Breakout Session 6: Track A

Enabling AI/ML-Readiness of Data from Dual-Energy X-ray Absorptiometry (DXA) Images via Optical Character Recognition (OCR) and Deep Learning

Dr. Evelyn Hsieh
Associate Professor of Medicine/Chief of Rheumatology, Yale School of Medicine/VA Connecticut Healthcare System

Mr. Dax Westerman
Senior Data Scientist, Vanderbilt University Medical School

Enabling AI/ML-Readiness of Data from Dual-Energy X-ray Absorptiometry Images via Optical Character Recognition and Deep Learning

BREAKOUT SESSION 6

Dax Westerman, MS, Sr. Data Scientist, Vanderbilt University Medical Center

Evelyn Hsieh, MD, PhD, Associate Professor, Yale School of Medicine (MPI)

Julie Womack, CNM, APRN, PhD, Associate Professor, Yale School of Nursing (MPI)

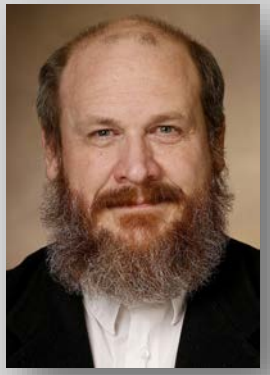
March 28, 2024

Study Team

Evelyn Hsieh
MPI

Julie Womack
MPI

Ruth Reeves
Associate Professor
Bioinformatics



Dax Westerman
Sr. Data
Scientist

NLP Support Service Core

Cynthia Brandt
Professor
Bioinformatics

Farah Kidwai-Khan
Sr. Data Scientist

Paul DiDomenico
Radiologist

Background

- Osteoporosis leads to >2 million fractures each year in the United States.
- A key limitation of artificial intelligence/machine learning research focused on osteoporosis has been the difficulty extracting bone mineral density data from dual-energy x-ray absorptiometry (DXA) images.
- Current methods rely on machine learning algorithms that draw information regarding DXA results from text summaries in the patient medical record.
- However, these text summaries can be inconsistent across clinical sites and are subject to transcription errors.

Motivation for the Study

- We proposed a proof-of-concept study to convert DXA images or scanned reports from the VA Connecticut Healthcare System to text documents, which will subsequently be compatible with the natural language processing (NLP)/machine learning pipelines that our team has developed for extracting T-score and bone mineral density data.
- The project leverages a partnership between Yale School of Medicine, Yale School of Nursing, VA Connecticut Healthcare System, Tennessee Valley Healthcare System VA, and Vanderbilt University Medical Center.

Information Extraction from Scanned Documents

- Clinical notes within the VA EHR usually involve narrative text (e.g., TIU notes), which permits direct application of Information Extraction via NLP techniques.
- Imaged documents require an approach like optical character recognition (OCR) to extract text for downstream processing by NLP. However, imaged documents with non-narrative structure require additional post-processing to interpret.

DXA scans (imaged) versus clinical notes (unstructured text)

DXA Scans	Clinical Notes
PDF or image format	Computable text document
Non-narrative topology	Top-down narrative workflow
Concepts and numerical strongly reply on position relative to anchor concepts (e.g., values in a specific table column) or position in document	Concept meaning derives from interpreting in context of narrative text
Range of pixel color values requiring non-generalizable preprocessing techniques before leveraging OCR	Directly computable from source

Introduction to Proposed Solution

- To capture and classify concepts from DXA images, we have developed the following approach:
  - For tabular data, we employ a method which leverages a Detection Transformer (DETR) model and Table Transformer (TATR) model to identify and describe table position and structure.
  - Object Character Recognition (OCR) is then employed to extract the text (narrative and tabular).
  - Extracted text is then fed a natural language processing (NLP) system.
- Specifically, we will deploy **M**edical **I**nformation **R**etrieval **R**epresenting **O**ptically **R**ecognized **EHR**s (MIRROR EHR):
  - An OCR-to-NLP pipeline for imaged clinical documentation providing text recognition and normalization, classification of project-specific concepts, and relationship assignment among recognized concepts.
  - Adapt MIRROR EHR to DXA imaged reports for capturing bone mineral density and T-score/Z-score information.

