Attributes and Membership Disclosure Risks in Robust HIPAA – Compliant Medical Databases

Sanjiv M. Narayan, Rohan V. Krishnamurthi, Edward Tatchim

I. Introduction

Data can be either useful or perfectly anonymous but never both — Paul Ohm.[1] Nowhere is this more true than in health and medicine, where the trade-off between utility and privacy is acute. Biomedical innovation have lengthened average lifespan by decades and contributed to prosperity since World War II. Nevertheless, the sharing, secondary use or even mis-use of biomedical data have raised ethical issues beyond the Tuskegee catastrophe,[2] and growing privacy concerns. Medical cyber-attacks are increasingly frequent[3] and there is evidence that even trusted organizations have transferred personal health information (PHI) to industry.[4-7] Biomedical data pose unique challenges. While one's name, address or financial data can be edited or even deleted, one's DNA profile and biomedical identifiers may persist even to the next generation.

We hypothesized that the Health Information Portability and Accountability Act (HIPAA) – the U.S. privacy standard – does not prevent attribute and membership disclosure risk.[8,9] We tested this hypothesis in the validated MIMIC-IV dataset from MIT/Beth Israel Deaconess Hospital, comprising HIPAA-compliant de-identified data in over 180,000 patients in Massachussetts.[10,11]

Team members completed rigorous training on Human Subjects research for CITI certification, obtained formal approval from Physionet (https://mimic.mit.edu/) then accessed MIMIC_IV. Using advanced privacy engineering techniques gained during our Master's programs at UC Berkeley, we first identified patient profiles ("Phenotypes") and disclosure risks via k-anonymity and t-closeness. We then show how generalizations can improve these privacy metrics yet retain utility, and test multidimensional partitioning via Mondrian. *We only used de-identified data, and made no attempts at linkage with other datasets or re-identification.*

I.I Team Members

Sanjiv M. Narayan

- Doctor of Medicine (MD), University of Birmingham, UK
- Professor of Medicine, Computational Medicine T23 Stanford University

Edward Tatchim

- Bachelors of Science from the University of Maryland, College Park, MIT Applied Data Science Program, Data Engineer at GEICO Tech Solutions

Rohan V. Krishnamurthi

- Bachelors of science from Northeastern University (2022), MITx Micromasters in Statistics and Data science (2024), Data engineer at HarbourVest Partners, LLC (current)

II. Background

II.I Privacy Engineering

Privacy engineering (*DS233*) is a course focused on surveying privacy mechanisms

that are applicable to systems engineering. With a particular focus on the rise of inference threats due to the advancement of artificial intelligence (*AI*) and machine learning (*ML*). Our project uses mechanisms such as k-anonymity, t-closeness, Mondrian, and others to quantify the current level of privacy in MIMIC-IV. From here, we explore various means of improving its privacy.

II.II. Data

MIMIC-IV is a publicly released database sourced from electronic health records of the Beth Israel Deaconess Medical Center and managed by the dedicated Physionet group at MIT. As shown in Table 1, MIMIC-IV covers data in 180,733 patients from 2008 to 2019, and offers critical insights for healthcare advancements with patient measurements, orders, diagnoses, procedures, treatments, and deidentified free-text clinical notes. For this proof-of-concept study, we used a subset represented in this dataset.

	Hospital admissions	ICU admissions
Number of stays	431,231	73,181
Unique patients	180,733	50,920
Age, mean (SD)	58.8 (19.2)	64.7 (16.9)
Female Administrative Gender, n (%)	224,990 (52.2)	32,363 (44.2)
Insurance, n (%)		
Medicaid	41,330 (9.6)	5,528 (7.6)
Medicare	160,560 (37.2)	33,091 (45.2)
Other	229,341 (53.2)	34,562 (47.2)
Hospital length of stay, mean (SD)	4.5 (6.6)	11.0 (13.3)
In-hospital mortality, n (%)	8,974 (2.1)	8,519 (11.6)
One year mortality, n (%)	106,218 (24.6)	28,274 (38.6)

Table 1: Demographics for patients admitted to the intensive care unit (ICU) in MIMIC-IV[1]

