Applying the power of retrieval-augmented generation (RAG) in health care

How RAG turns complexity into fast, reliable insights

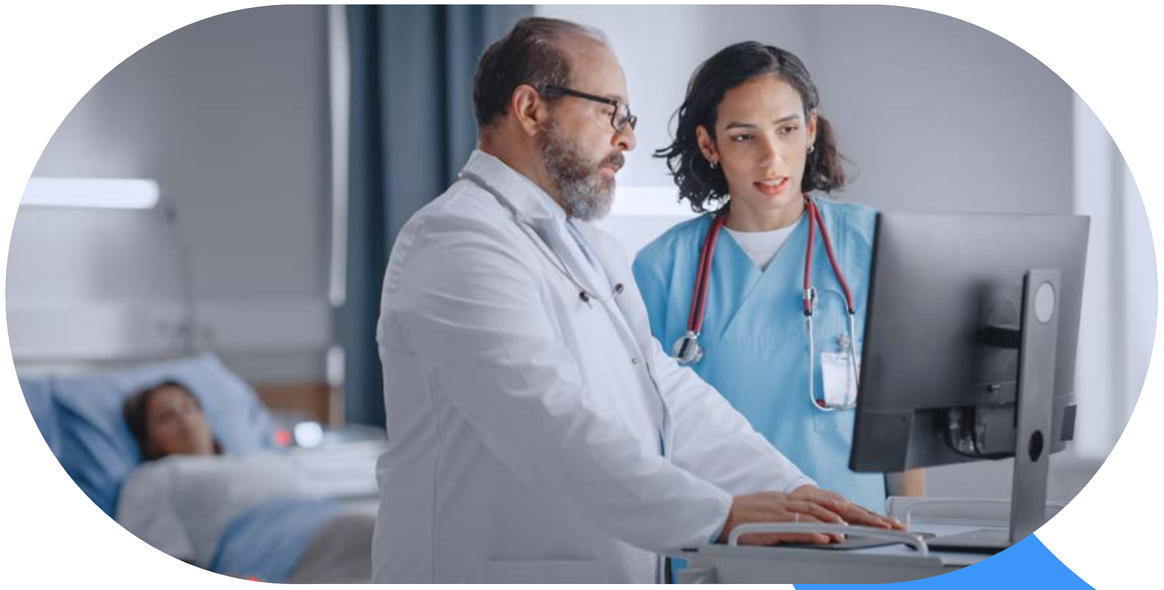
sas

contents

Health care teams spend staggering amounts of time hunting for answers buried in medical histories, billing policies, coverage details and regulatory guidance. That strain slows decisions, affects outcomes and quietly erodes the time of already overextended staff. This has been an accepted part of the workflow for years, but what if a better option is finally within reach? AI can now surface trustworthy, context-relevant information in seconds, offering relief to teams and revealing how critical work can move further, faster and with integrity.

01

Elevating AI from instrumental to indispensable



Worldwide, it is estimated that over 80% – 90% of enterprise data growth is trapped in unstructured formats. Traditional enterprise search is slow, fragmented and restricted by data silos and rigid tools. RAG architecture solves these challenges. We all know that one friend who always seems to have the answer and can be trusted to get it right. Think of RAG as that colleague in the workplace – only faster, more reliable and always ready with the information you need.

What is RAG?

Retrieval-augmented generation (RAG) is a method that combines two AI capabilities – retrieval and generation – to strengthen the quality of AI outputs. Rather than relying solely on pretrained AI models, RAG pairs semantic search with large language models (LLMs) to retrieve relevant information from unstructured data. By drawing from approved, preselected sources, often internal and proprietary, RAG delivers citation-backed, source-grounded responses. The results are:

- More trustworthy, explainable insights.
- More accurate insights.
- Faster delivery.

How RAG creates better outcomes

Health care professionals make high-stakes decisions every day, and they need to trust the information in front of them. General purpose AI tools are less transparent because they are trained using data pulled from many sources over the internet. This data can vary widely in quality, which means it's less reliable and requires more manual verification.

