Generative Masking and Asymmetric Hash Learning for Large-Scale Physiological Signal Indexing in Digital Healthcare Systems

Logan D. Hansen
Department of Computer Science, George Mason University, Fairfax, VA, USA.
ldhansen@gmu.edu

Ishaan Hegde
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
ishaan1997@missouri.edu

Nirk Heiley
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
bailey171@unh.edu

Chetan A. Pillai
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
chetan1998@binghamton.edu

Abstract

The exponential growth of ambulatory physiological monitoring has given rise to vast repositories of electrocardiogram, photoplethysmogram, and other biosignal waveforms, placing unprecedented pressure on retrieval and indexing infrastructure in digital healthcare systems. Traditional indexing approaches struggle to preserve clinically meaningful similarity relationships at scale while maintaining low-latency query performance. This paper addresses that gap by proposing a unified system architecture that couples generative masking with asymmetric hash learning for large-scale physiological signal indexing. Generative masking, informed by statistical priors over physiological dynamics, produces compact latent representations that suppress noise while amplifying diagnostically salient morphology. These representations are then mapped into binary hash codes through an asymmetric learning scheme where the query and database sides are allowed to follow distinct encoding pathways, and self-supervised semantic excavation aligns hash distances with hidden functional similarity. We describe the end-to-end system stack spanning edge preprocessing, cloud-based hash indexing, and federated governance layers, and we analyze how the interplay between masking and hashing resolves key structural trade-offs among retrieval precision, storage efficiency, and inference latency. Beyond technical performance, the paper examines fairness implications arising from population-specific masking priors, robustness to distributional drift in wearable sensor streams, sustainability considerations connected to model compression and edge deployment, and regulatory aspects of privacy-preserving similarity search under evolving health data protection frameworks. The discussion foregrounds infrastructural design principles that can guide the integration of generative and hashing components into future digital health platforms, ensuring that large-scale

physiological signal indexing remains clinically trustworthy, equitable, and operationally sustainable.

Keywords

generative masking, asymmetric deep hashing, physiological signal indexing, digital healthcare systems, retrieval infrastructure, fairness, sustainability.

1. Introduction

Contemporary digital healthcare infrastructures are being reshaped by the proliferation of wearable and nearable devices that continuously capture physiological signals such as electrocardiograms, photoplethysmograms, electroencephalograms, and impedance-based respiration traces. The volume, velocity, and variety of these signals have outpaced the evolution of backend retrieval mechanisms that allow clinicians, researchers, and automated decision support pipelines to locate relevant signal segments among millions or billions of stored recordings. The indexing problem is not merely one of database engineering; it is deeply entangled with the semantic fidelity required to support clinical search tasks, such as retrieving morphologically similar arrhythmia episodes, identifying temporal precursors of decompensation events, or finding matches for rare heart rate variability patterns across heterogeneous cohorts and device types. Traditional signal indexing strategies based on handcrafted feature templates, dynamic time warping, or locality-sensitive hashing of fixed-length embeddings often fail to preserve the nuanced, context-dependent similarity that physiological interpretation demands [1, 2]. Meanwhile, the rise of foundation-model-style training on raw biosignals has opened new representational possibilities, but adapting such representations into searchable indices that can be efficiently queried at population scale remains a largely unsolved systems challenge [3].

A promising path forward lies in the joint design of generative masking and asymmetric hash learning as a cohesive indexing pipeline. Generative masking refers to a class of self-supervised pre-training strategies in which portions of an input signal are stochastically obscured, and a model is trained to reconstruct the missing content by exploiting statistical regularities of the underlying physiological processes. When the masking schedule is informed by prior knowledge of signal morphology, such as the expected intervals between R-peaks in an electrocardiogram or the systolic rise time in a photoplethysmogram, the resulting latent codes acquire a structured invariance that helps separate informative variability from sensor noise and motion artifacts. Asymmetric hash learning complements this by projecting high-dimensional latent codes into compact binary strings in a manner that decouples the query-side encoding from the database-side encoding. This asymmetry permits the query pathway to be optimized under stringent computational constraints at the edge, while the database-side pathway can leverage heavier offline precomputation to maximize retrieval precision. By embedding a self-supervised semantic excavation mechanism within the hash learner, the system aligns the Hamming space with clinically meaningful similarity relationships even when ground-truth similarity labels are absent from the training corpus.