De-identification met the HIPAA Safe Harbor provision.[8]

II.III. Preliminary EDA

MIMIC-IV has several patient quasi-identifiers including age and gender.[11] MIMIC-IV also contains ICD-10-CM for highly sensitive attributes at the patient-level including usage of alcohol (F10.20), other psychoactive drugs (F19.20), and anxiety (F41.1). We focused on these attributes, considered among the most sensitive ICD10 codes.[12] The dataset attributes most relevant to our project are –

- subject_id: patient id
- gender: patient's gender
- age: patient's age
- ecg_no_within_stay: enumerates ECGs within a given ED/hospital stay
- all_diag_all: unique ICD10 codes for hospital discharge diagnoses after concatenating emergency room and hospital diagnostic codes.
- ecg_taken_in_ed: noolean variable inidicating if the ECG was taken in ED
- ecg_taken_in_hosp: boolean variable indicating if ECG was taken in hospital
- ecg_taken_in_ed_or_hosp: boolean variable indicating if the ECG was taken either in the ED or in the hospital (i.e. no outpatient ECG)

III. Biomedical Privacy Landscape

Anonymizing medical data is a well-established practice aimed at balancing the need for privacy with utility for research and education. Although privacy protections have focused on anonymization via HIPAA[8,9] and the General Data Protection Regulation (GDPR)[13] in the European Union, this does not prevent linkage with other datasets for re-identification, nor does it prevent risks of attribute or membership disclosure.

As the healthcare landscape evolves, so have disclosure risks. For instance, a cyber-attack

at 23andMe in 2023 disclosed membership in individuals with Ashkenazi Jewish heritage.[14] Medical data are diverse, encompassing numerical metrics, multi-media disease representations, anatomical features and elements such as gender identity. This adds complexity to privacy protections.

An important concept is that fully granular data may not be required for medical care or research. For instance, the care of a 31 year old with a disease may be similar to that for a 38 year old, so that generalizing age to decades may not impact care. More broadly, such approaches could improve privacy while retaining clinical utility.

IV. Methods

IV.I Data cleansing and preparation

Personal health information (PHI) in MIMIC-IV was removed as required by HIPAA, and replaced with randomly generated ciphers, resulting in deidentified integer identifiers for patients, hospitalizations, and ICU stays. Look-up tables were used to randomly assign patients with a unique identifier (subject_id) and hospitalizations with a unique identifier (hadm_id). Dates were perturbed by shifting into the future (for instance, 2180) using a patient-level offset, that preserves the interval between time points for the same patient, but not between patients. The authors combined 2 published algorithms to remove PHI from the database and replace it with 3 consecutive underscores ("___").

IV.II. K-Anonymity

Several equivalence classes (EC) in MIMIC-IV had very low k-anonymity. Figure 1 illustrates the distribution of ages, an EC which provided only k=1 anonymity due to outliers <20 and >85 years. This was also true for seemingly 'bland' quasi-identifiers e.g. ecg_no_in_ed (k=1; numbers of ECG recorded in the emergency department). Clearly, adding any attribute to these ECs would not improve privacy (via the inverse subset theorem).

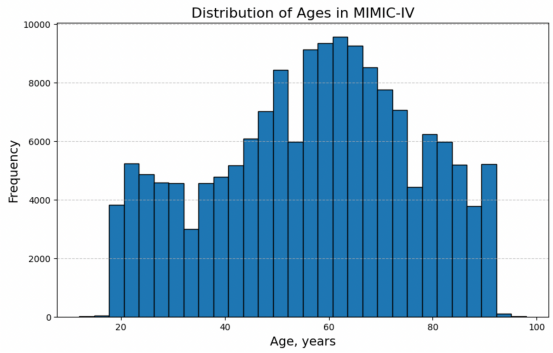

Figure 1: age provided k=1 anonymity (i.e. disclosure of an individual) due to outliers.

To improve privacy, we generalized several ECs, putting age into decade- bins (resulting in k=5328, i.e. each decade is shared for 5328 individuals), binned ECG and others. The generalized table had a worst-case k-anonymity for any single EC of k=377.

