Requirement Specification Document

# 1. Project Overview

Project Name: Healthcare Claim Liability OCR Extraction Project  
  
Objective: The purpose of this project is to automatically extract claim-related attribute information from unstructured documents such as PDFs stored in ImageRight and from external websites, in order to populate a structured dataset for claim liability modeling and analysis.

# 2. Data Sources

2.1 Internal Documents (ImageRight PDFs):  
- 10 different claims, each having 6 PDF document types.  
- Each PDF document contains 4-5 pages with multiple claim-related information.  
  
2.2 External Links:  
- 40+ external websites provided for additional claim attribute extraction.  
- Attributes to be extracted from these links match the list of required fields.

# 3. Attributes to Extract

- complaint

- plaintiff demand

- were providers involved in care

- mediation notes

- number of emails at a point in time

- specific injury

- plaintiff economic damage alleged

- plaintiff non-economic damage alleged

- plaintiff punitive damage alleged

- venue caps on damage

- plaintiff dependent

- settlement of co-defendant

- are we the last defendant

- plaintiff annual income

- cause of loss

- documentation issues

- missing notes

- alternate records

- number of co-defendant

- plaintiff history of substance abuse or behavioral disorder impacting the treatment

- does plaintiff have a life care plan

- allegation

- misadventure court

- defendant provider

- prior NPDB support reports

- defendant prior claim frequencies

- media coverage of claim

- are we the deep pocket

- high risk procedure

- plaintiff sympathy score

- judge scorecard

- RM Laila limits available

- defendant provider honesty

- birth injury

- permanent impairment

- plaintiff gender

- high equity setting

- close date

- provider’s facility ISO

- coverage limit per incident

- claimant date of birth

- claim description

- provider gender

- injury severity

- insurance coverage available to co-defendant

# 4. Solution Vision

The solution will leverage OCR technologies to extract text from PDF documents, including scanned or image-based files. Extracted information will be processed using NLP and rule-based extraction techniques, mapping the extracted values to the defined attributes. For external links, web scraping combined with text parsing and NLP will be used. The final output will be a structured tabular dataset containing all claims and their respective attributes.

# 5. Data Flow & Architecture Overview

- Upload PDF documents to a central repository (ImageRight or GCS bucket).  
- OCR extraction of text content from PDFs.  
- NLP / regex / rule-based extraction of attribute values.  
- External website scraping and attribute extraction.  
- Aggregation of extracted data into a structured table (one row per claim).  
- Output stored in CSV or database for downstream claim liability modeling.

# 6. Assumptions & Dependencies

- All PDFs follow a consistent format with labeled or structured data.  
- External websites are accessible and allow scraping or API access.  
- OCR engine (e.g., Tesseract, PyMuPDF, or commercial OCR) is available.  
- Required NLP and extraction libraries are installed and available.  
- All attributes are expected to appear in at least one document per claim.

# 7. Future Enhancements

- Enhance extraction accuracy using LLMs or AI-based entity recognition.  
- Integrate directly with claim processing systems for automated ingestion.  
- Add alerting or validation for missing or inconsistent attribute values.  
- Extend external link extraction to additional data sources.

# 8. Sign-Off Criteria

- Confirmation that all required attributes are extracted accurately.  
- Structured output table validated with sample claims.  
- OCR and external link extraction workflow approved by stakeholders.  
- Documentation reviewed and signed off by project manager and client.