Position: Real-World Clinical NLP Robustness Requires Messy, Multimodal, Longitudinal, Privacy-Preserving Corpora

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

Abstract

Real-world healthcare data is inherently complex: noisy, incomplete, heterogeneous, and temporally irregular, making clinically robust AI development difficult. Traditional models, trained on idealized datasets, fail to capture these operational realities, limiting real-world performance. We propose the ***Messy Clinic*** paradigm: a dataset and blueprint for building AI systems grounded in authentic clinical complexity. This framework enables seamless multimodal data integration across EHRs, imaging, genomics, and wearables, using privacy-preserving methods such as federated learning, synthetic data, and differential privacy. It provides (1) structured methodologies for longitudinal multimodal management, (2) AI techniques resilient to noise and missing data, and (3) governance mechanisms for consent, accountability, and bias mitigation. By embracing the "Messy Clinic," healthcare AI can shift from artificial idealism to real-world robustness, accelerating research, enabling personalized care, and aligning AI with the true complexity of clinical practice.

1 Introduction

Real-world healthcare data is highly complex and variable, creating major modeling challenges. EHRs frequently contain missing data caused by clinical workflows, inconsistent documentation, and patient non-adherence (Cleary, 2024; Durojaiye et al., 2025). Clinical narratives and claims data also include substantial noise and transcription errors that obscure true clinical signals (Garza et al., 2023; Girman et al., 2022). Inconsistencies in terminologies and formats across institutions further hinder interoperability and multisource integration (Mollerus et al., 2025; Kokash et al., 2025). Additionally, irregular visits and asynchronous monitoring complicate longitudinal modeling and sequential analysis (Qian et al., 2025; Zhang and Li, 2025). These realities require robust, noise-tolerant, and temporally flexible modeling frameworks (Yifei et al., 2025; Zilliz, 2024).

AI models trained on idealized datasets often fail under the variability of real-world clinical practice, exhibiting brittleness and poor generalization (El Arab et al., 2025). This limitation is especially evident in Large Language Models (LLMs), whose performance varies across workflows and modalities (Artsi et al., 2025). Addressing this challenge requires embracing the inherent *messiness* of healthcare data—including noise, heterogeneity, missingness, and temporal irregularity—as a design principle rather than a defect (Owen and Samson, 2025). Models trained on sanitized datasets are highly vulnerable to domain shifts and biased predictions (Yang et al., 2023). Robustness-oriented approaches such as **MedFuzz** expose vulnerabilities through clinically realistic adversarial perturbations (Potts, 2024), while observational supervision, domain-specific augmentation (Bhaskhar, 2023), and architectures such as **MARIA** for incomplete records (Caruso et al., 2024) further support clinically grounded learning.

This perspective redefines data quality in healthcare AI. Rather than removing noise and incompleteness, the ***Messy Clinic*** paradigm treats them as clinically meaningful signals that models must learn to interpret (Zhang et al., 2024). The traditional "clean data first" philosophy has produced systems poorly suited to real clinical complexity (Wu et al., 2025), whereas a "robustness to messy data first" strategy prioritizes adaptability under uncertainty. Because clinical reasoning itself relies on incomplete and conflicting evidence, trustworthy AI systems must reflect this reality. Importantly, this shift does not reject curation, but advocates realistic modeling and simulation of clinical messiness during training, reframing the ***Messy Clinic*** paradigm as a philosophy of system design rather than a collection of post hoc noise-handling methods.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1163–1180
November 6-10, 2026 ©2026 Association for Computational Linguistics

To operationalize this paradigm, we propose three directions. First, we advocate unified architectures integrating heterogeneous and multimodal healthcare sources, including EHRs, clinical notes, imaging, genomics, and wearable streams (Zhu et al., 2024). Second, we emphasize federated learning (FL) for collaborative yet privacy-preserving model development (Dhade and Shirke, 2024). Third, we incorporate advanced temporal modeling methods capable of handling irregular sampling and asynchronous clinical events (Yifei et al., 2025). Together, these contributions provide a technically grounded foundation for scalable, privacy-aware, and clinically robust healthcare AI while serving as a conceptual framework rather than a fixed implementation blueprint.

2 Characteristics of Real-World Healthcare Data

Real-world healthcare data is inherently *messy*, reflecting heterogeneous clinical workflows, fragmented collection practices, and dynamic care environments rather than the controlled conditions of curated research datasets. This complexity introduces persistent challenges for healthcare AI, including incompleteness, noise, heterogeneity, and temporal irregularity.

Incompleteness: Missing data is pervasive in EHRs due to documentation variability, institutional differences, inconsistent patient engagement, and unavailable tests (Cleary, 2024; Rhodes, 2016; Maletzky et al., 2022). Missingness is often non-random, disproportionately affecting specific populations and introducing structural bias that standard imputation cannot resolve (Durojaiye et al., 2025). Longitudinal EHRs further combine structured and unstructured records collected at irregular intervals across multiple care settings, increasing sparsity and fragmentation (Yifei et al., 2025). Such gaps significantly impair model training and temporal inference.

Noise and Errors: Clinical datasets frequently contain transcription mistakes, duplicates, coding inconsistencies, misclassifications, and contradictory entries generated through both manual and automated processes (Garza et al., 2023). Claims data may distort diagnoses or treatments because of reimbursement incentives and coding discrepancies (Girman et al., 2022), propagating label noise into downstream models. Since perfectly clean datasets are rarely attainable (Zilliz, 2024), especially in under-resourced settings, robust learning methods capable of tolerating noisy and adversarial inputs are essential (Fries et al., 2022; Cordeiro and Carneiro, 2020).

Inconsistency and Heterogeneity: Healthcare data varies substantially across schemas, terminologies, coding systems, and institutional workflows (Mollerus et al., 2025). Differences in EHR vendors, standards (e.g., ICD-10, SNOMED CT, HL7), and regional practices create interoperability barriers and limit model generalization. Although ontology-based harmonization partially alleviates these issues, scalable automated alignment remains unresolved (Kokash et al., 2025). Continuous changes in clinical practices and software also introduce temporal heterogeneity and model drift.

Temporal Irregularities: Clinical records are event-driven and asynchronously sampled, with highly variable intervals between visits, interventions, and outcomes (Qian et al., 2025). Such irregularity complicates modeling disease progression and treatment response (Zhang and Li, 2025). Standard time-series assumptions of uniform sampling often fail, while interactions among laboratory data, medications, and wearable streams introduce additional spatiotemporal complexity requiring dynamic attention and multi-scale reasoning architectures (Yifei et al., 2025).

Global Heterogeneity and Low-Resource Settings. Many biomedical AI systems assume access to mature, standardized EHR infrastructure, despite large portions of the world lacking reliable longitudinal records or consistent diagnostic coding. Although initiatives such as *MedLog* (Noori et al., 2025) aim to improve global interoperability, AI paradigms dependent on clean and standardized EHRs remain globally limited. The ***Messy Clinic*** framework instead treats incompleteness, non-standard ontologies, and absent modalities as intrinsic modeling challenges rather than preprocessing failures, enabling AI systems better suited for diverse and low-resource healthcare setups.