This paper presents a system-level study of how generative masking and asymmetric hashing can be integrated into a robust, scalable, and governable retrieval infrastructure for digital healthcare. Rather than proposing a single fixed algorithm, we develop a conceptual architecture that exposes the critical design dimensions and trade-offs that system architects must navigate. We analyze how the choice of masking granularity interacts with hash code length, how population-specific generative priors can encode health disparities into the index,

and how edge-cloud partitioning influences latency, energy consumption, and privacy. The discussion is grounded in real-world deployment scenarios drawn from hospital-wide telemetry archives, population-scale remote patient monitoring programs, and multinational clinical trial data harmonization efforts. Governance, fairness, robustness, and sustainability are woven into each layer of the analysis because an indexing system deployed in healthcare cannot be evaluated on retrieval accuracy alone; it must also satisfy ethical, regulatory, and operational constraints over long lifecycles.

2. Background and Challenges in Large-Scale Physiological Signal Indexing

Physiological signals present unique indexing challenges compared to image, text, or general time series data. Electrocardiogram and photoplethysmogram recordings exhibit quasi-periodic structures in which clinical relevance often resides in subtle amplitude variations, interval fluctuations, or transient waveform patterns that last only a few milliseconds. Inter-individual variability is high, yet clinically meaningful patterns such as ST-segment elevation or atrial fibrillation episodes must be recognized across diverse populations, acquisition devices, and lead configurations. Compression artifacts, electrode displacement, and ambient light noise introduce non-stationary distortions that can drastically alter the similarity landscape under distance metrics such as Euclidean or correlation-based measures. Consequently, a retrieval system must embed robustness to domain shift and noise directly into its indexing representations rather than relying on brittle preprocessing heuristics [4].

Deep hashing has emerged as a powerful approach for large-scale similarity search in computer vision and multimedia, substantially reducing storage footprints and accelerating query times by mapping high-dimensional features to compact binary codes where Hamming distance approximates semantic distance. Early attempts to adapt deep hashing to physiological signals have demonstrated feasibility but typically rely on supervised labels, symmetrical encoder architectures, or fixed-length segment representations that do not fully capture the hierarchical temporal structure inherent in biosignals [5, 6]. The symmetric design forces the query and database encoders to be identical, which limits the ability to optimize query-side latency independently of database-side accuracy. Moreover, when hash codes are learned directly from raw signal windows without a principled representation learning stage, the resulting codes often overfit to irrelevant noise and fail to generalize across acquisition conditions.

Self-supervised learning for time series has progressed rapidly, with contrastive and generative pre-training objectives producing representations that rival or exceed supervised baselines on downstream tasks such as arrhythmia classification, sleep staging, and blood pressure estimation [7, 8]. A recurring insight is that masking-based pre-training, where the model must reconstruct occluded temporal segments, is particularly effective for periodic signals because it compels the network to learn the underlying dynamical rules governing the signal’s evolution. Yet embedding such representations into an indexing pipeline introduces the question of how to preserve their rich semantic structure in a binarized code while simultaneously aligning the code with a query workload that may involve partial segments, multi-channel signals, or asynchronous sampling rates. This gap motivates a closer look at the interaction between generative masking and asymmetric hash learning.

3. Generative Masking with Statistical Priors for Signal Representation

The generative masking component of the proposed architecture operates as a self-supervised pre-training stage designed to produce a general-purpose signal encoder that captures both the

local morphological detail and the global rhythm structure of physiological waveforms. In a typical setup, a raw multi-channel signal segment is fed into a transformer or convolutional backbone, and a fraction of the input time steps or frequency components are masked according to a stochastic schedule. The model is trained to predict the original signal values at the masked positions by minimizing a reconstruction loss. However, uniform random masking, while simple, is suboptimal for signals with strong periodic and event-driven structure because it may leave the model with too little information during certain phases of the cardiac or respiratory cycle, or may mask diagnostically trivial baselines while leaving complex arrhythmic beats fully visible during training, thereby distorting the learned invariance profile.

