Auditing Patient Privacy Risk in Synthetic Rare Disease Germline Data

Sikha Pentyala[1], Ziwei Pan[2], Luca Foschini[2], Martine De Cock[1,3], Jineta Banerjee[2]

[1]University of Washington Tacoma [2]Sage Bionetworks [3]Ghent University

Corresponding author: sikha@uw.edu

Introduction. Rare diseases collectively affect over 300 million people worldwide, yet their underlying causes remain poorly understood. Genomic advances in rare disease research are limited due to small, heterogeneous cohorts that are further constrained by the difficulty of sharing data across institutions because of privacy regulations [11, 2, 3]. A central barrier to sharing this genomic data is that it is inherently identifying [6, 5].These challenges are heightened for individual-level germline variant data stored in Variant Call Format (VCF) files, where variant combinations unique to patients act as quasi-identifiers.

Synthetic Data Generation (SDG) offers a promising approach by generating artificial data that preserves the statistical characteristics of real datasets without replicating personal information [8, 7]. However, synthetic data may not be inherently privacy-preserving, especially when generators are tuned for high fidelity. Therefore such SDG need to be audited for privacy risk alongside fidelity and utility prior to release. While recent work has introduced privacy auditing frameworks for synthetic single nucleotide variants (SNVs) level genomic data [10, 13], they operate on fixed positioned genotypes, and cannot accommodate mixed variants SNVs and insertions/deletions (indels) and generator-introduced positional perturbations.

To address this gap, we adapt and develop a model-agnostic privacy auditing framework for rare disease germline variant data, combining three complementary metrics. Our framework requires only the real data, the synthetic data to be audited, and public population allele frequencies.[1] Specifically, we design (a) a record-level proximity metric based on Distance to Closest Record (DCR) using Gower's distance [4], which quantifies how close each synthetic record is to its nearest real record, with unusually small distances indicating potential memorization or data leakage; (b) a membership inference attack (MIA) adapted from the genomic beacon likelihood-ratio framework [14, 12], which tests whether an adversary can determine if a specific individual was in the training cohort given the synthetic data and public allele frequencies; and (c) rare variant exposure scoring, which measures how much of each real patient's unique variant fingerprint is reproduced in the synthetic data, providing a continuous, per-patient measure of identifiability risk. Extending the notion of rare-variant collision risk in Rojo et al. [13] to the VCF setting, we introduce per-patient re-identification and exposure scores under both exact and position-tolerant (fuzzy) matching.

Methods. Let $P = \{p_1, \ldots, p_N\}$ denote the set of N real patients and $S = \{s_1, \ldots, s_M\}$ the set of M synthetic patients. For any patient p_i (or s_j), let $V(\cdot)$ denote their variant set, where each variant is a tuple $v = (\text{CHROM}, \text{POS}, \text{REF}, \text{ALT})$, representing the chromosome, genomic position, reference allele, and alternate allele. We compute 3 privacy metrics over the real and synthetic data as below.

- *Distance to Closest Record (DCR).* Instead of comparing raw variant sets, we summarize each patient's VCF into a profile record of $d = 33$ features: variant type ratios (SNV, indel, complex), Ti/Tv ratio, variant tier counts (common, recurrent, unique), mean quality, tumor type, and per-chromosome variant fractions. For each synthetic patient s_j, we compute $\text{DCR}(s_j) = \min_{p_i \in P} d_{\text{Gower}}(s_j, p_i)$, where Gower's distance averages $|a_k - b_k|/\text{range}_k$ over numerical features and exact matches (0/1) for categorical features, yielding a score in $[0, 1]$. Low DCR indicates the generator is copying rather than generalizing. We additionally report the nearest-neighbor distance ratio $\text{NNDR}(s_j) = d_j^{(1)}/d_j^{(2)}$, where $d_j^{(1)}$ and $d_j^{(2)}$ are the distances to the first and second closest real records; low NNDR suggests a synthetic record matches one real patient disproportionately.
- *Membership Inference Attack (MIA).* We assess the extent to which the synthetic data reveals whether a given real patient (target candidate) was used to train the generator. For each candidate patient p_i, we compute a likelihood-ratio test (LRT) score over their rare variants (allele frequency $f < 0.05$). For each such variant v, we define $P_0 = 1 - (1 - f)^{2N}, P_1 = P_0 + (1 - P_0) \cdot m$, where P_0 is the probability that v appears in the synthetic data if p_i was *not* in training, P_1 is the probability if p_i *was* in training, and $m \in (0, 1)$ is an assumed memorization rate. The LRT score accumulates $\log(P_1/P_0)$ for variants present

[1]https://gnomad.broadinstitute.org/

in the synthetic data and $\log((1-P_1)/(1-P_0))$ for those absent in the synthetic data.[2] We test every real patient as a candidate member and calibrate using two approaches: (1) an empirical null [12], comparing against pseudo-non-members generated by sampling variants from population allele frequencies, yielding AUC, where higher values of AUC indicates successful MIA; and (2) a theoretical null [14], yielding closed-form per-patient p-values under a Gaussian approximation where a lower p-value indicates memorization and thus membership.