Overall Project Methods

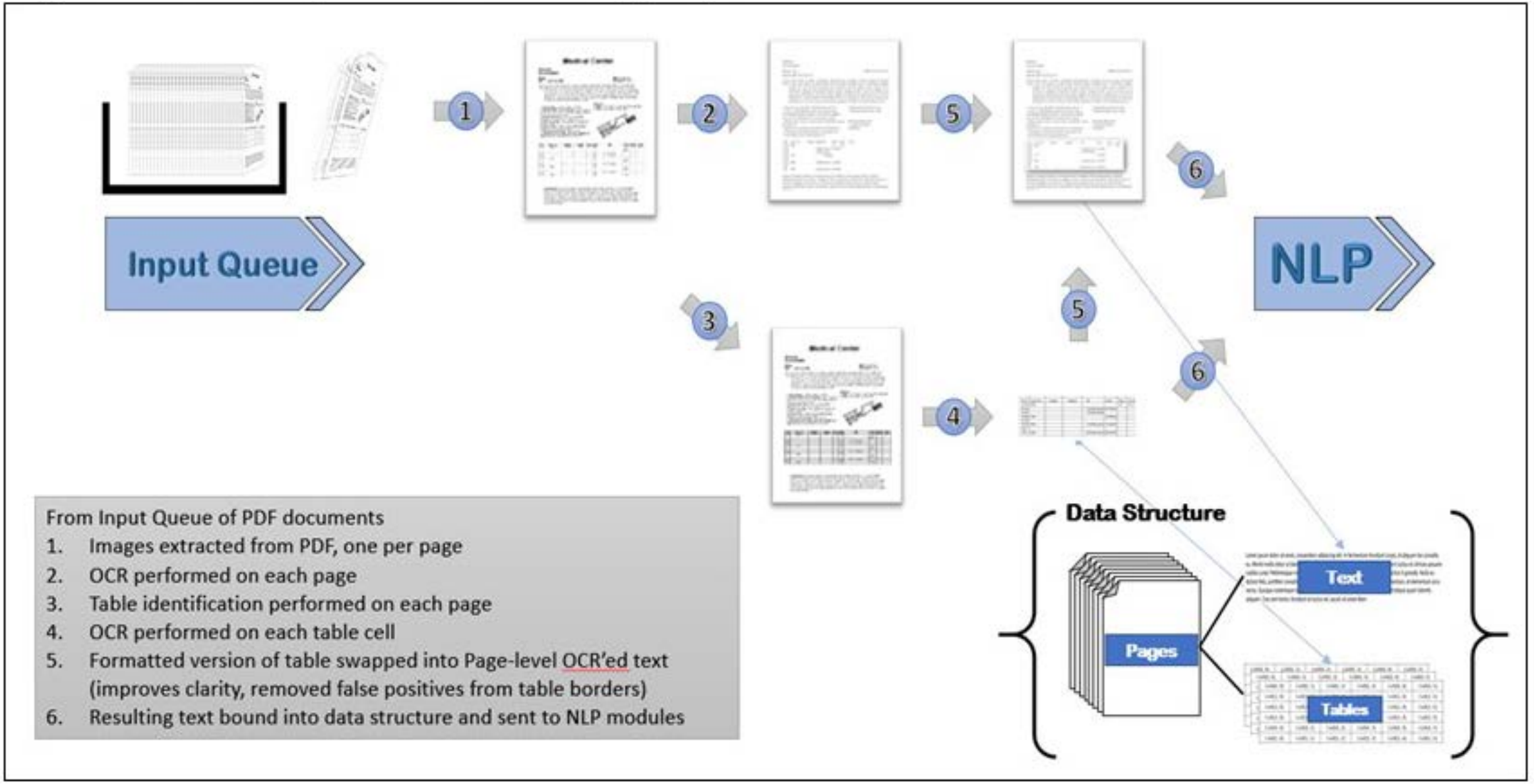
- Recalibrate MIRROR EHR to process DXA reports combining content expertise and end-user input.
  - Sample among 50 DXA reports generated by the two VACHS DXA machines as well as PDFs from external DXA machines to assess the ability of MIRROR's existing OCR module to navigate the document geography of this report type, recognize tables, interpret text within cells, and operate over non-tabular data.
- Evaluate the modified MIRROR EHR version in a new corpus of DXA reports.
  - Using the annotation tool (Prodigy), label the scanned DXA document dataset (130 dev, 70 test). Two trained annotators will identify target concepts (e.g., bone mineral density, T-scores, Z-scores) and DXA-relevant topological identifiers (e.g., table types).
  - Comparison Method: Label the imaged document for tokens that are incorrectly rendered in the text output of the OCR module and provide the correct text string. Use the string distance metrics to calculate the percentage of exact token matches, non-exact token matches and derive equivalence scores, where equivalency is defined as the selected string distance equal to the iteration number.

