The Evidence Synthesis Landscape

Under Pressure

Dr Seye Abogunrin
Global Access Evidence Lead

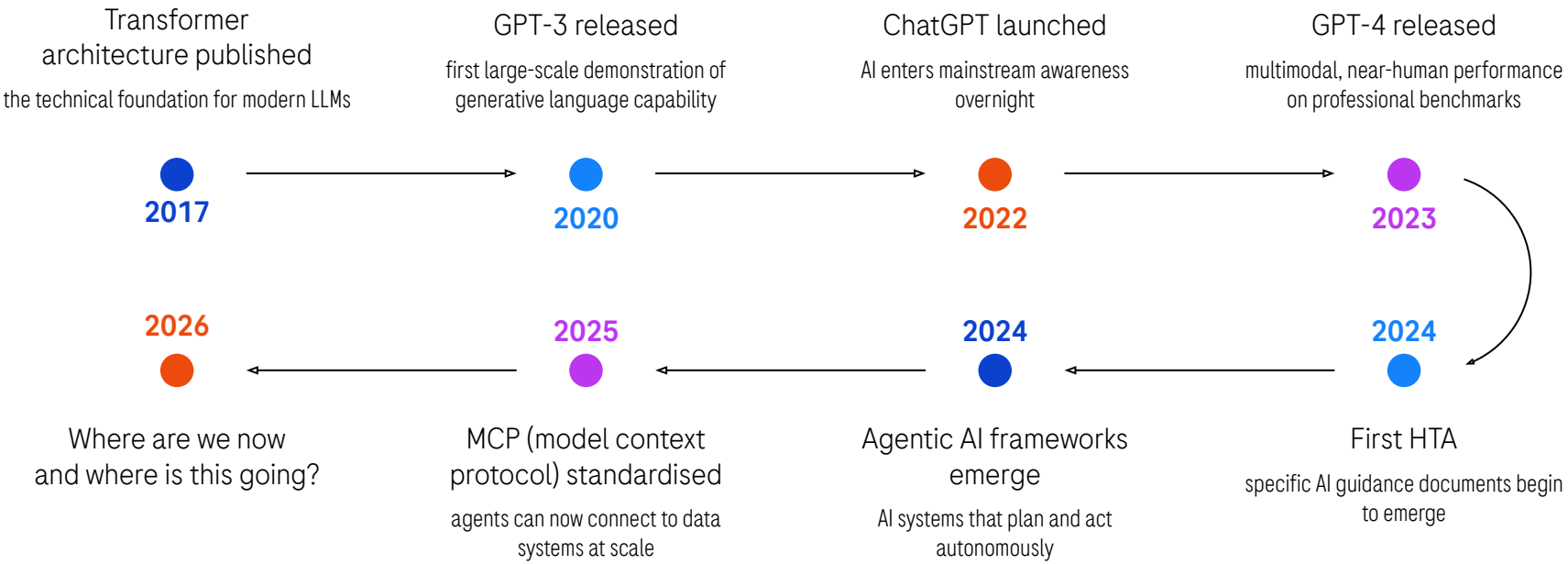
24 April 2026 | public use

M-XX-00023658

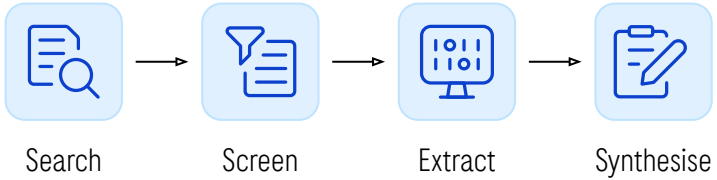
From Curiosity to Disruption

A Timeline of Key AI Milestones

A World That Changed Quickly



It Wasn't Broken... Until the Goalposts Moved



Before

- Manual database searching
- Dual independent screening
- Structured data extraction forms
- Human-led synthesis and meta-analysis
- Submission timelines of 12–18 months

ChatGPT

The Disruption

- AI can screen thousands of abstracts in minutes
- LLMs can extract structured data from full texts
- Generative models can draft evidence summaries
- Agentic systems can coordinate entire review workflows
- Timelines could compress – but at what methodological cost?

Should You Be Using AI for Your HTA Literature Reviews?

Should you be using AI for your literature reviews when preparing an HTA submission?

– And if so:

which AI, for which steps, under what conditions, with what documentation?

The answer is no longer simply yes or no.
It depends on where you are in the journey – and whether your methods, governance, and reporting are ready.

What We Will Cover Today

01 What makes an HTA SLR different and why it matters for AI adoption

02 A framework: three stages of automation

03 Roche's journey – what we tried, what worked, what didn't

04 What HTA-compliant AI use looks like in practice

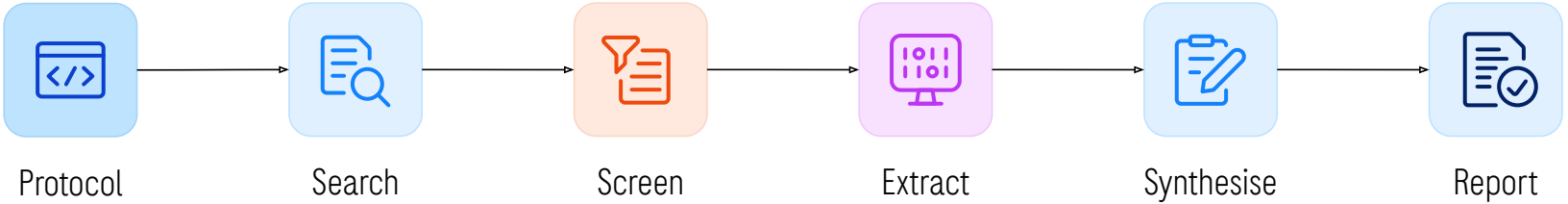
05 How to prepare for your next submission – a practical checklist

06 Open discussion – your questions, your context

What is an HTA SLR and what Makes it Different?

The Systematic Literature Review

A Recap



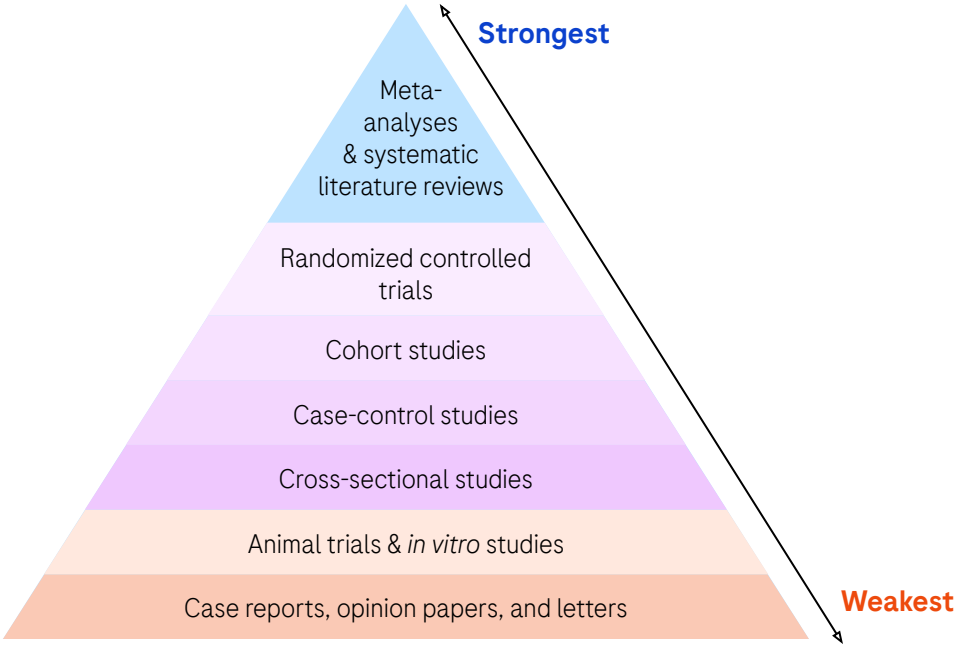
A systematic literature review (SLR) is a structured, reproducible method for identifying, appraising, and synthesising evidence on a defined research question.

Key characteristics:

- Pre-specified protocol and eligibility criteria
- Comprehensive, documented database searching
- Independent, dual screening of identified records
- Structured data extraction and quality assessment
- Transparent reporting of methods and results

Why are SLRs conducted?

Hierarchy of scientific evidence



- Identify, appraise, and synthesize all relevant evidence to a specific research question
- Summarize current and ongoing knowledge, and identify gaps in research
  Appraise all relevant evidence about a disease area (e.g. randomized trials of specific disease)
- Provide the highest quality of scientific evidence in a specific area of study
- Systematic reviews are crucial to healthcare decision making

But Sytematic Literature Reviews Mean Different Things to Different Stakeholders

	Academic SLR	Clinical SLR	HTA SLR
Primary purpose	Generate new knowledge	Inform clinical practice	Support reimbursement decisions
Key audience	Research community	Clinicians	HTA bodies, payers
Data collected	Determined by research question	Efficacy, safety	Efficacy, safety, QoL, costs, healthcare resource use, epidemiology
Output used for	Publication	Guidelines	Submission dossier
Stakes if wrong	Retraction	Clinical harm	Rejected submission or incorrect funding decision
Methodological scrutiny	Peer review	Peer review	Peer review

The HTA SLR

A Distinctive Methodological Challenge

HTA reviews must be exhaustive across multiple databases, grey literature, and unpublished sources. Missed studies can invalidate a submission.

HTA bodies require evidence against specific comparators – often multiple, jurisdiction-specific. The review must anticipate and address these explicitly.

Unlike academic reviews where reproducibility is aspirational, HTA submissions may be scrutinised, queried, or re-run by the receiving body. Every methodological choice must be documented, defensible and reproducible.

