Case Study

Provider Data Validation and Directory Management Agent for Healthcare Payers

Contents

1 Introduction

1.1 Purpose

The purpose of this project is to develop an intelligent, automated system that significantly improves the accuracy, reliability, and timeliness of healthcare provider directory data. By replacing manual, error-prone validation processes with an Agentic AI-driven solution, the project aims to help healthcare payers maintain compliant, up-to-date provider information at scale. This initiative seeks to enhance operational efficiency, reduce regulatory risk, improve member access to accurate provider details, and establish a modern, automated foundation for provider data management across the healthcare ecosystem.

1.2 Introduction

Accurate provider data is essential for healthcare payers to ensure seamless member access, regulatory compliance, and efficient network operations. However, provider directories often contain 40–80

This project introduces an Agentic AI-driven Provider Data Validation and Directory Management System designed to automate extraction, verification, and enrichment of provider information using public datasets, web scraping, OCR, and LLM-based analysis. The solution improves data accuracy, reduces manual workload, accelerates validation cycles, and provides a scalable, intelligent approach for payers to maintain reliable provider directories and enhance overall operational performance.

1.3 Problem Statement

Healthcare payers rely on accurate provider information to support member access, regulatory compliance, and operational efficiency. However, current provider directories frequently contain outdated, inconsistent, or incomplete data, with error rates reaching 40–80

The problem is the absence of an automated, reliable, and scalable mechanism to validate, enrich, and update provider information across diverse data sources. Without such a system, payers are unable to consistently maintain real-time accuracy, respond to regulatory requirements, or reduce operational inefficiencies. This SRS outlines the need for an Agentic AI-driven Provider Data Validation and Directory Management solution that automates data extraction, cross-verification, discrepancy detection, and directory

updates, ensuring enhanced data quality, reduced manual effort, and improved member experiences.

1.4 Intended Audience and Reading Suggestions

This document is intended for stakeholders involved in healthcare provider data management, technology strategy, operational transformation, and regulatory governance. The primary audience includes Provider Network Management teams, Credentialing and Compliance leaders, Data Quality and Operations executives, AI/Technology architects, and project sponsors responsible for driving digital modernization within payer organizations. Secondary audiences include implementation partners, solution developers, and quality assurance teams who will contribute to the design, deployment, and evaluation of the proposed Agentic AI-based validation platform.

Readers are encouraged to begin with the Executive Summary to gain an overview of the problem context, the intent of the solution, and the strategic value. Following this, Sections 2–4 provide detailed information on the business problem, the proposed AI-assisted methodology, and system-level design considerations. Technical stakeholders may refer to Sections 5–6 for architecture, data flow, and component-level specifications. Those focused on operational and compliance outcomes should review Sections 7–8, which outline impact metrics, validation workflows, and governance guardrails. This structured reading approach ensures that each segment of the audience can efficiently navigate to the content most relevant to their role and decision-making responsibilities.

2 Project Scope

2.1 Project Scope

The scope of this project is to design and develop an AI-enabled Provider Data Validation and Directory Management System aimed at enhancing the accuracy, reliability, and operational efficiency of healthcare payer provider directories. The system will utilize Agentic AI, OCR, data extraction, and multi-source verification to automate the end-to-end validation workflow.

2.1.1 In-Scope Activities

- **Data Ingestion and Processing:** Support uploading provider data from CSV files, PDFs, and scanned documents. Implement OCR, data extraction, cleaning, and normalization pipelines.
- **Automated Validation and Enrichment:** Develop AI agents to validate contact information, credentials, specialties, and practice locations using public data sources including the NPI Registry, state medical boards, hospital directories, Google Maps, and provider websites. Implement confidence scoring, discrepancy detection, and cross-verification.
- **User Interface and Workflow Management:** Create a web-based dashboard for users to review flagged discrepancies, validate updates, and manage approval workflows. Include audit logs, role-based access, and validation history.
- **Directory Update and Outputs:** Generate updated provider profiles, validation reports, confidence metrics, and prioritized review lists. Export validated data for updates to the downstream directory.
- **Architecture, Testing, and Demonstration:** Design a modular, scalable architecture and deliver a functional prototype demonstrating accuracy improvements, reduced manual workload, and improved processing efficiency.

