which exhibit similar or identical performance to the most informative set. Due to the extreme sparsity of the dataset, there is little difference between a set of k randomly selected features for each record and the set of k most informative ones when the information is perfect, and somewhat worse when the information has errors.

4.2.1. Perfect information. We find that with as few as 5 randomly selected features available to the adversary as $aux(r)$, it is possible to uniquely identify $> 95\%$ of records with high certainty (no other record with > 0.8 similarity).

4.2.2. 20% error tolerance. With randomly selected features and an error margin ($+-20\%$), we require 10 features to de-anonymize $> 90\%$ of records in the database. This falls to 6 features for a tolerance of 60%.

Figure	$aux(r)$	$Sim(r, r')$	k for $> 90\%$	k for $> 80\%$	k for $> 60\%$
6	Most Informative	Perfect	4	4	3
6	Most Informative	20% Tolerance	5	4	3
6	Most Informative	Presence-only	-	-	10
7	Random	Perfect	5	4	3
7	Random	20% Tolerance	10	8	6
7	Random	Presence-only	-	16	11
8	Acquaintance	Perfect	4	3	3
8	Acquaintance	20% Tolerance	10	8	6
8	Acquaintance	Presence-only	-	-	-
9	Insider	Perfect	6	4	3
9	Insider	20% Tolerance	7	5	4
9	Insider	Presence-only	-	-	-
10	App	Perfect	4	4	3
10	App	20% Tolerance	11	9	7
10	App	Presence-only	-	-	-

TABLE 3. DE-ANONYMIZATION RESULTS

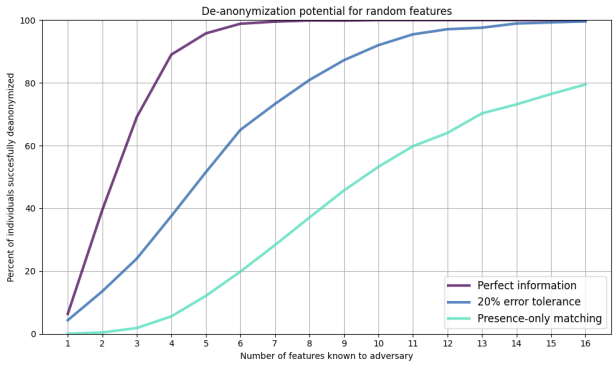

Figure 7. De-anonymization with k randomly selected features

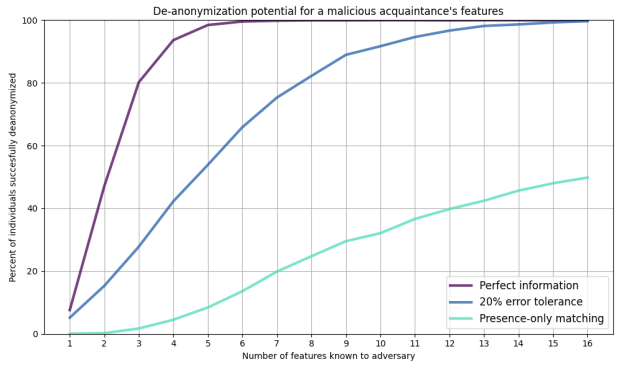

Figure 8. De-anonymization with malicious colleague's information

4.2.3. Presence-only information. With presence-only matching, the percentage of de-anonymized records stays below the threshold values, though it is not unsuccessful, reaching as high as $50\%+$. We find similar cases for the different subsets of information that we considered, revealing that presence-only information fails to de-anonymize a record in the absence of at least one or two features of real data.

Finally, we test our attack using different subsets of auxiliary information representing the information realistically available to different adversaries.

4.3. Malicious colleague's information

4.3.1. Perfect information. Under the assumption that a malicious colleague has perfect information about a subset of features, we find that 4 features are sufficient to identify $> 90\%$ of records in the Max Health database with high certainty, and 3 features for $> 80\%$ certainty. This is comparable to the performance observed with the most informative features, which is particularly concerning since this is the class of information that is most accessible not just to adversaries in contact with the victim but also those observing the victim's internet presence.