Submissions operate under strict deadlines. The pressure to deliver a comprehensive, high-quality review in compressed timeframes is where AI tools first gained traction – and where the risks of cutting corners are highest.

Why AI Is Entering This Space and Why It Matters That We Get It Right

The drivers of AI adoption in HTA SLRs:

- Exponential growth in published literature
- Increasing complexity of evidence landscapes
- Compressed submission timelines
- Resource and cost pressures on review teams
- Advances in AI capability that make automation genuinely feasible

The reason methodological rigour cannot be sacrificed:

- HTA decisions directly affect patient access to treatments
- Scrutiny of methodologies used in submissions is increasing, not decreasing
- AI errors in evidence synthesis may be systematic, not random – and harder to detect
- The field is watching: how industry handles this will shape future guidance

The Evidence Synthesis Challenge

The Cost of the Status Quo

The scale of the problem:

- A typical SLR in the pharmaceutical industry costs in excess of $141,000 in the US alone
- The 10 largest pharma companies and 10 largest academic centres spend an estimated $16–18 million annuallyon SLRs
- A 2018 case study found the average time to complete a systematic review was 66 weeks and 881 person-hours
- Reviews frequently go out of date before – or shortly after – publication

What AI can offer:

- 17 of 25 studies in a published pragmatic review found >50% time reduction with AI automation
- Abstract review time reduced by a factor of 5 to 6
- Workload reductions of up to 10-fold at 95% recall
- One AI tool replicated an entire NMA update in under 10 minutes

The gap that remains:

"Despite these demonstrated benefits, no studies have yet quantified the economic cost savings associated with automation in evidence synthesis – a gap our published work has helped to identify." (Abogunrin et al., Front Pharmacol, 2025)

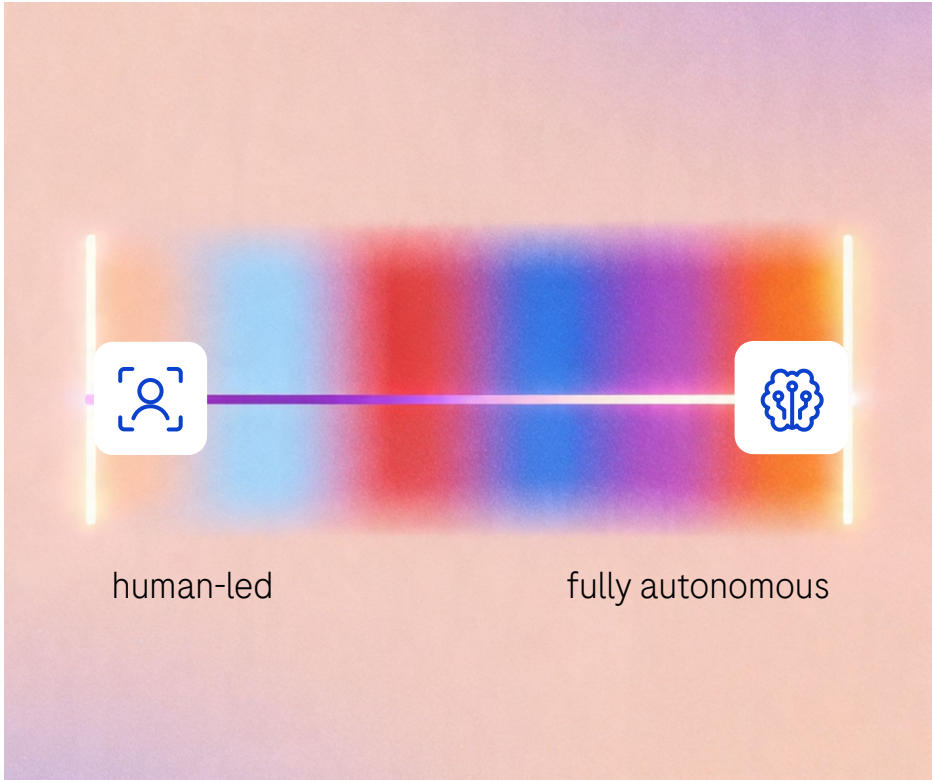
Three Stages of Automation

Not All AI Use Is Equal

A Framework for Thinking About Automation

Organisations adopting AI for systematic literature reviews are not making a single decision. They are making a series of decisions – about which steps to automate, to what degree, and under what conditions.

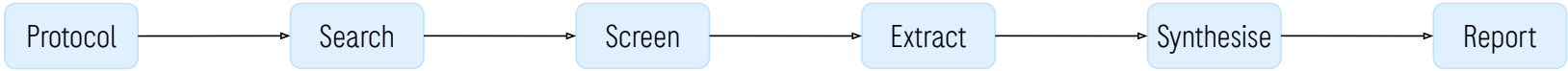
A useful starting point is to distinguish between three broad stages of automation – each with different implications for methodology, governance, and HTA acceptability.



Partial Automation

Stage 1

AI as a Tool, Humans in Control



Definition:

AI assists at specific, well-defined steps in the review process. Human reviewers remain involved at every stage and make all consequential decisions.

Typical applications at this stage:

- Natural language processing to support search strategy development
- Automated de-duplication of search results
- AI-assisted title and abstract screening (human adjudicates all uncertain cases)
- AI-generated data extraction drafts reviewed and corrected by humans

What this looks like in practice:

- AI reduces workload and flags records for human attention
- Sensitivity and specificity of the AI tool are validated against a human-reviewed reference set
- All AI-assisted decisions are documented and auditable
- The human reviewer remains the decision-maker of record

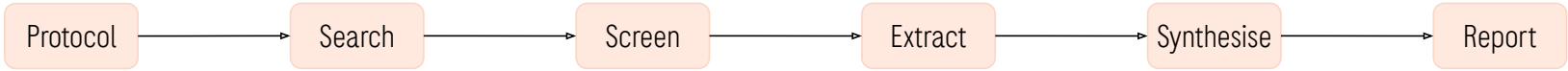
HTA acceptability

Increasingly accepted where validation is documented and human oversight is explicit

Selective Full Automation

Stage 2

Removing Humans from Specific Steps



Definition:

Specific, well-validated steps in the review process operate without human involvement. Human oversight is applied at the input and output level, not within the step itself.

Typical applications at this stage:

- Fully automated first-pass screening with validated recall thresholds
- Automated data extraction for structured, well-defined data fields
- AI-driven identification of relevant comparators or outcomes across a corpus
- Automated quality assessment against pre-specified criteria

What this looks like in practice:

- The automated step has been validated against a reference standard with documented performance metrics
- Human review occurs at defined quality checkpoints – not at every record
- Inter-rater reliability between AI and human reviewers has been established and reported
- Clear criteria exist for when human escalation is triggered

Key question for HTA submissions:

Can you demonstrate that the sensitivity of your automated step is sufficient to ensure no material evidence has been missed?

HTA acceptability:

Emerging – dependent on validation evidence and transparency of reporting

Autonomous Automation

Stage 3

The Agentic Horizon

Definition:

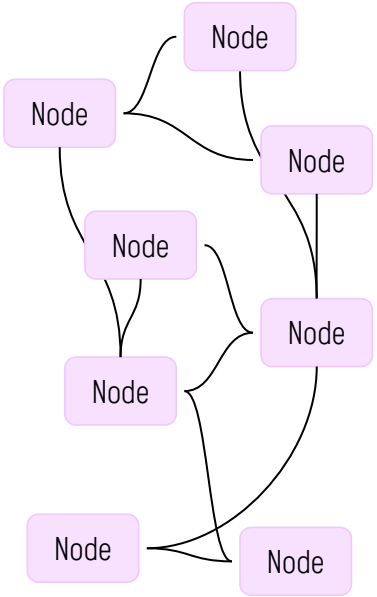
AI systems autonomously coordinate and execute multiple steps of the review process, with little or no human involvement in individual decisions. Humans define the task, set the parameters, and review the final output.

What this could look like:

- An AI agent formulates the search strategy, executes searches across databases, retrieves and screens records, extracts data, and generates a draft synthesis — with human review only at the output stage
- Multiple specialised agents work in coordination: a search agent, a screening agent, an extraction agent, a synthesis agent
- Model context protocol (MCP) servers connect agents to licensed literature databases, internal data systems, and output templates

The questions this raises:

- Who is methodologically responsible for decisions made autonomously by an agent?
- How do you audit a process where the decision-maker is not human?
- What does reproducibility mean when an AI system's outputs may vary across runs?
- How do you penalise, correct, or retrain an agent that makes systematic errors?



HTA acceptability

Not yet established — but the conversation is beginning

Where Are You and Where Are You Going?