Image-Specific Information Extraction Methods

- The concepts of interest exist within table structures within the imaged text documents.
- Extraction of table data from DXA scans requires an enhancement to the OCR → NLP pipeline to include detection and interpretation of table structures.
- Given the variability in position, dimensionality, and format among DXA images, we elected to employ object detection[1] and structure recognition[2] transformer models to extract data values.[3]

1. https://huggingface.co/microsoft/table-transformer-detection
2. https://huggingface.co/microsoft/table-transformer-structure-recognition-v1.1-all
3. Smock B, Pesala R, Abraham R. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. IEEE; 2022.

Previous MIRROR-EHR Workflow

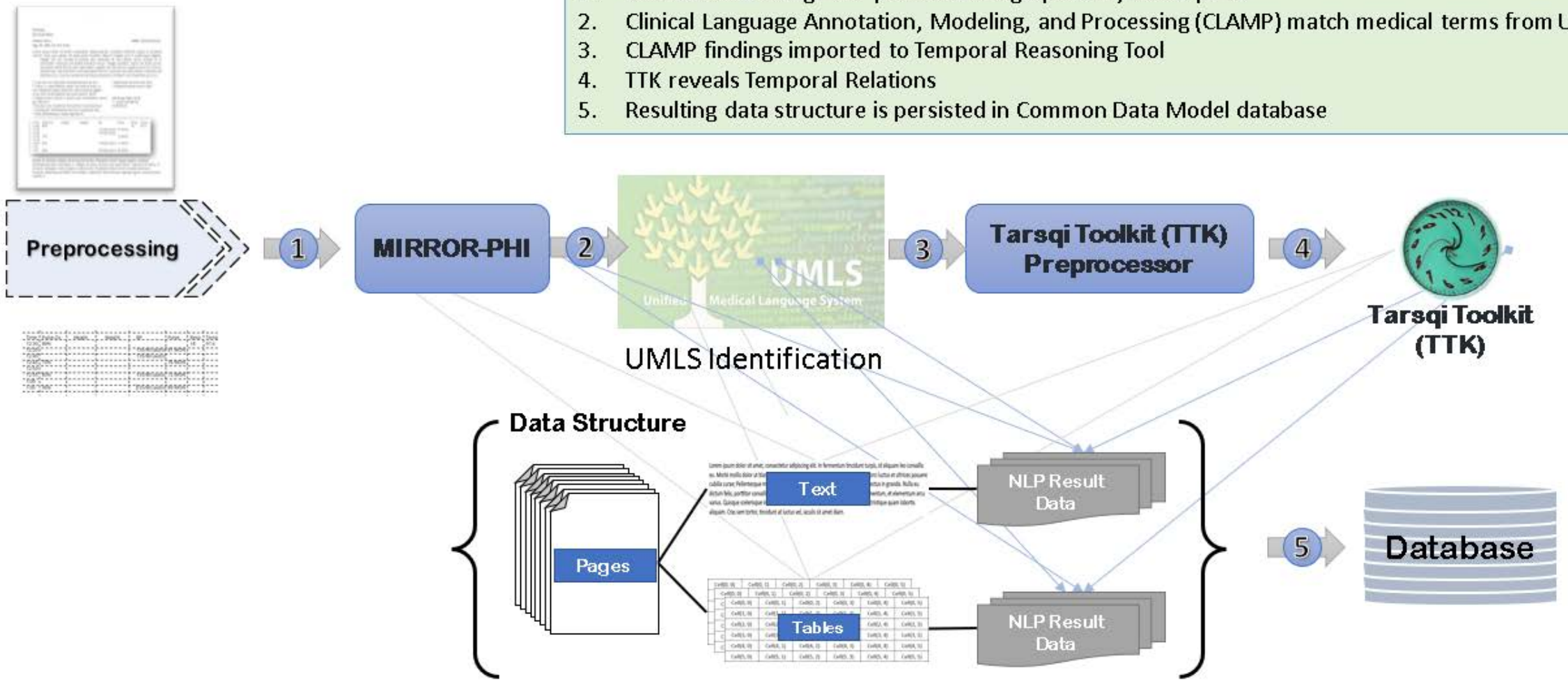


Previous MIRROR-EHR Workflow