4.3.2. 20% error tolerance. Introducing a $+-20\%$ error tolerance requires 9 features to identify $> 90\%$ of records and 8 features for $> 80\%$ certainty. This shows greater impact from noise compared to perfect information scenarios, reflecting that these features are fairly tightly bound.

4.4. Medical insider's information

4.4.1. Perfect information. A medical insider, such as a nurse or lab technician, may have access to a broader range of features. Similarly, an adversary within the medicla system or who has gained access to data shared within EHRs or an ABDM-like ecosystem, has access to many features. With perfect information, we find that 5 of these features are sufficient to identify $> 90\%$ of records with high certainty, and 4 features for $> 80\%$ certainty.

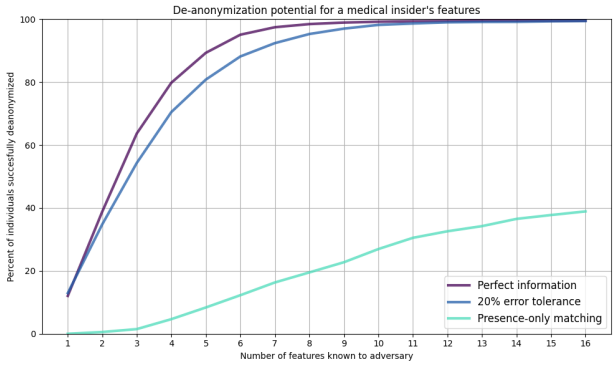

Figure 9. De-anonymization with medical insider's information

4.4.2. 20% error tolerance. With a $+-20\%$ error tolerance, the medical insider requires 7 features to de-anonymize $> 90\%$ of records and 5 features for $> 80\%$ certainty. This demonstrates moderate robustness against noise and reveals that there is an even higher risk compared to the more-publicly accessible class of information previously considered.

4.5. Smart medical service's information

4.5.1. Perfect information. A smart medical service, such as a health tracking app, may have access to a curated set of health-relevant features. With perfect information, we find that 4 features are sufficient to identify $> 90\%$ of records with high certainty, and 3 features for $> 80\%$ certainty. This is again comparable to the performance of the most-informative features.

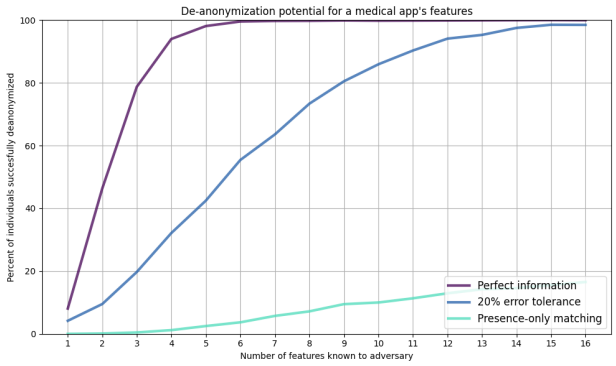

Figure 10. De-anonymization with smart medical service's information

4.5.2. 20% error tolerance. With a $+-20\%$ error tolerance, the smart medical service cannot reach $> 90\%$ de-anonymization, but achieves $> 80\%$ with 9 features and $> 60\%$ with 7 features.

5. Discussion

Our findings reveal significant privacy vulnerabilities in EHRs, particularly when realistic adversaries possess even small amounts of external information. Our de-anonymization experiments demonstrated that many of the stakeholders in the EHR ecosystem, when able to "guess" as few as 4 to 11 features, could potentially re-identify a substantial proportion of patients in the dataset. Any malicious acquaintance, or any online medical service, when given seemingly innocuous details such as lifestyle habits or basic medical history, could exploit the sparsity of EHRs to compromise privacy. This vulnerability is robust to various constraints, including the extent of the adversary's knowledge, random errors in their "guesses", and limitations based on the nature of information plausibly available to them. The ease with which realistic adversaries can breach privacy highlights the urgent need for caution when introducing personal and sensitive data into the digital healthcare ecosystem.