Stage 1

Partial Automation
You are here if:

- AI tools assist your team but humans review all outputs
- You have begun documenting AI use in your methods sections
- You are validating tools informally or on a project-by-project basis

Stage 2

Selective Full Automation
You are here if:

- Specific steps operate without per-record human review
- You have formal validation evidence for your automated steps
- You are actively working on how to report this in submissions

Stage 3

Autonomous Automation
You are here if:

- You are exploring or piloting agentic workflows
- You are asking governance and accountability questions you do not yet have answers to
- You are ahead of the guidance – and you know it

Roche's Journey: Build vs Buy vs Hybrid

The KiaKia Vision

KiaKia Vision

Making literature reviews across Roche

F Findable **A** Accessible **I** Interoperable **R** Reusable

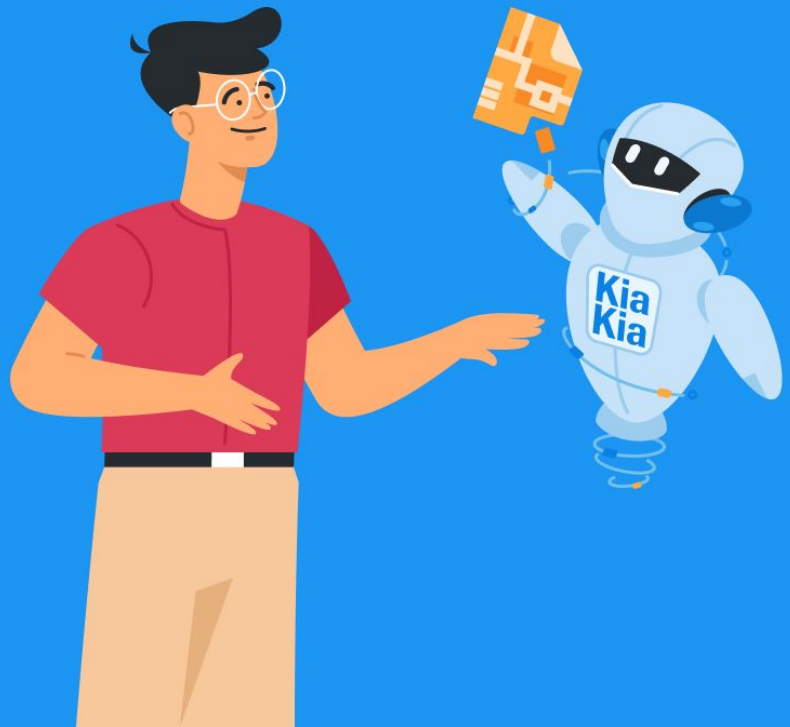
Accelerating the literature review process with optional full or partial automation

- KiaKia will make literature reviews **easier, reusable, more accessible and efficient** to all users
- KiaKia is **transforming** how we conduct literature reviews and will continue to bring **more value in the future.**

KiaKia means **fast or quick or rapid** in the Yoruba language, which symbolises what we aim to achieve with this solution - **rapid review of scientific literature using automation methods**

KiaKia

Where automation encounters literature reviewing



KiaKia is Roche's internal tool for conducting literature reviews.

It supports all steps in the process:

- Protocol writing
- Screening (study selection)
- Data extraction from full-texts into structured format
- Quality appraisal ^
- Report-writing*

All stages of the process include AI-powered automation and the tool can be used for reviews of various topics e.g., clinical, economic, epidemiology, health-related quality of life, etc..

^ Extraction of information required to conduct quality appraisal or risk of bias assessment

* Q4 2026

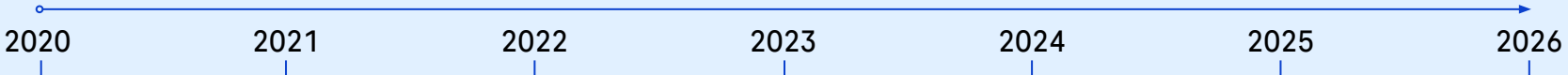
KiaKia development: people, data, and architecture

Team

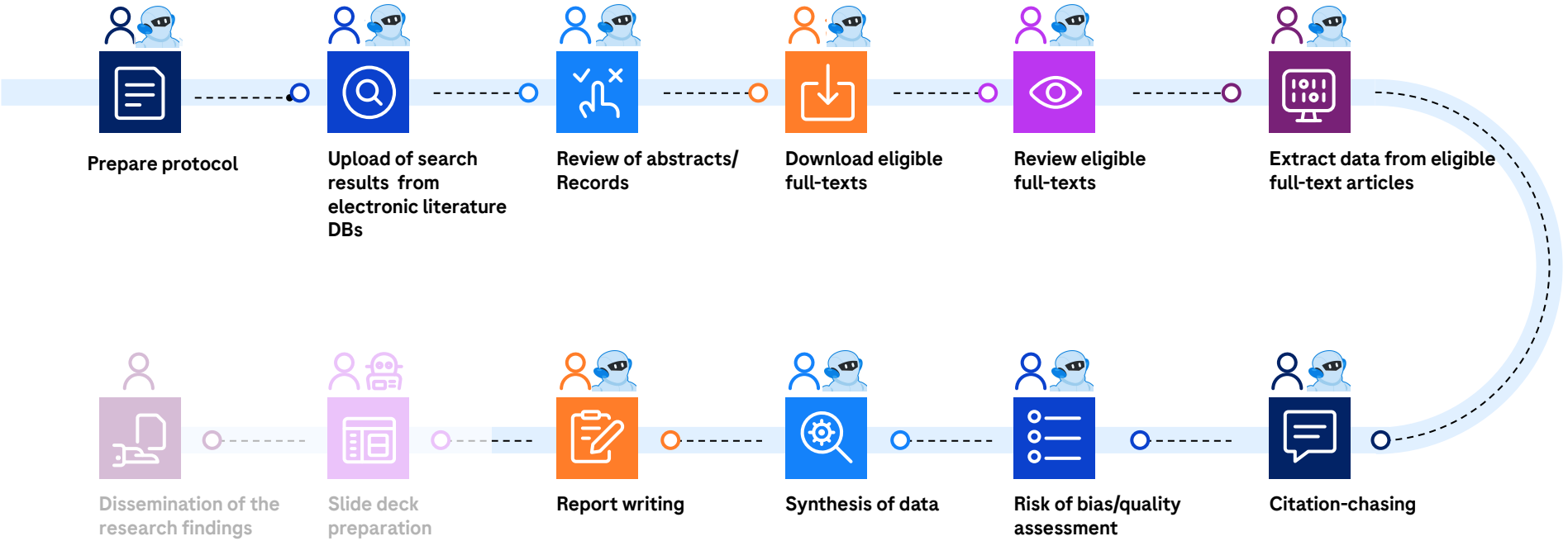
- Business analysts
- Developers
- LLM-specialist data scientists
- Evidence synthesis SMEs

evelopment principles

- Internal data sources, e.g. Romine, Roche eLibrary,
  - Explored citation sources, including Tavily, others
  - No fabricated references
- Programmatic use of multiple LLMs via Roche enterprise access (PortKey)
  - Supports cost control and model switching
  - Meets Roche privacy and IP requirements
- Methodological approach for assessing AI; evolved from traditional ML to LLMs ; > 50 experiments across 12 datasets
- Designed for integration across Roche infrastructure



We are automating all steps of a literature review.



Project management: assign team members, including vendors, manage tasks (screening and conflict resolution), manage changes to the protocol, dashboard to view screening progress, screening decisions, downloads (lists of records, studies attrition diagram)

KiaKia

Others

What can you do in KiaKia today?

Prepare protocol
- Draft and approve
- Manage changes
- Clone existing protocols
- AI-assisted protocol generation

Upload of search results from electronic literature DBs (run outside KiaKia)
- List of titles & abstracts

Review of abstracts/ Records
- List of eligible & ineligible titles & abstracts
- Auto deduplication
- Ranking
- AI suggestions

Download eligible full-texts
- DOI to full texts or store link to local copy

Review eligible full-texts
- List of eligible and ineligible full-texts (.docx & PRISMA diagram)

Extract data from eligible full-text articles
- Data points
- AI pre-extraction of eligible full-texts

Citation-chasing
- List of titles & abstracts and eligible full-texts

Risk of bias/quality assessment
- Summaries of assessment and scoring, and report elements

Synthesis of data
- Qualitative and quantitative analysis report elements

Report writing
- Summaries of findings, bibliographies and associated full-texts

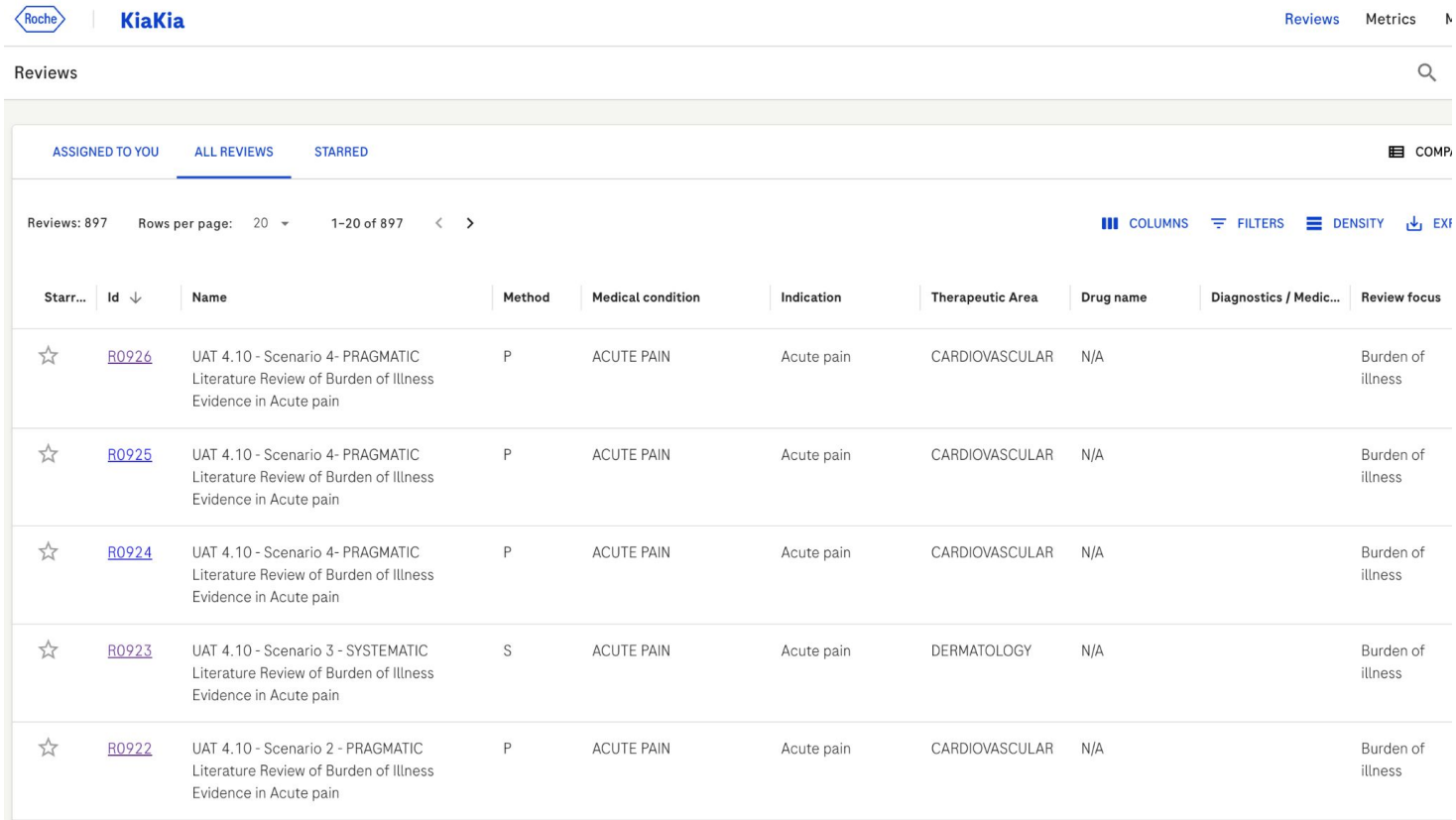
Slide deck preparation
- Visual summaries of findings

