Position: Real-World Clinical AI Requires Multimodal, Longitudinal, and Privacy-Preserving Corpora

Azmine Toushik Wasi[1,2*] **and Shahriyar Zaman Ridoy**[1,3]
[1]Computational Intelligence and Operations Laboratory (CIOL)
[2]Shahjalal University of Science and Technology [3]North South University
Correspondence to: azmine32@student.sust.edu

Abstract

Healthcare data is inherently noisy, incomplete, heterogeneous, and temporally irregular, posing major challenges for clinically robust AI. Existing models trained on curated datasets often fail to reflect operational conditions, limiting reliability and generalizability. We posit that real-world clinical AI requires multimodal, longitudinal, and privacy-preserving corpora to achieve trustworthy, scalable impact, yet current pipelines lack the methodological and ethical infrastructure to meet these needs. To address this, we present the *Messy Clinic* framework, a dataset blueprint that embraces the authentic complexity of clinical environments. It integrates multimodal sources, while embedding privacy-preserving techniques such as federated learning, differential privacy, and synthetic data generation. The framework also provides approaches for handling irregular inputs, noise-resilient modeling, and governance mechanisms addressing bias, consent, and accountability. Together, *Messy Clinic* establishes a foundation for reliable healthcare AI, supporting biomedical discovery, personalization, and reproducibility.

1 Introduction

Real-world healthcare data poses unique modeling challenges due to its complexity and variability. Incompleteness is particularly prevalent in EHRs, where missing values arise from clinical workflows, patient non-adherence, or inconsistent documentation [10, 15]. Noise and transcription errors are also widespread, especially in medical claims and narratives, often obscuring clinical signals [18, 20]. Further, inconsistency in terminologies and data formats across institutions hinders interoperability and data fusion [29, 22], amplifying heterogeneity across sources. Temporal irregularities from unscheduled visits and asynchronous monitoring complicate sequential and longitudinal analyses [32, 47]. Collectively, these challenges demand modeling frameworks that are robust to noise, tolerant of incompleteness, and flexible to irregular temporal dynamics [45, 49].

AI models trained on curated, homogeneous datasets often fail under real-world clinical variability, exposing methodological brittleness and weak generalization [16]. Large Language Models (LLMs) further exhibit instability across data types and workflows, limiting adaptability in practice [4]. We posit that fidelity to clinical *messiness*, noise, heterogeneity, missingness, and temporal irregularities, is essential for trustworthy AI [30]. Narrowly representative datasets heighten sensitivity to domain shifts, yielding biased predictions [44]. Recent advances reflect this shift: *MedFuzz* introduces adversarial perturbations to stress-test LLMs [31]; "observational supervision" and domain-specific augmentation enhance resilience [5]; and robust models like *MARIA* directly accommodate incomplete, inconsistent records [6]. Together, these approaches highlight real-world fidelity as a principle, not an impediment, for clinical AI. Crucially, this perspective reframes data quality in healthcare AI, treating noise, incompleteness, and heterogeneity as essential features to be modeled rather than artifacts to eliminate [46]. The traditional "clean data first" paradigm produces brittle models inca-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: EurIPS Wprkshop on Multimodal Representation Learning for Healthcare.

pable of handling clinical complexity [41]. By contrast, a "robustness to messy data first" approach emphasizes developing models that thrive amid the ambiguity and variability inherent in real-world healthcare. Recognizing that clinical reasoning operates on incomplete and sometimes conflicting evidence, this strategy encourages training that characterizes and simulates data messiness, fostering adaptable and trustworthy systems. Such alignment with clinical authenticity enhances applicability across diverse populations, settings, and unforeseen scenarios, ultimately transforming AI from a laboratory prototype into a robust, translational clinical tool.

To operationalize our position that trustworthy, scalable clinical AI requires multimodal, longitudinal, and privacy-preserving corpora, we present key contributions of the ***Messy Clinic*** paradigm. First, we propose a unified data architecture that integrates heterogeneous, longitudinal, and multimodal healthcare data, including structured EHRs, unstructured clinical notes, imaging, genomics, and wearable streams, into a coherent, analyzable framework [48]. This enables richer patient representations and directly supports our stance that AI must reflect the authentic complexity of clinical environments. Second, we adopt federated learning (FL) to enable collaborative model development across institutions while preserving patient privacy [13], reinforcing the necessity of privacy-preserving data infrastructures. Third, we advance temporal modeling by incorporating methods for irregular sampling and asynchronous clinical events, addressing the longitudinal nature of real-world patient records [45]. Collectively, these contributions establish a technically robust, ethically grounded, and privacy-aware foundation for AI, operationalizing our position by ensuring that models are resilient, interpretable, and aligned with the real-world complexities of healthcare.

2 Problem Background

2.1 Characteristics of Real-World Healthcare Data

Real-world clinical data is intrinsically *messy*, arising from diverse sources, heterogeneous formats, varied collection methods, and the dynamic nature of clinical environments. Unlike curated research datasets with standardized structures, real-world data reflects the uncontrolled, fragmented realities of clinical workflows, creating persistent challenges that hinder effective data-driven modeling and analysis in healthcare AI.

♦ **Incompleteness:** Missing data is pervasive in EHRs and clinical datasets [10, 34, 26], arising from clinician documentation habits, institutional differences, patient engagement, or logistical limitations. Non-random missingness can disproportionately affect certain demographics, introducing structural bias often unaddressed by standard imputation [15]. Longitudinal EHRs mix structured and unstructured entries recorded irregularly across care settings [45], complicating temporal modeling and undermining AI training and inference.

♦ **Noise and Errors:** Clinical data suffers from random and systematic errors, including transcription mistakes, misclassifications, duplicates, and contradictions [18]. Claims data may misrepresent diagnoses or treatments due to coding discrepancies and incentives [20]. Label noise propagates through models, distorting predictions, and complete noise elimination is impractical [1, 49]. Robust, fault-tolerant modeling and adversarial testing are therefore essential.

♦ **Inconsistency and Heterogeneity:** Clinical data varies across schemas, terminologies, coding standards, and protocols [29]. Even within hospitals, disparate charting and diagnostic systems hinder integration; across institutions, differences in EHR vendors, regulations, and standards like ICD-10, SNOMED CT, or HL7 exacerbate heterogeneity [22]. Temporal heterogeneity due to evolving practices and devices further risks model drift, challenging generalizability and real-time AI deployment.

