Paper RE07

From Chaos to Clarity: Can AI/ML Truly Decode Unstructured Clinical Data?

Uma Adiseshan, Zifo RnD Solutions, Boston, USA

ABSTRACT

Unstructured clinical data, including physician notes, patient-reported outcomes, and especially clinical images, contains high-value information for downstream analysis, biomarker discovery, and long-term research reuse. However, privacy regulations (e.g., HIPAA in the United States and GDPR in the European Union) restrict sharing of sensitive data unless protected health information (PHI) is properly removed. In clinical images, PHI may exist in both pixel overlays (labels, IDs) and metadata (e.g., DICOM headers), creating a challenging de-identification problem at scale.

This paper presents a hybrid Generative AI (GenAI) approach combining automated detection/redaction with targeted human review to enable scalable, compliant PHI redaction while preserving clinical utility. In an ophthalmology trial use case (approximately 12,000 images across fundus, OCT, and external modalities), the hybrid workflow reduced average processing time to approximately 15 seconds per image (AI plus QA), required human review for approximately 3% of images, achieved PHI redaction specificity of 98.79, and enabled substantial manual effort reduction. The approach also supports downstream machine learning feature extraction and secondary analyses by producing standardized, de-identified image archives.

INTRODUCTION

Unstructured data is increasingly central to clinical development and evidence generation. While structured datasets remain foundational for analysis and submissions, real-world clinical insights often reside in unstructured artifacts, including free text, scanned documents, and imaging. Medical images are particularly valuable for diagnosis, research, collaboration, and ML-driven feature extraction, but are difficult to share broadly because they frequently embed PHI in image pixels and/or metadata.

Regulatory frameworks such as HIPAA (US) and GDPR (EU) require that identifying information be removed before data can be published or shared. HIPAA Safe Harbor identifies 18 categories of identifiers that must be removed for de-identification. Consequently, clinical teams need PHI redaction solutions that are accurate, auditable, scalable, and designed to preserve diagnostic regions so downstream clinical interpretation and analytics remain valid.

BACKGROUND: PHI REDACTION CHALLENGES IN CLINICAL IMAGES

PHI redaction in images is challenging due to data complexity across imaging modalities and device outputs, metadata risks in addition to pixel-level overlays, volume and scale in modern trials and archives, variation across devices and sites, regulatory compliance requirements plus auditability needs, and the need to maintain data utility (ensuring redaction does not degrade clinically meaningful regions).

Traditional approaches include manual redaction and conventional image processing. Rule-based automation can help but often struggles with variability, handwriting, low-quality borders, and partially occluded text. A hybrid GenAI approach (automated detection/redaction plus human oversight) can balance scalability, accuracy, and adaptability, especially for ambiguous or rare edge cases.

OBJECTIVE

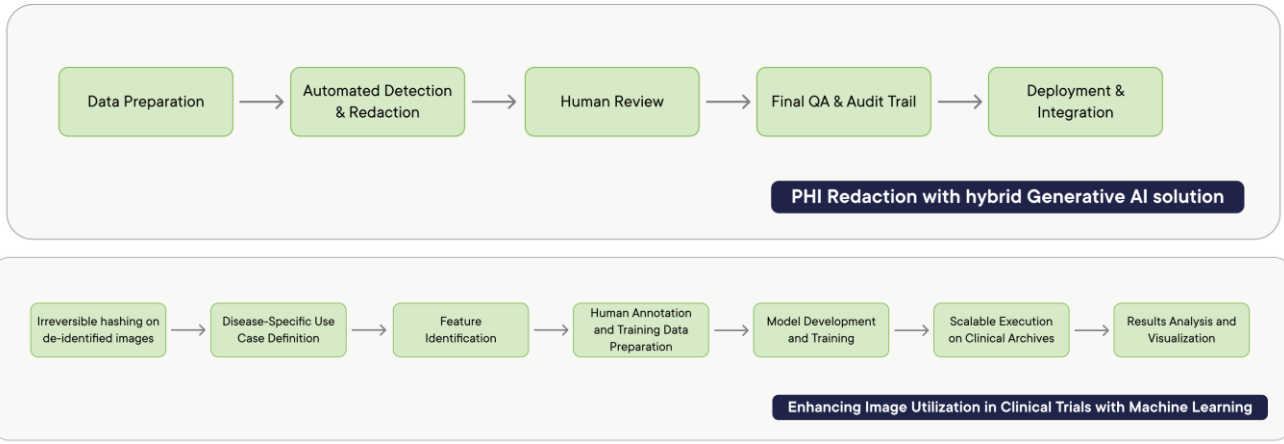
- To create a rigorous, automated, and compliant PHI redaction solution using GenAI.
- To make images downstream ready for secondary analysis and long-term storage.