Dissemination of the research findings

Project management: assign team members, including vendors, manage tasks (screening and conflict resolution), manage changes to the protocol, dashboard to view screening progress, screening decisions, downloads (lists of records, studies attrition diagram)

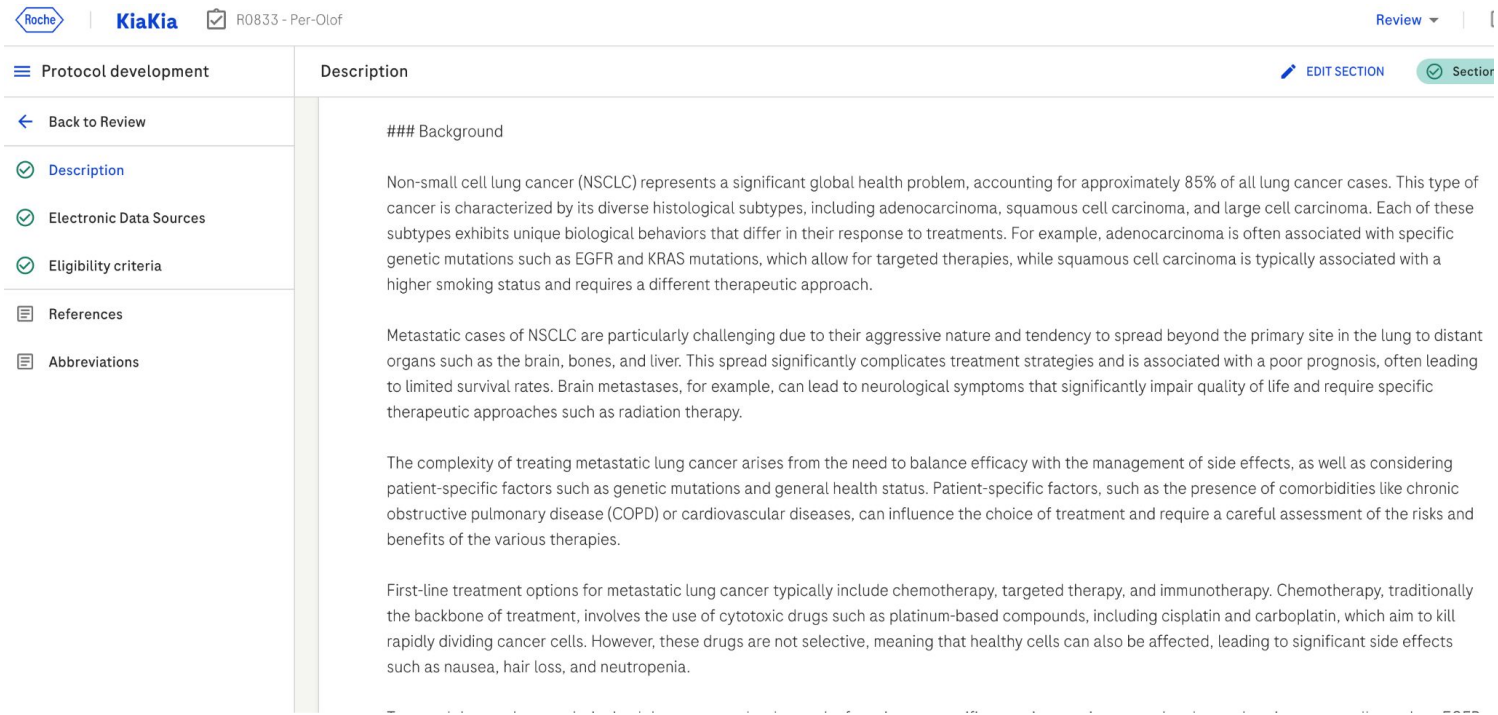
KiaKia Others

A glimpse into KiaKia

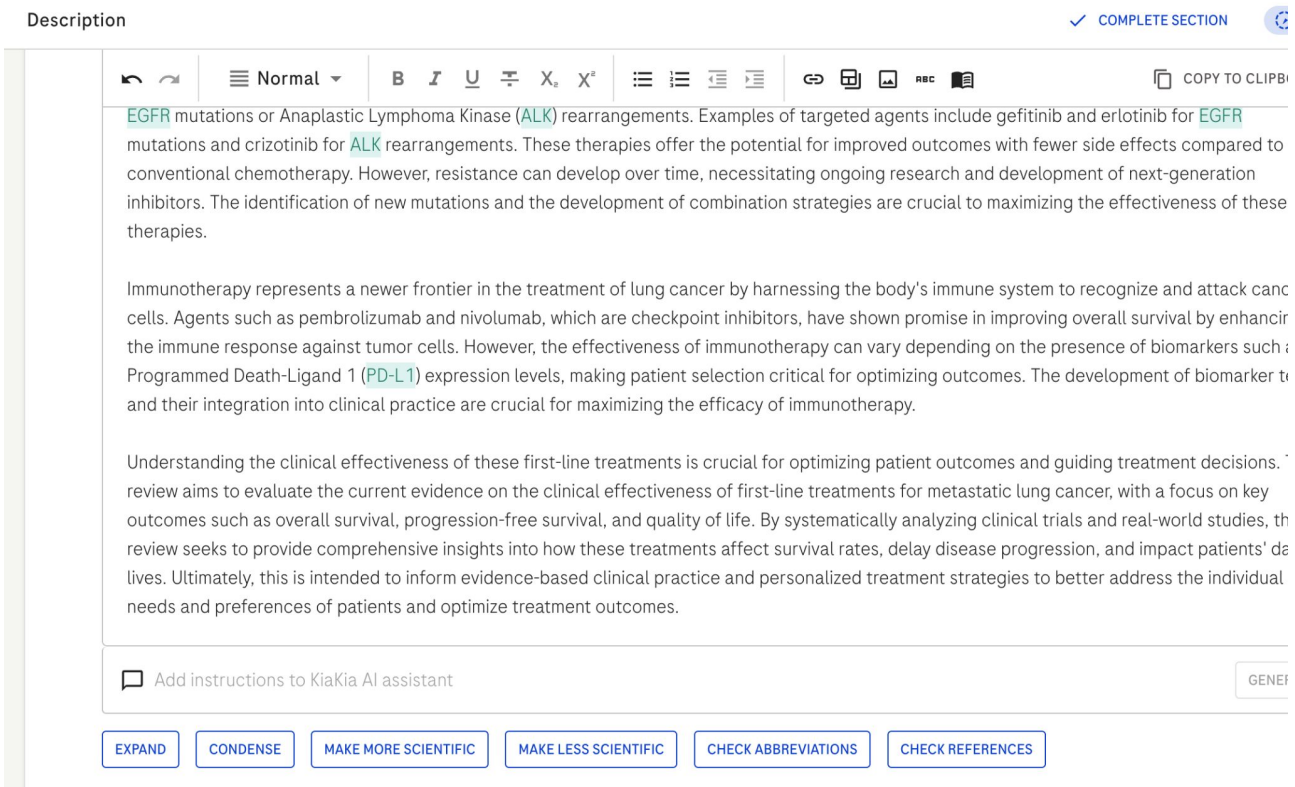


Starr...	Id	Name	Method	Medical condition	Indication	Therapeutic Area	Drug name	Diagnostics / Medic...	Review focus
☆	R0926	UAT 4.10 - Scenario 4- PRAGMATIC Literature Review of Burden of Illness Evidence in Acute pain	P	ACUTE PAIN	Acute pain	CARDIOVASCULAR	N/A		Burden of illness
☆	R0925	UAT 4.10 - Scenario 4- PRAGMATIC Literature Review of Burden of Illness Evidence in Acute pain	P	ACUTE PAIN	Acute pain	CARDIOVASCULAR	N/A		Burden of illness
☆	R0924	UAT 4.10 - Scenario 4- PRAGMATIC Literature Review of Burden of Illness Evidence in Acute pain	P	ACUTE PAIN	Acute pain	CARDIOVASCULAR	N/A		Burden of illness
☆	R0923	UAT 4.10 - Scenario 3 - SYSTEMATIC Literature Review of Burden of Illness Evidence in Acute pain	S	ACUTE PAIN	Acute pain	DERMATOLOGY	N/A		Burden of illness
☆	R0922	UAT 4.10 - Scenario 2 - PRAGMATIC Literature Review of Burden of Illness Evidence in Acute pain	P	ACUTE PAIN	Acute pain	CARDIOVASCULAR	N/A		Burden of illness