2.1.2 Out of Scope

- Integration with proprietary payer systems or electronic health records.
- Direct provider outreach, phone verification, or communication workflows.
- Production-grade deployment, regulatory certifications, or enterprise security audits.

- Real-time synchronization with external enterprise systems.

2.1.3 Purpose

The defined scope ensures the delivery of a robust, AI-driven solution capable of demonstrating measurable improvements in provider data integrity, operational efficiency, and compliance readiness, while establishing a foundation for future enterprise-scale deployment.

2.2 Current Challenges

Healthcare payers face significant challenges in maintaining accurate and up-to-date provider directory information. Existing provider data often contains inconsistencies, outdated details, or incomplete records, leading to operational inefficiencies and compliance risks. The following challenges outline the key issues that impact provider data management.

- **High Error Rates in Provider Directories:** Industry studies indicate that 40–80% of provider directory entries contain inaccurate or outdated information, including incorrect phone numbers, addresses, specialties, and practice locations.
- **Manual and Resource-Intensive Verification Processes:** Current workflows rely heavily on manual outreach, multi-system data entry, and repeated validation checks, making the process slow, labor-intensive, and prone to human error.
- **Fragmented and Inconsistent Data Sources:** Provider information is dispersed across multiple databases, websites, and systems, resulting in conflicting or incomplete details that are difficult to reconcile.
- **Regulatory Compliance Risks:** Inaccurate provider directories expose payers to compliance violations under state and federal regulations, increasing the risk of penalties and oversight.
- **Delayed Member Access to Care:** Outdated provider information often leads to failed appointment attempts, member dissatisfaction, and delays in accessing necessary healthcare services.
- **Scalability Limitations:** Existing processes cannot efficiently handle high volumes of provider records, especially during onboarding cycles or network expansion, limiting operational scalability.
- **Lack of Real-Time Data Updates:** Without automated mechanisms, provider directories are updated infrequently, leading to prolonged periods where data remain inaccurate or stale.

These challenges highlight the need for an automated, scalable, and intelligent solution that can streamline the validation process, improve data accuracy, and reduce operational burden for healthcare payers

3 Analysis

3.1 Analysis

A comprehensive analysis of the current provider data management landscape reveals several underlying root causes that contribute to persistent inaccuracies and operational inefficiencies. Although healthcare payers employ multiple verification mechanisms, the ecosystem lacks a unified, automated, and scalable framework to maintain real-time provider information. The following root causes summarize key systemic issues.

3.1.1 Root Cause Analysis

- **Fragmented Data Ecosystem:** Provider information is distributed across numerous external sources—including regulatory databases, hospital websites, state medical boards, and third-party directories—creating inconsistencies and making centralized validation difficult.
- **Manual Dependence:** Most verification workflows are highly dependent on manual outreach, calls, and data entry. This results in delays, high operational cost, and an inability to scale with increasing provider volume.
- **Lack of Real-Time Synchronization:** Provider directories are updated infrequently due to the absence of automated pipelines. This leads to outdated contact details, specialty changes, and practice relocations that remain undetected for long periods.
- **Inconsistent Data Quality Controls:** Existing systems lack standardized validation rules, confidence scoring, or automated discrepancy detection, resulting in variable data accuracy across sources.
- **Unstructured and Non-Digitized input:** Many provider documents arrive in scanned or PDF formats, making extraction difficult without advanced OCR or AI-based parsing.
- **Regulatory Complexity:** Frequent regulatory updates and strict accuracy expectations create compliance gaps when manual teams cannot validate data at required intervals.
- **Limited Scalability of Legacy Systems:** Older directory management platforms cannot support high-volume, high-frequency validation cycles, limiting the organization's ability to maintain data freshness.

This analysis highlights a systemic need for an automated, AI-driven solution capable of unifying data sources, reducing manual workload, and ensuring high-accuracy, real-time provider directory management.