3 Components of *Messy Clinic*

Here, we outline the foundational data, methodological, and ethical pillars that holds the design and implementation of the ***Messy Clinic*** paradigm.

3.1 Multimodal Data Integration

Multimodal data integration is a cornerstone of the ***Messy Clinic*** dataset, recognizing that a holistic

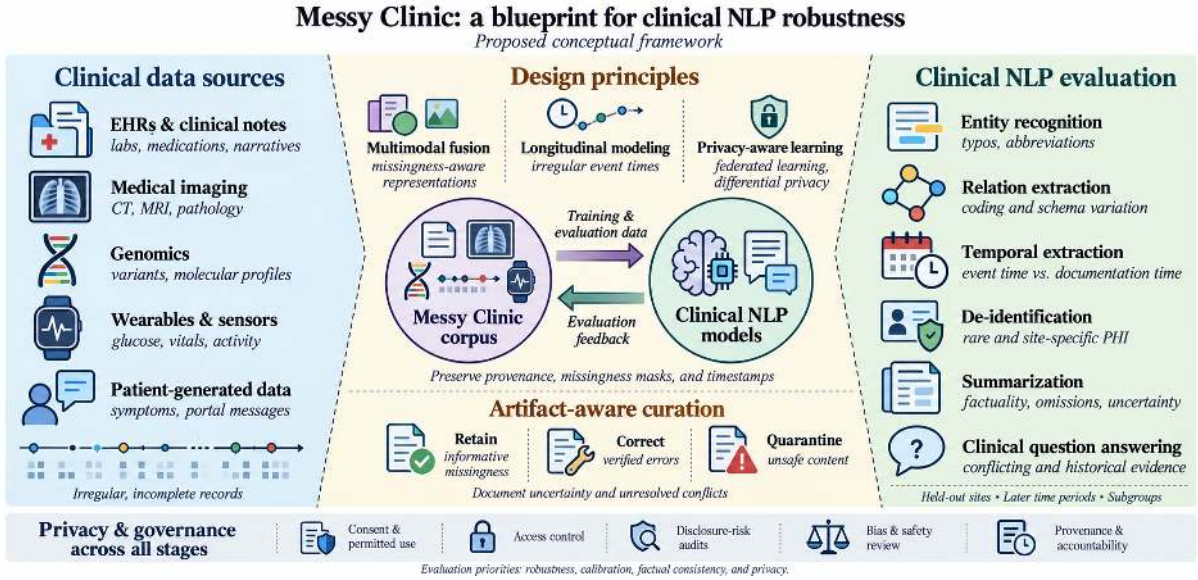


Figure 1: Conceptual blueprint of the *Messy Clinic* paradigm for clinical NLP robustness. Heterogeneous clinical sources inform multimodal, longitudinal, and privacy-aware modeling. Artifact-aware curation retains informative missingness, corrects verified errors, and quarantines unsafe content while preserving provenance and uncertainty. Six clinical NLP tasks motivate evaluation across sites, time periods, and subgroups. Privacy and governance controls span all stages. The diagram represents a proposed framework, rather than a validated implementation.

understanding of patient health requires combining information from diverse sources. Table 5 summarizes the key multimodal data types and their characteristics within the context of healthcare AI.

3.1.1 Electronic Health Records (EHRs)

Electronic Health Records (EHRs) form the foundation of real-world clinical data, encompassing structured fields (diagnoses, labs, medications), unstructured narratives (notes, reports), and multivariate time-series data (vitals, physiological trends) (Zhu et al., 2024). These modalities provide a comprehensive yet fragmented view of patient health. Modern AI systems leverage Large Language Models (LLMs) and Knowledge Graphs (KGs) to extract and structure information from text and time-series inputs. Frameworks like **EMERGE** employ Retrieval-Augmented Generation (RAG) aligned with medical KGs such as **PrimeKG** to enhance semantic fidelity (Zhu et al., 2024), while UCLA's **MEME** synthesizes fragmented records into "pseudonotes" resembling clinical documentation (Chiang, 2025). Such approaches, coupled with NLP and deep learning, help unify siloed, multimodal EHR data into coherent, interpretable patient representations (Georgiou, 2025).

3.1.2 Medical Imaging

Medical imaging spans modalities such as MRI, CT, X-ray, pathology, ophthalmology, and dermatology images (Zilliz, 2025). Integration challenges stem from variations in acquisition equipment, protocols, and patient demographics (Drukker et al., 2023). Annotation bias from inter-observer variability further introduces inconsistencies (Drukker et al., 2023). Since interpretation depends on clinical context, overlooking contextual data can reduce model reliability (Liu et al., 2025). Addressing dataset bias and variability is vital for fair, generalizable AI systems and to avoid perpetuating healthcare inequities (Drukker et al., 2023).

3.1.3 Genomic Data

Genomic data, though structured, requires integration with other modalities for comprehensive analysis. Mapping genomic information into shared embedding spaces enables multimodal querying, for instance, identifying patients with specific mutations, abnormal tumor markers, and imaging-confirmed lesions, highlighting its central role in precision medicine and multimodal research acceleration (Zilliz, 2025).

3.1.4 Wearable and Sensor Data

Wearables provide continuous monitoring of metrics such as heart rate, glucose, blood pressure, and activity, offering granular, real-time insights beyond clinic visits (Littrell, 2025). Future devices will analyze vast daily data streams to flag early disease signals like cardiac or metabolic risks (Littrell,

2025). AI systems aggregate and link wearable outputs with EHRs, exemplified by Apple's HealthKit, which integrates continuous data into clinical platforms for ongoing health monitoring (Georgiou, 2025).

3.1.5 Patient-Generated Health Data (PGHD) and Portal Communications

PGHD, self-reported symptoms, lifestyle data, and app-based inputs, constitutes a growing information source but often falls outside traditional health data regulations (e.g., HIPAA), posing governance and privacy concerns (Winter, 2021). Patient portal messages provide valuable real-time insights yet increase clinician workload. Generative AI tools assist in drafting responses, improving efficiency but risking errors such as medication typos or outdated advice due to automation complacency or confirmation bias (Littrell, 2025). Thus, stringent oversight and safety validation remain essential (Littrell, 2025).

3.1.6 Strategies for Cross-Modal Fusion and Representation Learning

Effective multimodal integration demands advanced cross-modal fusion and representation learning for coherent, actionable insights (Zilliz, 2025). Attention-based architectures dynamically weight modalities, prioritizing reliable inputs amid noise or missing data. Intermediate fusion models like **MARIA** use masked self-attention and modality dropout for robustness and unbiased learning (Caruso et al., 2024). Multimodality also functions as internal validation, redundant modalities enable error detection via cross-referencing. AI-driven narrative generation further enhances semantic coherence by transforming fragmented EHR data into unified pseudonotes (Zhu et al., 2024; Chiang, 2025), surpassing traditional ETL pipelines (Georgiou, 2025). Embedding these integration and harmonization techniques within the ***Messy Clinic***framework is key to unlocking multimodal data's full potential, enabling richer patient representations and more accurate, 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.