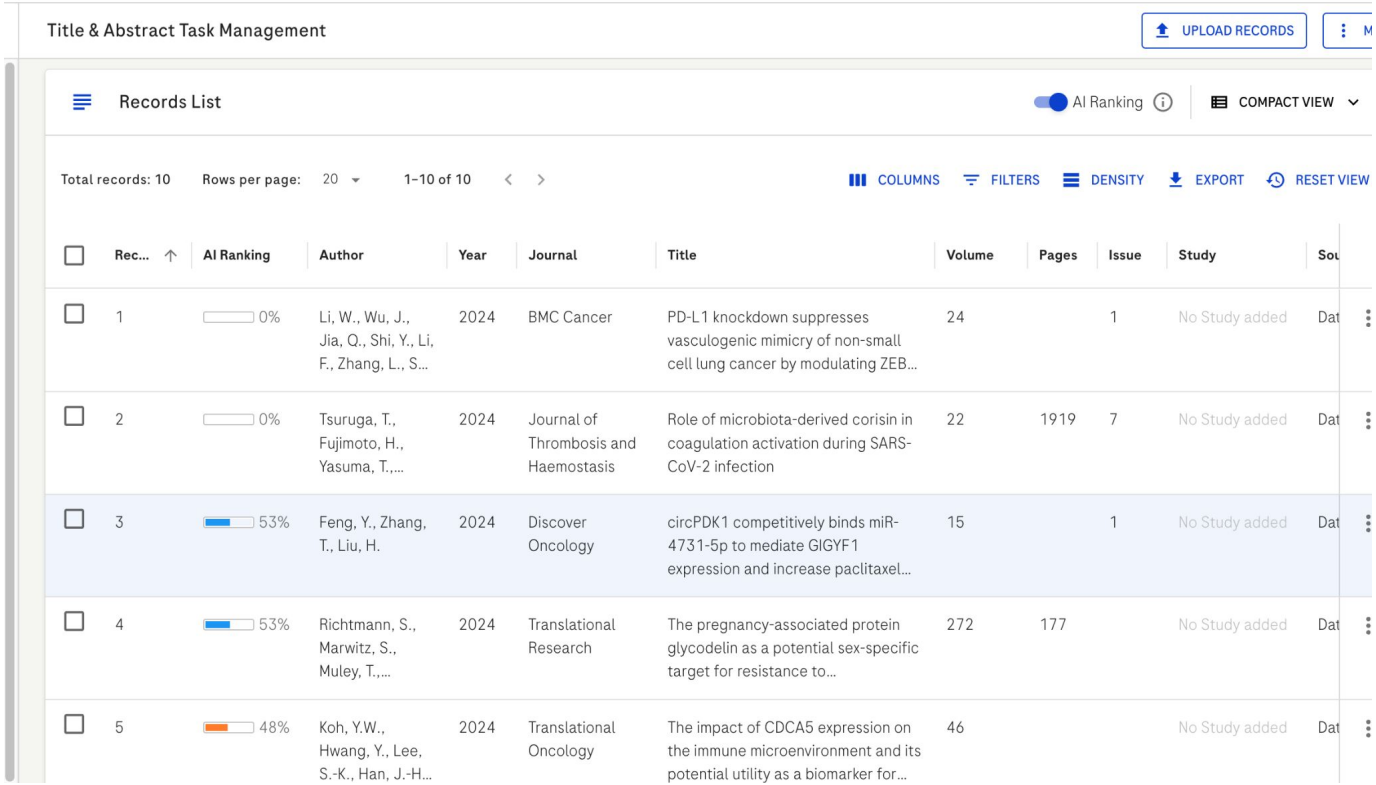
A glimpse into KiaKia



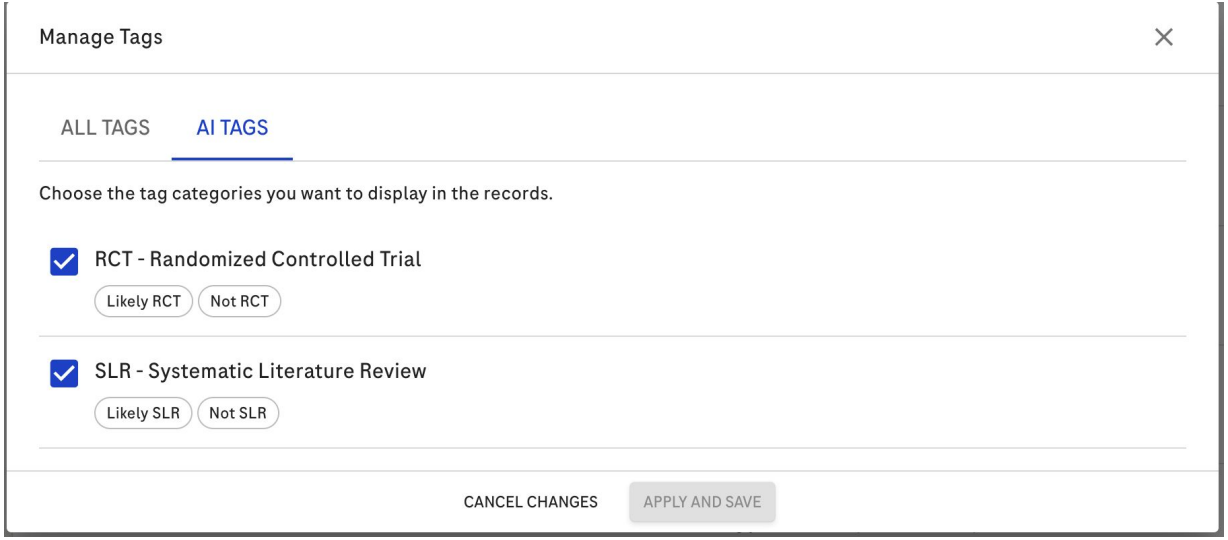
A glimpse into KiaKia (AI-assisted protocol writing)



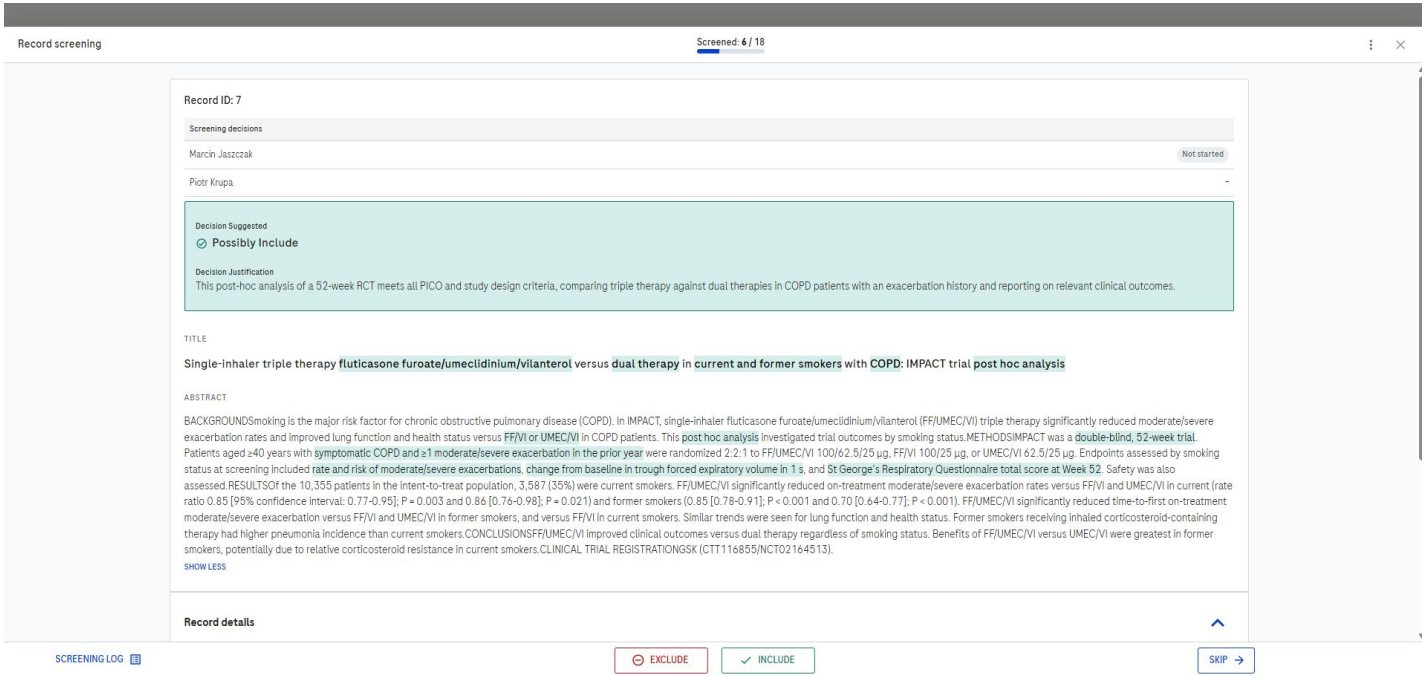
A glimpse into KiaKia (AI-ranking algorithm)