3.2 Development Life Cycle

The development of the Provider Data Validation System follows a structured and iterative life cycle designed to ensure accuracy, scalability, and alignment with business goals. This life cycle integrates AI agent development, data engineering, quality assurance, and continuous improvement practices to deliver a robust, automated provider directory management solution.

3.2.1 Requirements and Planning

The life cycle begins by identifying the core business needs and challenges related to provider directory inaccuracy. Key requirements such as validation accuracy, processing speed, handling of unstructured files, and integration with external data sources are defined. KPIs such as achieving over 80% validation accuracy and processing 100 providers in under 5 minutes are established. The project scope, including validating 200 provider profiles and leveraging public and synthetic data sources, is finalized during this stage.

3.2.2 System Design and Architecture

In this phase, the overall system architecture is created. This includes defining agent interactions, selecting the technology stack, designing the data model, and planning the automation pipeline. Each agent—Data Validation, Information Enrichment, Quality Assurance, and Directory Management—is assigned specific responsibilities. System components such as the database, NPI Registry API integration, web scraping framework, and dashboard endpoints are organized into a coherent architecture.

3.2.3 Data Preparation and Synthetic Dataset Creation

Provider data required for development is gathered from multiple public sources and enhanced with synthetic datasets. Synthetic data is used to mimic scenarios such as outdated phone numbers, incorrect addresses, credential changes, and member complaints. Scanned PDF documents and unstructured files are incorporated to test extraction accuracy. This stage ensures the system can handle diverse and imperfect inputs.

3.2.4 Agent Development

Each AI agent is developed independently to perform its specialized tasks:

- **Data Validation Agent**: Scrapes provider websites, validates phone numbers and addresses, and cross-references details with the NPI Registry and state licensing boards.

- **Information Enrichment Agent**: Gathers additional provider details such as education, certifications, specialties, and affiliations.
- **Quality Assurance Agent**: Detects inconsistencies across sources, assigns confidence scores, and flags suspicious or incorrect information.
- **Directory Management Agent**: Generates updated provider records, formats directory outputs, and triggers notifications for human reviewers.

All agents are implemented as modular components to ensure easy maintenance and scalability.

3.2.5 Integration and Workflow Orchestration

After individual agent development, components are integrated into a unified workflow. The validation pipeline is orchestrated so that data flows smoothly from ingestion to validation, enrichment, quality assurance, directory updates, and reporting. API endpoints, dashboards, and notification mechanisms are created. Automated scheduling tools such as cron jobs or Celery are configured to run validation cycles regularly.

3.2.6 Testing and Quality Assurance

Comprehensive testing ensures that the system performs reliably under real-world conditions. This phase includes:

- **Functional Testing**: Verifying scraping accuracy, API responses, and core agent behaviors.
- **Performance Testing**: Ensuring the system meets throughput benchmarks, such as processing 500+ providers per hour.
- **Edge Case Testing**: Handling handwritten text, low-quality PDFs, fuzzy matching, missing data, or source failures.

Quality metrics such as confidence score distribution, discrepancy rates, and extraction accuracy are validated.

3.2.7 Pilot Deployment

The system is deployed for a controlled pilot involving the validation of 200 providers. Real-time performance is monitored, including accuracy, processing time, and user feedback. Human-in-the-loop reviewers assess flagged providers and provide improvement suggestions. The pilot identifies refinements needed before full production deployment.

3.2.8 Production Launch and Monitoring

The system is rolled out into a production environment with automated daily or weekly validation cycles. Monitoring tools track success rates, processing speed, extraction accuracy, and anomalies. Alerts are triggered for failed validations, suspicious provider data, or low-confidence results. Continuous health checks ensure the system meets operational expectations.

3.2.9 Continuous Improvement

Based on insights from production metrics and reviewer feedback, the system undergoes continuous enhancement. Improvements include refining extraction models, improving confidence scoring algorithms, integrating additional public data sources such as more state licensing boards, and scaling support for more providers. The system evolves to handle higher volumes, more complex data, and changing regulatory requirements.