3.2.1 Capturing Complex Temporal Dependencies and Asynchronous Data Streams

The inherent temporal nature of healthcare data, with clinical events occurring at irregular and asynchronous intervals, presents a significant challenge for traditional analytical models. These models often struggle to capture the complex temporal and spatiotemporal dependencies embedded in data from sources like wearable sensors and EHRs (Yifei et al., 2025). Advanced machine learning frameworks, particularly deep learning and graph-based models, are essential for effectively modeling and analyzing healthcare time-series data. These frameworks can leverage both temporal and spatial dependencies to enhance diagnostic and prognostic capabilities. Research efforts are actively focused on developing such advanced ML systems to address irregularities in medical data, with applications ranging from epidemic spread modeling to spatiotemporal disease prediction (Chakraborty, 2025).

3.2.2 Addressing Irregular Time Intervals and Temporal Gaps in Patient Records

The *irregular* nature of patient data, characterized by non-uniform visit frequencies and event timestamps, poses major challenges for standard time-series models. Existing tensor factorization methods often fail to capture non-linear temporal dynamics and remain sensitive to missing EHR entries (Yifei et al., 2025). Robust approaches must therefore explicitly model missing observations arising from heterogeneous clinical practices and episodic patient encounters.

3.2.3 Advanced Techniques for Modeling Temporal EHR Data

REPAR (RNN Regularized Robust PARAFAC2) addresses these challenges by combining recurrent neural networks for temporal regularization with low-rank constraints for robustness to missing data (Yifei et al., 2025). Its hybrid optimization framework supports binary, count, and numerical variables while capturing complex non-linear temporal dependencies. Experimental results show improved reconstruction quality and predictive performance for tasks such as in-hospital mortality and sepsis prediction, even under high missing-data ratios (Yifei et al., 2025). Such methods are critical for extracting meaningful temporal phenotypes from noisy longitudinal EHRs.

Distinction from Noise Augmentation Approaches. Artificially corrupting clean datasets does not replicate real-world clinical messiness. Random or Gaussian noise injection (Cordeiro and Carneiro, 2020) fails to capture structured artifacts such as conditional testing patterns, resource-driven missingness, ontology inconsistencies across institutions, and diagnostic delays that introduce systematic temporal offsets. These artifacts encode genuine clinical signal, meaning naive corruption may discard information robust models should instead exploit. When augmentation is necessary, clinically grounded simulation using generative approaches such as GFlowNets (Bengio et al., 2023) better models the conditional processes governing uncertainty and missingness, while methods like **MedFuzz** (Potts, 2024) provide adversarial evaluation under clinically plausible perturbations. Nevertheless, the ***Messy Clinic*** paradigm prioritizes authentic clinical distributions because real-world structural dependencies cannot be fully recreated through synthetic perturbation alone.

4 Data Acquisition and Curation

4.1 Data Sourcing and Initial Ingestion

4.1.1 Identifying and Establishing Access

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 (Dhade and Shirke, 2024). This is especially important in healthcare, where stringent data protection rules make direct data sharing difficult (Xu et al., 2021).

4.1.2 Initial 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 (Girman et al., 2022). 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 (Wang et al., 2024a).

4.2 Data Harmonization and Quality Assurance

Within the ***Messy Clinic*** framework, harmonization and embracing messiness are complementary rather than contradictory. Harmonization addresses interoperability by aligning heterogeneous schemas, standardizing ontologies (e.g., SNOMED CT, LOINC), and correcting systematic coding inconsistencies so data from multiple institutions can be jointly analyzed. Embracing messiness operates at the modeling stage: instead of excluding patients with irregular timestamps, missing modalities, or noisy records, we use architectures specifically designed to learn from such conditions, including MARIA's masked self-attention for missing modalities (Caruso et al., 2024) and REPAR's RNN-regularized tensor factorization for irregular longitudinal EHRs (Yifei et al., 2025). Thus, harmonization improves data accessibility, while robustness methods preserve and model real clinical complexity rather than eliminating it.

4.2.1 Mitigating Data Heterogeneity

Data heterogeneity, arising from inconsistent terminologies, formats, and collection standards across institutions, poses a major barrier to interoperability and integration. EHR structures often differ by infrastructure and documentation practices, fragmenting datasets. Data harmonization thus aims to *reconcile* diverse data sources into consistent, comparable formats to support reliable analysis and

decision-making (Kokash et al., 2025).

4.2.2 Standardization

Biomedical ontologies such as SNOMED CT, ICD-10, HPO, and MONDO enable standardized medical concepts and cross-institutional interoperability (Kokash et al., 2025). Large Language Models (LLMs), trained on biomedical corpora, further enhance harmonization by aligning ontologies, standardizing EHR data, and correcting inconsistent coding (Kokash et al., 2025). A two-step alignment strategy, ontology-based vector embeddings plus LLM validation, achieved 78–92% mapping precision and reduced manual workload in the MPRINT Hub project (Kokash et al., 2025). LLMs also extract entities from notes and time-series data, aligning them with knowledge graphs for semantic consistency (Zhu et al., 2024), while repositories like OntoCR consolidate patient data into OMOP CDM for semantic interoperability.

LLMs act not only as generators but as key enablers of harmonization within the ***Messy Clinic*** framework. They bridge disparate coding systems, terminologies, and schemas, automating semantic reconciliation tasks once done manually and error-prone. Beyond basic cleaning, LLMs enable scalable, accurate integration where rule-based tools fail. This highlights the need for medical-domain LLMs fine-tuned for ontology alignment and schema matching, supporting a new subfield of *Data Harmonization LLMs* (Santos et al., 2025).

4.2.3 Data Pre-Processing

High data quality is essential for dependable healthcare AI. LLMs now play key roles in data annotation and cleansing; e.g., "Multi-News+" employs chain-of-thought prompting and majority voting to filter noisy data efficiently (Choi et al., 2024). Preprocessing, Gaussian blur for imaging, spell-checking for text, spectral filtering for audio, serves as a first defense (Zilliz, 2024). Multimodal models enhance robustness by reweighting inputs dynamically: poor-quality imaging can be offset by labs or text records (Zilliz, 2024). Noise-tolerant training (pixel dropout, word swaps, contrastive learning) further improves generalization. To manage missingness, MARIA uses masked self-attention and modality dropout to focus on available data without bias (Caruso et al., 2024), while tensor factorization methods like REPAR integrate RNN regularization and low-rank constraints to handle incomplete longitudinal EHRs effectively (Yifei et al., 2025). These strategies collectively strengthen model resilience in messy, incomplete datasets.

4.2.4 Quality Control and Validation

Ensuring data quality in healthcare AI is an ongoing process requiring constant monitoring and human oversight (Girman et al., 2022). LLM variability and limited generalizability underscore the importance of a *human-in-the-loop* framework (Artsi et al., 2025). Studies show clinicians may miss AI-generated errors due to automation bias, reinforcing the need for human validation (Littrell, 2025). Regulatory guidance (e.g., FDA) now defines data quality as a life-cycle obligation involving iterative cleansing and review. Post-deployment, continuous human monitoring remains essential to identify errors and ensure adherence to safety standards (Artsi et al., 2025; Littrell, 2025). Ultimately, reliable healthcare AI depends on ongoing data validation, error correction, and transparent accountability mechanisms integrated throughout the data and model lifecycle.