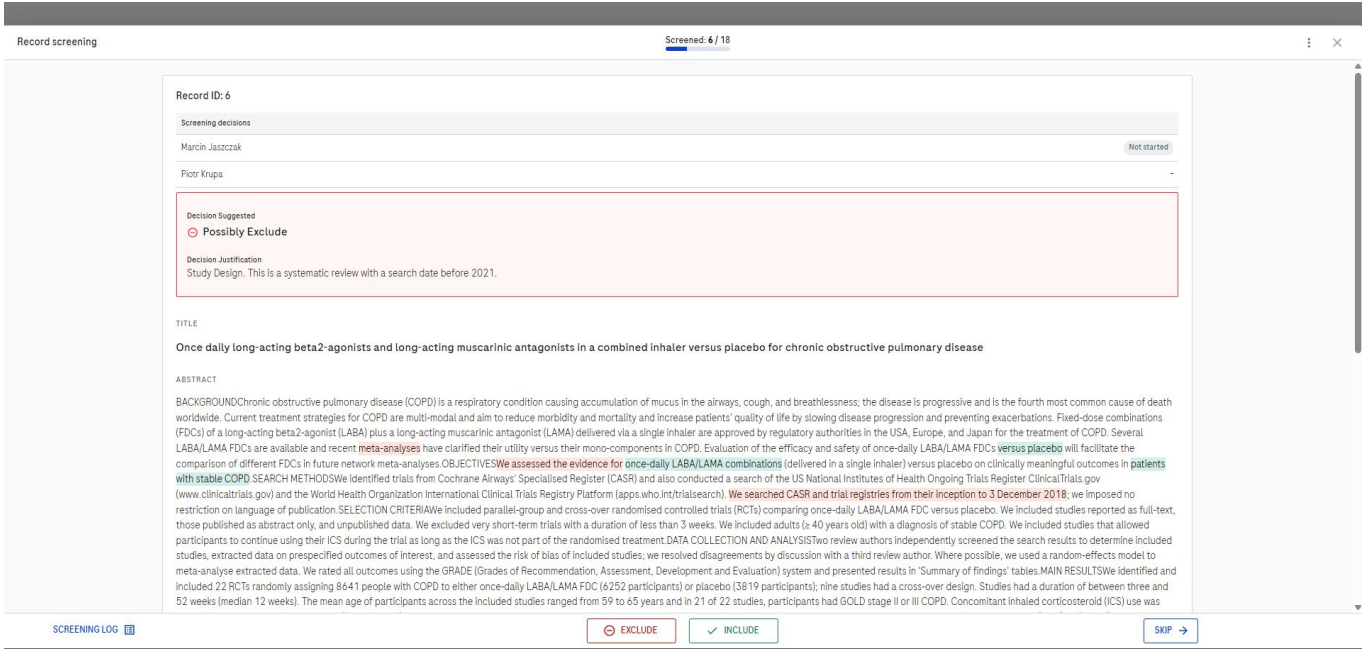
A glimpse into KiaKia (AI-assisted tagging)



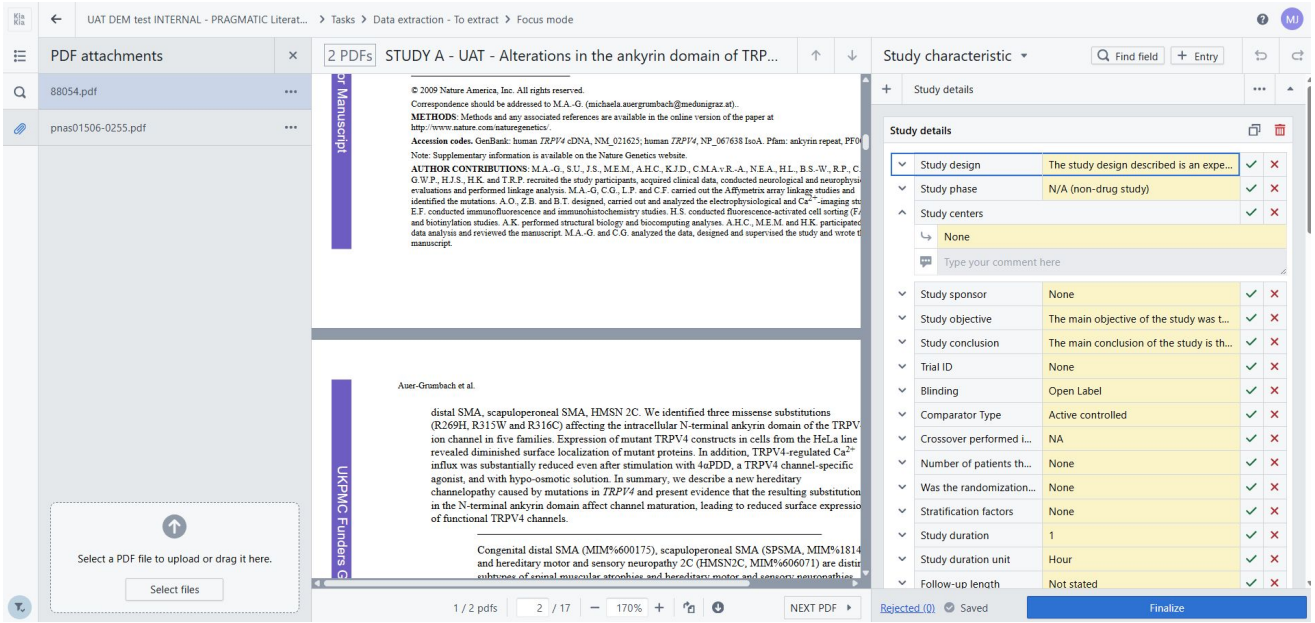
A glimpse into KiaKia (AI-assisted screening suggestions)



A glimpse into KiaKia (AI-assisted screening suggestions)



A glimpse into KiaKia (AI-assisted data extraction)



KiaKia AI experiments: example results

					Excluded		Included (would move to the next step)	
ID	Disease	Total number of records	N records used to train	N records used to test	True negatives	False negatives	True positives	False positives
1	mNSCLC (2L+)	5285	80	5045	40.46%	0.02%	4.06%	55.46%
2	mCRPC (cl)	1025	40	925	31.24%	0.00%	1.95%	66.81%
3	eNSCLC	2338	80	2138	41.86%	0.47%	4.07%	53.60%
4	COVID-19	5721	80	5521	17.61%	0.04%	10.52%	71.83%
5	DLBCL	3386	80	3186	9.98%	0.31%	9.73%	79.97%
6	SCLC	10044	80	9844	46.35%	0.01%	1.34%	52.30%
7	mNSCLC (1L)	17242	82	16962	32.26%	0.12%	8.27%	59.35%
8	eNSCLC (non-RCT)	702	80	532	22.37%	0.75%	14.47%	62.41%
9	eNSCLC (RCT)	519	80	319	24.45%	1.57%	27.59%	46.39%
10	mCRPC (eco)	1126	40	926	24.30%	0.00%	2.05%	73.65%

KiaKia AI experiments: example results

ID	Disease	Total number of records	N records used to train	N records used to test	Precision	Recall	WSS@95	%conflicts (human vs SVM)	Time to complete automated Screening	Time to complete human TIABS	Δ time spent for automated vs human TIABS
1	mNSCLC (2L+)	5285	240	5045	0.27	0.58	0.86	8,2	9.6	88.1	78.5
2	mCRPC (cl)	1025	100	925	0.25	0.67	0.90	4,5	6.2	17.1	10.9
3	eNSCLC	2338	200	2138	0.26	0.57	0.85	9,4	7.8	39.0	31.1
4	COVID-19	5721	200	5521	0.36	0.47	0.81	14,4	7.8	95.4	87.5
5	DLBCL	3386	200	3186	0.27	0.61	0.72	20,6	7.8	56.4	48.6
6	SCLC	10044	200	9844	0.17	0.84	0.88	5,9	7.8	167.4	159.6
7	mNSCLC (1L)	17242	280	16962	0.32	0.72	0.76	15,4	11.3	287.4	276.0
8	eNSCLC (non-RCT)	702	170	532	0.43	0.68	0.71	18,6	7.3	11.7	4.4
9	eNSCLC (RCT)	519	200	319	0.83	0.67	0.71	13,8	7.8	8.7	0.8
10	mCRPC (eco)	1126	200	926	0.18	0.95	0.84	9,0	7.8	18.8	10.9

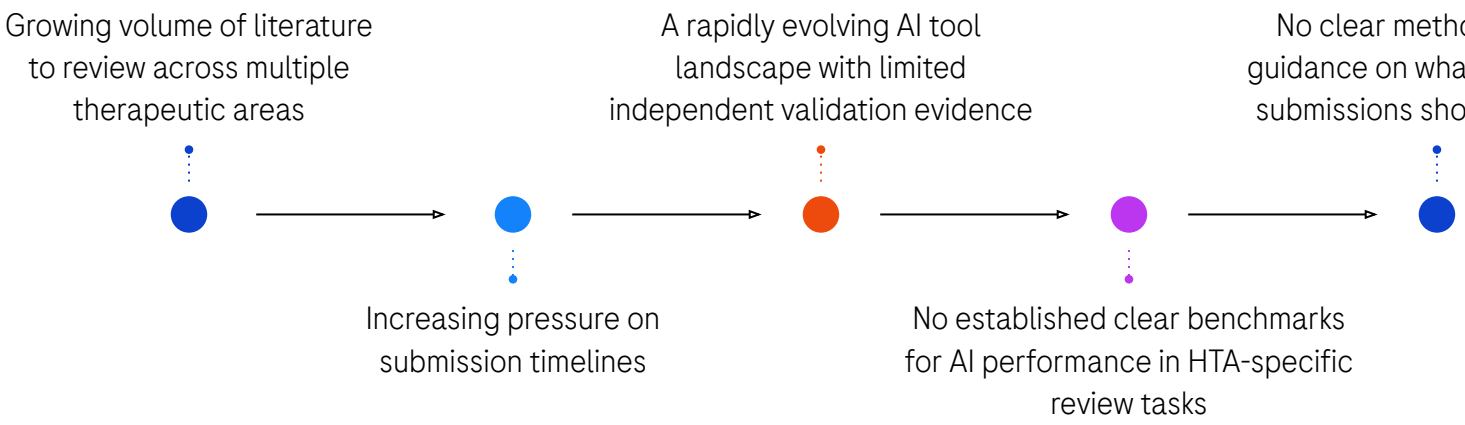
KiaKia AI experiments: example results

ID	Disease	Total number of records	N records used to train	N records used to test	Precision	Recall	WSS@95	%conflicts (human vs SVM)	Time to complete automated Screening	Time to complete human TIABS	Δ time spent for automated vs human TIABS
1	mNSCLC (2L+)	5285	80	5045	0.13	0.90	0.67	24.4	1.6	85.4	83.8
2	mCRPC (cl)	1025	40	925	0.12	0.89	0.81	12.3	0.9	16.1	15.2
3	eNSCLC	2338	80	2138	0.10	0.82	0.59	33.1	1.6	37.0	35.4
4	COVID-19	5721	80	5521	0.20	0.72	0.58	32.7	1.6	93.4	91.8
5	DLBCL	3386	80	3186	0.24	0.53	0.71	23.0	1.6	54.4	52.9
6	SCLC	10044	80	9844	0.07	0.92	0.78	15.5	1.6	165.4	163.8
7	mNSCLC (1L)	17242	82	16962	0.17	0.89	0.63	27.0	1.6	284.1	282.5
8	eNSCLC (non-RCT)	702	80	532	0.34	0.83	0.58	26.7	1.6	10.2	8.6
9	eNSCLC (RCT)	519	80	319	0.65	0.84	0.57	17.9	1.6	6.7	5.1
10	mCRPC (eco)	1126	40	926	0.12	1.00	0.77	15.8	0.9	16.1	15.2

Roche's Journey

Where It Started and Why

Like many organisations, Roche did not begin with a fully formed strategy for AI in evidence synthesis. We began with a problem – and a growing sense that the tools available to solve it were changing faster than our methods for evaluating them.