One caveat of our attack model is that, in the case of adversaries using semi-public or "guessable" information, the adversary must guess up to 10 uncorrelated data points to successfully execute the attack. For example, information about a person's smoking history might be encoded in multiple columns, but each of those columns is highly correlated to one another; as a result, the adversary must obtain information on seemingly independent characteristics such as smoking history, travel habits, etc. to create a uniquely-identifiable combination of features.

The implications of our specified neighbourhood attack further amplify the privacy risks associated with EHRs, as we demonstrate that an individual's data can be compromised even if their own information remains entirely private from any adversary. By leveraging familial relationships encoded in EHRs, an adversary can de-anonymize not only the target patient but also their relatives, creating a cascading effect that extends the breach across entire family networks. This attack is particularly concerning in the Indian context because families often share medical services and personnel, and because familial medical history is often too critical a component of patient records to remove.

The risk of escalation is further increased by the integration of these EHRs into the Indian health ecosystem through the government's Ayushman Bharat Digital Mission (ABDM). The inter-connectedness of datasets means that the de-anonymization of a single EHR can have far-reaching consequences. Under the ABDM model, each record in the EHR is embedded with a unique identifier (the Ayushman Bharat Health Account (ABHA) ID), which links to the patient's record with every other healthcare provider and which is in turn linked to the patient's government identity (Aadhar ID). This means that a successful de-anonymization attack on one EHR could potentially expose an individual's identity across the entire Indian digital ecosystem. This breach could reveal not only sensitive medical information but also other critical personal data, including financial identities and linked demographic details.

The harms that we outline are therefore severe. The simple de-anonymization of an EHR that we have demonstrated can be escalated in both breadth (to the victim's family) and in depth (to the victim's non-medical data). Once an adversary collects the small amount of background information needed to identify the victim in the "weakest link" healthcare provider's EHR, they can then use that to identify the victim across the healthcare ecosystem, escalate the de-anonymization to their family members, and obtain critical private details for every family member with a digital health record. Such a scenario would not only compromise individual privacy but also erode public trust in the digital healthcare system.

As Cynthia Dwork's work establishes, the inherent sparsity of personal datasets makes meaningful privacy practically unattainable. Even when an individual's data is not explicitly included in a dataset, their privacy can still be compromised through correlations and auxiliary information available to adversaries. This is particularly concerning in the context of Indian healthcare, where the scale and critical nature of medical data exacerbate these vulnerabilities. With one of the largest healthcare ecosystems in the world, India's rapid digitalization of health records creates vast datasets that are both high-dimensional and sparse, making them prime targets for de-anonymization attacks. The sensitivity of medical data, combined with the potential for widespread privacy breaches, poses significant risks to patient trust and the integrity of the healthcare system.

The only viable way to safeguard sensitive health data in the face of these vulnerabilities is through the implementation of strongly restrictive access paradigms and strong purpose limitation mechanisms. Access controls must ensure that only authorized entities can access specific subsets of data, while purpose limitation must strictly define and enforce the permissible uses of the data, preventing function creep and unauthorized secondary uses. However, this is not currently the case in Indian EHR systems such as Max's, nor in the government's electronic health project, the Ayushman Bharat Digital Mission (ABDM). To preserve privacy in the digital healthcare ecosystem, it is imperative to adopt stricter access controls, implement purpose limitation at both the policy and technical levels, and ensure compliance through rigorous auditing and accountability measures. Without these changes, the promise of digital health in India will remain overshadowed by significant privacy risks.