To address these limitations, the masking process can be guided by statistical priors derived from population-level signal morphology. For instance, the prior may encode the expected timing of R-wave peaks, the typical duration of QT intervals, or the phase distribution of systolic and diastolic portions of photoplethysmogram pulses. Masking probabilities can then be elevated in regions phase-locked to these fiduciary points, encouraging the model to learn inter-beat dependencies and cross-channel redundancies. A recent framework that operationalizes this concept for photoplethysmogram signals has demonstrated that such informed masking yields representations that are more robust to motion artifacts and improve downstream classification accuracy while using shorter segments [12]. When extended to multi-modal settings that jointly process electrocardiogram and photoplethysmogram signals, the same principle can be applied by designing a joint prior over the temporal correspondence between electrical and mechanical cardiac events, thereby producing a shared latent space that is naturally indexed by a single hash code.

The structural trade-off in this stage lies between the expressivity gained by fine-grained, physiologically informed masking and the computational cost of estimating the priors across diverse subpopulations. Overly narrow priors that are tuned to a healthy adult cohort, for example, risk degrading representation quality for pediatric, geriatric, or pathological subpopulations whose signal morphology deviates from the assumed template. Conversely, a maximizing approach that blends multiple demographic- and condition-specific priors within a mixture-of-experts generative architecture can improve fairness but increases model size and training instability. These trade-offs are not only technical but also ethical, because representation quality feeds directly into retrieval equity. When the generative encoder systematically under-represents rare morphologies, downstream hash indices will de-prioritize those patterns, potentially causing missed clinical findings in already underserved populations.

4. Asymmetric Hash Learning with Semantic Excavation

Once a robust latent representation has been established via generative masking, the indexing subsystem must map each signal segment to a compact binary code that permits efficient approximate nearest neighbor search. This is accomplished through an asymmetric hash learning framework where the query encoder and the database encoder are architecturally decoupled. The database encoder can be a high-capacity transformer that ingests full-length, multi-channel recordings and generates highly discriminative binary codes offline, making use of batch processing and high-throughput hardware. The query encoder, in contrast, may be a lightweight convolutional or distilled model optimized for real-time execution on edge devices such as smartwatches, patches, or mobile phones, where power and memory constraints are severe. Asymmetric design allows the two pathways to be tuned for their respectively weighted objectives without forcing a compromise that would degrade both sides.

A central challenge in hashing for physiological signals is the absence of explicit similarity labels that capture the multidimensional notion of clinical relevance. Two electrocardiogram segments may be similar in terms of rhythm but dissimilar in waveform morphology, or may share the same underlying pathology despite divergent manifestations. To learn hash functions that respect such latent structure, the hash learner incorporates a self-supervised semantic excavation module that mines pairwise similarity relationships from the generative latent space itself. The core idea is to construct a semantic graph over a large pool of unlabeled signal segments by measuring localized neighborhood consistency in the generative embedding space across multiple scales and augmentations, and then use this graph to train the hash functions through a margin-based ranking loss that pulls similar pairs closer in Hamming space while pushing dissimilar pairs apart. The margin is dynamically scaled according to the estimated degree of semantic similarity, preventing hard threshold decisions that would collapse fine-grained distinctions [7].

The combination of asymmetric encoding and margin-scalable semantic constraints yields several infrastructural advantages. First, it allows the indexing layer to adapt to changing clinical taxonomies without fully retraining the generative backbone, because the hash learner can be re-anchored to a revised semantic graph based on updated diagnostic guidelines or newly recognized biomarker patterns. Second, the decoupled architecture naturally supports heterogeneous hardware deployments: hospital servers, regional cloud nodes, and personal devices can each execute the encoder appropriate to their capabilities while sharing a unified hash space. Third, because the binary codes are compact and can be searched via Hamming distance using highly optimized bitwise operations, the overall system achieves sublinear query complexity even when the index contains tens of millions of signals, a critical property for nationwide health data lakes and global clinical trial platforms.

