Digitization, standardisation, and harmonisation of multi-facet and multimodal Radiopharmaceutical Therapy (RPT) and IVF documents with multimodal language models, OMOP CDM, and contextual modeling data schema

Zonunmawia[1,2] Anurag Moyde[2,3] Sagarika[2,4] Gautam Ahuja[1,2,5] Siddhant Poudyal[1,2] Bableen Kaur[1,2] Vinita Sharma[2,4] Himani Balutia[1,2] Navya Tripuri[2,3] Juhi Bharti[6] Gayatri Swetha Gunupuru[6] Reeta Mahey[6] Neeta Singh[6] Mohan Shashikant Kamath[7] Neena Malhotra[6] Parveen Kundu[8] Rajat Vashistha[9] Rintu Kutum[1,2,3,5,10]

[1]Koita Centre for Digital Health at Ashoka, Ashoka University [2]Augmented Health Systems Laboratory, Ashoka University [3]Department of Computer Science, Ashoka University [4]Department of Biology, Ashoka University [5]Mphasis AI & Applied Tech Lab at Ashoka, Ashoka University [6]Department of Obstetrics and Gynaecology, AIIMS New Delhi [7]Department of Reproductive Medicine & Surgery, CMC Vellore [8]Rohtak Nuclear Medcare Imaging Therapy and Research Center [9]NucTheraIn Pvt. Ltd. [10]Trivedi School of Biosciences, Ashoka University

Correspondence: rintu.kutum@ashoka.edu.in, rintu.kutum@augmented-health-systems.org

The project is funded through Trivedi School of Biosciences, Ashoka University, Mphasis AI & Applied Tech Lab, and KCDHA. **RPT:** Principle investigator: **Dr. Rintu Kutum** (Ashoka University). Co-PI: **Dr. Parveen Kundu** (RNM, Rohtak) and **Dr. Rajat Vashistha** (NucTheraIn Pvt. Ltd); **IVF:** Principle investigator: **Dr. (Prof.) Neena Malhotra** (AIIMS New Delhi). Co-PI: **Dr. (Prof.) Mohan S. Kamath** (CMC Vellore), and **Dr. Rintu Kutum** (Ashoka University). Authors would like to acknowledge the high performance computing infrastructure support and student internship opportunities from Mphasis AI & Applied Tech Lab at Ashoka, KCDHA, Trivedi School of Biosciences, and Department of Computer Science (capstone project support), and Ashoka University.

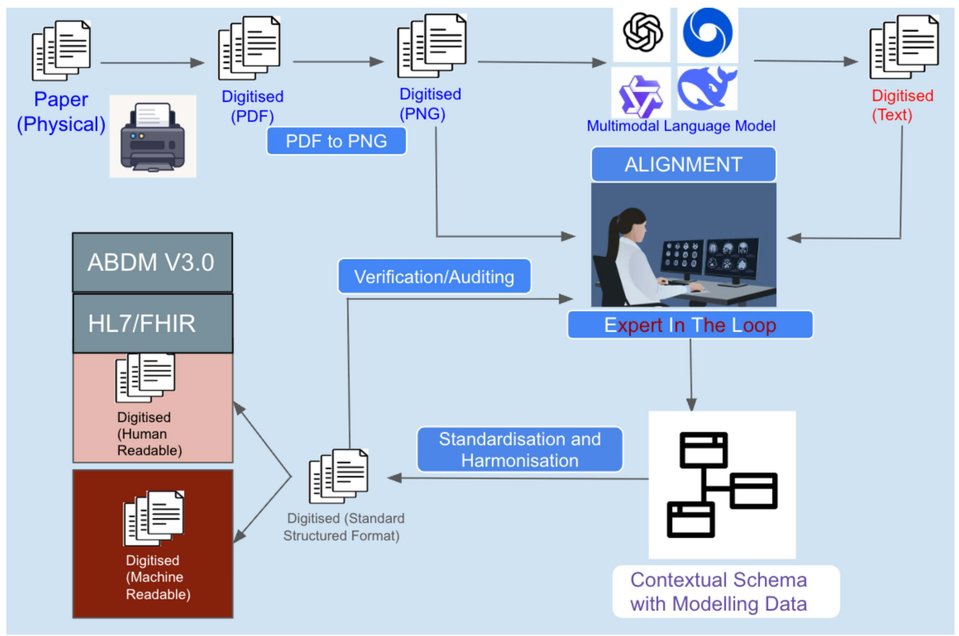
Background

Clinical documents are the primary means of collecting and storing patient-centered medical information for diagnosis, prognosis, and disease analysis. However, documents produced across different modalities, healthcare centres, and countries create a highly heterogeneous data ecosystem, varying in format, structure, layout, and the presence of hard-to-capture handwritten text. This variability is further amplified by the lack of standardized curation practices, including inconsistent medical terminology.
This creates a strong requirement for a robust framework that can transform physical or scanned clinical documents into standardized, schema-conformant digital records to enable efficient modelling and analysis for various downstream tasks.

Methods

We developed a **multi-stage, AI-assisted Human-in-the-Loop (HITL)** system that captures the workflow of converting unstructured clinical documents into structured, validated records.