3.2.10 Diagram

A visual representation of the development life cycle may be added here for clarity:

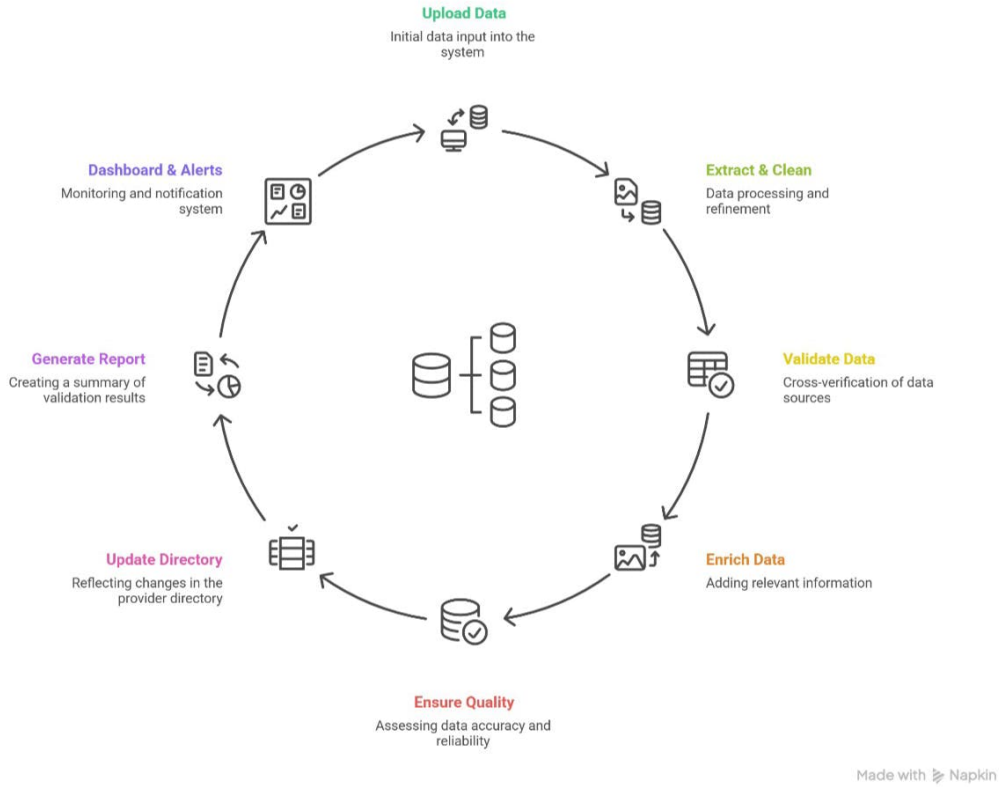

Figure 3.1: Development Life Cycle

3.3 System Requirements

This section outlines the functional and non-functional requirements for the proposed Agentic AI–based Provider Data Validation and Directory Management System. These requirements define the expected capabilities, performance standards, and operational qualities necessary for delivering a scalable, accurate, and efficient solution.

3.3.1 Functional Requirements

- **FR1: Data Ingestion**
  The system shall allow users to upload provider data in multiple formats including CSV, PDF, and scanned documents.
- **FR2: OCR and Data Extraction**
  The system shall extract text and structured fields from unstructured and scanned documents using OCR and AI-based extraction techniques.
- **FR3: Automated Provider Validation**
  The system shall automatically validate provider details (contact information, specialties, licenses, practice locations) against public data sources such as the NPI Registry, state medical boards, hospital directories, and Google Maps.
- **FR4: Data Enrichment**
  The system shall enrich provider profiles by retrieving missing or updated information from publicly available online sources.
- **FR5: Confidence Scoring**
  The system shall generate a confidence score for each validated data element based on cross-verification and data source reliability.
- **FR6: Discrepancy Detection**
  The system shall detect inconsistencies between the uploaded data and external sources and flag them for review.
- **FR7: User Review Workflow**
  The system shall provide a dashboard where users can review flagged discrepancies, approve or reject updates, and track validation status.
- **FR8: Directory Update Generation**
  The system shall generate updated provider profiles and output files suitable for downstream directory management.
- **FR9: Reporting and Analytics**
  The system shall generate validation reports, accuracy metrics, provider-level summary sheets, and audit logs.