5. System Integration and Scalable Indexing Infrastructure

The end-to-end indexing system integrates the generative masking and asymmetric hash learning components into a layered architecture that spans edge devices, fog nodes, and centralized cloud repositories. At the edge, wearable sensors stream raw photoplethysmogram and electrocardiogram data into a lightweight preprocessing unit that performs real-time segmentation, artifact detection, and signal quality assessment. Segments that pass a quality gate are encoded by the compact query-side hashing encoder, yielding a binary query code that can be transmitted over bandwidth-constrained channels, or stored locally for on-device similarity matching in personal health vaults. When a medical professional initiates a search for similar signal patterns, the query code is dispatched to a cloud-hosted index server that maintains the binary database codes alongside encrypted metadata pointers to the original signal storage locations. The server computes Hamming distances against the database codes, retrieves the top candidates, and returns ranked results to the requesting client, where final verification and visualization occur.

The scalability of this architecture rests on several design choices. The binary code length is a critical hyperparameter that directly trades retrieval precision against memory and communication overhead. Empirical analyses in analogous large-scale image retrieval systems indicate that codes of 64 to 256 bits typically offer a favorable balance for datasets in the tens of millions range, but for physiological signals where inter-class differences can be subtle, longer codes may be necessary and can be accommodated because the per-signal storage cost remains modest. The index structure itself can be implemented using multi-index hashing or product quantization over sub-codes, enabling efficient candidate pruning without

exhaustive Hamming distance computation. A further scalability lever is the temporal segmentation granularity employed during generative masking; coarser segments reduce the number of index entries but risk blending distinct clinical events, while over-fine segmentation inflates the index size with redundant entries that increase false positive rates.

System integration also demands careful consideration of governance and data provenance. Each indexed segment must carry audit metadata indicating the generative model version, the hash encoder version, and the demographic prior set used during training, so that future retrospective analyses can account for model drift and cohort-specific representation biases. In federated deployment scenarios, where multiple healthcare providers jointly contribute to a shared index without exposing raw patient data, the asymmetric hash learning framework can be extended so that each site locally computes database codes from its own population, and only the binary codes and associated encounter metadata are contributed to the federated index. This preserves patient privacy while enriching the global retrievability of rare conditions, provided that the generative priors are harmonized across sites to avoid fragmentation of the semantic space. The federated governance layer must also implement differential privacy mechanisms to prevent membership inference from crafted query sequences, a topic that intersects with evolving regulations such as the European Health Data Space and the updated Health Insurance Portability and Accountability Act interpretations.

6. Fairness, Robustness, and Policy Implications

Embedding an indexing system within digital healthcare workflows amplifies the societal consequences of model behavior, making fairness, robustness, and regulatory compliance first-class design requirements rather than afterthoughts. The generative masking stage is especially sensitive to disparities because its training data and masking priors are typically sourced from research cohorts that may under-represent certain demographic groups, geographic regions, or disease severity levels. When the prior distribution for photoplethysmogram morphology is estimated predominantly from lighter skin tones, for example, the masking strategy can systematically neglect the signal variability induced by higher melanin concentration and its interaction with optical sensor physics, leading to lower reconstruction fidelity and poorer latent representations for darker-skinned individuals. These representation gaps propagate through the hash learner and manifest as reduced retrieval recall for clinically equivalent signal patterns in affected populations [13, 14].

Addressing such fairness concerns requires a multi-pronged strategy at the infrastructure level. Generative priors should be stratified across relevant demographic and physiological axes with transparent reporting of coverage gaps. The semantic excavation graph, which steers hash code alignment, must be constructed using diversity-maximizing sampling and should be periodically audited using fairness metrics such as equal opportunity in retrieval recall across groups. Furthermore, the asymmetric architecture provides a practical mechanism for localized bias correction: a specialty clinic serving a predominantly underrepresented population can deploy a query encoder fine-tuned on that population’s signal characteristics while still querying a globally shared database index, provided the fine-tuning does not distort the binary code space in a way that breaks cross-population semantic alignment. Achieving this requires coordinated regularization that penalizes large shifts in the code distribution relative to the global anchor.