6. Conclusion

This paper demonstrates a viable avenue of attack to uniquely identify individuals in datasets which claim to be securely anonymized. It creates a formal method of specification to show that this is a quality inherent to the nature of personal data, especially in the medical domain. It further specifies methods to propagate the de-anonymization to more data belonging to the individual and their relatives.

While we successfully show that there are privacy risks with the existing approach to medical anonymization, and suggest mitigation techniques from a systems perspective, there are no algorithmic methods of securing privacy on large, sparse, personal datasets. This has implications on the vulnerability of EHRs in the Indian case as well as globally, particularly in a rapidly innovative AI-driven world.

Further investigation is necessary to establish the true risks to privacy when EHRs are used in tightly linked and cross-referenced systems such as ABDM. We hope that this paper highlights the need for a new approach to the storage of sensitive personal data at scale.

Acknowledgments

The authors would like to thank the Centre for Digitalization, AI & Society and the Trivedi School of Biosciences at Ashoka University for facilitating the research process. We are sincerely grateful to Dr. Abhaya Indrayan for constructive comments and the team at Max Healthcare for providing the prototypical EHR.

References

[1] Michael Barbaro and Tom Zeller. A face is exposed for aol searcher no. 4417749, 2006. Accessed: 2025-01-01.

[2] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977.

[3] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3:1376, 2013.

[4] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[5] Jing Dong, Matthew Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

[6] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[7] Cynthia Dwork, Nitin Kohli, and Deirdre K Mulligan. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2):1–30, 2019.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, page 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[9] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[10] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.

[11] Khaled El Emam, Alexander Yakovlev, Angela Neisa, and Elizabeth Jonker. Evaluating the privacy risks of de-identified data. *Journal of the American Medical Informatics Association*, 15(5):627–637, 2008.

[12] Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053, 2024.

[13] Katherine Gariépy-Saper and Nicholas Decarie. Privacy of electronic health records: A review of the literature. *Journal of the Canadian Health Libraries Association*, 42(1):74–84, April 2021.

[14] Arpita Ghosh and Robert Kleinberg. Inferential privacy guarantees for differentially private mechanisms. *CoRR*, abs/1603.01508, 2016.

[15] Santosh G. Honavar. Electronic medical records – the good, the bad and the ugly. *Indian Journal of Ophthalmology*, 68(3):417–418, March 2020.

[16] J. Jayapradha, M. Prakash, Youseef Alotaibi, Osamah Ibrahim Khalaf, and Saleh Ahmed Alghamdi. Heap bucketization anonymity—an efficient privacy-preserving data publishing model for multiple sensitive attributes. *IEEE Access*, 10:28773–28791, 2022.

[17] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[18] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[19] Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636. ACM, 2009.

[20] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, 2008.

[21] W Nicholson Price II and I Glenn Cohen. Privacy in the age of big medical data. *Nature Medicine*, 25:44–56, 2019.

[22] Sunil Kumar Srivastava. Adoption of electronic health records: A roadmap for india. *Healthcare Informatics Research*, 22(4):261–269, October 2016.

[23] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*, 25(2-3):98–110, 1997.

[24] Apurva Venkat. Max healthcare improves patient safety with e-health record system. How-To, October 2017. Accessed: 2025-01-01.

[25] Manisha Wadhwa. Towards a new indian model of information and communications technology-led growth and development. Technical report, CSD Working Paper Series, March 2020. Working paper.

[26] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren. Information security in big data: Privacy and data mining. *IEEE Access*, 2:1149–1176, 2014.

