



A Buyer's Guide to Evaluating Medical Records AI



A Buyer's Guide to Evaluating Medical Records AI

Artificial intelligence is rapidly changing what organizations can do with medical records. Tasks that once required hours of manual review can increasingly be accelerated through automated summarization, information extraction, chronology development and other forms of AI-assisted analysis. For insurance carriers, law firms, third-party administrators and other organizations that routinely work with complex medical information, the potential is significant.

The challenge for buyers is that AI can be easier to demonstrate than to evaluate.

In a typical product demonstration, the experience can appear straightforward. A set of medical records goes in. A summary, chronology or answer comes out. The output may be organized, readable and delivered almost instantly. Compared with manually reviewing hundreds or thousands of pages, the value can seem self-evident.

But an impressive output answers only one part of the buying question.

It does not necessarily tell the buyer whether the AI had access to the right information, whether important information was omitted, whether chronology and context were preserved, whether a conclusion can be traced to its source, how the system performs on complex real-world record sets, what happens when new records arrive, or how the technology will fit into the organization's existing workflow. It also does not answer increasingly important questions about security, governance and the use of sensitive information.

Those considerations become particularly important when AI is assisting professionals whose decisions may affect claims, litigation, financial exposure or medical evaluation.

The right question, therefore, is not simply: **How good is the AI?**

It is: How well does the entire solution help professionals work with complex medical information accurately, transparently, securely and within the realities of their existing workflow?

That distinction is consistent with the broader direction of AI governance and evaluation. The National Institute of Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework encourages organizations to incorporate trustworthiness considerations throughout the design, development, use and evaluation of AI products, services and systems.¹ The implication for buyers is important. Evaluating AI should extend beyond the sophistication of the underlying model or the quality of a single output. The surrounding system matters too.

For medical records AI,
that requires a different
kind of buyer's framework.

Medical Records Require a Different Evaluation Lens

Medical records are not simply a collection of documents waiting to be summarized. They are longitudinal histories assembled over time through encounters with different providers, facilities and specialties. A single matter may involve physician notes, diagnostic imaging, laboratory results, operative reports, medications, physical therapy records, billing information and administrative material spanning months or years.

The professionals reviewing those records may also be looking for very different things. A claims professional may need to understand treatment associated with an injury and determine whether similar symptoms were documented previously. An attorney may need to identify the evidence supporting a particular chronology before a deposition. An independent medical examiner may need to reconstruct years of treatment before evaluating the condition presented today.

AI has enormous potential to make that work more efficient. But research into AI-assisted medical chart review also demonstrates why efficiency cannot be the only evaluation criterion.

A 2026 study of an AI chart-review tool integrated into an electronic health record collected physician feedback on AI-generated chart summaries. Positive feedback was common, but physicians also identified missing information, confusing content, token limitations and some hallucinations. The researchers concluded that AI-assisted chart review could augment manual review while emphasizing that users may need access to source information to verify AI-generated summaries.²

Other research reinforces why the quality of an AI-generated output cannot be judged solely by how polished it appears. A 2025 study published in JAMA Internal Medicine compared physician- and large language model-generated discharge summaries across 100 hospital encounters. The AI-generated narratives were rated as more concise and coherent than physician-generated narratives, but they were also less comprehensive and contained more errors.⁷ Overall quality and reviewer preference were comparable, and the potential for harm remained low, underscoring both the promise of the technology and the continuing importance of professional review.⁷

The lesson is not that AI should be avoided. It is that the value of AI depends in part on the environment surrounding the output: what information the system receives, how that information is analyzed, how results are presented and how easily a professional can independently examine the underlying evidence.

A medical records AI solution should therefore be evaluated not solely on the answer it produces, but on the process that surrounds the answer.

The following questions provide a framework.

1. What Problem Is the AI Actually Designed to Solve?

“AI-powered” is an increasingly common description, but it tells a buyer surprisingly little about what a product actually does.

One medical records AI solution may summarize individual documents. Another may extract diagnoses, procedures or billing information. Another may create a chronology. Some may allow users to ask questions across uploaded files. Others may attempt to organize information across an entire matter.