Nevertheless, combining attributes still resulted in unacceptable risks. For example, while "ECG taken in ED" provided k=75864, "ECG taken in hospital" provided k=77427, "gender" provided k=76011 and binned_ECG provided k=377, combining them provided k=1 meaning that a single individual could be identified by their combination.

IV.III. T-closeness

T-closeness helps to identify which quasi-identifiers are the least secure in relation to our selected sensitive attributes [alcohol use, drug use, anxiety]. The average t-closeness

score across all results is 0.02146, which we use as a benchmark to categorize scores as high (red/orange) or low (green). Higher scores indicate a greater risk of individual identification based on the quasi-identifiers. For instance, a t-closeness of 0.0934 for [binned_age, binned_ecg] with respect to the attribute alcohol suggests that if an attacker knows the age range and number of ECGs for a patient, they have a higher likelihood of identifying that individual.

QID/SA	Alcohol Use	Drug Use	Anxiety
binned_age, gender	0.0178	0.0047	0.048
binned_age, gender, ecg_taken_in_hosp	0.0448	0.0089	0.0137
binned_age, gender, ecg_taken_in_ed	0.0175	0.006	0.0049
binned_age, binned_ecg	0.0934	0.0126	0.0221
gender, binned_ecg	0.0224	0.0017	0.0034

Figure 2: t closeness table of QIDs vs SAs

From the table, it is evident that drug use is the most secure sensitive attribute. Therefore, our efforts to enhance data protection will focus on anxiety and alcohol. Using the binned_ecg as a quasi-identifier demonstrates lower privacy levels, suggesting that utilizing individual ECG attributes (ecg_taken_in_ed, ecg_taken_in_hosp) may provide greater privacy.

Figure 3: QID: [binned_age, gender] and SA:alcohol

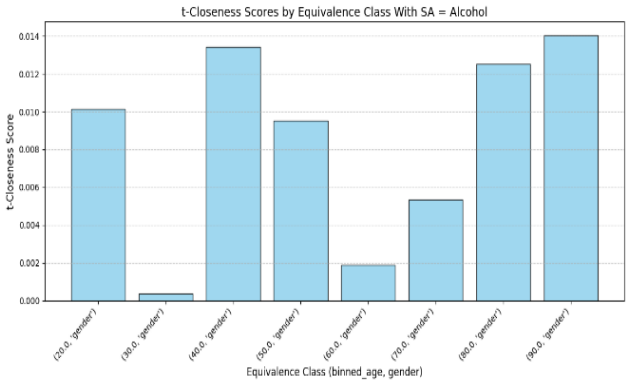

Figure 4: same as figure X with gender anonymized to 'gender'

The figures above illustrate a potential method for anonymizing our dataset. While transforming gender from a binary identifier (male/female) to a uniform value (gender) significantly reduces utility, it also halves the t-closeness scores, leading to a substantial improvement in privacy.

IV.IV. Mondrian

Mondrian allowed us to programmatically perform multidimensional cuts to find the best privacy-utility balance for the MIMIC-IV dataset. Although suboptimal when compared to k-optimize and incognito algorithms, Mondrian is an n-log n algorithm and much more efficient as it doesn't go over all possible cuts to find the best k value.

In practice, we used two quasi-identifiers "binned_age" and "binned_ecg." These two columns had 159,608 records in the initial dataset prior to partitioning. See data points displayed on the diagram below

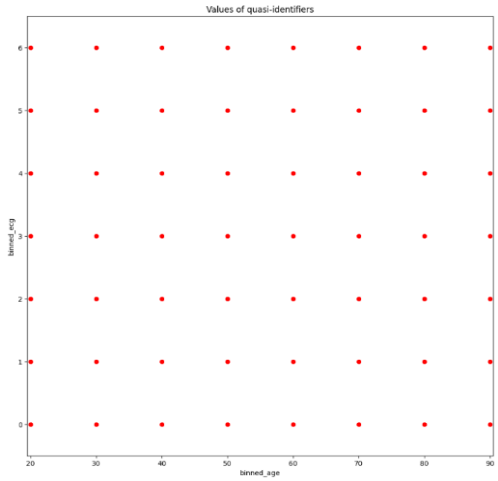

Figure 5: binned_age binned_ecg plot

After defining a function to randomly select partitions with allowable cuts based on another median-defined function and desired k of 80, Mondrian yielded 26 partitions with 7 unique outcomes, and a k of 471.