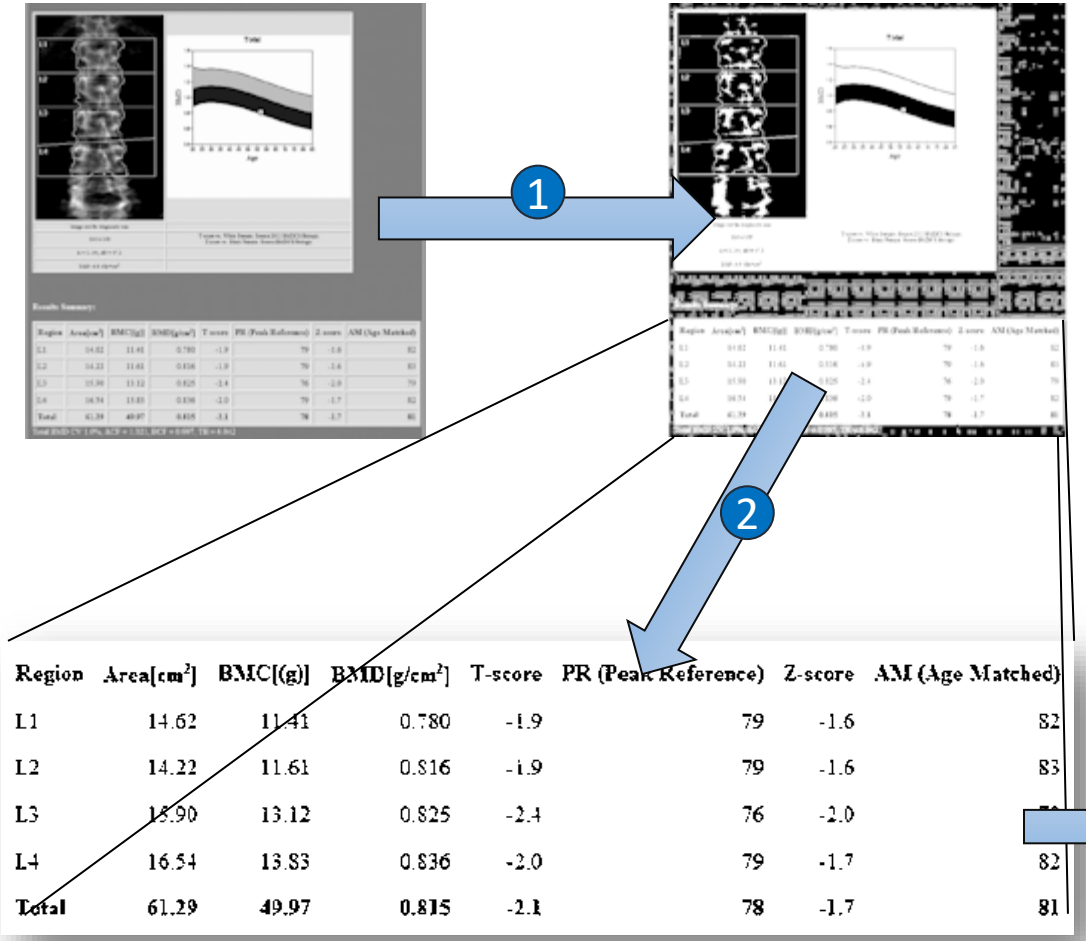
MIRROR EHR NLP Modules

Page and Table text arrives from OCR outputs into preprocessing for NLP

1. MIRROR-PHI recognizes patient demographic Key-Value pairs
2. Clinical Language Annotation, Modeling, and Processing (CLAMP) match medical terms from UMLS
3. CLAMP findings imported to Temporal Reasoning Tool
4. TTK reveals Temporal Relations
5. Resulting data structure is persisted in Common Data Model database



Overview of Updated MIRROR-EHR Workflow for DXA Imaging

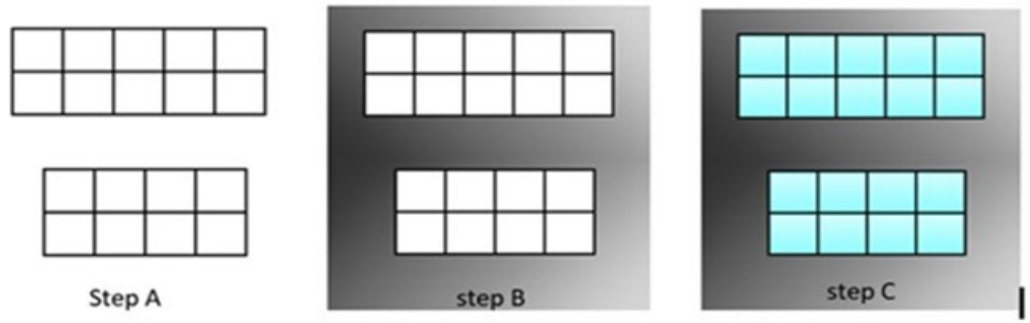


Region	Area[cm²]	BMC[(g)]	BMD[g/cm²]	T-score	PR (Peak Reference)	Z-score	AM (Age Matched)
L1	14.62	11.41	0.780	-1.9	79	-1.6	82
L2	14.22	11.61	0.816	-1.9	79	-1.6	83
L3	15.90	13.12	0.825	-2.4	76	-2.0	79
L4	16.54	13.83	0.836	-2.0	79	-1.7	82
Total	61.29	49.97	0.815	-2.1	78	-1.7	81