RESEARCH QUESTION

Can a hybrid Generative AI approach be reliably deployed for scalable, automated PHI redaction in unstructured clinical images while maintaining regulatory compliance and data utility?

METHODS

A high-level GenAI workflow was implemented as a robust framework suitable for AI/ML use cases.



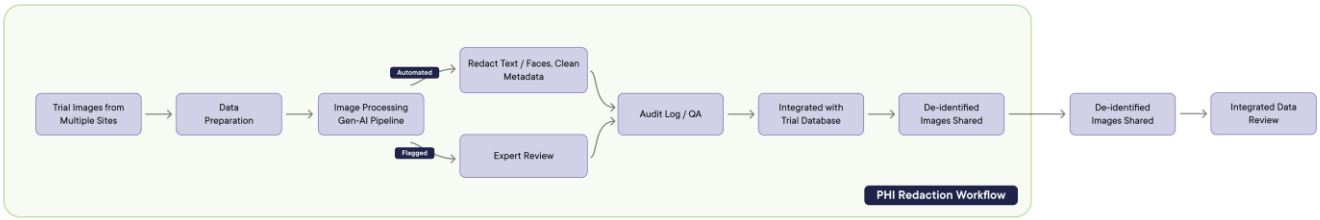
Data Preparation
- Harmonize diverse image inputs (fundus, OCT, external) through a unified pipeline.
- Convert images to DICOM where appropriate for standardization and consistent metadata handling.

Automated Detection and Redaction
- Apply AI-driven identification of PHI markers across modalities (including handwritten or partially obscured text), while avoiding clinically relevant features.
- Redact PHI in both pixel space (text overlays/burned-in annotations) and DICOM metadata (headers/tags).
- Build the model using HIPAA Safe Harbor rules as the basis for PHI categories.

Human Review, Final QA, and Auditability
- Trigger human review for ambiguous outputs and quality verification.
- Maintain audit logs / QA outputs to ensure traceability and compliance.
- Institutionalize expert review and ongoing performance audits to capture rare edge cases and refine the model over time.

Enabling Downstream ML and Secondary Use
Post-redaction, a machine-learning pipeline can extract additional image-based features from clinical trial archives to support secondary analyses, biomarker discovery, and hypothesis validation.

Use Case:
For this particular trial 12,000 images (fundus, OCT, external) were analyzed. The average time taken for manual PHI redaction
from ophthalmology images is 3 minutes. The AI model was built using HIPAA safe harbor rules for PHI redaction. A ML-based
pipeline was built for segmented image outputs. The below visual diagram shows the process workflow for PHI Redaction and downstream analysis.



RESULTS

Key Deliverables

- Unified ETL pipeline for diverse image harmonization
- DICOM conversion for standardization
- PHI redaction from image pixels and DICOM metadata
- AI-based de-identification for efficiency and accuracy
- AI-driven quality checks for compliance
- ML pipeline to extract additional image-based features from clinical trial archives