When RAG supports decision workflows alongside technologies such as natural language processing, agentic AI tools and LLMs, users can trust that every response is grounded in approved, reliable sources. When clinicians want added assurance, they can click directly into underlying documents to examine the source data for every finding. This level of transparency replaces hours of manual searching and second-guessing, giving teams more time to focus on patients instead of piecing together information.

From concept to clinic

If you've ever asked an AI chatbot a question, you've seen how powerful these tools can be. However, in health care, where decisions carry critical weight, it's essential to verify details and ensure the information holds up under scrutiny.

With RAG, not only are the answers more trustworthy, but the sources are clearly cited, transparent and easily investigated. As a result, users gain confidence in the ability of RAG-enabled AI systems to address a wide spectrum of clinical areas and their everyday challenges.

Medical records

- Confirm clinical notes support the level of service.
- Identify signatures.
- Flag identical patient records.

Clinical policy bulletins (CPB)

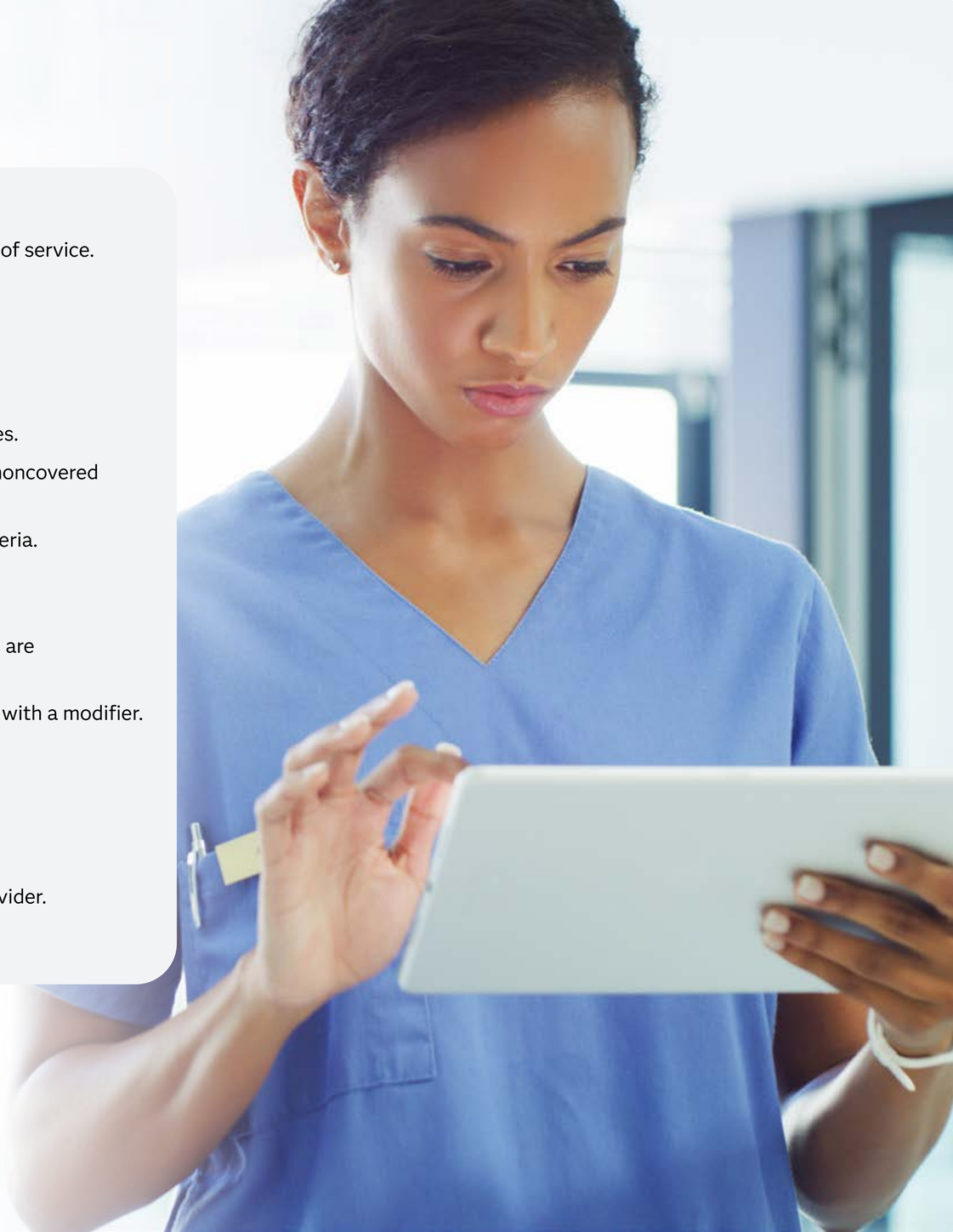
- Determine applicable CPB for procedures.
- Confirm a diagnosis code is covered or noncovered and if pre-authorization is required.
- List procedures covered and clinical criteria.