♦ **Temporal Irregularities:** Clinical data is event-driven, with irregular and asynchronous intervals between visits, interventions, and outcomes [32]. For example, one patient may be seen monthly, another only during emergencies, complicating longitudinal modeling of disease progression, treatment effects, or responses [47]. Traditional time-series models assuming uniform sampling or stationarity are inadequate. Interactions among multiple temporal sequences, labs, medications, or wearable outputs, add spatiotemporal complexity, necessitating architectures with dynamic attention and multi-scale temporal reasoning [45]. Improper handling risks misleading inferences or overfitting to temporal artifacts.

2.2 Over-Reliance on Cleaned and Idealized Data for Clinical AI

Training clinical AI models on overly cleaned or idealized datasets can lead to significant limitations in real-world applications. Such datasets often exclude the inherent messiness of clinical data, including missing values, noise, and inconsistencies, which are prevalent in actual healthcare settings [11]. This exclusion can result in models that perform well under controlled conditions but fail

to generalize to the diverse and dynamic nature of real-world clinical environments. For instance, models trained on homogeneous data may exhibit biases, underperforming for underrepresented populations or rare conditions [12]. Additionally, the absence of real-world complexities in training data can lead to models that are brittle, lacking robustness against the variability encountered in everyday clinical practice [17]. Therefore, while data cleaning is essential, it is equally important to preserve the authentic characteristics of clinical data to develop AI systems that are both effective and equitable across diverse patient populations.

3 *Messy Clinic* Framework

Inherent complexity, noise, and variability of real-world clinical data challenge conventional AI approaches, which often rely on idealized datasets. To develop models that are robust, generalizable, and ethically sound, it is essential to embrace these data characteristics rather than obscure them. In this section, we outline the foundational data, methodological, and ethical pillars that underpin the design and implementation of the *Messy Clinic* paradigm in healthcare AI.

3.1 Multimodal Data Integration

Multimodal integration is central to the *Messy Clinic* dataset, providing a holistic view of patient health across diverse sources.

⊛ **Electronic Health Records (EHRs).** *EHRs* include structured fields (diagnoses, labs, medications), unstructured narratives (clinical notes, radiology reports), and time-series data like vitals [48]. Modern AI leverages *LLMs* and knowledge graphs to extract and align heterogeneous inputs. Frameworks like *EMERGE* use Retrieval-Augmented Generation (RAG) with curated KGs such as *PrimeKG* [48], while UCLA's *MEME* generates synthetic *pseudonotes* for text-based models [8]. Combined with NLP and deep learning, these approaches unify interpretable patient representations from messy, multimodal EHR data [19].

⊛ **Medical Imaging.** Medical imaging spans MRI, CT, X-rays, pathology, ophthalmology, and dermatology [50]. Integration is challenged by equipment variability, protocols, population differences, and annotation bias [14]. Accurate interpretation relies on contextual integration with other clinical data, critical for fair, reliable, and generalizable AI [24, 14].

⊛ **Genomic Data.** Structured genomic data provides insights into disease susceptibility, progression, and treatment response. For multimodal analysis, it must be mapped into shared embedding spaces with other modalities, enabling queries like identifying patients with specific mutations, tumor markers, and imaging features [50].

⊛ **Wearable and Sensor Data.** *Wearables* track continuous physiological signals, heart rate, glucose, blood pressure, activity, offering granular monitoring beyond clinical visits [23]. AI aggregates these streams with EHRs for real-time analysis, exemplified by platforms like Apple HealthKit [19].

⊛ **Patient-Generated Health Data (PGHD) and Patient Portals.** *PGHD*, including self-reported symptoms and lifestyle data, grows rapidly but often lies outside HIPAA regulations, raising privacy concerns [40]. Patient portal messages add clinician workload, and AI-assisted drafting can introduce errors, highlighting the need for human oversight [23].

⊛ **Cross-Modal Fusion and Representation Learning.** Effective integration uses attention mechanisms and fusion layers to weigh modalities dynamically, prioritizing high-quality inputs amid noise or missing data [50]. Intermediate fusion methods like *MARIA* combine modality-specific features into shared representations with masked self-attention and modality dropout for robustness [6]. Redundancy allows cross-checking of data streams, and NLP-generated pseudonotes improve semantic coherence across fragmented EHRs [48, 8, 19]. These techniques enrich patient representations and enhance context-aware AI predictions.

3.2 Longitudinal Data Management

Longitudinal data management is crucial for the ***Messy Clinic*** dataset, as healthcare data inherently evolves over time, reflecting disease progression, treatment responses, and changes in status.

⊛ **Capturing Complex Temporal Dependencies and Asynchronous Data Streams.** Healthcare data is inherently temporal, with clinical events occurring at irregular, asynchronous intervals, challenging traditional analytical models [45]. Advanced ML frameworks, including deep learning and graph-based models, leverage temporal and spatial dependencies to enhance diagnostic and prognostic capabilities, with applications from epidemic modeling to spatiotemporal disease prediction [7].

⊛ **Addressing Irregular Time Intervals and Temporal Gaps.** Patient data exhibits irregular visit frequencies and non-uniform timestamps. Standard tensor factorization approaches often fail to capture non-linear temporal patterns and handle missing entries [45]. Robust models must explicitly

address these gaps arising from episodic patient encounters and variations in clinical practice.
⊛ **Advanced Techniques for Temporal EHR Modeling.** REPAR models complex temporal EHR data using RNNs for temporal regularization and low-rank constraints to handle missing data. Its hybrid optimization supports binary, count, and numerical inputs, achieving superior reconstruction and predictive performance for tasks like in-hospital mortality and sepsis prediction [45]. REPAR enables extraction of dynamic phenotypes from noisy, irregular temporal records, improving clinical decision-making.

3.3 Data Acquisition and Curation

Messy Clinic dataset should be constructed by sourcing diverse real-world data, including hospital *HIS*, outpatient clinics, public health agencies, and commercial providers, with Federated Learning enabling privacy-preserving collaboration [13, 42]. Raw data undergoes ETL processes, including de-identification and mapping to standardized terminologies such as *SNOMED CT*, *ICD-10-CM*, and *LOINC*, while NLP extracts structured information from unstructured notes [20, 38]. Harmonization integrates multimodal sources like EHRs, imaging, genomics, and wearables, with ontology-based alignment and LLM-assisted entity extraction ensuring semantic consistency [22]. Quality assurance leverages noise reduction, cross-modal redundancy, robust architectures, and human-in-the-loop validation to address missingness, irregularity, and errors [49, 6, 23]. Privacy-preserving methods, including Federated Learning, Synthetic Data Generation, Differential Privacy, and cryptographic techniques, protect patient confidentiality while enabling collaborative AI development [3]. Detailed methodologies and representative examples are provided in Appendix B.