1. **Document Upload and Layout-Aware OCR**
   Clinical documents (PDF or image) are uploaded and processed using a vision-based OCR model that preserves spatial layout, enabling extraction from mixed printed, typed, and handwritten regions.
2. **AI-Driven Pre-Annotation Using Lightweight Language Models**
   Extracted text is passed through a small language model using prompt-engineered templates to produce initial entity labels and terminology mappings, reducing manual workload.
3. **Expert Validation Through an Interactive HITL Dashboard**
   Domain experts inspect and refine AI-generated annotations, correct bounding boxes, and validate contextual meaning using a visual interface.
4. **Schema Validation via LinkML**
   Validated annotations are checked against LinkML schemas to ensure structural correctness and terminology bindings; inconsistencies are flagged for review.
5. **Interoperable Structured Output**
   Schema-conformant records are exported in formats compatible with national and international standards:
   - Ayushman Bharat Digital Mission (ABDM)
   - HL7 FHIR (Fast Healthcare Interoperability Resources)



Results

A. Limitations of Current Vision Models

- Poor recognition of handwritten text
- Inconsistent or inaccurate layout extraction
- Fragmented outputs that hinder reliable schema mapping

B. Why Human-in-the-Loop is Required

HITL ensures:

- Correction of OCR errors and mis-labeled terms
- Validation of contextual meaning
- Alignment with clinical terminology standards

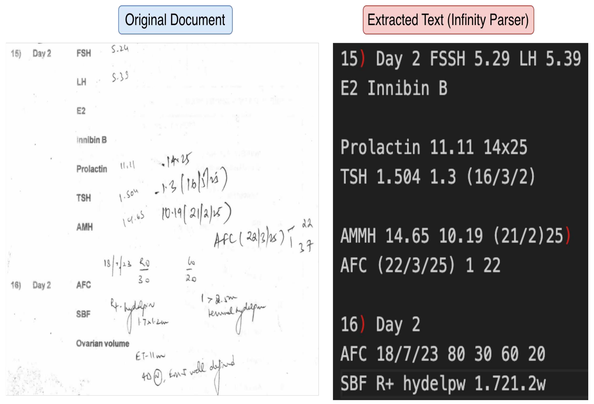


Figure: IVF Document to Extracted Text

D. Standardized Terminology Mapping

Validated annotations map to standardized medical concepts, improving harmonization and enabling downstream analytics.

Investigation-

Exacted *PET-CT PSMA done on 22-08-19-suggestive of tumor in prostate with involvement of B/L seminal vessels and metastasis to multiple pelvic lymph nodes and multiple skeletal sites.*

Corrected *PET-CT PSMA done on 22-08-19-suggestive of tumor in prostate with involvement of B/L seminal vessels and metastasis to multiple pelvic lymph nodes and multiple skeletal sites.*

Annotated <PET-CT PSMA><PET CT of whole body using 18F PSMA (prostate specific membrane antigen) / Procedure / Procedure / SNOMED / 37310747> done on 22-08-19-suggestive of <tumor><Tumor / Observation / LOINC System / LOINC / 1033700> in <prostate><Prostate / Spec Anatomic Site / Body Structure / SNOMED / 4165732> with involvement of <B/L seminal vessels><Both seminal vesicles / Spec Anatomic Site / Body Structure / SNOMED / 4280949> and <metastasis><Metastasis / Measurement / Metastasis / Cancer Modifier / 36769180> to <multiple><Multiple / Meas Value / Qualifier Value / SNOMED / 4112104> <pelvic lymph nodes><Pelvic lymph nodes / Meas Value / NAACCR Value / NAACCR / 40198614> and <multiple><Multiple / Meas Value / Qualifier Value / SNOMED / 4112104> <skeletal sites><Metastasis to bone / Condition / Disorder / SNOMED / 40320638>.

Figure: Standardized annotated terms

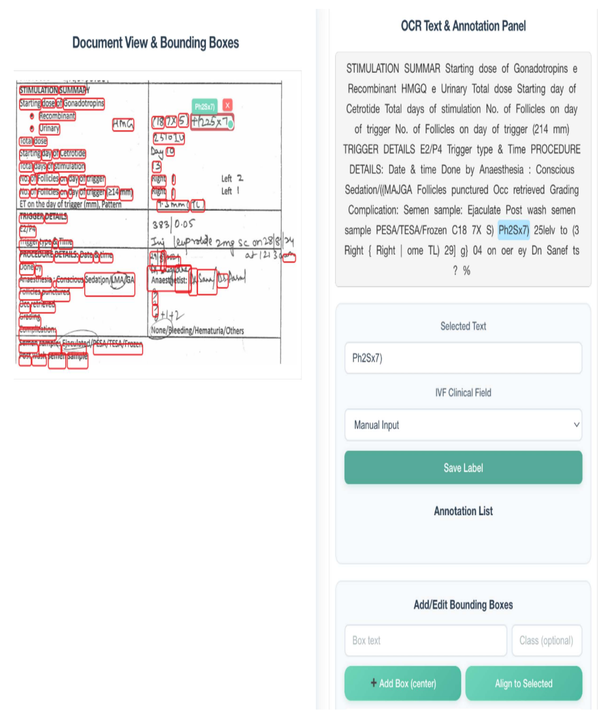


Figure: Human-in-the-Loop System

C. Features

- Upload clinical documents in PDF or image format
- Layout-aware OCR text extraction
- AI-assisted terminology pre-annotation
- Expert review correction of bounding boxes and labels
- LinkML-driven schema conformance
- Export to ABDM, FHIR, or analytics-ready formats

Conclusions

The current work showcase the necessity for the development of robust framework for end-to-end transformation of clinical documents to standardised interoperable data through AI-integrated Human in the Loop System conforming to globally consistent schema to enable open, collaborative, effective, and consistent transfer, analysis and modelling of healthcare data.