Design Choices

Data Assumptions

- This solution assumes that datasets range from hundreds of MBs to a few TBs and that core DICOM metadata fields, e.g., Modality and StudyDate, are present in most files.

Indexing Strategy

- Only top-level DICOM metadata fields were extracted, and dataset sequences were excluded, as this is sufficient for typical cohort-level queries.
- Pixel arrays are skipped during reading to reduce memory and I/O overhead, enabling scalability to large datasets. The original file path is saved to allow later retrieval of raw image data if needed.
- Metadata is saved to Parquet files partitioned by Modality and StudyDate. These columns were chosen because clinical queries commonly filter by modality type (e.g., CT, MR) and scan dates. Modality was selected over SOPClassUID because it is more clinically interpretable and less granular.
- Partitioning by Modality and StudyDate ensures that query performance remains efficient even at larger scales, by allowing DuckDB to prune irrelevant partitions during query execution.
- Missing metadata fields are set to None, ensuring incomplete DICOM files do not interrupt indexing.
- Input directories are scanned recursively to index DICOM files organized across nested folder structures.

Querying Strategy

- DuckDB was selected as the SQL engine because it can directly query partitioned Parquet datasets without needing to load them into memory or ingest into a database.
- At query time, the partitioned Parquet files are virtually combined into a single dicom_index table abstraction.

Potential Improvements

- Parallelize DICOM reading and metadata extraction to improve indexing speed at scale.
- Shard Parquet files by row count or size to optimize storage and query performance for very large datasets (hundreds of TBs).
- Add support for indexing DICOM files directly from cloud storage (e.g., AWS S3, Google Cloud Storage).
- Allow users to configure which DICOM fields are extracted through CLI arguments or a configuration file.
- Introduce additional validation or fallback handling for datasets with missing or inconsistent metadata fields.