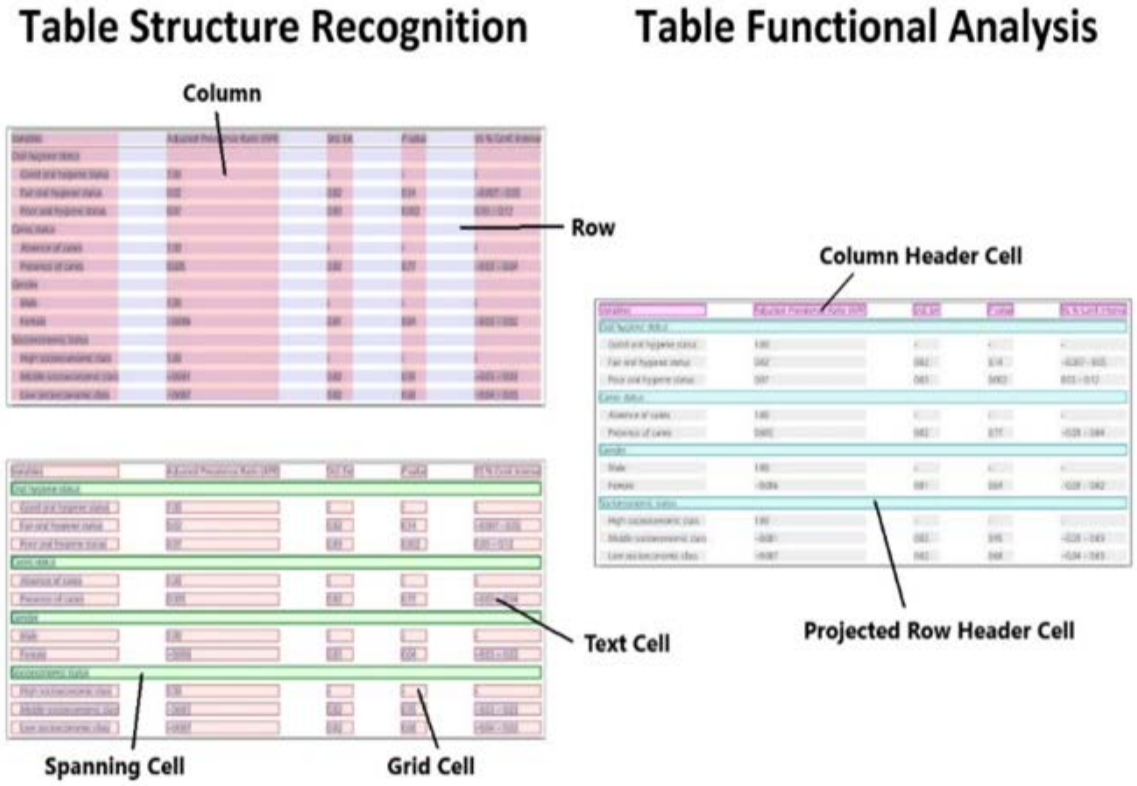

```
[
    {
        0: ['Region',
            'Area [cm2]',
            'BMC[(g)]',
            'BMC[g/cm2]',
            'T Score',
            'PR (Peak Reference)',
            'Z- Score',
            'AM (Age Matched)'],
        1: ['L1', '14.62', '11.41', '0.780', '-1.9', '79', '-1.6', '82'],
        2: ['L2', '14.22', '11.61', '0.816', '1.9', '79', '-1.6', '83'],
        3: ['L3', '15.90', '13.12', '0.825', '-2.4', '76', '-2.0', '79'],
        4: ['L4', '16.54', '13.83', '0.836', '-2.0', '79', '-1.7', '82'],
        5: ['Total', '61.29', '49.97', '0.815', '-2.1', '78', '-1.7', '81']
    }
]
```

1. DXA Image undergoes pre-process steps for OCR
2. Tables identification and recognition
3. Extraction of data for downstream analysis

Achievements



Previous Table Detection Approach



Current Table Detection Approach

- The previous method to for table detection in imaged documents relied heavily on the table borders and their display. The DXA scan format did not lend itself well to this method.
- In reviewing more novel methods, we incorporated a deep-learning approach that was less sensitive to document variability in table definition, resulting in improved OCR interpretation of table values.

Adapted from Smock B, Pesala R, Abraham R. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition

Best Practices

- Recognize the need to separate two distinct data-streams for training and validating an image interpretation system.
  - Structure within a document such as tabular data or sectioning is highly informative to the image interpretation training process
  - Information pertinent to the clinical use-case (e.g. Z-scores) are distinct but dependent on fairly robust identification of document structure
- Subject matter expert (SME) data-labelling labor is best reserved for data-stream Type 2.
- Document structure labelling requires consulting with SMEs in conjunction with image analysis labor.

Lessons Learned

- Extraction of concepts and values from tables in imaged documents often depends on the table structure being evident and well-defined. Tables in DXA images often have thickened borders and color ranges similar to that of the image background color. This can confound table identification methods that depend on the borders to define text bounding boxes for OCR.
- Images containing artifacts (e.g., radiology, colored charts, etc.) with differing color palettes than the base text document can confound typical OCR pre-processing (e.g., binarization, skeletonization, etc.).

Future Plans

- Segmentation and identification of artifacts (e.g., radiology images, numerical charts, tables) in radiology images as separate entities using object detection methods to isolate areas of interest for OCR processing
  - Minimizing confounding effects found in pre-processing steps
  - Leveraging document topology to inform downstream interpretability

- The final tool will directly enhance the quality and impact of our team's future research focused on osteoporosis and fracture outcomes, by allowing high quality identification and extraction of DXA-based measures from the electronic health record.