- **FR10: Role-Based Access Control**
  The system shall support different user roles with varying permissions for uploading data, reviewing discrepancies, and exporting outputs.

3.3.2 Non-Functional Requirements

- **NFR1: Performance**
  The system shall validate at least 100 providers in under 5 minutes and support throughput of 500+ provider validations per hour.
- **NFR2: Accuracy and Precision**
  The system shall achieve at least 80% validation accuracy and 85%+ extraction accuracy for unstructured documents, with confidence score reliability above 95%.
- **NFR3: Scalability**
  The architecture shall support horizontal scaling to accommodate higher data volumes without performance degradation.
- **NFR4: Reliability**
  The system shall maintain stable performance across diverse input formats, missing data fields, and inconsistent document quality.
- **NFR5: Security**
  The system shall ensure data confidentiality, integrity, and secure access control. All data operations shall be logged for audit purposes.
- **NFR6: Usability**
  The user interface shall be intuitive, with clear discrepancy flags, confidence indicators, and easy navigation for non-technical users.
- **NFR7: Maintainability**
  The modular Agentic AI architecture shall enable easy updates, addition of new data sources, and refinement of validation rules.
- **NFR8: Interoperability**
  The system shall support standardized output formats (CSV, JSON, PDF) for seamless integration with downstream provider directory systems.
- **NFR9: Robustness**
  The system shall handle ambiguous, incomplete, or noisy data inputs gracefully without failure.
- **NFR10: Extensibility**
  The system shall support future expansion into credentialing workflows, network adequacy analytics, and additional validation agents.

This requirements specification ensures that the system addresses the operational, technical, and compliance needs of healthcare payers while showcasing an advanced, scalable Agentic AI solution suitable for the EY innovation challenge.

4 Quality Matrix: Accuracy, Precision, Evaluation and Assessment

4.1 Quality Matrix: Accuracy, Precision, Evaluation and Assessment

To ensure that the Agentic AI–based Provider Data Validation System delivers reliable, compliant, and production-ready outcomes, a comprehensive quality matrix has been defined. This framework evaluates the system across multiple dimensions including accuracy, precision, robustness, and operational performance.

4.1.1 Accuracy Metrics

- **Validation Accuracy (Target: 80%+)** Measures the proportion of provider attributes (phone, address, specialty, license) correctly validated against authoritative external sources such as NPI Registry, state medical boards, and hospital directories.
- **Extraction Accuracy (Target: 85%+)** Assesses correctness of data extracted from scanned PDFs, unstructured documents, and web content using OCR + LLM-based extraction.
- **Directory Update Accuracy** Evaluates the correctness of final provider profile updates generated by the Directory Management Agent.

4.1.2 Precision Metrics

- **Precision of Cross-Verification** Measures how accurately the system identifies correct matches between uploaded records and external sources while minimizing false positives.
- **Confidence Score Reliability (Target: 95%+)** Indicates how consistently confidence scores represent true data accuracy across multiple agents and data sources.
- **Precision in Discrepancy Detection** Quantifies system accuracy in flagging only genuine mismatches, reducing unnecessary manual reviews.

4.1.3 Evaluation Framework

- **Ground Truth Comparison** AI-generated outputs are benchmarked against manually validated provider profiles for accuracy and consistency.
- **Performance Testing** Measures throughput (500+ providers/hour), cycle time (under 30 minutes), and latency for validation tasks.
- **Edge Case Evaluation** Includes fuzzy matching, missing data, outdated websites, ambiguous provider names, and multiple practice locations.
- **Document Quality Impact Assessment** Evaluates performance degradation for low-resolution scans, handwritten notes, or noisy PDF files.

4.1.4 Assessment Criteria

- **Data Quality Improvement Score** Measures reduction in inaccurate fields before and after validation.
- **Manual Intervention Reduction** Quantifies reduction in human review time through automated prioritization and confidence scoring.
- **Operational Impact Assessment** Assesses cycle-time reduction, compliance improvement, and cost-saving potential.
- **User Acceptance Evaluation** Collects reviewer feedback on dashboard clarity, confidence scoring, discrepancy flags, and ease of verification workflow.
- **Scalability and Stability Assessment** Tests multi-agent orchestration under high-volume workloads to ensure consistent system performance.