4.3 Privacy-Preserving Mechanisms

4.3.1 Federated Learning

Building a ***Messy Clinic*** dataset requires robust privacy-preserving methods to protect patient confidentiality while enabling large-scale AI development. Federated Learning (FL) allows multiple healthcare institutions to collaboratively train AI models without sharing raw data, exchanging only model updates or parameters (Dhade and Shirke, 2024; Rafalski, 2025). In this framework, hospitals act as client nodes training local models, a central server aggregates updates, and a global model evolves through iterative refinement using aggregation algorithms like Federated Averaging (Dhade and Shirke, 2024; Rafalski, 2025). FL enables access to heterogeneous, real-world datasets while preserving privacy and substantially improving model performance (Xu et al., 2021). Challenges include non-independent and non-identically distributed (non-IID) data across clients, which can be addressed via approaches such as Agnostic FL, q-Fair FL, and Multi-task Learning (Xu et al., 2021). Communication efficiency is also critical, requiring client selection, model compression, and update reduction. Despite its privacy advantages, FL remains vulnerable to security threats like model inversion and poisoning attacks, necessitating additional safeguards for robust and secure collaborative AI (Syenza, 2025; Rafalski, 2025).

4.3.2 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 (Apellániz et al., 2025). 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 (Apellániz et al., 2025). 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 (Apellániz et al., 2025). 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 (Apellániz et al., 2025). Consequently, SDG represents a promising, privacy-preserving strategy to augment healthcare datasets, enabling accelerated research and model development without exposing sensitive information (Lucente, 2025).

4.3.3 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 (Rafalski, 2025). 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 (Xu et al., 2021). 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 (Xu et al., 2021).

4.3.4 Multi-Party Collaboration

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 (Xu et al., 2021). HE allows computations directly on encrypted data, ensuring data remains confidential during transmission and processing (Xu et al., 2021). 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 (Rafalski, 2025).

4.3.5 Privacy-Preserving Methods

Clinical free text—including discharge summaries, progress notes, and radiology reports—contains dense PHI, requiring stronger privacy protections than structured data (Stubbs and Uzuner, 2015). Traditional de-identification combines rule-based matching with NER; transformer models improve recall on rare PHI categories (Wu et al., 2022). DP-SGD provides formal privacy guarantees against memorization and model inversion via gradient clipping and calibrated noise (Abadi et al., 2016; Carlini et al., 2021). Federated fine-tuning enables decentralized training across institutions without sharing raw clinical notes (Dhade and Shirke, 2024). Generative models can synthesize clinical notes conditioned on structured EHRs for privacy-preserving data augmentation (Tang et al., 2023). Together, these methods extend the ***Messy Clinic*** privacy framework to unstructured clinical text while preserving utility.

5 Ethical and Governance Framework

Developing a multimodal, longitudinal, and privacy-preserving ***Messy Clinic*** dataset demands a robust ethical and governance framework to ensure trust, compliance, and harm mitigation.

5.1 Consent Models

Traditional Consent. Static, paper-based consent cannot accommodate dynamic AI research or continuous data reuse. Participants often remain unaware of how their data evolves in use (Kaye et al., 2015).

Dynamic Consent. Dynamic consent employs digital tools for ongoing participant engagement, en-

abling real-time consent modification through online or mobile platforms (Kaye et al., 2015). It enhances transparency, regulatory compliance, and participant empowerment (Andrade, 2025). Two-way interaction allows data uploads by participants and feedback or updates from researchers (Kaye et al., 2015).

Broad Consent. Broad consent enables storage and secondary use of identifiable information or biospecimens for future, unspecified research under defined oversight (Maloy and III, 2020). While flexible for large-scale research, it raises ethical concerns over adequacy of non-specific consent given unpredictable data uses (Wendler, 2013).

Challenges with PGHD and AI. The surge of Patient-Generated Health Data (PGHD) from wearables complicates consent since such data often lie beyond traditional health data laws (e.g., HIPAA) and are governed by tech company policies (Winter, 2021). Some firms bypass consent for AI innovation, creating tension between research utility and informed consent principles (Winter, 2021). Building trust requires balancing individual rights, community interests, and corporate data control (Winter, 2021).

5.2 Data Governance and Accountability

Robust governance ensures ethical, secure management of decentralized healthcare data within federated learning systems (Syenza, 2025). It maximizes privacy protection, supports institutional collaboration, and aids compliance with complex legal and reimbursement standards (Syenza, 2025). Federated settings, however, pose risks of re-identification, data leakage, and unregulated secondary use (Syenza, 2025), while large-scale AI training conflicts with data reuse restrictions under GDPR (Winter, 2021).

A comprehensive framework spans **procedural, relational, and structural** mechanisms (Syenza, 2025). Procedural controls include privacy safeguards, data-sharing agreements, and transparent monitoring. Relational mechanisms involve stakeholder engagement, dynamic consent, and public participation to maintain social legitimacy. Structural elements establish oversight bodies, data stewards, and consumer representation for accountability. Despite GDPR and HIPAA, fragmented ethical oversight still produces accountability gaps, underscoring the need for unified, evidence-based governance to uphold privacy and trust in federated systems (Syenza, 2025).

5.3 Bias Mitigation and Fairness

Integration of AI into healthcare carries the risk of generating biased results, which can lead to misleading clinical conclusions or overestimation of model performance (Yang et al., 2023).

Sources of Bias in Healthcare AI. Bias in medical imaging datasets can cause uneven diagnostic accuracy across patient groups, worsening healthcare disparities (Xu et al., 2024b). Key sources include homogeneous datasets lacking population diversity (El Arab et al., 2025), institutional and geographic biases due to limited collection sites and varying protocols (Xu et al., 2024b), annotation bias from inter-observer variability among radiologists, and selection bias from non-representative patient sampling, such as overrepresentation of advanced-stage cases (Xu et al., 2024b).

Strategies for Mitigation. Mitigating bias and variability in medical imaging datasets is critical for developing fair and reliable AI systems (Xu et al., 2024b). Key strategies include data augmentation using GANs to diversify datasets, federated learning to leverage decentralized and diverse data while preserving privacy, and standardizing data collection and annotation protocols to reduce variability (Xu et al., 2024b). Additionally, explainable AI (XAI) enhances transparency for bias detection, while regulatory frameworks increasingly mandate bias mitigation (Xu et al., 2024b). Adversarial testing methods like MedFuzz further probe model weaknesses by simulating real-world challenges, improving robustness against biases (Potts, 2024).

6 Alternative Views

Here, we examine two substantive alternative views that challenge our position:

Alternative View 1: Scaling and Better Architectures Will Solve Deployment Challenges. A common view holds that continued advances in model scale, multimodal pretraining, and architecture will naturally resolve current deployment limitations (Zhou et al., 2025; Wang et al., 2024b). Although model capabilities have improved rapidly, deployed systems still fail because of fragmented data, workflow misalignment, poor monitoring, and lack of interoperability rather than insufficient reasoning ability alone (Maimaitiaili et al., 2025; Tejani et al., 2024). The ***Messy Clinic*** paradigm therefore argues that model scaling and data infrastructure are co-dependent: models trained only on curated benchmarks cannot learn the structural

missingness, temporal irregularities, and ontology drift present in clinical deployment (Artsi et al., 2025; El Arab et al., 2025; Yang et al., 2023). Our position is not anti-scaling; rather, it specifies the real-world data conditions required for scaling to translate into clinical benefit.