Robustness to distributional drift is equally critical. Wearable sensor characteristics evolve across hardware generations, signal acquisition protocols change, and population health profiles shift over time due to demographic transitions or emerging diseases. An indexing

system deployed in the field must continue to function reliably under such open-world conditions. The generative masking framework contributes to robustness by learning representations that are invariant to low-level sensor effects when the masking schedule is augmented with device-specific perturbations during training. The hash learner can be updated incrementally through continual learning protocols that introduce new semantic anchors without catastrophically forgetting previously learned similarity structures. Architecturally, this points toward a modular design where the generative backbone, the hash encoders, and the index structure are versioned independently, allowing fine-grained updates and rollbacks under the governance of a model registry that tracks provenance and performance metrics across time.

From a policy perspective, the deployment of large-scale physiological signal indices raises questions concerning consent, secondary use, and the right to explanation. Patients whose data contribute to an index must be informed that their de-identified signal patterns may be used to retrieve similar cases for clinical decision support, research, or commercial algorithm development. The opacity of learned binary codes poses a challenge for explainability, as clinicians may be reluctant to trust retrieval results that cannot be traced back to recognizable physiological features. Addressing this limitation calls for explanation modules that project Hamming neighbors back into the signal domain using a generative decoder, overlaying highlighted waveform regions that most influenced the hash distance. Such transparency mechanisms align with regulatory requirements for high-risk artificial intelligence systems under emerging frameworks and foster trust among clinical end-users.

7. Sustainability and Long-Term Operational Considerations

The environmental footprint of artificial intelligence systems in healthcare is gaining attention, and large-scale signal indexing is no exception. Training generative masking models on decades-long, multi-terabyte physiological archives consumes substantial energy, while maintaining always-on indexing services in cloud data centers adds to the operational carbon budget. Mitigating these impacts requires a holistic sustainability strategy that considers model architecture, hardware selection, and deployment topology. The generative masking frontend can be built on efficient transformer variants with structured sparsity and mixed-precision training, leading to significant reductions in floating-point operations per signal segment without compromising downstream hash quality. Knowledge distillation from a large masking teacher to a compact student encoder further compresses the query-side model for edge inference, slashing the per-query energy cost by orders of magnitude compared to deploying the full model on the cloud [15, 17].

The binary nature of the hash codes inherently supports energy-efficient search because Hamming distance computation relies on bitwise XOR followed by popcount operations, which are natively fast on commodity processors and require minimal power on custom accelerators. In contrast to dense vector similarity search, which necessitates high-dimensional floating-point dot products and often relies on power-hungry GPU clusters, the hash-based index can scale to billions of entries on conventional server CPUs, dramatically reducing the operational energy footprint. The asymmetric design amplifies this advantage by offloading the heavy generative inference to offline batch processing, while the online query path remains extremely lightweight. From a life-cycle perspective, the decoupling of the generative and hashing components allows the computationally intensive representation learning to be amortized over many months before a model refresh is required, while the hash encoder and index can be incrementally fine-tuned at lower cost.

Long-term sustainability also involves organizational and economic dimensions. A public health authority that deploys such an indexing system must budget not only for initial development but also for ongoing monitoring, bias audits, and interoperability maintenance. Open benchmarking suites that measure retrieval quality, fairness, energy consumption, and latency under realistic workloads can create accountability and facilitate cross-vendor comparisons. Institutions can pool resources through shared index-hosting cooperatives that amortize infrastructure costs while maintaining local control over data governance. These cooperative models, similar in spirit to federated learning networks for rare disease diagnosis, can ensure that the benefits of large-scale physiological signal indexing are not concentrated in well-resourced academic medical centers but extend to rural clinics and low-resource settings. Sustainability, in this broader sense, means designing the system so that it can adapt, endure, and remain equitable across the inevitable changes in technology, regulation, and population health.