Each capability can be valuable. They are not necessarily interchangeable.

The first step in evaluating a solution is therefore to move the conversation away from AI itself and toward the work that needs to be done. Where is professional time being consumed today? Is the primary challenge reading individual records, reconstructing a medical history, understanding treatment over time, locating supporting evidence, analyzing medical expenses, or answering different questions as a matter progresses?

That distinction matters because a solution can perform its intended task extremely well and still fail to address the buyer’s actual problem. An excellent individual-record summary, for example, may reduce the time required to read that document while leaving the reviewer responsible for connecting information across dozens of other records. A powerful search capability may locate every reference to a known diagnosis but provide less help when the reviewer is trying to determine what changed after a particular event.

Buyers should therefore resist beginning with a generic question such as, “What can your AI do?”

A better starting point is: What part of our medical-record workflow is this designed to improve, and what work will our professionals still need to perform themselves?

“AI-powered” describes the technology. It does not tell the buyer what problem has actually been solved.

2. What Information Is the AI Actually Working From?

An AI system can only analyze the information available to it. In medical-record workflows, that makes the quality and composition of the underlying record set part of the evaluation.

Consider a system asked to analyze the medical history associated with an injury. Has it received records from every relevant provider? Are records still outstanding? Does the set contain duplicate pages or administrative material? Was another treating provider identified in a note but never pursued? Has treatment continued since the records were originally assembled?

These are not necessarily failures of AI. They are characteristics of the information environment in which medical-record AI operates.

That creates an important distinction between the apparent completeness of an AI-generated answer and the completeness of the information from which the answer was generated. A concise, coherent analysis can still represent only the portion of the history that was available to the system.

Before asking what the AI knows, ask what the AI has seen.

Buyers should understand how records enter the system and what has to happen before the AI can meaningfully analyze them. Must records first be collected, downloaded and uploaded through another process? Who sorts and classifies them, identifies duplicate pages, reconciles different formats and organizes information arriving from multiple providers? How does the solution handle non-clinical material or records that arrive weeks or months after the initial analysis? And does the system provide visibility into exactly what information was included?

These questions matter because medical records rarely arrive as a clean, static data set. In practice, the work required to prepare and maintain the record set can itself become a significant part of the review process. An AI capability that accelerates analysis but leaves substantial reconstruction work surrounding it may solve only one portion of the buyer's problem.

This is particularly important when evaluating medical records AI as part of a larger workflow rather than as an isolated technology. The process of obtaining, preparing and maintaining the source information can materially affect the usefulness of everything the AI does afterward.

3. Does It Understand Individual Documents or the Case?

Individual medical records are created for individual encounters. Claims and legal matters rarely remain confined to one encounter.

A patient's history may involve multiple physicians, facilities, diagnoses, procedures and episodes of treatment. The significance of an individual record often becomes apparent only when it is considered alongside information documented elsewhere.

That creates a meaningful distinction between document-level intelligence and case-level intelligence.

A document-level tool may accurately explain what happened during a particular office visit. A case-level approach must go further. It must help the reviewer understand how that encounter relates to other events across the available medical history: what happened before it, what followed it, whether similar symptoms appeared elsewhere, which providers were involved and how treatment evolved.

Buyers should therefore ask how the system operates across a large collection of records. Does it simply produce a series of individual summaries that the reviewer must still assemble mentally? Or can it help create a consolidated view of the available medical story?

The answer becomes especially important as record volume increases. Reviewing twenty concise AI-generated summaries may be easier than reviewing twenty original records, but the professional may still be left with the work of determining how those twenty pieces fit together.

Buyers should also ask what exists beneath the output. Does the solution simply generate a narrative from the documents it receives, or does it create a structured foundation of information, such as encounters, providers, diagnoses, procedures, treatment and billing, that can be used for different purposes? That distinction becomes important when the next reviewer has a different question. A chronology, injury analysis, billing review and case summary may all draw from the same medical history, but a workflow built around separate static outputs can require the underlying information to be processed or reconstructed repeatedly.