Alternative View 2: Clinical AI Adoption Is Primarily an Organizational Problem. Another perspective attributes deployment failures mainly to clinician resistance, organizational inertia, and regulatory barriers rather than technical shortcomings (Belani, 2025; Karofsky, 2025). While these factors are important, many organizational failures originate from technical deficiencies such as opaque behavior, poor workflow integration, limited monitoring, and lack of explainability (medDARE AI, 2025; Fiddler AI, 2025). The ***Messy Clinic*** framework addresses these issues through governance pipelines, monitoring, federated accountability, and robustness to irregular clinical data (Syenza, 2025; Xu et al., 2021). Clinician trust ultimately depends on whether systems behave predictably under real-world conditions, making technical robustness a prerequisite for sustainable organizational adoption.

7 Concluding Remarks

We argue that robust clinical NLP requires messy, multimodal, longitudinal, and privacy-preserving corpora rather than sanitized datasets. The ***Messy Clinic*** paradigm treats missingness, noise, inconsistencies, and temporal irregularities as clinically meaningful signals that models must learn to handle. By integrating heterogeneous healthcare modalities with privacy-aware and human-centered design, the framework supports more reliable, equitable, and generalizable healthcare AI. Overall, ***Messy Clinic*** reframes clinical AI around authentic real-world complexity rather than idealized data assumptions.

Limitations

As a position paper, this work prioritizes architectural and methodological specification over empirical benchmarking, which is consistent with the position paper format at major NLP venues (Xu et al., 2024a; Sculley et al., 2025). The following limitations should be noted by readers.

Absence of a Reference Implementation. The ***Messy Clinic*** blueprint describes a target system architecture but does not include a deployed dataset, cost model, or institutional partnership structure. The practical barriers to assembling such a corpus, data use agreements, IRB approvals, federated infrastructure costs, and cross-institutional incentive alignment, are real and non-trivial. We regard scoping these as a necessary next step rather than a prerequisite for establishing the paradigm.

Empirical Validation. The framework's claims rest on a synthesis of existing empirical literature rather than on experiments conducted by the authors. Dedicated empirical studies comparing models trained under the ***Messy Clinic*** specification against those trained on cleaned datasets across multiple clinical tasks and sites are needed to quantify the robustness gains the paradigm predicts.

Scope of Clinical NLP Text Privacy. The original submission did not address privacy-preserving techniques specific to clinical free text; this has been addressed in Section 4.3.5 of the revised paper. Further work on multilingual de-identification and low-resource clinical NLP privacy remains a gap.

Infrastructure Assumptions. Several recommendations (e.g., federated LLM fine-tuning, GFlowNet-based messiness simulation) require computational resources that may not be available in low-resource healthcare settings. The framework should be interpreted as a specification of what is technically achievable and desirable, with implementations scaled to available infrastructure.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Diana Andrade. 2025. Dynamic consent: A new gdpr standard for clinical trials. `https://clinicaltrialvanguard.com/article/dynamic-consent-a-new-gdpr-standard-for-clinical-trials/`.

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

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

Gaurav Belani. 2025. Deploying ai models in clinical workflows: Challenges and best practices. Accessed July 1, 2025.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J. Hu, Mo Tiwari, and Emmanuel Bengio. 2023. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55.

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

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

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

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

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

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

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

Filipe R. Cordeiro and Gustavo Carneiro. 2020. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?

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

Karen Drukker, Weijie Chen, Judy Gichoya, Nicholas Gruszauskas, Jayashree Kalpathy-Cramer, Sanmi Koyejo, Kyle Myers, Rui C. Sá, Berkman Sahiner, Heather Whitney, Zi Zhang, and Maryellen Giger. 2023. 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(06).

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

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

Fiddler AI. 2025. Model monitoring framework. Accessed July 1, 2025.

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Ruisi Su, Samuele Garda, Sunny MS Kang, Stella Biderman, Matthias Samwald, Stephen H. Bach, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Perinan, Théo Gigant, Patrick Haller, Jenny Chim, Jose Posada, John Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. Bigbio: a framework for data-centric biomedical natural language processing.

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, and Meredith N. Zozus. 2023. Error rates of data processing methods in clinical research: A systematic review and meta-analysis of manuscripts identified through pubmed.

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

Cynthia J Girman, Mary E Ritchey, and Vincent Lo Re III. 2022. 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.

Eric Karofsky. 2025. Beyond the molecule: How human-centered design unlocks ai's promise in pharma. Accessed July 1, 2025.

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

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

Austin Littrell. 2025. Pros and cons of ai patient portal messages. Accessed: 2025-07-17.

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

Ida Lucente. 2025. Generative ai in healthcare: Use cases, benefits, and challenges. https://johnsnowlabs.com/generative-ai-healthcare/.

Musitapa Maimaitiaili, Yiershatijiang Jiamaliding, Guangle Dai, Hui Xiao, Warisijiang Kuerbanjiang, and Yuexiong Yi. 2025. Artificial intelligence platform architecture for hospital systems: Systematic review. *Journal of Medical Internet Research*, 27:e79788.

Alexander Maletzky, Carl Böck, Thomas Tschoellitsch, Theresa Roland, Helga Ludwig, Stefan Thumfart, Michael Giretzlehner, Sepp Hochreiter, and Jens Meier. 2022. Lifting hospital electronic health record data treasures: Challenges and opportunities. *JMIR Medical Informatics*, 10(10):e38557.

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

medDARE AI. 2025. Five barriers slowing ai adoption in hospitals. Accessed: 2026-01-26.

Milvus Team. 2025. What are the applications of multimodal search in healthcare? https://milvus.io/ai-quick-reference/what-are-the-applications-of-multimodal-search-in-healthcare.

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

Ayush Noori, Adam Rodman, Alan Karthikesalingam, Bilal A. Mateen, Christopher A. Longhurst, Daniel Yang, Dave deBronkart, Gauden Galea, Harold F. Wolf III, Jacob Waxman, Joshua C. Mandel, Juliana Rotich, Kenneth D. Mandl, Maryam Mustafa, Melissa Miles, Nigam H. Shah, Peter Lee, Robert Korom, Scott Mahoney, Seth Hain, Tien Yin Wong, Trevor Mundel, Vivek Natarajan, Noa Dagan, David A. Clifton, Ran D. Balicer, Isaac S. Kohane, and Marinka Zitnik. 2025. A global log for medical ai.

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

B Potts. 2024. 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*.

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

Kacper Rafalski. 2025. Federated learning: A privacy-preserving approach to collaborative ai model training. https://www.netguru.com/blog/federated-learning. Updated March 17, 2025. 26 min read.