- *Rare Variant Exposure Scores.* For each real patient p_i, we define a fingerprint $\mathcal{U}(p_i)$ as the set of variants carried exclusively by p_i in the cohort (i.e., appearing in exactly one patient). We compute a pairwise overlap score between each synthetic patient s_j and each real patient p_i as $\omega_m(s_j, p_i) = \frac{|V(s_j) \cap_m \mathcal{U}(p_i)|}{|\mathcal{U}(p_i)|}$, where $V(s_j)$ is the variant set of synthetic patient s_j, and $m \in \{\text{exact}, \text{fuzzy}\}$ denotes the matching mode. Exact matching requires identical variant tuples; fuzzy matching relaxes POS to a tolerance of $\pm\delta = 500$ base pairs, accounting for perturbations introduced by the generator. From this overlap matrix we derive two complementary metrics: (1) a *re-identification score* $R_m(s_j) = \max_{p_i} \omega_m(s_j, p_i)$, measuring how closely a synthetic record resembles any real patient's fingerprint, and; (2) an *exposure score* $E_m(p_i) = \max_{s_j} \omega_m(s_j, p_i)$, measuring how much of a real patient's fingerprint is recoverable from the synthetic data. Together, these capture leakage risk from both the synthetic and real patient perspectives.

Results. We evaluated the above on synthetic data generated from germline variant calls (SNV + indels) from a cohort of 61 (N) Neurofibromatosis type 1 (NF1) patients [1], generating 61 (M) synthetic patients using `SYNTHPOP` [9],a non-differentially-private generator, which integrates domain knowledge such as patient profiles and variant classifications. Table 1 summarizes the privacy audit. Profile-level metrics indicate strong separation: median DCR 0.138 (5th pct. 0.114; none < 0.05) and NNDR 0.968 rule out proximity or clustering, suggesting generalization. However, variant-level leakage persists. The worst case $R_{\text{fuzzy}} = 0.187$ ($\sim$19% overlap; 9.0% exact) and the $\sim 2\times$ gap between fuzzy and exact means (0.096 vs. 0.028) reveal leakage missed by position-exact audits. Moreover, E_{fuzzy} mean (0.160) exceeds R_{fuzzy} mean (0.096), indicating broadly distributed exposure. MIA achieves AUC = 1.0 with all 61 members significant at $p < 0.05$, even at $m = 0.1$.[3] With thousands of rare variants per patient (AF < 0.05), the cumulative LRT perfectly separates members from non-members, making membership fully identifiable from the synthetic VCF. Table 2(a) shows the most exposed patients ($E_{\text{fuzzy}} \approx 0.18$–$0.19$) despite widely varying fingerprint sizes (115–2,941), so exposure is not driven by size.Discrepancies between R and E highlight complementary views, e.g., $R_{\text{exact}} = 0$ but $E_{\text{exact}} = 0.070$ indicates non-dominant leakage. Table 2(b) shows that despite high variant-level exposure, DCR remains 0.132–0.160 (> 0.05), confirming distinct privacy dimensions. MIA detects all members ($p < 0.001$) but cannot distinguish which patients are most exposed or how much leaks, which R and E provide. No single metric suffices– DCR, MIA, and variant-level exposure reveal complementary risks, motivating multi-dimensional privacy auditing before releasing synthetic germline VCF data.

Acknowledgments.

This material is based upon work supported by the National Science Foundation under Grants Nos. 2451163 and 2523406, and by NSF NAIRR awards 240091 (TACC) and 240485 (TACC, AWS). This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581. The views expressed are those of the authors and do not necessarily reflect NIH policies. Additional support was provided by the University of Washington eScience Institute.

[2] We assume (a) each variant is independent and (b) public allele frequencies approximate the population distribution from which the cohort is drawn.

[3] We sweep $m \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and report worst-case results.

Table 1: Privacy audit summary. DCR: record proximity (lower = higher risk); MIA: membership inference (AUC = 0.5: no power, higher = higher risk); R/E: fingerprint leakage (0 = no leakage, higher = higher risk).