3.4 Ethical and Governance Framework

Development of the ***Messy Clinic*** dataset requires robust ethical and governance frameworks to ensure trust, compliance, and fairness. **Consent models** range from traditional static consent to dynamic digital consent, empowering participants with ongoing control, while broad consent allows flexible future use; PGHD and AI introduce additional consent complexities [21, 27, 40]. **Data governance** in federated learning addresses privacy, accountability, and regulatory compliance through procedural, relational, and structural mechanisms, though challenges remain around data reuse and model transparency [37]. **Bias mitigation** targets sources such as population homogeneity, institutional, annotation, and selection biases, using strategies like GAN-based augmentation, federated learning, standardized protocols, explainable AI, and adversarial testing [43, 31]. Collectively, these approaches uphold ethical AI deployment, with detailed discussion provided in Appendix C.

4 Concluding Remarks and Recommendations

In conclusion, we assert that *trustworthy and scalable clinical AI requires multimodal, longitudinal, and privacy-preserving corpora, supported by robust ethical governance, to reflect the complexity of real-world clinical environments.* Creating ***Messy Clinic*** dataset can operationalize this vision by embracing the inherent messiness of patient data, including missing entries, inconsistencies, and irregular temporal patterns, treating it as a feature rather than a limitation. By integrating multimodal sources such as EHRs, imaging, genomics, wearables, and patient-generated data, ***Messy Clinic*** improves representation fidelity, reduces predictive errors, and produces actionable insights. Advanced fusion architectures, attention mechanisms, and LLM-driven harmonization reconcile fragmented sources and mitigate noise, while continuous human-in-the-loop quality control ensures ethical compliance and patient safety. Privacy-preserving techniques, including federated learning, differential privacy, and synthetic data generation, protect sensitive information throughout. Collectively, the ***Messy Clinic*** blueprint provides a concrete pathway for developing AI systems that are resilient, interpretable, socially responsible, and capable of improving clinical outcomes and promoting equity in healthcare.

To realize the ***Messy Clinic*** dataset's potential, institutions should follow comprehensive technical, ethical, and governance recommendations (details in Appendix D). Privacy-preserving methods, federated learning, cryptography, and differential privacy, enable collaborative AI training while protecting patient data. Advanced longitudinal modeling captures complex temporal patterns, irregular visits, and incomplete records. Ethical frameworks with dynamic consent, stakeholder engagement, and bias mitigation uphold trust and regulatory compliance. Adversarial testing and multi-site validation assess robustness in real-world scenarios. Continuous monitoring, iterative quality control, and human-in-the-loop oversight maintain reliability and reduce errors.

References

[1] Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[2] Diana Andrade. Dynamic consent: A new gdpr standard for clinical trials, January 2025.

[3] Patricia A Apellániz, Juan Parras, and Santiago Zazo. Improving synthetic data generation through federated learning in scarce and heterogeneous data scenarios. *Big Data and Cognitive Computing*, 9(2):18, 2025.

[4] Yaara Artsi, Vera Sorin, Benjamin S Glicksberg, Panagiotis Korfiatis, Girish Nadkarni, and Eyal Klang. Large language models in real-world clinical workflows: A systematic review of applications and implementation. *medRxiv*, pages 2025–06, 2025.

[5] Nandita Bhaskhar. *Robust, Data-Efficient, and Trustworthy Medical AI*. Stanford University, 2023.

[6] Camillo Maria Caruso, Paolo Soda, and Valerio Guarrasi. Maria: a multimodal transformer model for incomplete healthcare data. *arXiv preprint arXiv:2412.14810*, 2024.

[7] Tanujit Chakraborty. Advancing machine learning systems for temporal healthcare data, 2025. Accessed: 2025-07-17.

[8] Mark Chiang. Ucla researchers develop ai system to convert fragmented ehr data into readable clinical narratives, July 2025. Accessed: 2025-07-17.

[9] Juhwan Choi, Jungmin Yun, Kyohoon Jin, and YoungBin Kim. Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. *arXiv preprint arXiv:2404.09682*, 2024.

[10] FAYE Cleary. *Challenges of studying and predicting chronic kidney disease progression and its complications using routinely collected electronic healthcare records*. PhD thesis, London School of Hygiene & Tropical Medicine, 2024.

[11] M. A. Dakka, T. V. Nguyen, J. M. M. Hall, S. M. Diakiw, M. VerMilyea, R. Linke, M. Perugini, and D. Perugini. Automated detection of poor-quality data: case studies in healthcare. *Scientific Reports*, 11(1), September 2021.

[12] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32), August 2022.

[13] Pallavi Dhade and Prajakta Shirke. Federated learning for healthcare: a comprehensive review. *Engineering Proceedings*, 59(1):230, 2024.

[14] Karen Drukker, Weijie Chen, Judy Gichoya, Nicholas Gruszauskas, Jayashree Kalpathy-Cramer, Sanmi Koyejo, Kyle Myers, Rui C Sá, Berkman Sahiner, Heather Whitney, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging*, 10(6):061104–061104, 2023.

[15] Oyewole Christopher Durojaiye, Charlotte Fiori, and Katharine Cartwright. Delivery of outpatient parenteral antimicrobial therapy (opat) in an ever-changing national health service (uk): Benefits, barriers, and opportunities. *Antibiotics*, 14(5):451, 2025.

[16] Rabie Adel El Arab, Mohammad S Abu-Mahfouz, Fuad H Abuadas, Husam Alzghoul, Mohammed Almari, Ahmad Ghannam, and Mohamed Mahmoud Seweid. Bridging the gap: From ai success in clinical trials to real-world healthcare implementation—a narrative review. In *Healthcare*, volume 13, page 701. MDPI, 2025.