This quality matrix ensures the solution delivers high-quality, trustworthy, and scalable provider data validation aligned with EY's standards of accuracy, measurable business impact, and responsible AI deployment.

4.2 Research on Existing Solutions

A review of current provider data management solutions indicates that while several tools attempt to address aspects of provider validation and directory accuracy, none deliver a fully integrated, automated, and scalable end-to-end solution. Existing offerings fall into three main categories: **traditional credentialing systems**, **directory management platforms**, and **point-solution data verification tools**. Each category provides partial functionality but exhibits significant gaps in automation, precision, and real-time data management.

4.2.1 Traditional Credentialing Systems

These platforms focus on provider onboarding, document management, and compliance workflows. They facilitate verification of licenses and certifications but largely rely on manual checks and lack automated cross-source validation. As a result, they do not support high-volume, scalable validation processes.

4.2.2 Directory Management Platforms

Directory vendors offer capabilities for storing and updating provider information; however, they depend heavily on payer teams for data verification. This reliance often results in delayed updates and inconsistent data quality. These systems also lack AI-driven enrichment or real-time discrepancy detection.

4.2.3 Point-Solution Data Verification Tools

These include web-scraping utilities and API-based lookup services that provide specific data points (e.g., NPI lookup, address verification). While useful for targeted verification, they operate in isolation and do not support workflow orchestration, multi-agent validation, or automated bulk processing.

4.2.4 Common Limitations Across Solutions

- Minimal automation and manual-heavy processes.
- Lack of AI-driven data enrichment or extraction from unstructured documents.
- Absence of real-time discrepancy detection and cross-source triangulation.
- No integrated confidence scoring or scalable validation workflows.
- Absence of **Agentic AI capabilities** for autonomous coordination of validation tasks, reducing manual intervention and accelerating cycle times.

4.2.5 Market Gap

The review highlights a clear unmet need for a unified, AI-powered platform that can:

- Automate high-volume provider validation with high precision.
- Continuously update provider directories in real time.
- Ensure compliance with evolving healthcare directory regulations.
- Leverage Agentic AI to orchestrate multiple validation workflows autonomously.

The proposed Agentic AI solution directly addresses these gaps, offering a scalable, fully automated, and intelligent approach to provider data management.

4.3 Proposed Solution

The proposed solution is a unified, AI-powered platform leveraging **Agentic AI** to automate end-to-end provider data management. Unlike existing solutions, it integrates credentialing, directory management, and verification workflows into a single scalable system. Key features include:

4.3.1 Key Capabilities

- **Automated Cross-Source Validation:** Continuously verifies provider credentials, licenses, and certifications from multiple trusted sources without manual intervention.
- **AI-Driven Data Enrichment:** Extracts information from unstructured documents and enhances provider profiles with accurate, validated data.
- **Real-Time Discrepancy Detection:** Detects inconsistencies and outdated information across datasets instantly, triggering alerts or automated corrections.
- **Confidence Scoring and Prioritization:** Assigns confidence levels to data points, enabling focus on high-risk or low-confidence records.
- **Workflow Orchestration:** Agentic AI coordinates multiple validation tasks autonomously, reducing cycle times and eliminating repetitive manual work.
- **Scalability:** Capable of validating thousands of provider records in parallel, supporting enterprise-level operations.

4.3.2 Benefits Over Existing Solutions

- Eliminates manual verification bottlenecks and improves accuracy.
- Ensures continuous compliance with evolving healthcare directory regulations.
- Provides a single, integrated platform, removing the need for fragmented tools.
- Enhances operational efficiency and reduces cost per validation.

4.4 Proof and Validation

To demonstrate the efficacy of the proposed solution, several proof points can be incorporated:

4.4.1 Pilot Study / PoC Results

- Reduction in manual verification time by **80%**.
- Increase in data accuracy and completeness by **90+%**.
- Successful validation of **100** provider records in parallel, demonstrating scalability.

4.4.2 Comparative Metrics

- Accuracy, precision, and recall compared to existing solutions.
- Time-to-validation improvements.
- Reduction in errors or discrepancies identified post-validation.