Benefits

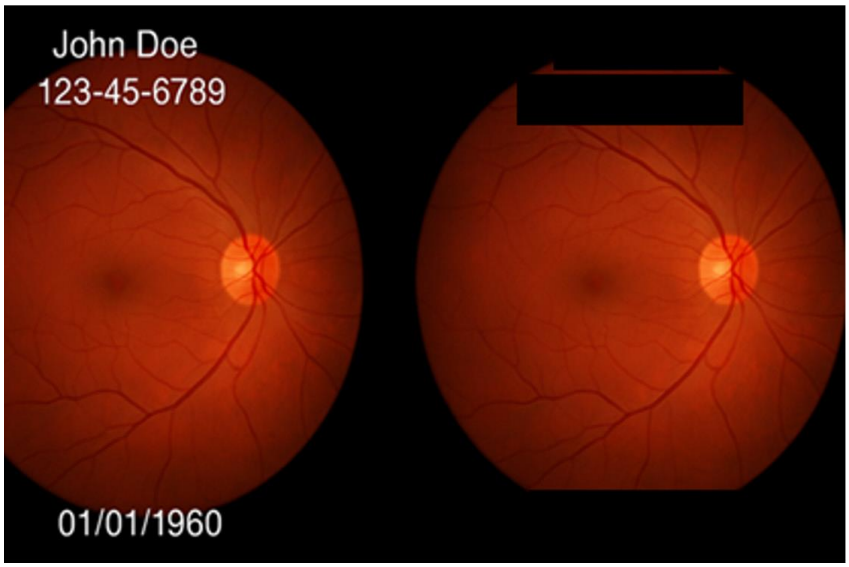
- Approximately 50% faster PHI redaction after image reception from clinical sites
- Empowered clinical analysis with image data and extracted features
- Enabled re-examination of archived images to identify new biomarkers, validate hypotheses, and maximize the value of clinical image data

Redaction Performance

Images processed: approximately 12,000
Number of images with missed PHI: 23
PHI redaction specificity: 98.79
Human review required: 3%
Average time per image (AI plus QA): 15 seconds
Manual redaction time saved: approximately 100 hours

Before and After Redaction (Visuals)

A generic example illustration of before and after PHI redaction in ophthalmology images. On the left, PHI (like patient name and ID) appears over the fundus image; on the right, that information has been redacted or blurred.



DISCUSSION

Special Considerations for Ophthalmology

- Preserving diagnostic regions: systematic testing ensured redaction never altered the retina or optic nerve.
- Varied image formats: automated, lossless conversion to DICOM plus metadata mapping/validation and PACS integration.
- Noisy borders / poor OCR / de-ID complexity: preprocessing, denoising, ensemble OCR, deep learning-based text detection, manual review triggers, and border/boundary data augmentation.
- Transparent dataset documentation: shared datasets were accompanied by summaries of de-identification processes and residual risks.

Risk Analysis and Continuous Improvement

Regular performance audits and expert reviews were institutionalized to monitor AI accuracy, detect rare failure modes, and support iterative refinement. Proactive risk analysis and adaptable processes help maintain effectiveness as data sources evolve and as regulatory expectations change.

CONCLUSION

Regulatory Considerations

Regulatory alignment can be strengthened by incorporating established frameworks such as the FDA Predetermined Change Control Plan (PCCP) and Good Machine Learning Practices (GMLP). The approach emphasizes data quality management, versioning, and validation protocols; strategies for missing data and completeness controls; traceability via audit trails and data lineage; and structured risk assessments using diverse datasets and explainable AI approaches.

A hybrid GenAI plus human-in-the-loop approach can provide a practical path from chaos to clarity for unstructured clinical imaging data by enabling scalable PHI redaction while maintaining compliance and preserving analytic value. In the ophthalmology case study, the approach supported high redaction specificity, reduced per-image processing time, and limited human review to a small subset while enabling downstream ML feature extraction and secondary analysis on de-identified archives.

As AI/ML methods continue to mature, combining explainable AI, ongoing validation, human oversight, and regulatory-aligned lifecycle controls can help unlock the full potential of unstructured clinical data without compromising trust, privacy, or data utility.

REFERENCES

1. HHS - Guidance on De-identification of Protected Health Information: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
2. FDA - Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: https://www.regulations.gov/document/FDA-2022-D-2628-0037
3. FDA - Good Machine Learning Practice for Medical Device Development: Guiding Principles: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles
4. ChatGPT (used for refining/cleaning text in the original poster workflow): https://chatgpt.com/

ACKNOWLEDGMENTS

The authors acknowledge the broader study and operations teams involved in implementing the PHI-redaction workflow and supporting review, QA, and compliance activities.

RECOMMENDED READING

- HIPAA De-identification guidance (HHS)
- FDA guidance on PCCP for AI-enabled software functions
- FDA GMLP guiding principles

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Uma Adiseshan
Company: Zifo RnD Solutions
Email: uma.a@zifornd.com