7. Appendix A

TABLE 4: Max Health dataset features

Type	Category	Feature	Column Examples	Examples	Type
Survey	Personal	Personal	Year of birth	1996	Continuous
			Gender	Male	Boolean
Survey	Lifestyle	Sleep	On average how many hours sleep do you get per day (in mins)	420	Continuous
Survey	Lifestyle	Physical Activity	Does your work involve vigorous intensity activity for at least 10 minutes continuously?	Yes	Boolean
			In a typical week, on how many days do you do vigorous intensity activities for work?	4	Continuous
			How much time do you spend doing vigorous-intensity activities at work on a typical day?	30	Continuous
Survey	Lifestyle	Food habits	In a typical week, on how many days do you eat fruit?	7	Continuous
			How many servings of fruit do you eat on one of those days?	2	Continuous
Survey	Lifestyle	Alcohol History	Do you consume any alcoholic products?	Yes	Boolean
			How old were you when you started consuming alcohol regularly?	25	Continuous
			How frequently have you had at least one standard drink?	1-4 Days per week	Categorical
			During each of the past 7 days, how many drinks did you have each day? 1. Monday	0	Continuous
Survey	Lifestyle	Travel History	How often do you travel or go on vacation?	Once a year	Categorical
			Have you travelled domestically in last one year?	Yes	Boolean
Survey	Lifestyle	Social & Personal History	Do you currently smoke any tobacco products daily?	No	Boolean
			In the past, did you ever smoke any tobacco products daily?	Yes	Boolean
			How old were you when you first started smoking?	25	Continuous
			How many of these products do you use a day? 1: Manufactured cigarettes	0	Continuous
Survey	Lifestyle	Allergies	Do you have any allergies?	No	Boolean
			Drug Allergies	0	Categorical
			Food Item Allergies	0	Categorical
Survey	Medical	Current Medications	Drug Name	Metformin	Categorical
			Strength	500mg	Categorical
			OD/BD/TDS/Weekly/SOS	BD	Categorical
			Duration	Daily	Categorical
Survey	Medical	Past Medical History	Diabetes Mellitus	Yes	Boolean

Continued on next page

TABLE 4 - continued from previous page

Type	Category	Feature	Column Examples	Examples	Type
			Year of Start	2016	Continuous
			Year of End	0	Continuous
			Duration	Ongoing	Categorical
Survey	Medical	Past Surgical History	Surgery Name	Appendectomy	Categorical
			Year of Surgery	1988	Continuous
			Place of Surgery	Kota, Rajesthan	Categorical
Survey	Medical	Family History	Diabetes Mellitus (Father)	No	Boolean
			Year of Start (Father)	0	Continuous
			Year of End (Father)	0	Continuous
			Duration (Father)	NA	Categorical
Observation	Medical	Anthropometric Data	Height (cm)	180	Continuous
			Weight (kg)	83	Continuous
Observation	Medical	Investigation Comments	Dexascan	No	Boolean
			Comments/Reports	0	String
Observation	Measurement	Vitals	Pulse	85	Continuous
			Blood Pressure	110/80	Categorical
			Temperature	Normal	Categorical
Observation	Measurement	Systemic Examination	Respiratory system	Normal	Categorical
			Cardiovascular system	Normal	Categorical
			Abdomen	Normal	Categorical
Observation	Measurement	General Examination	Build	Average	Categorical
			Nourishment	Proper	Categorical
			Pallor	No	Categorical
Observation	Measurement	Blood Test Report	Absolute monocyte count	0.34	Continuous
			Absolute neutrophil count	3.74	Continuous
Observation	Measurement	Radiographs/ Spirometry	X-Ray (Chest/PA)	NA	Continuous
			PFT	NA	Continuous