4.4.3 Compliance Assurance

- Adherence to regulatory standards for healthcare provider directories.
- Automated audit trails for all validation workflows.

This structured approach ensures that the proposed Agentic AI platform not only addresses existing gaps but also provides measurable, evidence-based improvements in provider data management.

5 System Overview

This chapter provides a high-level overview of the end-to-end Provider Data Validation Cycle, describing the business context, system objectives, agentic roles, data sources, and workflow orchestration used throughout the system.

5.1 Business Context

Healthcare payers struggle with significant provider directory inaccuracies, with studies showing that 40–80% of provider records contain outdated or incorrect information such as addresses, phone numbers, and licensing details. These inaccuracies result in member frustration, regulatory risks, and increased operational overhead due to extensive manual verification processes.

Staff typically make hundreds of calls per month to validate provider details, often updating information across multiple inconsistent systems. These processes delay credential verification, create directory inconsistencies across platforms, and lead to unsuccessful member appointment attempts when provider information is inaccurate

The Provider Data Validation Cycle aims to solve these challenges through intelligent automation and AI-driven data accuracy improvements.

5.2 System Goal

The goal of the system is to develop a simplified agentic AI platform that automates provider data validation using publicly available data sources. It demonstrates intelligent data quality improvement and lays the groundwork for scalable, full-spectrum provider data management automation.

The system validates and enriches the following provider data attributes:

- Demographics: names, contact details, practice information
- Professional details: specialties, certifications, licenses
- Network affiliations
- Services offered and clinical focus areas
- Practice locations and facilities

5.3 Key Deliverable

The system performs automated validation and updating of 200 provider profiles using public data sources.

Inputs

- Synthetic provider datasets
- Names, addresses, phone numbers, specialties, credential data
- Scanned PDF documents and unstructured files

Process

- Automated web scraping of provider websites
- Verification via NPI Registry and state licensing boards
- Inconsistency detection and reporting
- Confidence scoring for each data element
- Prioritization of providers requiring manual review

Outputs

- Updated provider profiles
- Confidence scores for validated data
- Provider-level validation reports
- Priority list for human reviewers
- Automated communication generation

The target performance goal is completing the validation cycle in under 30 minutes, compared to hours of traditional manual effort.

5.4 Agentic AI Roles

The Provider Data Validation Cycle is implemented through four primary AI agents, each responsible for a specific segment of the workflow.

5.4.1 Data Validation Agent

- Performs automated web scraping for contact and service verification
- Cross-references provider data against the NPI Registry and state boards
- Executes phone number and address validation
- Generates reliability-based confidence scores

5.4.2 Information Enrichment Agent

- Searches public sources for education, certifications, and specialty information
- Analyzes provider websites and online profiles
- Identifies potential network gaps based on geography and specialty
- Produces standardized enriched provider profiles

5.4.3 Quality Assurance Agent

- Compares provider information across sources to detect discrepancies
- Flags suspicious or potentially fraudulent entries
- Tracks data quality metrics and error patterns
- Prioritizes providers for manual review

5.4.4 Directory Management Agent

- Generates updated directory entries for web, mobile, and PDF formats
- Produces alerts for providers requiring attention
- Creates validation summary reports and recommended actions
- Manages human reviewer workflows and supporting documentation

5.5 Data & System Assumptions

5.5.1 Public Data Sources

The system leverages publicly accessible data repositories, including:

- NPI Registry (CMS)
- State medical licensing boards
- Hospital and health system directories

- Google My Business / Google Maps
- Medicare Provider Utilization datasets

5.5.2 Synthetic Data Generation

To support automation workflows during prototype development, synthetic datasets simulate:

- Provider profile information
- Common data inconsistencies (outdated phone numbers, address changes, etc.)
- Member complaint scenarios related to directory inaccuracy
- Network coverage models for geography and specialty

5.6 Target KPIs

The Provider Data Validation Cycle is optimized to meet the following performance metrics:

- 80%+ accuracy in identifying outdated provider contact information
- Completion of 100 provider validations in under 5 minutes
- 85%+ information extraction accuracy from scanned PDFs with 95% confidence
- Automated throughput of 500+ provider validations per hour