[17] Rabie Adel El Arab, Mohammad S. Abu-Mahfouz, Fuad H. Abuadas, Husam Alzghoul, Mohammed Almari, Ahmad Ghannam, and Mohamed Mahmoud Seweid. Bridging the gap: From ai success in clinical trials to real-world healthcare implementation—a narrative review. *Healthcare*, 13(7):701, March 2025.

[18] Maryam Y Garza, Tremaine Williams, Songthip Ounpraseuth, Zhuopei Hu, Jeannette Lee, Jessica Snowden, Anita C Walden, Alan E Simon, Lori A Devlin, Leslie W Young, et al. Error rates of data processing methods in clinical research: a systematic review and meta-analysis of manuscripts identified through pubmed. *International Journal of Medical Informatics*, 195:105749, 2025.

[19] Michael Georgiou. Ai in medical diagnostics: Solving fragmented data challenges for accurate disease detection, May 2025. Accessed: 2025-07-17.

[20] Cynthia J Girman, Mary E Ritchey, and Vincent Lo Re III. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. *Pharmacoepidemiology and Drug Safety*, 31(7):717, 2022.

[21] Jane Kaye, Edgar A Whitley, David Lund, Michael Morrison, Harriet Teare, and Karen Melham. Dynamic consent: a patient interface for twenty-first century research networks. *European journal of human genetics*, 23(2):141–146, 2015.

[22] Natallia Kokash, Lei Wang, Thomas H Gillespie, Adam Belloum, Paola Grosso, Sara Quinney, Lang Li, and Bernard de Bono. Ontology-and llm-based data harmonization for federated learning in healthcare. *arXiv preprint arXiv:2505.20020*, 2025.

[23] Austin Littrell. Pros and cons of ai patient portal messages, May 2025. Accessed: 2025-07-17.

[24] Jiahui Liu, Xiaohao Cai, and Mahesan Niranjan. Medical image classification by incorporating clinical variables and learned features. *Royal Society Open Science*, 12(3):241222, 2025.

[25] Ida Lucente. Generative ai in healthcare: Use cases, benefits, and challenges, June 2025.

[26] Alexander Maletzky, Carl Böck, Thomas Tschoellitsch, Theresa Roland, Helga Ludwig, Stefan Thumfart, Michael Giretzlehner, Sepp Hochreiter, Jens Meier, et al. Lifting hospital electronic health record data treasures: challenges and opportunities. *JMIR medical informatics*, 10(10):e38557, 2022.

[27] John W. Maloy and Pat F. Bass III. Understanding broad consent. *Ochsner Journal*, 20(1):81–86, 2020.

[28] Milvus Team. What are the applications of multimodal search in healthcare?, 2025.

[29] Féline Mollerus, Cecil Lynch, and Hilgo Bruining. Data interoperability for a systems approach to developmental conditions. *Neuroscience & Biobehavioral Reviews*, page 106245, 2025.

[30] Anthony Owen and Eddie Samson. Validating ai/ml-based systems: Frameworks for testing model accuracy, fairness, robustness, and generalization in qa. 2025.

[31] B Potts. Medfuzz: Exploring the robustness of llms on medical challenge problems. *Microsoft Research https://www. microsoft. com/en-us/research/blog/medfuzz-exploring-the-robustness-of-llms-on-medical-challenge-problems*, 2024.

[32] Linglong Qian, Hugh Logan Ellis, Tao Wang, Jun Wang, Robin Mitra, Richard Dobson, and Zina Ibrahim. How deep is your guess? a fresh perspective on deep learning for medical time-series imputation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[33] Kacper Rafalski. Federated learning: A privacy-preserving approach to collaborative ai model training, March 2025. Updated March 17, 2025. 26 min read.

[34] Harry B Rhodes. *Factors influencing the quality of EHR performance: An exploratory qualitative study*. Capella University, 2016.

[35] Shahriyar Zaman Ridoy, Jannat Sultana, Zinnat Fowzia Ria, Mohammed Arif Uddin, Md Hasibur Rahman, and Rashedur M Rahman. An efficient text cleaning pipeline for clinical text for transformer encoder models. In *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, pages 1–9. IEEE, 2024.

[36] Aécio Santos, Eduardo H. M. Pena, Roque Lopez, and Juliana Freire. Interactive data harmonization with LLM agents. In *Novel Optimizations for Visionary AI Systems Workshop at SIGMOD 2025*, 2025.

[37] Syenza. Federated learning governance in healthcare: A framework for ethical and effective implementation, July 2025.

[38] Meng Wang, Xinyu Zhang, Xiaojun Zhang, Yue Lin, Kuan Yan, and Linhong Wang. Applications and challenges of artificial intelligence in emergency medicine: a literature review. *BMC Medical Informatics and Decision Making*, 24(1):172, 2024.

[39] David Wendler. Broad versus blanket consent for research with human biological samples. *The Hastings Center Report*, 43(5):3, 2013.

[40] Jenifer Sunrise Winter. Ai in healthcare: Data governance challenges. *Journal of hospital management and health policy*, 5(8), 2021.

[41] Qingyang Wu, Ying Xu, Tingsong Xiao, Yunze Xiao, Yitong Li, Tianyang Wang, Yichi Zhang, Shanghai Zhong, Yuwei Zhang, Wei Lu, and Yifan Yang. Surveying attitudinal alignment between large language models vs. humans towards 17 sustainable development goals, 2025.

[42] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of healthcare informatics research*, 5(1):1–19, 2021.

[43] Zikang Xu, Jun Li, Qingsong Yao, Han Li, Mingyue Zhao, and S Kevin Zhou. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1):286, 2024.

[44] He S. Yang, Daniel D. Rhoads, Jorge Sepulveda, Chengxi Zang, Amy Chadburn, and Fei Wang. Challenges and considerations of developing and implementing machine learning tools for clinical laboratory medicine practice. *Archives of Pathology & Laboratory Medicine*, 147(7):826–836, Jul 2023.

[45] Ren Yifei, Linghui Zeng, Jian Lou, Li Xiong, Joyce C Ho, Xiaoqian Jiang, and Sivasubramanium V Bhavani. Unraveling complex temporal patterns in ehrs via robust irregular tensor factorization. *AMIA Summits on Translational Science Proceedings*, 2025:451, 2025.

[46] Ning Zhang, Zhiwei Xu, Zehao Wang, Bin Zhang, Xinyu Liu, Xingyu Liu, Lizhen Cui, and Philip S. Yu. A comprehensive review of multimodal ai: theory, methodologies and applications. *Artificial Intelligence Review*, 2024.