The more durable approach is one in which the intelligence extracted from the records can support multiple questions and uses over the life of the matter.

The goal should not merely be to make each document easier to read. It should be to reduce the effort required to understand the information across the matter.

4. Does the AI Preserve Chronology and Context?

Medical information derives much of its meaning from relationships.

A diagnosis matters, but so does when it first appeared. A symptom may be significant, but its significance may change if similar complaints were documented years before the event at issue. A treatment decision becomes more informative when the reviewer can see what prompted it and what happened afterward.

This makes chronology more than a formatting convenience.

Buyers should examine whether a medical records AI solution helps users understand information over time. Can it distinguish among events occurring before and after a relevant date? Can it bring together treatment across providers? Can the user understand the sequence of symptoms, diagnoses, procedures and responses to treatment?

Context is equally important. Extracting a fact from a record can make information easier to find while simultaneously separating it from details necessary to interpret it correctly. A diagnosis may have been provisional. A medication may have been discontinued quickly. A symptom may appear once or repeatedly. Two providers may document the same issue differently.

The evaluation question is therefore not simply whether AI can locate information. It is whether the solution helps the professional understand what that information means within the broader medical history.

A useful demonstration should show this explicitly. Rather than asking a vendor to locate a known diagnosis, ask the system to explain how a condition developed over time and then show the records supporting that history.

That is a more meaningful test of medical-record intelligence.

5. Can You Get From the Answer Back to the Evidence?

Generative AI creates a new challenge for professional users: an answer can sound authoritative before it has been independently verified.

That makes traceability one of the most important criteria in evaluating medical records AI.

If a system reports that a patient experienced similar symptoms before an accident, can the reviewer immediately identify the source of that statement? Can the user examine the original note? Can the surrounding context be reviewed? Can the professional determine independently whether the AI's characterization is fair?

Research into AI-assisted chart review reinforces the importance of this connection. In the 2026 study described earlier, researchers noted that physicians may need workflows that enable access to source data so AI-generated summaries can be verified.²

The principle appears elsewhere in professional technology governance as well. Guidance surrounding clinical decision-support software has emphasized the importance of enabling healthcare professionals to independently review the basis for software-generated recommendations.³ Medical records AI used in claims or litigation is a different application and should not be conflated with regulated clinical decision-support software, but the underlying principle is useful: when professionals are expected to exercise independent judgment, transparency into the basis of a technology-generated output matters.

Buyers should therefore go beyond asking whether a product “provides citations.”

Ask the vendor to demonstrate the experience.

Choose an important finding. Follow the citation. Examine the original source. Look at the surrounding information. Determine how quickly a reviewer can move between generated intelligence and underlying evidence.

Ask the vendor to demonstrate the experience.

Comprehension requires more than an answer.
It requires a path back to the evidence.

6. How Is the AI Tested and What Happens When It Gets Something Wrong?

AI vendors frequently describe their technology as accurate. Buyers should ask a second question:

Accurate at what?

Accuracy in extracting a procedure code is different from accuracy in summarizing a physician note. Summarizing a note is different from reconstructing chronology across a multi-provider history. Answering a targeted question is different from determining which information is relevant when the user has not specified exactly what to find.

Performance claims are meaningful only when buyers understand what was measured.

NIST's AI risk-management resources emphasize testing, evaluation, verification and validation as important components of responsible AI deployment.¹ NIST's AI Risk Management Framework similarly emphasizes the importance of measuring and evaluating AI systems in context, including through testing, evaluation, verification and validation.⁴

For buyers, this means asking how the solution has been tested, what kinds of records were included, whether the testing reflects realistic use cases and how different types of errors are measured.

Omissions deserve particular attention. A hallucinated statement may be identifiable when it conflicts with the source. Missing information can be more difficult because the reviewer may not know that something important was absent from the AI-generated output. The 2026 chart-review study identified omissions more frequently than hallucinations and specifically noted the difficulty of detecting missing information.²

The 2025 discharge-summary study points to the same concern from another direction. AI-generated narratives were more concise and coherent than physician-generated summaries,

but they were less comprehensive and contained more inaccuracies and omissions.⁷ That tradeoff is particularly relevant when evaluating medical records AI. An output can be easier to read and appear highly polished while still leaving out information a professional would consider important.