See the partitions below:

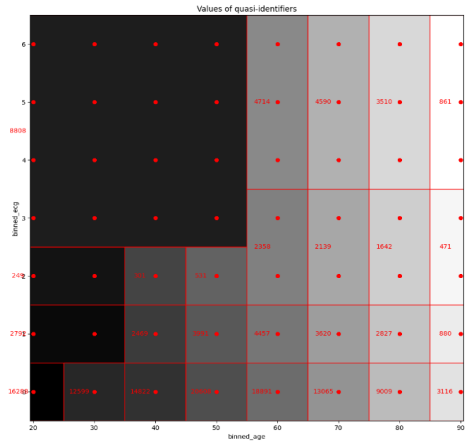

Figure 6: Mondrian output of partitions

From the above diagram, note that all values found within the cuts get generalized to a single value.

Next, we will see the discernability costs of the various k values in relation to these partitions. Figure 7 below depicts discernability costs against different k values:

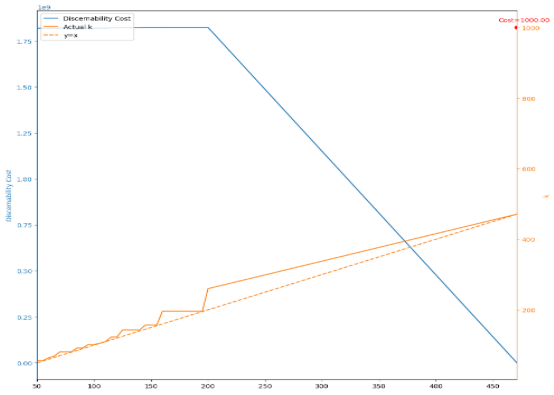

Figure 7: discernability costs vs k values

V. Discussion

Privacy protections in MIMIC-IV, a well-validated, de-identified HIPAA compliant database was suboptimal with substantial risks of attribute or membership disclosure. Using the validated metric of k-anonymity, we found that single quasi-identifiers including age and the numbers of ECGs recorded during the hospital visit could indicate a single individual. Generalizing the dataset improved privacy, yet combining variables in an equivalence class again enabled disclosure risks for a single individual. We conclude that even HIPAA-compliant datasets should ideally be pre-processed via extensive generalization, data suppression or tools such as differential privacy. In terms of utility, strategies such as generalizing age to decades may be acceptable, while other strategies must be assessed case by case.

We selected sensitive attributes of use of alcohol, use of psychoactive drugs and anxiety disorder, which in surveys are among the top 3 recognized sensitive attributes.[12] For these attributes, we found that gender, and the simple number of ECGs recorded during a hospitalization could disclose these sensitive

attributes like alcohol use, drug use, and anxiety.

The dataset achieved high k-anonymity for variables such as gender and ECG location (ED or hospital) but revealed vulnerabilities for variables like precise age and ECG count in ED, particularly for outliers. Generalizing these quasi-identifiers—e.g., banding ages or categorizing ECG counts—significantly improved privacy metrics.

T-closeness analysis highlighted alcohol and anxiety as vulnerable sensitive attributes compared to drug use, with binned quasi-identifiers providing better privacy outcomes. Strategies such as reducing quasi-identifier granularity and applying multidimensional partitions demonstrated potential for enhancing both k-anonymity and t-closeness. These findings underscore the importance of iterative anonymization techniques in achieving a balance between data utility and privacy in synthetic healthcare datasets.