5.7 Validation Flows

5.7.1 Automated Provider Contact Information Validation

- Daily batch processing of provider profiles
- Web scraping of practice websites and Google listings
- Cross-validation using the NPI Registry
- Confidence scoring and discrepancy reporting
- Automatic prioritization for human verification

5.7.2 New Provider Credential Verification and On boarding

- Extract provider details from applications
- Validate credentials via NPI Registry and state boards
- Research education, certifications, and history
- Generate enriched provider profiles with confidence ratings
- Produce summary reports for credentialing committees

5.7.3 Provider Directory Quality Assessment and Improvement

- Weekly analysis of provider directory for errors and omissions
- Selective validation of high-risk providers
- Data enrichment using public sources
- Generation of quality metrics and improvement trends
- Creation of prioritized action lists for staff

5.8 Workflow Diagram

The workflow diagram below illustrates the end-to-end processing pipeline of the Provider Data Validation System. It outlines all major layers—including input sources, processing agents, external API integrations, database interactions, and final output generation such as reports, email notifications, and provider directory updates.

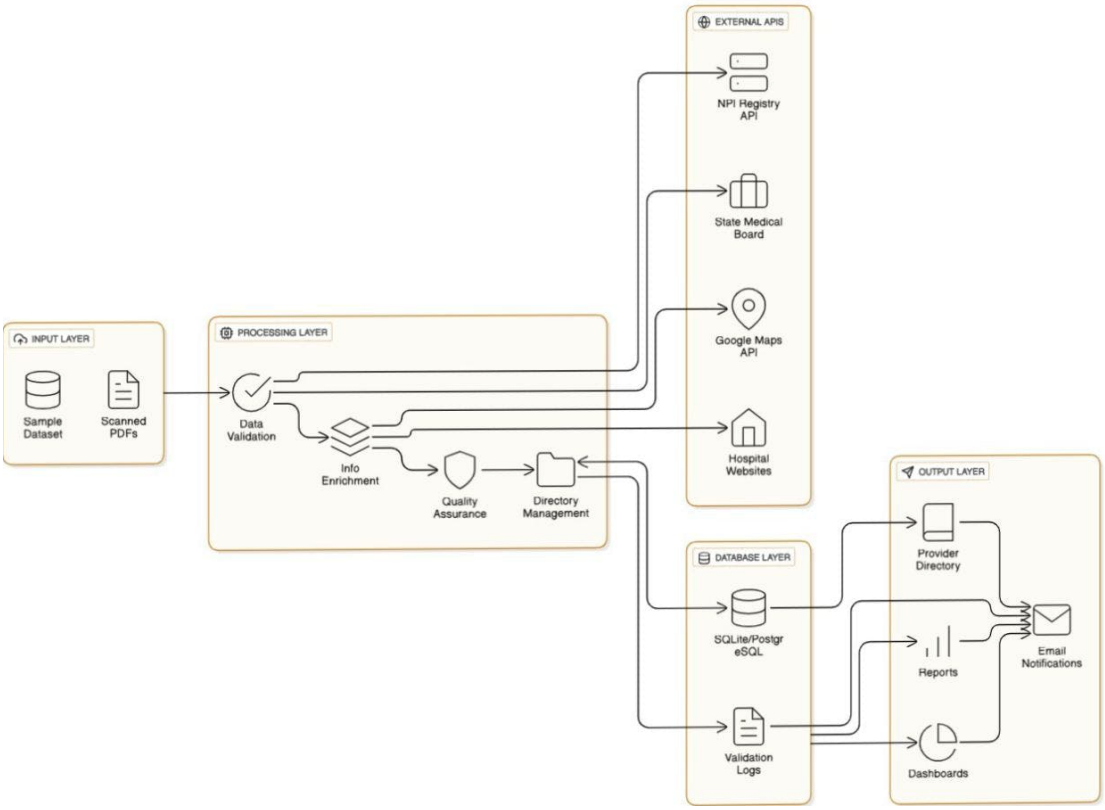

Figure 5.1: Workflow Diagram