Metric	Statistic	Value
Record Proximity (DCR, Gower's)		
	Median DCR	0.138
	5th pct. DCR	0.114
	NNDR median	0.968
	Frac. < 0.05	0.000
Membership Inference (Beacon LRT)		
	AUC (empirical)	1.0
	TPR @ 5% FPR	1.0
	Frac. $p < 0.05$ (theor.)	1.0
	Best m	0.1
Re-identification R_m		
	R_{exact} max / mean	0.090 / 0.028
	R_{fuzzy} max / mean	0.187 / 0.096
	Frac. $R_{\text{exact}} > 0.01$	0.426
Exposure E_m		
	E_{exact} max / mean	0.090 / 0.072
	E_{fuzzy} max / mean	0.187 / 0.160

Table 2: Top 5 most exposed patients, ranked by E_{fuzzy}. $|\mathcal{U}|$: unique variants in fingerprint. (a) Exposure and re-identification scores under exact and fuzzy matching. The gap between exact and fuzzy quantifies how much leakage exact-matching methods would miss. (b) Cross-metric comparison for the same patients.

(a)

Patient	$\lvert\mathcal{U}\rvert$	E_{exact}	E_{fuzzy}	R_{exact}	R_{fuzzy}
p_1	2941	0.0704	0.1874	0.0000	0.1874
p_2	2930	0.0788	0.1870	0.0399	0.1870
p_3	2520	0.0754	0.1853	0.0000	0.1369
p_4	115	0.0870	0.1827	0.0870	0.1478
p_5	1934	0.0822	0.1810	0.0822	0.0000

(b)

Patient	E_{fuzzy}	MIA p	DCR
p_1	0.1874	0	0.132
p_2	0.1870	0	0.143
p_3	0.1853	0	0.160
p_4	0.1827	0	0.134
p_5	0.1810	0	0.149

References

[1] Jineta Banerjee, Yang Lyu, Stavriani Makri, Alexandra Scott, Lindy Zhang, Ana Calizo, Kai Pollard, Kuangying Yang, John Gross, Jiawan Wang, Adam Levin, Allan Belzberg, Carlos Romo, Robert Allaway, Jaishri Blakeley, Angela Hirbe, and Christine Pratilas. 2024DataRelease_WholeExomeSeq_Rawfastq. *Synapse*, 2024. doi:10.7303/SYN53132831.1.

[2] Kym M Boycott, Ana Rath, Jessica X Chong, Taila Hartley, Fowzan S Alkuraya, Gareth Baynam, Anthony J Brookes, Michael Brudno, Angel Carracedo, Johan T den Dunnen, et al. International cooperation to enable the diagnosis of all rare genetic diseases. *The American Journal of Human Genetics*, 100(5):695–705, 2017.

[3] Yaniv Erlich, Tal Shor, Itsik Pe'er, and Shai Carmi. Identity inference of genomic data using long-range familial searches. *Science*, 362(6415):690–694, 2018.

[4] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[5] Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[6] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, and Stanley F. Nelson. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008.

[7] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. SoK: Privacy-preserving data synthesis. In *IEEE Symposium on Security and Privacy (SP)*, pages 4696–4713, 2024.

[8] Jorge M Mendes, Aziz Barbar, and Marwa Refaie. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, 7:1563991, 2025.

[9] Ben Nowok, Gilbert M Raab, J Snoke, and C Dibben. Synthpop: generating synthetic versions of sensitive microdata for statistical disclosure control. *R package version*, pages 1–3, 2016.

[10] Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. On utility and privacy in synthetic genomic data. In *Proceedings of the 29th Network and Distributed System Security Symposium (NDSS 2022)*. Network and Distributed System Security (NDSS), 2022.

[11] Francesca Pistollato, Fabia Furtmann, Lindsay J Marshall, Surat Parvatam, Jan Turner, Flora Tshinanu Musuamba, Giulia Russo, and Francesco Pappalardo. Advancing the frontier of rare disease modeling: a critical appraisal of in silico technologies. *npj Digital Medicine*, 8(1):676, 2025.

[12] Jean Louis Raisaro, Florian Tramer, Zhanglong Ji, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, Heidi Sofia, Dixie Baker, Paul Flicek, et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *Journal of the American Medical Informatics Association*, 24(4):799–805, 2017.

[13] Alejandro Correa Rojo, Yves Moreau, and Gökhan Ertaylan. PRISM-G: an interpretable privacy scoring method for assessing risk in synthetic human genome data. *bioRxiv*, pages 2025–10, 2025.

[14] Suyash S Shringarpure and Carlos D Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2015.