8. Conclusion

This paper has examined the integration of generative masking and asymmetric hash learning as a foundational paradigm for large-scale physiological signal indexing in digital healthcare systems. By weaving together the representation learning strengths of statistical-prior-informed masking with the efficiency of asymmetric binary encoding and self-supervised semantic excavation, the proposed architecture addresses the dual requirement of preserving clinical nuance and enabling sublinear-scale retrieval. The analysis has moved beyond narrow performance metrics to consider the infrastructural, ethical, and policy dimensions that will determine whether such systems can be safely and equitably deployed. Key design trade-offs include the tension between population-specific generative priors and cross-group fairness, the balance between query-side compression and database-side precision, the resilience of the index under hardware and demographic drift, and the sustainability of both training and inference workloads. Future research should develop standardized evaluation protocols that simultaneously measure retrieval accuracy, fairness across population strata, energy per query, and compliance with emerging regulatory frameworks. As digital healthcare systems continue to integrate real-world signal data streams at unprecedented scale, indexing architectures grounded in generative and hashing principles will play an increasingly central role in making that data not only storable but searchable, interpretable, and clinically actionable.

References

1. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

2. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65–69.

3. Xu, J., Li, Z., Huang, Q., & Yang, Y. (2019). Deep semantic hashing for fast large-scale medical image retrieval. IEEE Access, 7, 109775–109785.

4. Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C. K., Li, X., & Guan, C. (2021). Time-series representation learning via temporal and contextual contrasting. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(3), 1–24.

5. Zhu, F., Ye, F., Fu, Y., Chen, L., & Li, J. (2020). Generating realistic electrocardiogram signals with a generative adversarial network. Applied Sciences, 10(12), 4348.

6. Sarkar, P., & Etemad, A. (2020). Self-supervised ECG representation learning for emotion recognition. IEEE Transactions on Affective Computing, 13(3), 1541–1554.

7. Yu, Z., Wu, S., Dou, Z., & Bakker, E. M. (2022). Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint. Neurocomputing, 483, 87-104.

8. Li, Y., Zhang, J., Wang, Y., & Liu, C. (2022). Transformer-based deep learning approaches for physiological signal analysis: A review. IEEE Reviews in Biomedical Engineering, 16, 232–250.

9. Mehari, T., & Strodthoff, N. (2022). Self-supervised representation learning from 12-lead ECG data. Computers in Biology and Medicine, 141, 105114.

10. Liu, Y., Zhang, J., & Huang, Q. (2022). Privacy-preserving deep hashing for medical image retrieval. IEEE Transactions on Information Forensics and Security, 17, 1890–1903.

11. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3, 119.

12. Guo, Z., Chen, T., Jiao, Y., Pan, Y., Hu, X., & Ferrario, M. (2026). SIGMA-PPG: Statistical-prior Informed Generative Masking Architecture for PPG Foundation Model. arXiv preprint arXiv:2601.21031.

13. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144.

14. Pfohl, S., Zhang, H., Xu, Y., Foryciarz, A., Ghassemi, M., & Shah, N. H. (2022). A holistic approach to algorithmic fairness in healthcare. Journal of the American Medical Informatics Association, 29(7), 1193–1201.

15. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

16. Lomonaco, V., Pellegrini, L., Cossu, A., Carta, A., Graffieti, G., Hayes, T. L., ... & Maltoni, D. (2021). Continual learning for medical applications: A survey. Artificial Intelligence in Medicine, 119, 102166.

17. Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1–43.

18. Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In Artificial Intelligence in Healthcare (pp. 295–336). Academic Press.

19. Bender, D., & Sartipi, K. (2013). HL7 FHIR: An agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 326–331.

20. Chen, R., Lu, M., Chen, T., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5(6), 493–497.

21. Yue, Y., Khanal, A., Lyu, T., Weissman, S., & Liang, C. (2025, May). EHR Phenotyping Methods for Measuring Treatment Adherence Among People Living With HIV in All of Us: Towards Disparities and Inequalities in HIV Care Continuum. In AMIA Annual Symposium Proceedings (Vol. 2024, p. 1294).

22. Shui, Y., Jin, R., Dou, Z., & Gao, Z. (2026). ProtoGuard-SL: Prototype Consistency Based Backdoor Defense for Vertical Split Learning. arXiv preprint arXiv:2604.03595.