8. Appendix B

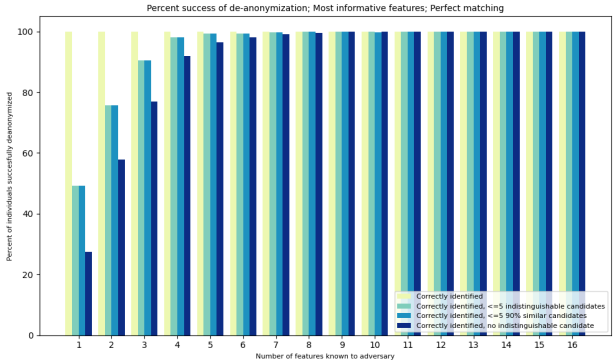

Figure 11. Performance at different thresholds for feature count 1

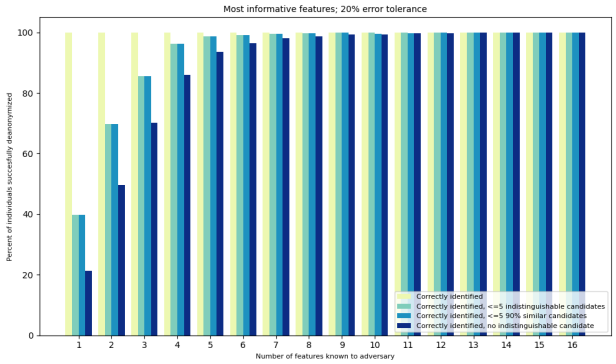

Figure 12. Performance at different thresholds for feature count 2

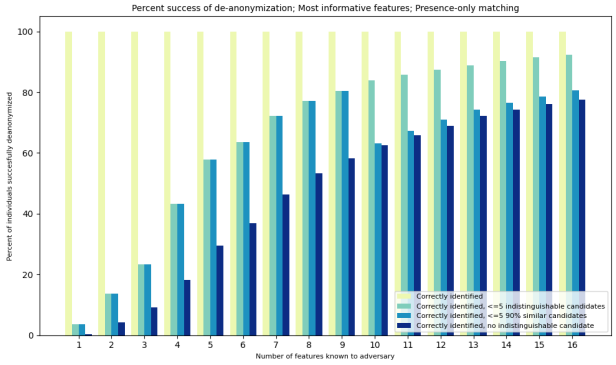

Figure 13. Performance at different thresholds for feature count 3

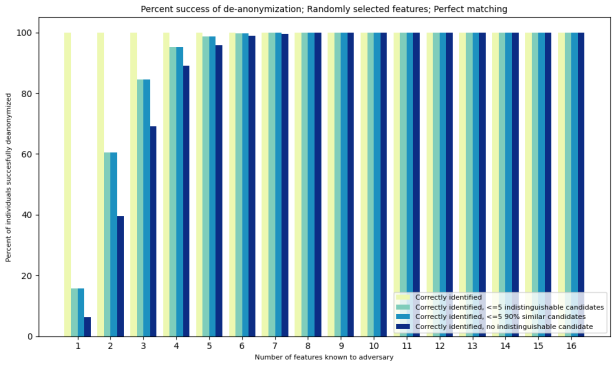

Figure 14. Performance at different thresholds for feature count 4

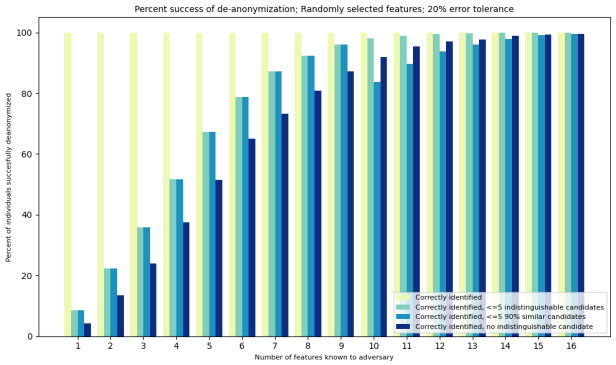

Figure 15. Performance at different thresholds for feature count 5

