# Healthcare Data Research & Collection Plan

## 🔹 Phase 1: Data Source Research & Collection (4-5 Days)

🎯 Objective: Identify, collect, and store diverse healthcare documents in multiple formats.

### Step 1: Define Scope & Requirements (Day 1)

Understand the types of documents needed:  
- Insurance: Policies, Explanation of Benefits (EOBs), Claims  
- Medical Records: Visit summaries, test results, prescriptions  
- Billing & Payments: Medical bills, payment statements  
Identify formats: PDF, scanned images, XML/JSON, text files.  
Ensure data diversity across U.S. payers, providers, and pharmacies.

### Step 2: Research Public Data Sources (Days 2-3)

Search Open Data Repositories:  
- HealthData.gov  
- CMS.gov (Medicare & Medicaid Data)  
- MIMIC-III (ICU records)  
- Kaggle Datasets - Healthcare  
- NIH Open Data  
- AHRQ Healthcare Data  
- Synthetic Health Data (Synthea)  
  
Look for De-identified & Synthetic Datasets:  
- Search for synthetic insurance claims, EOBs, and test results.  
- Identify hospital and pharmacy datasets.  
  
Check Government & Academic Research Papers:  
- Look for sample documents & case studies in research papers.  
- Use Google Scholar, IEEE, ResearchGate.

### Step 3: Data Collection & Storage (Days 4-5)

- Download relevant datasets & sample documents.  
- Convert unstructured data (e.g., images, PDFs) into structured formats.  
- Store everything in Google Drive, categorized by source & type.

## 🔹 Phase 2: Data Categorization & Organization (2-3 Days)

🎯 Objective: Structure the collected datasets for easy access & AI training.

### Step 4: Organize Data into Categories (Day 6)

Create three main folders in Google Drive:  
- Insurance → Policies, EOBs, Claims  
- Medical Records → Visit summaries, test results, prescriptions  
- Billing & Payments → Medical bills, payment statements  
  
Inside each folder, further classify based on:  
- Data format: PDFs, images, text, XML/JSON  
- Data source: Hospital, insurance company, pharmacy

### Step 5: Annotate Metadata (Day 7)

Create a spreadsheet (Google Sheets) to track metadata.  
Columns:  
- Document Type: Insurance, Medical, Billing  
- Source: CMS, Kaggle, NIH, etc.  
- Format: PDF, Image, Text, JSON  
- Availability: Public/Synthetic

## 🔹 Phase 3: AI Readiness & Initial Insights (3-4 Days)

🎯 Objective: Identify AI applications & suggest model training strategies.

### Step 6: AI/ML Use Case Identification (Days 8-9)

Potential AI Applications:  
- OCR (Optical Character Recognition): Extract text from PDFs/images  
- NLP (Natural Language Processing): Analyze medical records  
- Claims Fraud Detection: Identify anomalies in claims  
- Automated Summarization: Generate visit summaries  
- Billing Prediction Models: Estimate healthcare costs  
  
Model Training Approach:  
- Use pretrained NLP models (BERT, GPT) for text data.  
- Use OCR tools (Tesseract, Amazon Textract) for scanned documents.

### Step 7: Summarize Challenges & Gaps (Day 10)

- Identify challenges (e.g., missing data, unstructured formats).  
- Suggest improvements (e.g., more sources, better annotation tools).

## 🔹 Final Submission (Day 11-12)

🎯 Deliverables:  
- Google Drive Repository with structured datasets.  
- Metadata Spreadsheet (document details, sources, formats).  
- AI Readiness Report (use cases, challenges, improvements).  
- 5-minute Presentation (overview of research, findings, and AI potential).

## 🔹 Evaluation Criteria Breakdown

✅ Data Research & Collection (40%) → Quality & diversity of sourced data.  
✅ Organization & Categorization (30%) → Well-structured repository & metadata.  
✅ AI Readiness & Insights (20%) → Practical AI applications & model strategy.  
✅ Documentation & Presentation (10%) → Clear summary & concise reporting.