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

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

D. Sculley, William Cukierski, Phil Culliton, Sohier Dane, Maggie M Demkin, Ryan Holbrook, Addison Howard, Paul T Mooney, Walter Reade, Meg Risdal, and Nate Keating. 2025. Position: AI competitions provide the gold standard for empirical rigor in genAI evaluation. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Syenza. 2025. Federated learning governance in healthcare: A framework for ethical and effective implementation. https://news.syenza.com/federated-learning-governance-healthcare/.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining?

Ali S. Tejani, Tessa S. Cook, Mohannad Hussain, Teri Sippel Schmidt, and Kevin P. O'Donnell. 2024. Integrating and adopting ai in the radiology workflow: A primer for standards and integrating the healthcare enterprise (ihe) profiles. *Radiology*, 311:e232653.

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

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *Preprint*, arXiv:2401.06805.

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

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

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T. C. Poon, Natalie Fitzpatrick, Adam P. Levine, Luke T. Slater, Alex Handy, Andreas Karwath, Georgios V. Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, William Whiteley, Cathie Sudlow, Angus Roberts, and Richard J. B. Dobson. 2022. A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *npj Digital Medicine*, 5(1).

Qingyang Wu, Ying Xu, Tingsong Xiao, Yunze Xiao, Yitong Li, Tianyang Wang, Yichi Zhang, Shanghai Zhong, Yuwei Zhang, Wei Lu, and Yifan Yang. 2025. Surveying attitudinal alignment between large language models vs. humans towards 17 sustainable development goals. *Preprint*, arXiv:2404.13885.

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

Xinyi Xu, Zhaoxuan Wu, Rui Qiao, Arun Verma, Yao Shu, Jingtan Wang, Xinyuan Niu, Zhenfeng He, Jiangwei Chen, Zijian Zhou, Gregory Kang Ruey Lau, Hieu Dao, Lucas Agussurja, Rachael Hwee Ling Sim, Xiaoqiang Lin, Wenyang Hu, Zhongxiang Dai, Pang Wei Koh, and Bryan Kian Hsiang Low. 2024a. Position paper: Data-centric AI in the age of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11895–11913, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

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

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

Juexiao Zhou, Haoyang Li, Siyuan Chen, Zhangtianyi Chen, Zhongyi Han, and Xin Gao. 2025. Large language models in biomedicine and healthcare. *npj Artificial Intelligence*, 1:44.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. 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.

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

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

A Conceptual Scope and Integration Principles

This appendix consolidates the substantive clarifications and proposed protocols developed during author response. The *Messy Clinic* is a design paradigm from which we derive a corpus specification and a governance framework. It is not a released dataset, an implemented system, or a benchmark with completed experiments. The protocols below are proposals for future validation; the case study is illustrative.

The conceptual contribution comprises three elements: (i) preserving and modeling clinically informative artifacts rather than applying uniform cleaning; (ii) distinguishing artifact categories and assigning category-specific handling strategies; and (iii) specifying how harmonization, multimodal modeling, temporal reasoning, privacy, and governance interact. Existing modeling algorithms, interoperability standards, and privacy mechanisms are candidate implementations of these principles rather than new methods introduced by this paper.

Sequential harmonization and modeling. Structural harmonization aligns schemas, terminology, and units using standards such as SNOMED CT, LOINC, and OMOP CDM. Modeling then addresses residual missingness, irregular sampling, and disagreement. These stages have different objectives: structural consistency does not imply that clinical uncertainty has disappeared.

Cross-modal redundancy. Overlapping evidence across notes, laboratory measurements, and imaging reports provides opportunities to identify discrepancies. Agreement is a consistency signal,

not proof of correctness: modalities may share errors, refer to different times, or be unavailable. Disagreement should be surfaced with provenance and temporal context rather than resolved automatically.

Privacy and governance across stages. Federated access and learning can support collaboration without centralizing raw records. Differential privacy and secure aggregation may supplement this design where appropriate. Consent checks, access controls, audit trails, and human oversight apply throughout the pipeline. Federated learning alone does not guarantee privacy, and retaining site heterogeneity requires explicit evaluation of site-dependent bias.

B Operational Corpus Specification

A proposed corpus meets the full Messy Clinic specification when it satisfies all six requirements below. The numerical minima are proposed design criteria, not empirically established thresholds.

1. **Multi-institutional provenance:** records from at least two independent healthcare institutions, retaining metadata on site-specific coding conventions, schemas, and workflows.

2. **Multimodal coverage:** at least two modalities for a shared patient cohort, drawn from structured EHR, clinical text, imaging, physiological time series, and genomics. Complete modality coverage for every patient is not required.

3. **Authentic temporal structure:** at least two encounters per patient, with event-driven timestamps and irregular sampling preserved. Any derived resampling should retain access to original timestamps and missingness masks.

4. **Artifact-aware preservation:** retain informative missingness and residual uncertainty, with annotations of source mechanisms where feasible. Correct or quarantine harmful errors according to Appendix C; preservation is not a requirement to propagate known errors.

5. **Privacy protection:** document applicable de-identification, federated access, differential privacy, or other protections, including audit procedures. Synthetic augmentation must be assessed for disclosure risk; synthesis alone is not a privacy guarantee.

6. **Governance documentation:** specify the consent model, data-use agreements, access responsibilities, and demographic representation and bias audits.

Corpora meeting requirements 1–4 may be considered data-level compliant. Full specification compliance also requires requirements 5–6; it does not itself establish readiness for clinical deployment.

Construction protocol. Define the cohort, prediction times, permitted uses, and outcome windows before ingestion. Establish site-level agreements and local privacy controls; ingest records while retaining provenance, original timestamps, and missingness masks; harmonize terminology and units; annotate artifacts and corrections; and create patient-disjoint development and evaluation partitions. Reserve external sites and later time periods where available. Document exclusions, modality coverage, missingness distributions, and subgroup representation. In federated implementations, ingestion and harmonization occur locally, with only authorized model updates or approved aggregates exchanged.

Suggested record metadata. An implementation should retain a site-scoped pseudonymous patient identifier, encounter identifier, source site, modality, event time, documentation time when available, original code and units, harmonized representation, modality-availability mask, artifact category, correction or quarantine status, and permitted-use metadata. This is a proposed metadata checklist rather than a released executable schema. Cross-site patient linkage, if needed, requires a separately approved privacy-preserving procedure.

C Clinical Data Artifact Taxonomy and Handling

Table 1 distinguishes eight artifact categories. Multiple categories can apply to one record; handling should preserve provenance and document uncertainty about the generating mechanism.

Decision procedure. First screen for potentially harmful content requiring quarantine or human review. For remaining artifacts, determine whether the signal is systematically informative; if so, retain it with clinical or resource-context metadata.