Buyers should therefore ask vendors to define what they mean by accuracy. What tasks were evaluated? What constitutes an error? Are omissions measured separately from incorrect statements? Were results tested on individual records or complex multi-provider cases? Who evaluated the outputs, and against what standard? Testing should reflect the actual work the technology is being purchased to perform, not simply the performance of the underlying model on a generalized benchmark.

Buyers should also ask what happens when the system encounters ambiguity or conflicting information. Does it acknowledge uncertainty? Does it show the competing evidence? Does it produce an answer anyway?

The objective should not be to find an AI system that claims never to make mistakes. The more useful objective is to understand how performance is evaluated, how limitations are managed and how the surrounding workflow helps professionals recognize when further review is warranted.

Confidence should come from evidence about performance, not confidence in the presentation of the output.

7. What Happens When the Case Changes?

Medical-record review is rarely a one-time event.

Treatment continues. Additional records arrive. A previously unknown provider is discovered. An attorney requests another set of documents. A claim moves into litigation. A new expert becomes involved.

An AI solution that works well with a static set of records may behave differently when the underlying information changes.

Buyers should therefore ask what happens after the initial analysis. Can new records be incorporated into the existing case? Can the analysis or summary be refreshed? Does the user have to upload and reconstruct everything again? Can the system maintain an evolving view of the available medical history?

This issue also affects efficiency in ways that may not be obvious during an initial demonstration. Medical records frequently arrive in multiple deliveries over weeks or months. If every new batch requires a separate workflow, another upload or another disconnected output, the reviewer may repeatedly have to reintegrate new information into the medical story. Some of the reconstruction AI was intended to eliminate simply reappears later in the process.

Buyers should therefore distinguish between technology that analyzes a record set once and technology designed to maintain an evolving understanding of the matter as new information becomes available.

The relevant unit of work is not always a document or even a record set. In many claims and legal environments, it is an evolving matter.

The technology should be evaluated accordingly.

8. How Does the AI Fit Into the Workflow and How Is the Information Governed?

Even highly capable technology can create friction if using it requires professionals to step outside the systems and processes where their work already occurs.

Buyers should map the entire workflow. How are records transferred into the AI solution? Is another upload required? Does information move to another vendor? Where does the output go? Can it connect with existing claims, case-management or legal systems? Will users need to maintain another application or repeat information already captured elsewhere?

These questions have operational implications, but they also intersect with governance.

Medical records contain sensitive information. Organizations evaluating AI therefore need to understand where that information is processed, who can access it, how long it is retained, whether it is used to train models, what third parties are involved and what controls govern the environment.

Buyers should also distinguish between the AI model and the complete technology environment surrounding it. A solution may rely on third-party models, infrastructure or other technology providers, each of which can introduce additional questions about data access and handling. NIST's Generative AI Profile specifically identifies third-party AI integrations as an area where data privacy, information security and transparency risks may arise and points to procurement due diligence and third-party risk management as relevant controls.¹

The practical implication is straightforward: "secure" is not a sufficient answer. Buyers should understand the path their information takes through the solution and which organizations or technologies may have access to it along the way.



For insurance organizations, AI governance is becoming increasingly relevant to regulatory oversight. The National Association of Insurance Commissioners' *Model Bulletin on the Use of Artificial Intelligence Systems by Insurers* **provides a framework for regulatory expectations** around governance, risk management, testing and oversight of AI-supported decisions and addresses insurers' use of third-party AI systems.⁵

Law firms face their own responsibilities. The American Bar Association's Formal Opinion 512 states that lawyers using generative AI must consider existing professional obligations including competence, protection of client information, communication, supervision and candor.⁶ The opinion reinforces a broader point relevant to technology procurement: adopting AI does not transfer professional responsibility to the technology.

A buyer's security and governance review will necessarily be more detailed than a white paper can prescribe, and requirements will vary by organization. But vendors should be able to provide precise answers rather than broad assurances.