[47] Yuanyun Zhang and Shi Li. Chronoformer: Time-aware transformer architectures for structured clinical event modeling. *arXiv preprint arXiv:2504.07373*, 2025.

[48] Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3549–3559, 2024.

[49] Zilliz. How do multimodal ai models handle noisy data?, 2024. Accessed: 2025-07-17.

[50] Zilliz. What are the applications of multimodal search in healthcare?, 2025. Accessed: 2025-07-17.

A Supplementary Material

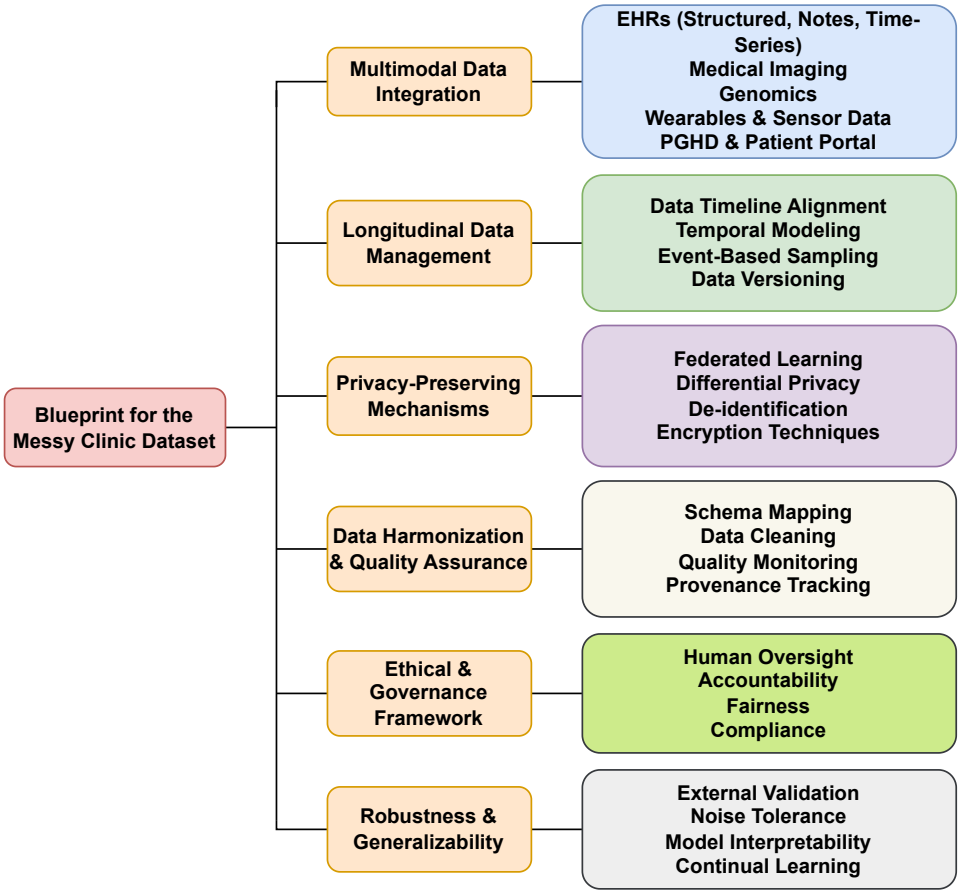

Figure 1: Blueprint for the ***Messy Clinic*** Dataset.

Table 1: Multimodal Data Types and Their Characteristics in Healthcare

Data Type	Typical Format	Key Characteristics / Challenges	Relevance to "*Messy Clinic*"	Sources
Electronic Health Records (EHRs)	Structured (diagnoses, labs, meds), Unstructured Text (clinical notes), Time-Series (vitals)	Incompleteness, inconsistent terminology, temporal irregularities, fragmentation across systems, coding discrepancies	Core clinical context; foundational for patient history & outcomes; source of significant "messiness"	[48]
Medical Imaging	Image files (MRI, CT, X-ray, pathology)	Acquisition variability (equipment, protocols), annotation bias (inter-observer), need for clinical context, high dimensionality	High-dimensional visual data for diagnosis & prognosis; sensitive to subtle variations and biases	[28]
Genomic Data	Structured (gene mutations, sequences)	Complex interpretation, ethical considerations, integration with phenotypic data, privacy of highly sensitive information	Foundational biological insights into disease susceptibility & treatment response; requires careful privacy handling	[28]
Wearable & Sensor Data	Continuous streams, time-series	Real-time, high volume, potential for noise/artifacts, data heterogeneity across devices, regulatory ambiguity for PGHD	Continuous physiological monitoring; offers dynamic, granular patient view outside clinic; captures real-world behavior	[19]
Patient Portal Messages / PGHD	Unstructured text, self-reported data	Unstructured nature, potential for errors/hallucinations in AI drafts, automation complacency risk, regulatory gaps for PGHD	Direct patient perspective; valuable for AI-assisted communication & workflow; highlights human oversight needs	[23]

B More Details on Data Acquisition and Curation

B.1 Data Sourcing and Initial Ingestion

⊛ **Identifying and Establishing Access to Diverse Real-World Data Sources** The foundational step in building the ***Messy Clinic*** dataset involves identifying and securing access to diverse real-world data sources, including structured and unstructured records from hospital Health Information Systems (HIS), outpatient clinics, public health agencies, research consortia, and commercial providers. For instance, methodologies exist to consolidate patient data across hospital HIS and outpatient settings [**?**]. A major challenge is that such data is often geographically dispersed and fragmented across multiple institutions. Federated Learning (FL) offers a crucial solution by enabling collaborative model training without centralizing sensitive raw data, thus providing access to diverse datasets otherwise restricted by privacy regulations and logistical barriers [13]. This is especially important in healthcare, where stringent data protection rules make direct data sharing difficult [42].

⊛ **Initial Extract, Transform, and Load (ETL) Processes** Once identified, raw data must undergo initial Extract, Transform, and Load (ETL) processes to move it from disparate source systems into a usable, centralized or distributed repository. This foundational step is crucial for managing the initial heterogeneity of raw data. The data life cycle, for instance, includes "curation of data to the clinical data repository; transformation and de-identification of data where necessary, creation of a data warehouse; and production of a study-specific dataset for analysis," emphasizing its iterative nature [20]. The transformation phase is critical and involves mapping local variables and proprietary codes to standardized nomenclatures and classifications, such as SNOMED CT, ICD-10-CM, and LOINC. Methodologies describe the use of EN/ISO 13606 archetypes as templates for identifying and transforming data, followed by loading the structured data into a database. Furthermore, Natural Language Processing (NLP) can be leveraged to extract structured data from unstructured sources, such as clinical notes, during this transformation [38].