Artifact category	Source mechanism	Illustrative example	Handling strategy
Meaningful clinical missingness	Conditional testing; a test judged unnecessary	Troponin not measured during a dermatology visit	Retain availability indicators and model the clinical context; do not assume that absence means a negative result.
Resource-driven missingness	Equipment, cost, or staffing limitations	MRI unavailable at a rural facility	Retain resource-context metadata; assess subgroup and site confounding.
Documentation noise	Typos, duplicates, copy-paste, template artifacts	Duplicate medication entries	Correct verified errors; preserve an audit trail. Use noise-tolerant modeling for uncertain cases.
Coding errors	Misapplied codes or inconsistent granularity	Type 2 diabetes code assigned to a Type 1 patient	Validate with ontologies and supporting records; review uncertain reconciliations.
Cross-institutional schema drift	Different vendors, units, and local terminology	Laboratory values in mg/dL versus mmol/L	Harmonize units and concepts using SNOMED CT, LOINC, and OMOP; retain original representations.
Temporal heterogeneity	Changes in protocols, software, or guidelines	Sepsis criteria changing over time	Annotate time periods and protocol versions; evaluate temporal drift.
Demographic or selection bias	Non-representative sampling and referral patterns	Late-stage cases overrepresented at a tertiary center	Retain records, quantify selection and subgroup effects, and audit calibration and performance.
Adversarial or automation artifacts	Generated-text errors and automation complacency	AI-drafted message with an incorrect dosage	Quarantine or exclude unsafe content; require human review and retain the audit record.

Table 1: Proposed taxonomy of clinical data artifacts and category-specific handling, consolidated from the author response. Examples are illustrative.

Correct detectable errors only when confidence is supported by validation, retaining the original record and transformation history. The response proposed correction precision above 0.90 as an illustrative threshold; application-specific validation is needed, especially for safety-critical fields. Annotate temporal shifts, demographic bias, and unresolved discrepancies for downstream auditing. After structural harmonization, retain remaining uncertain artifacts and evaluate noise-tolerant models. These actions can be combined: a corrected field may still need a bias flag or temporal annotation.

Simulation as a fallback. When authentic data cannot be accessed, clinically grounded simulation should represent conditional testing, resource constraints, documentation delays, and modality availability. Independent Gaussian corruption does not encode these mechanisms. Synthetic perturbations should be reported separately from naturally occurring artifacts, and their usefulness assessed against authentic held-out data where possible.

D Implications for Clinical NLP Evaluation

Table 2 translates the specification into six clinical NLP evaluation settings. It complements conventional curated benchmarks with explicit tests of record fragmentation, institutional variation, temporal uncertainty, and conflicting evidence.

Multilingual and low-resource settings. Evaluation should account for local abbreviations, language mixing, institution-specific identifiers, and variable documentation infrastructure. Report performance by language and site rather than relying on aggregate English-language results. De-identification must be validated for each intended setting. Context from structured data and imaging reports may support text interpretation, but introduces additional coverage and privacy requirements.

E Falsifiable Research Agenda

Table 3 consolidates the four hypotheses proposed in the rebuttal. Thresholds are proposed targets for preregistration, not observed effects. Failure to obtain statistical significance does not by itself establish equivalence or falsify a hypothesis; effect sizes, confidence intervals, statistical power, and predefined practical margins must inform interpretation.

Evaluation controls. Match cohort eligibility, model capacity, training budget, outcome windows, and prediction-time information across compar-

Clinical NLP task	Relevant artifact conditions	Proposed evaluation change
Named entity recognition	Typos, abbreviation variation, copy-paste artifacts	Evaluate authentic noisy notes and cross-site terminology. Report entity-level precision, recall, and F1 by site and artifact category.
Relation extraction	Coding inconsistencies and schema drift	Evaluate drug–disease and symptom–diagnosis relations across sites; report relation F1 and error types under terminology variation.
Temporal event extraction	Irregular timestamps and documentation delays	Distinguish event time from documentation time; evaluate temporal relations and uncertainty under delayed or incomplete documentation.
De-identification	Rare PHI, site-specific identifiers, multilingual notes	Report PHI-category recall and precision, including rare categories and held-out sites or languages; audit residual disclosure risk.
Note understanding and summarization	Missing prior notes, incomplete medication lists, fragmented histories	Evaluate factual consistency, omissions, unsupported statements, and communication of uncertainty as input completeness changes.
Question answering and clinical language models	Contradictory notes and outdated entries	Evaluate evidence-grounded answers, temporal conflict handling, and appropriate abstention; distinguish current from historical evidence.

Table 2: Clinical NLP task mapping from the author response, with proposed evaluation outcomes. These are evaluation recommendations, not reported experimental results.

isons. Complete-case filtering can change cohort composition; report selection effects and, where feasible, include matched-cohort sensitivity analyses. Use patient-disjoint partitions, evaluate multiple seeds, and report uncertainty and subgroup calibration. Public datasets can support initial single-site studies; multi-institutional and external-site claims require additional appropriately governed sites or suitable multi-site resources. Simulated site partitions must not be described as independent hospitals.

Tasks and metrics. For mortality, sepsis onset, and 30-day readmission, report AUROC and AUPRC alongside calibration and clinically relevant operating points. For length of stay, define the prediction target: continuous prediction requires regression metrics such as MAE; thresholded classification permits AUROC and AUPRC. Add the NLP outcomes in Table 2 to distinguish clinical-text robustness from general clinical prediction performance.

Candidate methods and comparators. Missing-modality models such as MARIA and irregular temporal models such as REPAR are candidate implementations of the framework (Caruso et al., 2024; Yifei et al., 2025). The response also identified GRU-D, mTAND, SeFT, AID-MAE, and Denoise2Impute as relevant candidates. Their inclusion in a future study requires task-compatible implementations and verified bibliographic entries; this appendix makes no new comparative performance claim about them.

F Illustrative Pipeline and End-to-End Case Study

Table 4 provides a text-based architecture summary of the four stages proposed during the author response. Privacy and governance controls span the stages rather than operating only after prediction.

Consider a hypothetical patient with type 2 diabetes and intermittent chest discomfort who receives care at a rural clinic (Site A) and a tertiary hospital (Site B), and uses a continuous glucose monitor (CGM). The following workflow illustrates handling decisions without reporting an actual prediction or clinical outcome.

Stage 1: ingestion. Site A holds HbA1c measurements, a medication list, and clinical notes. Site B holds an ECG report, a cardiology note, and CGM observations. A troponin measurement is unavailable at Site A because of resource limitations. Medication lists disagree following a change between visits. Authorized records are processed locally in the proposed federated workflow.

Stage 2: harmonization. Map the sites' diabetes codes to a common concept representation and normalize compatible laboratory units. Retain the original values, coding systems, and units. Align encounters and available event times. Flag the medication discrepancy with its provenance; a candidate reconciliation must not silently overwrite conflicting evidence.

ID	Hypothesis	Proposed protocol	Result challenging the claim
H1	Authentic missingness preserves useful signal beyond clean-then-corrupted training.	Compare matched models trained on native missingness versus complete-case data with synthetic corruption; evaluate mortality, sepsis, and readmission across at least three tasks and two external sites where accessible.	No meaningful AUROC gain under adequately powered comparisons. The rebuttal proposed $p > 0.05$ as a warning criterion, not evidence of equivalence.
H2	Cross-modal redundancy improves detection of clinical inconsistencies.	Inject clinically plausible medication–diagnosis mismatches or laterality errors; compare multimodal detection with the strongest unimodal baseline at matched false-positive rates.	A detection-rate gain of at most two percentage points under a predefined equivalence or practical-effect analysis.
H3	Multi-site federated training improves held-out-site generalization over single-site training.	Use at least three non-IID training institutions. Compare federated training with single-site models; include centralized multi-site training where permitted as a separate comparator.	External-site AUROC more than one percentage point below the single-site baseline challenges the claim; equivalence also fails to demonstrate superiority.
H4	Irregularity-aware temporal models improve prediction under high missingness.	Compare REPAR-class methods with LSTM, GRU, and Transformer baselines at 20%, 40%, and 60% missingness; distinguish authentic from simulated missingness.	No meaningful gain at 40% or above under adequately powered comparisons and predefined effect margins.