The First Decision

How Does an Organisation Adopt AI for Evidence Synthesis?

Buy

- Procure commercially available tools validated for specific review tasks
- Faster to deploy – but dependent on vendor decisions, update cycles, and validation evidence you do not control
- Dependent on vendor's methodological decisions and update cycles
- Governance and validation evidence must be obtained from the vendor
- Best suited when: commercial tools meet methodological requirements and can be validated against internal reference standards
- The landscape is crowded: 10+ tools now exist, ranging from free research previews to enterprise platforms at $125K+/year

Hybrid

- Combine commercial tools for some steps with internally developed capabilities for others
- Allows flexibility to start quickly while building towards more tailored solutions
- Requires clear governance of the boundary between bought and built components
- Best suited when: the review workflow has both standard and highly specialised components
- What Roche ultimately arrived at – not by design, but by necessity
- We attempted to build all components ourselves. We had to pivot – adopting external tools for specific steps such as data extraction, where commercial solutions outperformed what we could build at pace

Build

- Develop internal AI capabilities and tools tailored to specific review workflows
- Requires data science expertise, infrastructuare, and sustained investment
- Full methodological control – but significant time, investment, and skills required
- Long lead time before value is realised
- Supports scalability, reduced cost over time and connectivity of data to Roche infrastructure
- Best suited when: existing tools do not meet specific methodological requirements, or when proprietary data or workflows need to be protected
- KiaKia (Roche's internal tool) began here – an ambition to build all components in-house

KiaKia

Conducting Due Diligence in a Rapidly Evolving Landscape

The challenge Roche faced:

The technology was evolving faster than our ability to evaluate it. A tool we assessed in Q1 was meaningfully different by Q3 – new features, new models, new pricing. Due diligence is not a one-time exercise in this space. It is an ongoing one.

What structured due diligence required:

- Assessment against consistent criteria – capability existence, not just capability claims
- Distinguishing vendor-reported performance from independently validated performance
- Evaluating organisational fit – data classification requirements, security certifications, governance compatibility
- Internal sign-off processes – in a large organisation, getting a new tool approved for use in a submission requires navigating technology governance, legal, compliance, and procurement in parallel
- Periodic re-evaluation – as tools update and the landscape shifts

A key insight

For our due diligence, the capabilities within tools that we encountered did not necessarily show that the features performed at a threshold needed for an HTA submission. Those are very different questions – and the gap between them is where most organisations get into difficulty.

The Barriers We Did Not Anticipate

Some We Did

Challenges we anticipated:	Challenges we did not fully anticipate:
Stakeholder scepticism about AI-generated outputs	**No benchmark:** There was no agreed performance standard against which to validate AI tools for HTA-specific review tasks. Academic benchmarks existed – but they did not reflect the specific document types, PICO structures, and comparator complexity of HTA SLRs
Methodological complexity of integrating AI into a validated SLR workflow	**No data:** Building or fine-tuning AI tools requires training data. HTA SLR datasets are proprietary, inconsistently structured, and not publicly available
Absence of clear methodological guidance on reporting requirements	**Methodological gap:** Guidance on how to report AI use in HTA submissions was – and largely remains – underdeveloped. We were making methodological decisions in a guidance vacuum
Resource and capability requirements for implementation	**Skills mismatch:** The skills required to develop, validate, and govern AI tools in this context sit at the intersection of data science, evidence synthesis methodology, and HTA expertise. That combination is rare

The Benchmark Problem

Publishing Our Way to a Solution

The problem we encountered:

When we began validating AI tools for HTA SLR tasks, we found no agreed benchmark – no reference standard specific to HTA review types against which AI performance could be assessed.

The problem we encountered:

- Developed primarily for clinical and biomedical research contexts
- Used datasets not representative of HTA PICO structures, comparator complexity, or document types
- Performance claims from vendors could not be meaningfully compared or independently verified

Our response – the IRR paper:

Rather than developing a fully proprietary internal benchmark and moving on, we worked with collaborators to publish on inter-reviewer reliability (IRR) in human SLRs as a comparator framework. (Hanegraaf et al., BMJ Open, 2024)

The rationale:

- If human reviewers do not achieve perfect agreement, AI should not be held to a standard humans themselves do not meet
- IRR provides a meaningful, measurable, and published comparator against which AI performance can be evaluated
- This work opened a field-level conversation about what a relevant performance standard for AI in SLRs should look like

What Are Reviewers Willing to Trade Off?

A Question We Had to Answer

The gap we identified:

Beyond the benchmark problem, a second question had no published answer: When adopting AI for SLR screening, what trade-offs are reviewers actually willing to make – between accuracy, speed, cost, and transparency?

Our response – the discrete choice experiment:

We designed and published a discrete choice experiment (DCE) with 187 participants to formally investigate adoption preferences. (Abogunrin et al., BMJ Open, 2025)

Key findings:

- Accuracy was the dominant preference – reviewers prioritise getting it right over speed or cost savings
- 55.6% of participants were unfamiliar with AI at the time of the study
- 15.5% said they would never seek AI assistance – indicating meaningful adoption resistance
- Most participants were academically employed, suggesting different adoption dynamics in industry
- Transparency of the AI tool was a significant factor in willingness to adopt

Why this matters for HTA:

If accuracy is what reviewers prioritise above all else – and they are right to do so in the HTA context – then the field's validation and reporting requirements need to reflect that hierarchy. Speed and cost savings are secondary benefits, not the primary justification for adoption.

A Deliberate Strategy

Publishing Gaps, Not Just Solutions

The principle:

When we encountered a methodological barrier that the field did not have a published answer to, we chose to surface it – through peer-reviewed publication, conference presentation, and collaborative working – rather than solving it internally and moving on.

Gap identified	Our response	Publication
No quantification of AI efficiency in SLRs	Pragmatic review of 25 studies	Frontiers in Pharmacology, 2025
No benchmark for AI screening performance in HTA SLRs	IRR as comparator framework	BMJ Open, 2024
No data on reviewer trade-off preferences for AI adoption	Discrete choice experiment, 187 participants	BMJ Open, 2025
No systematic map of AI tool adoption across health-related SLRs	SLR of 112 articles, 111 tools	IJTAHC, 2026

The outcome:

- Each publication contributed a reusable methodological resource to the field
- The work attracted collaborators, accelerated guideline development conversations, and created a basis for HTA body engagement
- It also built credibility – demonstrating that Roche's approach to AI in evidence synthesis was methodologically grounded, not merely commercially driven

Turning Barriers into Contributions

A Deliberate Choice to Share

Our approach when we hit a wall:

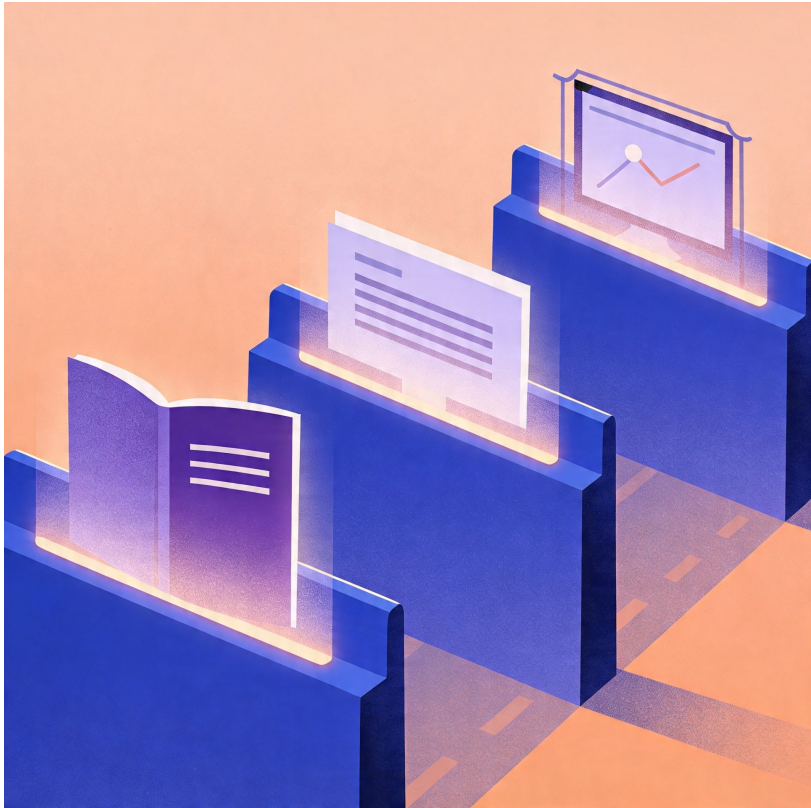
Rather than treating methodological barriers as internal problems to solve quietly, we chose to surface them – through publications, conference presentations, and collaborative working groups.