Future directions could include more sophisticated partitions or generalizations. From running Mondrian on a subset of the initial dataset, the best possible k for our dataset of binned_age and binned_ecg would be 471. Doing further analysis on the discernability of cost, we found that the cost is higher for values of k greater than 471 implying decreasing utility of dataset when $k > 471$. On the other hand, when $k < 471$, we noticed that the discernability cost is lower meaning the usability of the dataset increases when k values are within that lower bound neighborhood of values. While Mondrian reveals the current ideal privacy-utility reality of the dataset (at k=471), we recognize that there is room for reinforcing the privacy of our dataset while increasing its usability with a lower k value. To enforce this, we plan to explore and implement more robust algorithms like differential privacy, data suppression, and data masking.

VI. Limitations

This study provides novel findings, but has limitations. First, we limited our study to 1 dataset, MIMIC-IV (albeit well studied). We focused on a selected number of quasi-identifiers, and 3 sensitive attributes although these are among the most sensitive worldwide. We used well-used privacy metrics, and our future work will study more sophisticated tools. By design, we did not test linkage attacks, e.g. using name data or residence data by Zip code. However, we are concerned that that this is a real risk that would likely be simple for a real adversary. Accordingly, we are extending this work after the conclusion of our 233 didactic period to try to design improved privacy solutions.

VII. Contributions

Edward Tatchim: Analyzing related datasets and designing a privacy engineering solution to address risk disclosure triad, Mondrian algorithm

Rohan Krishnamurthi: Content composition, solution proposal, and t-closeness

Sanjiv Narayan: Project vision, scoping, and purpose, k-anonymity

VIII. Appendix

1: Table 1: Demographics for patients admitted to an intensive care unit (ICU) in MIMIC-IV

2: Figure 2: Age distribution of raw data, explaining k=1 for this and other equivalence classes.

3. Figure 3: t-closeness table of QIDs vs SAs

3: Figure 4: QID: [binned_age, gender] and SA:alcohol

5: Figure 5: binned_age binned_ecg plot

6: Figure 6: Mondrian output of partitions

7: Figure 7: discernability costs vs k values

IX. References

1. Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA LAW REVIEW*. 2010;57:1701-1778.
2. National_register. Protection of human subjects; Belmont Report: notice of report for public comment. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Fed Regist*. 1979;44:23191-23197.
3. Kruse CS, Frederick B, Jacobson T, Monticone DK. Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technol Health Care*. 2017;25:1-10. doi: 10.3233/THC-161263
4. Hodson H. Revealed: Google AI has access to huge haul of NHS patient data. *New Scientist*. 2016.
5. Wetsman N. How Twitter is changing medical research. *Nat Med*. 2019. doi: 10.1038/s41591-019-0697-7
6. Das S. NHS data breach: trusts shared patient details with Facebook without consent. https://www.theguardian.com/society/2023/may/27/nhs-data-breach-trusts-shared-patient-details-with-facebook-meta-without-consent. In: *the Guardian*. Manchester; 2023.
7. Gennaro M. Judge advances swath of medical privacy class action against Meta. https://www.courthousenews.com/judge-advances-swath-of-medical-privacy-class-action-against-meta/. 2023.
8. Moore W, Frye S. Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. *J Nucl Med Technol*. 2019;47:269-272. doi: 10.2967/jnmt.119.227819
9. Moore W, Frye S. Review of HIPAA, Part 2: Limitations, Rights, Violations, and Role for the Imaging Technologist. *J Nucl Med Technol*. 2020;48:17-23. doi: 10.2967/jnmt.119.227827
10. Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng C, Stanley H. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*. 2000;101:e215-e220.
11. Strodthoff N, Lopez Alcaraz JM, Haverkamp W. MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG https://physionet.org/content/mimic-iv-ecg-ext-icd-labels/1.0.1/. 2024.
12. ICD_codes. ICD codes https://www.icd10data.com/ICD10CM/Codes. 2024.
13. OECD, Marini A, Kateifides A, Bates J, DataGuidance. Data Guidance and Future of Privacy Forum. Comparing Privacy Laws: GDPR v. CCPA. https://fpf.org/wp-

content/uploads/2018/11/GDPR_CCPA_Comparison-Guide.pdf. 2018.

14. Helmore E. Genetic testing firm 23andMe admits hackers accessed DNA data of 7m users

https://www.theguardian.com/technology/2023/dec/05/23andme-hack-data-breach. In: *The Guardian*. Manchester, UK; 2023.