Table 3: Proposed research agenda derived from the rebuttal. Protocol clarifications separate single-site and centralized multi-site comparisons and distinguish non-significance from falsification.

Stage	Operation	Output and interaction
1. Ingestion	Access authorized site-local EHR, text, imaging, physiological, and genomic records.	Preserve timestamps, modality masks, provenance, and permitted-use metadata. Federated workflows retain raw records locally.
2. Harmonization	Align schemas, units, and concepts; reconcile inconsistencies with validation.	Produce aligned representations while retaining originals, unresolved conflicts, and residual missingness for modeling.
3. Modeling	Apply missingness-aware, temporal, and multimodal methods; assess cross-modal consistency.	Produce task outputs with uncertainty and discrepancy flags. Cross-modal evidence supports checking rather than automatic conflict suppression.
4. Governance and review	Enforce consent and access restrictions; audit bias, safety, and accountability.	Review outputs before use, record decisions, and feed corrections into the pipeline. Controls also apply upstream.

Table 4: Architecture summary consolidated from the rebuttal's pipeline sketch. The stages describe a proposed design, not a deployed system.

Stage 3: modeling. Annotate unavailable troponin as resource-driven missingness, distinguishing it from a clinically unnecessary test or a negative result. Use availability masks for missing modalities and a temporal model for irregular CGM observations. Treat medication disagreement as a potentially time-dependent change requiring review. Compare glucose observations with diabetes-related notes and compare the ECG report with the chest-discomfort narrative as consistency checks, without treating agreement as diagnostic validation.

Stage 4: governance. Enforce the patient's hypothetical permission for cardiovascular research and restriction on genomic use. Audit site and subgroup coverage and inspect the flagged medication history before an output reaches a clinician. Record consent checks, transformations, and review decisions.

Illustrative output. A proposed system would return a task-specific prediction or summary with missingness indicators, uncertainty, and the unresolved medication flag. Whether these features improve prediction, factuality, or safety remains an empirical question addressed by Appendix E. The example does not demonstrate clinical effectiveness.

G Failure Modes and Safeguards

Unequal modality quality. Fusion may overweight imaging that is systematically lower quality at under-resourced sites. Evaluate modality quality and availability by site and subgroup, test modality

ablations, and audit whether fusion worsens subgroup outcomes.

Missingness as a proxy for disadvantage. Resource-driven gaps can induce spurious associations between geography, access, and outcome. Keep resource context distinct from clinical judgment, compare stratified performance and calibration, and examine whether missingness features amplify inequities.

Selection and temporal bias. A model trained mainly on late-stage tertiary-center cases may be miscalibrated for early-stage community patients. Evaluate held-out sites and time periods, document referral and case-mix differences, and monitor distribution drift. Multi-site training is an intervention to test, not a guarantee of fairness.

Privacy and automation failures. Residual PHI, information leakage through updates, and incorrect generated medication instructions require dedicated auditing and access controls. Quarantine potentially harmful outputs, preserve evidence for review, and maintain clear responsibility for corrections. A human review step is a safeguard whose effectiveness must itself be assessed.

H Evidence Boundaries and Reproducibility

The contribution remains conceptual. No reference implementation, institutional consortium, cost model, new corpus, or completed comparative study is released here. Proposed data requirements and evaluation protocols are intended to make future studies assessable; they do not replace empirical validation.

MIMIC, i2b2/n2c2, OMOP CDM, and FHIR provide relevant resource or interoperability contexts discussed in the response. Structural standards and curated task benchmarks address different parts of the problem from robustness under residual clinical complexity. A reproducible implementation should document cohort construction, harmonization rules, artifact annotations, site and temporal splits, hyperparameters, privacy parameters, and governance restrictions.

Core technical claims should be supported by primary academic references. Vendor pages and deployment reports, where retained in the main paper, should be identified as contextual sources rather than experimental evidence. The rebuttal's proposed citation audit and literature additions require the accompanying bibliography to be updated; this appendix does not claim that an unavailable bibliography has been audited or corrected.

I Recommendations and Future Work Directions

To realize the full potential of the ***Messy Clinic*** dataset and its impact on healthcare AI, we propose the following recommendations:

- **Prioritize Multimodal Data Integration with Intelligent Fusion:** Develop AI architectures capable of fusing heterogeneous data types (EHRs, imaging, genomics, wearables, PGHD) using fusion layers and attention mechanisms that exploit cross-modal redundancy for noise reduction and robust representation learning, beyond simple data aggregation.
- **Develop AI-Native Data Harmonization Pipelines:** Incorporate Large Language Models (LLMs) and biomedical ontologies within data ingestion and curation workflows to automate semantic mapping, schema matching, and reconciliation of inconsistent terminologies across diverse clinical data sources, reducing manual intervention.
- **Implement Robust Longitudinal Data Modeling:** Employ and advance temporal modeling techniques such as RNN-regularized tensor factorization to capture non-linear temporal dependencies, irregular intervals, and missing data in longitudinal patient records, improving predictive analytics accuracy.
- **Establish Privacy-by-Design Frameworks:** Utilize federated learning as the core approach for collaborative AI training on sensitive data, supplemented by cryptographic methods including Secure Multi-Party Computation (SMPC), Homomorphic Encryption (HE), and Differential Privacy (DP) to balance privacy protection and model performance.
- **Implement Continuous Data Quality Assurance with Human Oversight:** Integrate iterative data quality assessments and cleansing throughout the data lifecycle, incorporating human-in-the-loop mechanisms to monitor AI outputs, detect errors, and ensure ethical compliance and patient safety.

Table 5: Multimodal Data Types and Their Characteristics in Healthcare

Data Type	Typical Format	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	(Zhu et al., 2024)
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	(Milvus Team, 2025)
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	(Milvus Team, 2025)
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	(Georgiou, 2025)
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	(Littrell, 2025)

- **Develop Ethical and Governance Frameworks:** Create consolidated governance structures addressing patient consent (favoring dynamic consent models where applicable), data ownership, accountability, and bias mitigation. Promote transparent stakeholder communication and public engagement to maintain trust in AI applications involving sensitive health data.
- **Incorporate Adversarial Testing and Real-World Validation:** Apply adversarial machine learning methods to evaluate AI models against realistic clinical complexities, noise, and biases. Conduct multi-site validation to verify model generalizability and robustness prior to clinical deployment.

By systematically implementing this blueprint, healthcare institutions can construct a ***Messy Clinic*** dataset that serves as a powerful foundation for developing next-generation AI models. These models will not only achieve superior predictive performance but will also be inherently more robust, generalizable, and trustworthy, ultimately translating into tangible improvements in patient care and public health outcomes.