B.2 Data Harmonization and Quality Assurance

⊛ **Mitigating Data Heterogeneity** Data heterogeneity, characterized by inconsistent clinical terminologies, varying data formats, and diverse collection standards across institutions, represents a major obstacle to effective data integration and interoperability. EHR formats, for example, can vary significantly across institutions due to differences in clinical terminologies, data collection standards, and underlying infrastructure. Data harmonization is therefore a critical practice aimed at *reconciling* various data types, levels, and sources into compatible and comparable formats to facilitate better decision-making and downstream analysis [22].

⊛ **Standardization and Schema Matching** Biomedical ontologies like SNOMED CT, ICD-10, Human Phenotype Ontology (HPO), and MONDO are essential for standardizing medical concepts, enabling precise communication, data integration, and interoperability across healthcare and research domains [22]. Large Language Models (LLMs), trained on extensive biomedical corpora, enhance this process by standardizing disparate EHR data, aligning ontologies, and addressing inconsistencies in medical coding across institutions [22]. For example, a two-step alignment strategy combines ontology-based vector embeddings with LLM validation to automate data mapping, achieving high precision (78–92%) and reducing manual effort, as demonstrated by the MPRINT Hub project [22]. Moreover, LLMs extract medical entities from time-series and clinical notes, aligning them with knowledge graphs to maintain consistency and prevent hallucinations [48]. Ontology-based repositories like OntoCR further support semantic interoperability by consolidating patient data into standardized models such as OMOP CDM, preserving clinical meaning throughout data reuse [**?**].

A crucial insight is that LLMs serve not just as generative tools but as vital enablers for harmonizing messy, heterogeneous healthcare data. Beyond their well-known capability to generate synthetic data, LLMs excel in interpretive, mapping, and semantic understanding tasks that are central to the ***Messy Clinic*** blueprint. Acting as intelligent translators and reconciliators, they bridge disparate terminologies, coding systems, and data formats, automating complex semantic reconciliation processes that have traditionally been manual, costly, and error-prone. This extends well beyond basic data cleaning, enabling scalable and accurate integration where rule-based approaches often fail. Consequently, strategic investments in healthcare LLMs should emphasize foundational data infrastructure functions like harmonization and interoperability. This calls for specialized medical LLMs, fine-tuning methods focused on ontology alignment and schema matching, and the potential rise of a new subfield, *Data Harmonization LLMs*, designed to intelligently structure and link diverse clinical information [36].

⊛ **Data Cleansing, Noise Reduction, and Handling Incompleteness** Data quality remains a cornerstone for reliable and high-performing AI models in healthcare, and effective strategies for data cleansing, noise reduction, and handling incompleteness are critical [35]. Large Language Models (LLMs) have shown considerable promise in this space, particularly for data annotation and cleansing tasks. For example, methods like “Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation” use chain-of-thought prompting and majority voting to identify and filter out noisy or irrelevant documents, improving dataset quality with reduced manual effort [9]. Preprocessing techniques, such as Gaussian blur for images, spell-checking for text, and spectral

filtering for audio, serve as front-line defenses against noise, streamlining downstream model processing [49]. Multimodal models further leverage cross-modal redundancy by dynamically weighting information from multiple sources through attention or fusion layers, enabling compensation for noisy or missing inputs; for instance, a low-quality X-ray image can be complemented by patient symptoms or lab results during diagnosis [49]. Robust training strategies that intentionally introduce noise during learning, such as pixel dropout, word swaps, and contrastive learning, encourage models to generalize well without requiring pristine datasets [49]. Addressing incompleteness, architectures like MARIA employ masked self-attention to selectively focus on available data while ignoring missing features, avoiding biased imputations. Regularization techniques such as modality and feature dropout simulate real-world missingness during training, bolstering model robustness [6]. For longitudinal healthcare data, tensor factorization approaches like REPAR integrate RNN regularization and low-rank constraints to handle missing entries effectively while modeling complex temporal patterns in EHRs, enhancing reconstruction accuracy and resilience to data gaps [45]. Together, these advances form a robust toolkit for managing the messiness of clinical data, enabling AI systems to deliver reliable and clinically meaningful insights despite imperfect inputs.

⊛ **Establishing Continuous Data Quality Control and Validation Loops** Data quality assessment in healthcare AI is not a one-time task but an ongoing, iterative process integral to the entire data lifecycle, requiring continuous monitoring and validation [20]. Human oversight plays a crucial role in this context, particularly given the documented variability in Large Language Models' (LLMs) performance and their limited generalizability in real-world clinical settings [4]. Empirical case studies, such as those analyzing AI-generated patient portal messages, reveal that even experienced physicians can overlook significant errors due to automation complacency, confirmation bias, or functional fixedness, highlighting the inherent risks of over-relying on AI without sufficient human review [23]. This underscores the necessity of embedding a *human-in-the-loop* framework as a core feature of AI-driven healthcare workflows to safeguard patient safety and maintain data integrity. Regulatory guidance, including from the FDA, echoes this perspective by framing data quality as a continuous, life-cycle process involving repeated quality checks and cleansing rather than a single preprocessing step. Additionally, the lack of robust post-deployment monitoring mechanisms for LLMs in clinical use remains a critical gap, with human oversight identified as a key mitigation strategy to catch and correct critical errors missed by AI [4, 23]. The practical implication is clear: responsibility for data accuracy and patient safety cannot be ceded solely to AI systems. Instead, continuous, iterative data quality control combined with structured human validation forms the backbone of trustworthy and ethical AI deployment in healthcare, ensuring ongoing error correction, performance validation, and adherence to safety standards throughout the data and model lifecycle.