The more sensitive the information and consequential the workflow, the less appropriate it is to treat AI procurement as a feature comparison alone.

9. Who (and What Expertise) Stands Behind the AI?

AI capabilities are evolving extraordinarily quickly. Models will improve. Context windows will expand. Features that appear differentiated today may become widely available tomorrow.

That makes the organization surrounding the technology an important part of the buying decision.

Medical records present challenges that extend beyond language-model capability. They have to be obtained, organized and maintained. They contain different document types and clinical terminology. They evolve over time. They are used differently by adjusters, attorneys, medical professionals and other reviewers. And they exist within operational environments where security, integration, governance and reliability matter.

Buyers should therefore ask what expertise shaped the solution. Did the vendor begin with an AI model and look for a medical-record use case, or did the technology emerge from an understanding of how medical records actually move through claims and legal workflows? Does the organization understand not only how to analyze records, but how those records are obtained, assembled, supplemented and ultimately used by professionals making decisions?

Does the vendor have experience working with medical records at scale? Does it understand how records are created and obtained? Does it understand claims and litigation workflows? Does it have the infrastructure to support enterprise customers? How will it continue evaluating and improving the technology as AI itself evolves?

The buyer is not simply selecting a model.

The buyer is selecting the organization responsible for putting that model to work.

The most interesting lesson is not that AI should be avoided. It is almost the opposite.

The Demo Is Not the Evaluation

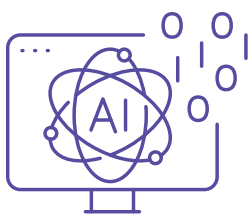
A product demonstration is designed to show technology working well. A buyer evaluation should be designed to discover how the technology works when the environment is less tidy.

That means testing realistic complexity.

Ask the vendor to demonstrate a substantial record set involving multiple providers and meaningful treatment history. Include duplicate or administrative material. Ask a question that requires relationships across time rather than a simple keyword search. Select an AI-generated finding and trace it back to its original source. Add another batch of records and see what happens.

If possible, use scenarios that resemble the organization's actual work rather than relying entirely on a vendor-curated example.

Buyers should also pay attention to what the system does when the answer is unclear. Conflicting records, incomplete histories and ambiguous documentation are part of real medical-record review. A technology that confidently resolves every ambiguity may be less informative than one that helps a professional see where ambiguity actually exists.



A strong demo shows what an AI system can do. A strong evaluation shows whether it can do it reliably in your environment.



What This Framework Looks Like in Practice

For more than 50 years, Compex has worked at the front end of the medical-record workflow, helping insurance carriers, law firms, third-party administrators and other organizations obtain the records required for claims and litigation. That experience has provided a particular perspective on medical records AI: intelligence is most useful when it is connected to the information and workflow surrounding it.

Compex's approach to Medical Records Intelligence reflects that perspective.

Asabell™ is Compex's medical records intelligence offering. Rather than operating only after a static set of documents has been collected and transferred to a separate AI process, Asabell can structure medical information as records move through the Compex delivery process, creating a case foundation that grows as information is received.

At the center of that approach is a structured foundation of medical information. Asabell organizes information such as encounters, providers, diagnoses, procedures, medications, body regions and billing data, with evidence citations connecting findings to the underlying medical records. That foundation can support different forms of analysis without requiring the record set to be reconstructed for every new question.

Case Summary is one expression of that foundation. It brings information across the available record set into a consolidated case-level view that includes a plain-language narrative, chronology of care, key diagnoses and treatment, organized provider views, billing context and evidence citations. Reviewers can move from a potentially important finding back to the underlying documentation rather than relying solely on the generated output.

The approach is also designed around the reality that medical matters evolve. As additional records are received and analyzed, Case Summary can be regenerated to reflect the expanded record set. The same structured information can support other review needs, helping reduce the repeated reconstruction that occurs when every new question or delivery starts another disconnected review.

That intelligence can be delivered alongside the records customers already receive, while case-management integrations and APIs can connect Compex with systems already used by claims and legal teams. Asabell can also analyze records obtained from other sources, so organizations do not have to change retrieval providers in order to use its intelligence capabilities.