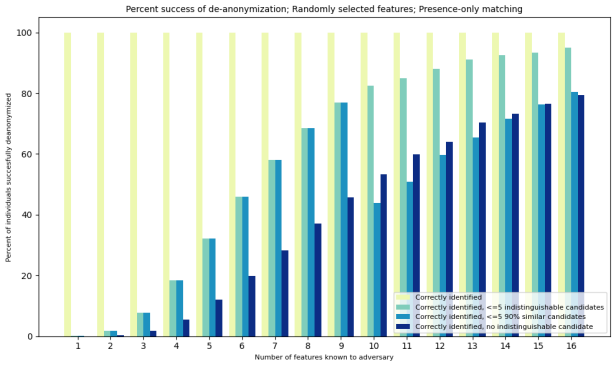

Figure 16. Performance at different thresholds for feature count 6

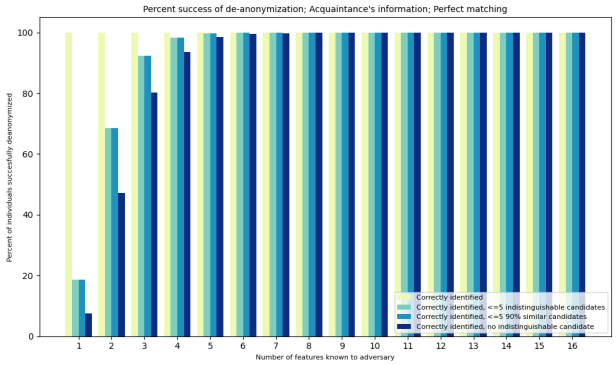

Figure 17. Performance at different thresholds for feature count 7

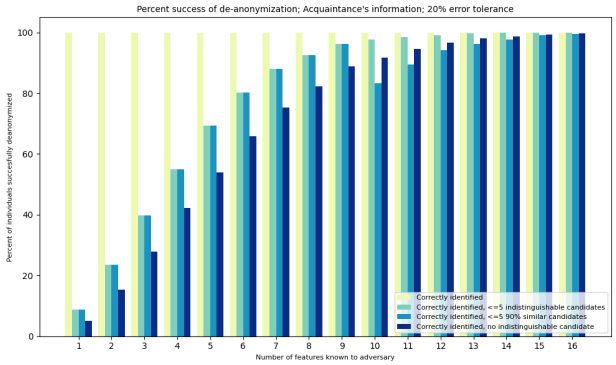

Figure 18. Performance at different thresholds for feature count 8

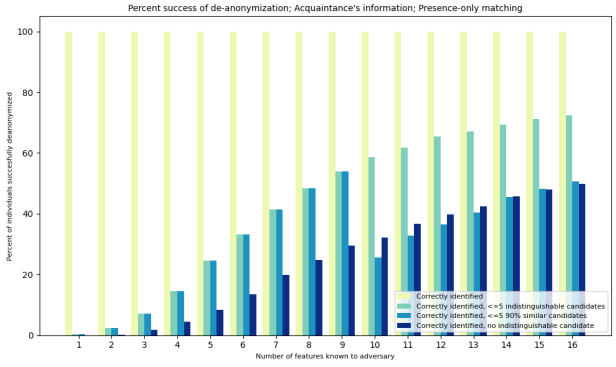

Figure 19. Performance at different thresholds for feature count 9

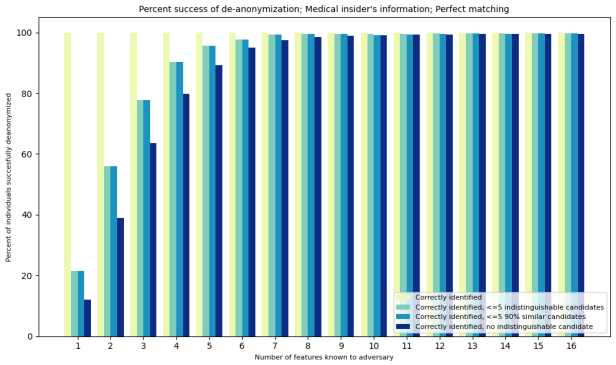

Figure 20. Performance at different thresholds for feature count 10