B.3 Privacy-Preserving Mechanisms

⊛ **Federated Learning** Building a ***Messy Clinic*** dataset requires strong privacy-preserving methods to protect patient confidentiality and comply with regulations while enabling large-scale AI development. Federated Learning (FL) offers a compelling solution by allowing multiple healthcare entities, such as hospitals and clinics, to collaboratively train AI models without sharing raw data [13]. Instead of centralizing sensitive information, only model updates or parameters are exchanged, ensuring that personal data remains securely on local devices. This approach inherently enhances privacy and security while still enabling the creation of powerful, collaborative AI models [33]. The FL framework typically includes:

- **Client Nodes (Hospitals):** Each hospital acts as a client node, training its own deep learning model locally using its specific patient data. Node-specific training data are never transmitted, maintaining confidentiality [13].
- **Central Server:** A centralized server orchestrates the process, aggregating local model updates [13].
- **Global Model:** A shared global model improves through collaborative learning, with the server distributing the global model details to clients, who then refine it locally and send back accumulated changes [13].
- **Aggregation Algorithms:** Techniques like Federated Averaging (FedAvg) are used to combine local updates into a global model [33].

Federated Learning (FL) offers significant privacy benefits by enabling collaboration across numerous medical institutions without sharing sensitive raw data, thereby preserving patient confidentiality while granting access to diverse, real-world datasets essential for robust AI development [13**?**]. This iterative learning process from heterogeneous medical data improves model performance substantially [42]. However, FL faces unique challenges: healthcare data is often non-independent and identically distributed (non-IID) across clients, complicating model training. Approaches like Agnostic Federated Learning (AFL), q-Fair Federated Learning (q-FFL), and Multi-task Learning (MTL) help address these statistical issues by accommodating client heterogeneity [42]. Communication efficiency is another concern, as large numbers of clients with variable network speeds require strategies such as client selection, model compression (pruning, quantization, distillation), and update reduction techniques to minimize bandwidth use [42]. Moreover, while FL reduces direct data exposure, it remains

vulnerable to security threats like model inversion and poisoning attacks, necessitating additional safeguards beyond FL's inherent protections to ensure robust privacy and security [37, 33].

⊛ **Synthetic Data Generation (SDG)** Synthetic Data Generation (SDG) offers a compelling approach to mitigate the scarcity and quality limitations of medical data, especially in resource-constrained settings, by creating artificial patient data that closely mirrors real-world distributions while preserving privacy [3]. Techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), including models like VAE-BGM, excel at capturing complex data structures to produce high-quality synthetic tabular data [3]. Integrating SDG within a Federated Learning (FL) framework further enhances privacy and generalizability by allowing local training of generative models across institutions and aggregating learned parameters, thereby generating synthetic datasets that reflect diverse, decentralized sources [3]. Importantly, SDG inherently protects privacy since synthetic samples are generated from latent spaces and do not replicate real patient records exactly, as confirmed by empirical evaluations showing non-zero minimum distances between synthetic and real data points [3]. Consequently, SDG represents a promising, privacy-preserving strategy to augment healthcare datasets, enabling accelerated research and model development without exposing sensitive information [25].

⊛ **Differential Privacy (DP)** Differential Privacy (DP) offers a rigorous framework to protect individual data contributions by adding carefully calibrated noise to datasets or model updates, ensuring that the inclusion or exclusion of any single individual minimally impacts analysis outcomes [33]. This mechanism provides plausible deniability, preserving overall data utility while masking personal information. Local DP variants further enhance privacy by injecting noise directly on user devices before data transmission. DP is extensively applied within Federated Learning (FL) to prevent indirect leakage of sensitive details from shared model updates [42]. Although the noise introduced by DP can sometimes reduce model accuracy, it remains a vital privacy safeguard, often combined with complementary methods to balance privacy protection and predictive performance effectively [42].

⊛ **Multi-Party Computation and Encryption** To enhance privacy in Federated Learning (FL), advanced cryptographic methods like Secure Multi-Party Computation (SMPC) and Homomorphic Encryption (HE) are employed. SMPC enables multiple parties to collaboratively compute results on their private data without exposing individual inputs, securing the aggregation phase against data inference [42]. HE allows computations directly on encrypted data, ensuring data remains confidential during transmission and processing [42]. Combining these techniques with Differential Privacy often yields a robust balance of privacy, security, and model accuracy in federated systems, addressing the complex challenges of healthcare data collaboration [33].

C More Details on Ethical and Governance Framework

The development and deployment of a ***Messy Clinic*** dataset, particularly one that is multimodal, longitudinal, and privacy-preserving, necessitates a robust ethical and governance framework. This framework is crucial for building public trust, ensuring regulatory compliance, and mitigating potential harms.

C.1 Consent Models

Traditional consent models, often static and paper-based, struggle to keep pace with the dynamic nature of AI-driven research and the continuous use of diverse datasets.

Traditional vs. Dynamic Consent. Traditionally, participants grant consent for specific research purposes, with details provided in patient information sheets and face-to-face interactions. However, this static approach often leaves participants unaware of how their data is used over time [21]. Dynamic consent emerges as a transformative approach, leveraging digital platforms to enable ongoing engagement and communication between individuals and data custodians. It allows participants to provide, withdraw, or modify their consent in real-time through online portals or mobile applications [21]. This method increases transparency, enhances trust by empowering participants with greater control, and improves compliance with evolving regulatory requirements [2]. Dynamic consent facilitates two-way communication, allowing participants to upload additional health data or researchers to inform participants about new opportunities or findings [21].

Broad Consent. Broad consent is an alternative approach that allows for an individual's consent for the storage, maintenance, and secondary research use of identifiable private information or biospecimens for a non-specific range of future research purposes, subject to certain restrictions and oversight [27] It offers flexibility, particularly when researchers do not anticipate having the time or resources to re-consent for every new study, and when the risk to participants from future research is not substantial [39]. However, it raises questions about whether generic consent with non-specific information suffices given the complexities of identifying future research uses [39].

Challenges with PGHD and AI. The rapid growth of Patient-Generated Health Data (PGHD) from wearables and mobile health applications complicates consent, as these data may fall outside existing health data regulations (e.g., HIPAA in the United States) and are often governed by technology companies' privacy policies [40]. Furthermore, the lure of AI innovation has, in some instances, led companies to bypass patient consent, highlighting a growing tension between health research involving big datasets and the principle of informed

consent [40]. Building public trust and ensuring fair, equitable, and transparent developments in health AI will require better understanding and balancing of individual and community rights with corporate interests in AI health data [40].