The differentiator, therefore, is not AI alone. It is the combination of medical-record retrieval expertise, structured intelligence, case-level understanding, source grounding and workflow.

That combination reflects a broader principle buyers should consider regardless of which vendor they evaluate: **the value of medical records AI depends not simply on the intelligence of the model, but on how effectively that intelligence is connected to the work professionals actually need to perform.**

Medical Records AI Buyer Evaluation Checklist

Before selecting a medical records AI solution, buyers should be able to answer the following questions:

What specific medical-record problem is the solution designed to solve?

Does it analyze individual documents, the complete available record set, or both?

- ☐ Does the solution create reusable structured information, or does each new analysis require the record set to be processed again?
- ☐ How does it preserve chronology across providers, encounters and dates?
- ☐ Can important AI-generated findings be traced to their supporting source records?
- ☐ What happens when source records contain incomplete, ambiguous or conflicting information?
- ☐ How are omissions, errors and hallucinations identified and evaluated?
- ☐ How has the solution been tested on realistic medical-record sets and use cases?
- ☐ What happens when additional records arrive?
- ☐ Can different professionals use the same record set to investigate different questions?
- ☐ Does the solution require another platform, upload or disconnected workflow?
- ☐ How does it integrate with existing claims, case-management or legal systems?
- ☐ How is customer information stored, accessed and retained?
- ☐ Is customer information used to train AI models?
- ☐ What third-party technologies or models process or have access to customer information?
- ☐ What medical-record and workflow expertise does the vendor bring beyond the AI technology itself?



Five Things to Ask a Vendor to Show You

Don't just ask whether a capability exists. Ask to see it.

- 1 Show me how you move from an AI-generated finding to the original medical record.
- 2 Show me how the system analyzes a case across multiple providers and records.
- 3 Show me how chronology and context are preserved.
- 4 Show me what happens when additional medical records are added to an existing matter.
- 5 Show me what the system does when the records are incomplete, conflicting or do not support a confident answer.

Buy the Workflow, Not Just the AI

Medical records AI will continue to improve rapidly. Models will become more capable, analysis will become faster and features that seem advanced today will increasingly become expected.

That makes it tempting to evaluate vendors based on the newest capability or most impressive output. But the durable value of medical records AI will depend on more than the sophistication of the model.

It will depend on whether the right information reaches the system. Whether complex histories can be understood across records and over time. Whether professionals can verify important findings. Whether uncertainty is visible rather than obscured. Whether sensitive information is governed appropriately. And whether the technology fits naturally into the workflow where professional judgment ultimately occurs.

The objective is not to remove that judgment from the process. It is to remove unnecessary information-management work from its path.

The best medical records AI will not be defined simply by the sophistication of the model. It will be defined by the quality of the workflow the model makes possible.

Medical Records, Understood.

Learn more about Compex and our approach to Medical Records Intelligence. [Request a quick demo.](#)

Sources

1. National Institute of Standards and Technology. [Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile](#). NIST AI 600-1, July 2024.
2. Kahl, Nicolas M., et al. ["Evaluation of electronic health record-integrated artificial intelligence chart review."](#) npj Health Systems, Vol. 3, Article 6, 2026.
3. U.S. Food and Drug Administration. [Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff](#). January 2026.
4. National Institute of Standards and Technology. [Artificial Intelligence Risk Management Framework \(AI RMF\)](#). NIST AI Resource Center.
5. National Association of Insurance Commissioners. [Model Bulletin: Use of Artificial Intelligence Systems by Insurers](#). Adopted December 2023.
6. American Bar Association Standing Committee on Ethics and Professional Responsibility. [Formal Opinion 512: Generative Artificial Intelligence Tools](#). July 29, 2024.
7. Williams, Christopher Y. K., et al. ["Physician- and Large Language Model-Generated Hospital Discharge Summaries."](#) JAMA Internal Medicine, Vol. 185, No. 7, 2025, pp. 818-825.

📞 (800) 326-6739

🌐 compexlegal.com

✉️ sales@compexlegal.com