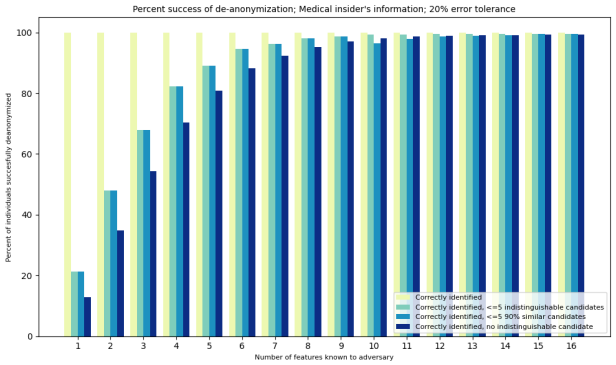

Figure 21. Performance at different thresholds for feature count 11

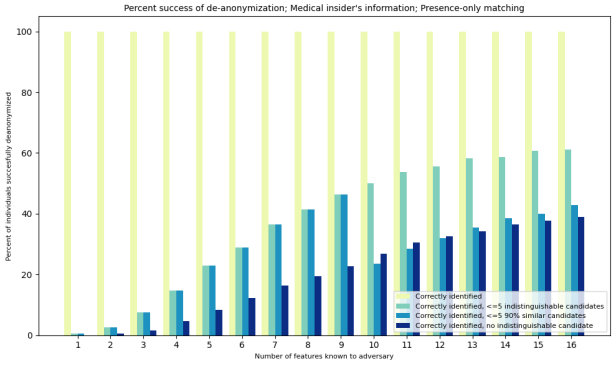

Figure 22. Performance at different thresholds for feature count 12

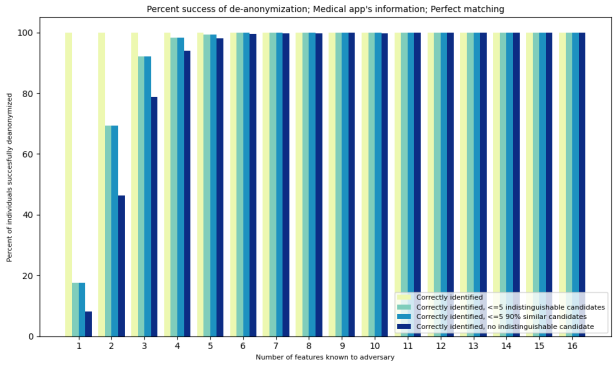

Figure 23. Performance at different thresholds for feature count 13

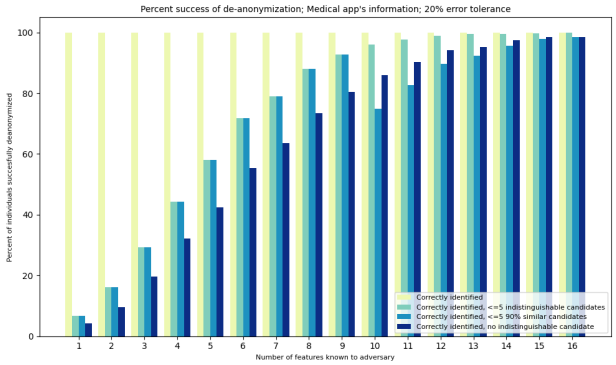

Figure 24. Performance at different thresholds for feature count 14

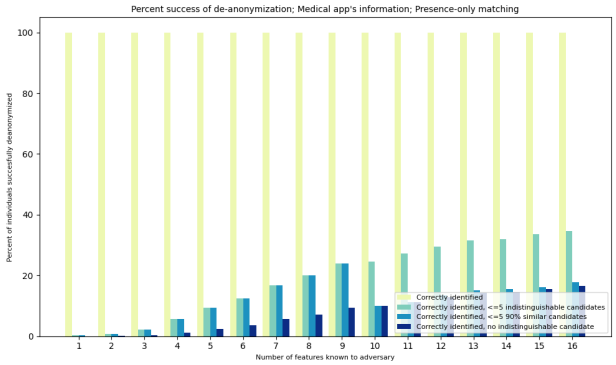

Figure 25. Performance at different thresholds for feature count 15