C.2 Data Governance and Accountability

Robust data governance is fundamental to managing decentralized healthcare data ethically, securely, and effectively within federated machine learning environments [37]. It plays a critical role in maximizing privacy protections while fostering trust among participating institutions, thereby facilitating broader access to multi-institutional datasets for more comprehensive research [37]. Governance frameworks also help organizations navigate the complex and evolving landscape of legal, ethical, and reimbursement standards essential to healthcare AI development [37]. However, federated learning introduces distinct governance challenges, including risks of patient re-identification, data leakage, and ethical concerns around secondary data use [37]. Moreover, the extensive data requirements of deep learning models, coupled with the opacity of algorithmic data handling, challenge existing regulations like GDPR, which restrict data reuse even as AI benefits from large-scale, multi-purpose data aggregation [40].

To address these challenges, a comprehensive governance approach employs a spectrum of mechanisms grouped into procedural, relational, and structural categories [37]. Procedural controls include privacy safeguards, formalized data-sharing agreements, continuous monitoring, and transparent evaluation protocols. Relational mechanisms emphasize stakeholder engagement, ongoing consent processes, public involvement, and capacity building, all crucial for maintaining the social license for secondary data use [37]. Structural elements involve establishing oversight bodies, designating data stewards, and integrating consumer representation to ensure accountability [37]. Transparent communication and active public participation are vital for sustaining trust and meeting ethical obligations in handling sensitive health data. Despite stringent regulatory frameworks such as GDPR and HIPAA, fragmented ethical approvals often result in accountability gaps across data custodians, underscoring the need for consolidated, evidence-based governance frameworks to uphold privacy and compliance in complex federated systems [37].

C.3 Bias Mitigation and Fairness

Integration of AI into healthcare carries the risk of generating biased or unrepresentative models, which can lead to misleading clinical conclusions or overestimation of model performance [44].

Sources of Bias in Healthcare AI. Bias in medical image datasets can lead to disparities in diagnostic performance across different patient demographics, potentially reinforcing existing healthcare inequalities [43]. Sources of bias include:

- **Homogeneous Datasets:** AI algorithms can exhibit algorithmic bias if trained on datasets that are not diverse or representative of the real-world population [16].
- **Institutional and Geographic Bias:** Medical images collected from a limited number of hospitals or regions may not generalize to other healthcare settings due to differences in imaging equipment, protocols, and population characteristics [43].
- **Annotation Bias:** Manual annotation of medical images can introduce inconsistencies due to inter-observer variability among radiologists [43].
- **Selection Bias:** The criteria used to select patients for imaging studies can introduce bias; for example, if a dataset primarily comprises patients with advanced-stage diseases, AI models may struggle to detect early-stage conditions [43].

Addressing bias and variability in medical image datasets is essential for ensuring fair and reliable AI applications in healthcare [43]. Strategies include:

- **Data Augmentation:** Techniques such as generative adversarial networks (GANs) can be used to generate synthetic data to diversify datasets [43].
- **Federated Learning:** This approach helps mitigate bias by training models on decentralized datasets, leveraging the diversity of data across institutions without centralizing sensitive information [43].
- **Standardization:** Implementing standardized data collection, annotation, and processing protocols across institutions can reduce variability and improve data quality [43].
- **Explainable AI (XAI):** Developing XAI models helps identify and address bias by making model decisions more transparent [43].
- **Regulatory Compliance:** Regulatory bodies are increasingly emphasizing the need for bias mitigation strategies in AI training and validation [43].

- **Adversarial Testing:** Methods like MedFuzz actively challenge AI models with scenarios designed to expose vulnerabilities related to biases and simplifying assumptions, providing deeper insights into real-world performance [31].

D Recommendations for Leveraging the *Messy Clinic* Dataset

To fully realize the potential of the ***Messy Clinic*** dataset in advancing healthcare AI, we propose the following recommendations:

- **Prioritize Multimodal Data Integration with Intelligent Fusion:** Develop AI architectures that integrate heterogeneous data types, including EHRs, imaging, genomics, wearables, and PGHD, using advanced fusion layers and attention mechanisms. Exploiting cross-modal redundancy can reduce noise, handle missing data, and generate robust, unified patient representations beyond conventional aggregation techniques.
  “‘
- **Develop AI-Native Data Harmonization Pipelines:** Leverage Large Language Models (LLMs) and biomedical ontologies during data ingestion and curation to automate semantic mapping, schema alignment, and reconciliation of inconsistent terminologies across diverse clinical sources. This reduces manual effort while ensuring standardized, interoperable datasets suitable for downstream modeling.
- **Implement Robust Longitudinal Data Modeling:** Apply and advance temporal modeling techniques, such as RNN-regularized tensor factorization or hybrid deep learning frameworks, to capture non-linear temporal dependencies, irregular sampling intervals, and missing entries in longitudinal patient records. Such approaches enhance predictive accuracy and enable nuanced modeling of disease progression and treatment response.
- **Establish Privacy-by-Design Frameworks:** Utilize federated learning as the foundation for collaborative AI training on sensitive healthcare data. Complement FL with cryptographic safeguards, including Secure Multi-Party Computation (SMPC), Homomorphic Encryption (HE), and Differential Privacy (DP), balancing robust privacy protection with high model performance.
- **Implement Continuous Data Quality Assurance with Human Oversight:** Integrate iterative data quality assessments, noise reduction, and cleansing across the data lifecycle. Human-in-the-loop mechanisms should monitor AI outputs, detect anomalies or errors, and ensure ethical compliance, ultimately safeguarding patient safety and maintaining trust in AI-driven decisions.
- **Develop Ethical and Governance Frameworks:** Establish comprehensive governance structures that address patient consent (favoring dynamic models when feasible), data ownership, accountability, and bias mitigation. Promote transparent stakeholder engagement and public participation to strengthen trust and align AI development with societal and regulatory expectations.
- **Incorporate Adversarial Testing and Real-World Validation:** Employ adversarial machine learning and stress-testing methods to evaluate model performance under realistic clinical noise, biases, and edge-case scenarios. Conduct multi-site validation studies to confirm generalizability and robustness prior to deployment in real-world healthcare settings.

By systematically implementing this blueprint, healthcare institutions can construct a ***Messy Clinic*** dataset that forms a robust foundation for next-generation AI systems. These models will achieve enhanced predictive performance while being resilient, generalizable, and trustworthy, ultimately translating into measurable improvements in patient care, operational efficiency, and public health outcomes.