Claims adjudication

- Identify how often two procedure codes are billed together.
- Include how often procedures are billed with a modifier.
- Detect templated billing.

Investigations

- Track attempts to contact providers.
- Identify previous investigations of a provider.
- Summarize investigations.



02

Real-world scenario: Clinical policy medical review

The challenge

A patient arrives with an osteoporosis diagnosis, and the clinician recommends 10 mg of bisphosphonate daily. But will the patient's health plan cover it? Behind the scenes, entire teams get to work sifting through patient records, clinical coverage data, prior authorization rules, payer contracts and dense CPBs. It's essential work, but it demands time, focus and significant manual effort to find all the answers.

The RAG-enabled approach

Imagine having all those critical answers within a single reach. With RAG and AI agents, users can organize and prepare data from patient records and other sources, then make it accessible through intuitive, natural-language queries. Instead of hours of manual research, a single prompt initiates an AI-driven analysis that draws from predetermined, reliable sources.

With RAG, a prompt such as "Recommended treatment for a patient's osteoporosis diagnosis is 10 mg daily of bisphosphonate. Is this covered by the patient's plan?" instantly sets that analysis in motion, examining predetermined, reliable source data on:

- Patient medical history preceding the osteoporosis diagnosis, including prior treatments, consultations and diagnostic tests.
- Standard recommended treatments for osteoporosis based on patient-specific data.
- Coverage details from the patient's insurance provider's CPB.
- Individual payer contracts.
- National and local coverage determinations (NCDs and LCDs).
- American Osteopathic Association content.

The impact

With RAG and LLMs working together, this entire process unfolds in minutes rather than hours, giving health care teams faster clarity and more time to focus on patient care.

What about data governance?

Whether your organization already has AI guardrails in place or is still shaping them, data governance naturally sits at the center of every conversation. It's also one of the reasons RAG stands out. RAG draws on proprietary, organization-specific data that has been vetted by subject matter experts, rather than relying on external sources that may be inconsistent or unverified. When outside data is needed, RAG incorporates it in a controlled, transparent way. Your data remains inside the RAG-managed setting, and it's never used to train the underlying models.

03

Real-world scenario: Disease outbreak management

The challenge

As reports of infectious diseases stream in from emergency rooms, general practitioners, schools and laboratories, a regional health agency faces the urgent task of turning fragmented data into a cohesive view of disease activity. Outbreaks pose an ongoing threat to human health, whether sparked by emerging pathogens or familiar ones, and timely public health intervention is critical.

The RAG-enabled approach

In a fast-moving outbreak, every minute spent hunting for data is a minute lost in responding to the threat. Instead of manually stitching together reports, spreadsheets, interview notes and historical records, RAG enables a unified view of what's happening across the region. When disease surveillance systems coordinate with a RAG system, possible applications include:

- Combining historical and current data into a near-real-time view of infection spread.
- Converting text interview forms and patient feedback into structured data.
- Merging disparate data sets with ease.
- Establishing baselines and thresholds more efficiently.
- Generating dashboards and reports quickly.
- Anticipating demand to optimize staffing and other medical resources.

The impact

When government health agencies bring RAG into disease surveillance workflows, they unlock the capacity to generate disease models and forecasts that are both more accurate and more timely. For instance, climate factors such as temperature and humidity can influence respiratory virus circulation significantly.

Weaving this information together with syndromic and clinical data strengthens disease forecasts, enhancing – but never replacing – the expertise of epidemiologists and public health leaders. More refined forecasts help health systems optimize medical resources by anticipating increases in emergency department utilization or hospitalization. Every hour saved in detection and response can translate into fewer infections, fewer hospitalizations and more lives protected.

"Health care needs AI with context, not another black box. With proven analytics and guardrails at the core, SAS Retrieval Agent Manager blends LLM reasoning with curated clinical knowledge to deliver outcomes you can trust."

Mark Wolff, Advisory Industry Consultant, Health Care and Life Sciences, SAS

04

Real-world scenario: Claims and payment integrity

The challenge

A claim is flagged as potentially fraudulent, and the claims investigator, already juggling a full caseload, chases down details that are scattered across systems that update on different schedules and operate under different rules.

Anyone monitoring or actively investigating fraud often must pull information from multiple sources – payer policies, billing rules, Centers for Medicare & Medicaid Services (CMS) manuals and more – to ensure they're working with the most current data. The slow, manual process increases the risk of missed information, overlooked inconsistencies and the potential for fraud, waste and abuse.

The RAG-enabled approach

When billing and payment data operate within a RAG-accessible system, investigators can ask for what they need and receive it instantly rather than piecing together information from various platforms. RAG makes it possible to:

- Automatically extract key details from claim submissions.
- Pull receipts, patient records, physician notes and prior investigation files.
- Merge payment and billing data from across the organization.
- Coordinate current payer policies, billing rules and CMS guidance.
- Synchronize payment data routinely, reducing the risk of decisions based on outdated information.

The impact

When RAG supports claims and payment integrity, investigations move faster and with more accuracy. The expertise of analysts and fraud investigating units is enhanced with a full view of each claim's history, making it easier to authenticate claims, protect resources and ensure payments are fair and accurate.

05

SAS Retrieval Agent Manager

Real-world scenarios show how much time, clarity and confidence organizations can gain when AI has access to trustworthy and timely information. Turning that potential into everyday practice requires tools designed to manage complex data environments and deliver intelligent insights at scale.

SAS Retrieval Agent Manager is a no-code platform that delivers context-aware responses from unstructured data using a RAG framework. With SAS Retrieval Agent Manager, health care organizations unlock enterprise knowledge, reduce manual data extraction and overcome the common limitations of standard RAG implementations, which are code-heavy, inflexible and hard to integrate.

KEY FEATURES

Intuitive no-code experience: Empowers nontechnical users to build and manage workflows through a no-code interface, with support for diverse document types, multilingual OCR and usage monitoring.

Trustworthy, transparent AI: Delivers citation-backed, source-grounded responses supported by built-in evaluation tools and human-in-the-loop oversight to ensure transparency, traceability and trust.

Modular, plug-and-play architecture: Supports multiple third-party LLMs and vector databases without vendor lock-in and offers native integration with file systems and Git for seamless enterprise onboarding.

High-performance, enterprise-ready integration: Provides fast, efficient AI through model-acceleration techniques and integrates smoothly into enterprise systems via robust APIs and secure, scalable deployment options.

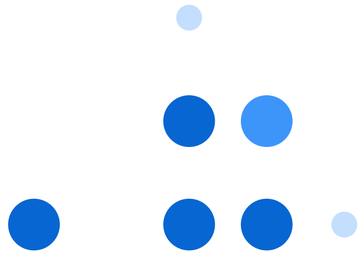
Autonomous agent orchestration: Coordinates agents that retrieve, reason and act across systems, automating complex, high-value agentic AI workflows with precision, integration and parallel execution.

Flexible deployment options: Provides robust support for on-premises deployments, giving organizations full control over their data while enabling easy extension to cloud or hybrid environments.

Model Context Protocol (MCP): Enables agents to go beyond retrieval, orchestrate APIs and automate enterprise processes with schema-driven, auditable and reusable tool calls.

Why SAS

SAS has spent 50 years at the forefront of data and analytics innovation, with a proven track record delivering trustworthy AI-powered services. Our solutions empower health care organizations to make transparent and explainable decisions that improve health outcomes, enhance operational efficiency and optimize resources and costs.



Take your RAG capabilities to the next level with SAS Retrieval Agent Manager. No coding, just insights.

Get started at **sas.com/ram**

To contact your local SAS office, please visit: **sas.com/offices**