Examples:

- The absence of a validated benchmark for AI screening in HTA SLRs → led to work on inter-rater reliability as a proxy validation approach
- The gap in reporting guidance → contributed to emerging frameworks for minimum reporting requirements in AI-assisted reviews
- The skills and governance challenge → informed our thinking on the role of external SMEs alongside internal AI development capability

Why this matters beyond Roche:

"If we solved these problems only for ourselves, the field would face them again and again. Publishing our barriers – not just our solutions – was a deliberate methodological contribution."



Getting the Capability Mix Right

What KiaKia Taught Us About Skills

What building KiaKia required — and what we underestimated:

- Data science capability to develop and validate AI components
- Evidence synthesis expertise to ensure methodological defensibility
- HTA domain knowledge to understand what submissions actually require
- Technology governance to navigate internal approval and compliance processes
- External SME input – to provide independent methodological credibility and check internal assumptions

Where the hybrid approach changed the equation:

- Building all components internally: high control, high resource demand, high risk of pace mismatch with commercial development
- Adopting external tools for specific steps (e.g. data extraction): faster, often better validated – but requires vendor due diligence, integration governance, and ongoing monitoring as tools update
- The boundary between built and bought components must be clearly governed – and documented for submissions

The practical lesson:

We attempted to build everything.
We learned that external tools, used with appropriate validation and governance, could outperform our internal builds for specific steps – particularly data extraction.
The decision to adopt external tools was not a retreat. It was a more honest allocation of our internal capability.

Honest Reflections

What We Learned and What We Would Change

What worked:

- Starting with well-defined, lower-risk steps – screening assistance before extraction automation
- Publishing barriers openly – it accelerated external collaboration and shaped field-level thinking
- Investing in inter-rater reliability as a proxy for benchmark validation
- Bringing external SMEs in early, not as reviewers of finished work but as co-designers

What we would do differently:

- Start the governance conversation earlier – we underestimated how much organisational alignment was required before the first tool went near a submission
- Invest in documentation infrastructure from day one – retrofitting audit trails is significantly harder than building them in
- Be more explicit about the distinction between Stage 1 and Stage 2 automation in our internal workflows – the boundary blurred in practice in ways that created methodological uncertainty
- Engage with HTA bodies earlier and more directly – the conversations we have had with bodies about what they want to see in submissions have been more productive than we expected

The overarching lesson:

AI adoption in evidence synthesis is not primarily a technology problem. It is a methodology, governance, and communication problem. The technology is the easiest part.

From Roche's Journey to a Framework You Can Use

Every organisation's journey will be different. The barriers you face, the tools available to you, and the HTA bodies you are submitting to will shape your path in ways that our experience cannot fully anticipate.

What we can offer are the questions worth asking – and a framework for what HTA-compliant AI use needs to look like, regardless of where you are starting from.

What Does HTA-Compliant AI Use Look Like?

What Does HTA-Compliant AI Use Actually Mean?

The question is not simply whether AI was used in an SLR. The question is whether the review – taken as a whole – remains reproducible, transparent, auditable, and methodologically defensible.

Four anchoring principles:

Reproducibility

Can the process be replicated and yield materially the same result?

Transparency

Is the role of AI documented at every step where it was used?

Human oversight

Where and how were human reviewers involved in consequential decisions?

Auditability and governance

Can every methodological choice be traced, explained, and defended?

Reproducibility

The Most Fundamental Requirement

In a traditional SLR, reproducibility means

- A documented, replicable search strategy
- Pre-specified eligibility criteria applied consistently
- Independent dual screening with disagreement resolution
- Structured data extraction against a pre-defined template

In an AI-assisted SLR, reproducibility additionally requires

- Documentation of the AI tool used, including version number
- The prompt, parameters, or configuration applied at each automated step
- Evidence that the tool performs consistently across runs — or characterisation of variability if it does not
- A record of the training data or fine-tuning approach used, where relevant
- Documentation of any tool updates that occurred during the review period

The emerging challenge:

Some generative AI tools produce outputs that vary across runs even with identical inputs. This is not disqualifying – but it must be characterised, documented, and accounted for in the methods section.

Transparency

Documenting the Role of AI at Every Step

Minimum transparency requirements by stage of review:

SLR Stage	If AI Used	What Must Be Documented
PICO simulation	AI-assisted simulation	Tool used, disclosure of AI use, prompts applied, human review of final PICO set(s)
Search strategy development	AI-assisted query generation	Tool used, disclosure of AI use, prompts applied, human review of final strategy
Title/abstract screening	AI-assisted or automated	Tool, disclosure of AI use, criteria applied, version, configuration, validation approach, recall rate, human adjudication process
Full-text review	AI-assisted or automated	Tool, disclosure of AI use, criteria applied, human override rate, human adjudication process
Data extraction	AI-assisted or automated	Tool, disclosure of AI use, extraction template, validation against human extraction, error rate
Evidence synthesis	AI-assisted synthesis	Tool, disclosure of AI use, parameters, human review of synthesised output, citation verification
Reporting	AI-assisted drafting	Tool, disclosure of AI use in drafting, human review and approval of final text

Human Oversight

Where It Is Non-Negotiable

The human-in-the-loop imperative:

Human oversight in an AI-assisted review is not about having a human present at every step. It is about ensuring that consequential decisions – those that materially affect the evidence base – are subject to human review and accountability.

Where human oversight is currently non-negotiable:

- Protocol development and eligibility criteria definition
- Final search strategy approval
- Adjudication of uncertain or borderline screening decisions
- Quality and risk of bias assessment of included studies
- Interpretation of synthesised findings
- Review and approval of the submission document

Where human oversight may be reduced with appropriate validation:

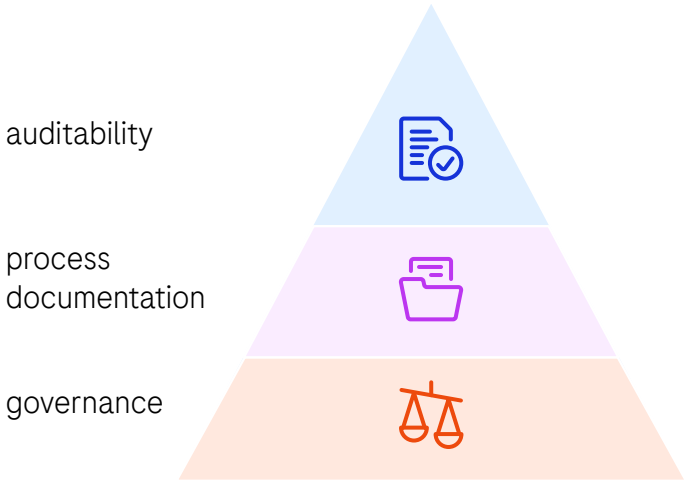
- First-pass screening of clearly irrelevant records
- De-duplication of search results
- Extraction of highly structured, well-defined data fields
- Formatting and organisation of extracted data

The key question for every automated step:

"If this step produces an error, will it be caught before it affects the final evidence base – and can I demonstrate that?"

Auditability and Governance

Building the Infrastructure for Trust



The governance question most organisations need to answer:

"If an HTA body questions a decision made by an AI tool in your submission – who in your organisation is accountable for that decision, and what is your response process?"

What auditability requires:

- A complete log of AI tool use across the review – tool, version, configuration, date of use
- Validation evidence for each automated step – performance metrics against a reference standard
- Documentation of human review decisions and any overrides of AI outputs
- A clear record of who approved each stage of the review and on what basis
- Version control for any AI tool that was updated during the review period

What governance requires:

- Organisational sign-off on which AI tools are approved for use in HTA submissions
- Clear roles and responsibilities for AI-assisted review teams
- A defined escalation pathway when AI outputs are uncertain or contested
- Regular review of approved tools as the technology and guidance landscape evolves

Are tools we have access to today ready to support HTA submissions?

The Commercial Tool Landscape

Crowded, Uneven, and Evolving Fast

Key findings from Roche's due diligence of fourteen tools – across 11 capability categories:

- No tool scored high on HTA acceptance – field-wide gap
- AI transparency is one of the weakest dimensions, particularly for regulatory audit trails and mandatory human oversight
- Independent validation is rare – most accuracy claims remain vendor-reported

- GRADE assessment: absent in every tool assessed
- Economic data extraction templates: pre-built templates only available in two tools (Nested Knowledge, Laser AI)
- Cost ranges widely – from free (ActiveSLR) to ~$125K-$167K/year (Nested Knowledge Enterprise); several tools do not disclose pricing

Commercial Tool Landscape Assessment

Methods Overview

Tools assessed:

ActiveSLR, DistillerSR, EasySLR, Elicit, Laser AI, Loon Lens, MadeAI, Nested Knowledge, otto-SR, Rayyan, scholara.ai, ScholarIQ, Scite, TERA.

Dimensions:

41 items mapped to 11 umbrella categories across two domains:*

Domain	Umbrella categories	Rating scales
Functional capabilities	**1.** Project setup **2.** Search **3.** Screening **4.** Extraction & appraisal **5.** Synthesis & reporting **6.** Data portability	Yes, Partial, No
Adoption & operations	**7.** Industry adoption **8.** AI transparency **9.** Independent validation **10.** User experience **11.** Total cost of ownership	High, Medium, Low

Data collection and analysis:

Step 1: Web-based research

- Ratings are grounded in vendor documentation, product demos, peer-reviewed literature, conference proceedings, and independent validation studies where available. Information was retrieved from PubMed, bioRxiv/medRxiv, ClinicalTrials.gov, and vendor websites.
- Sources were categorized a four-tier hierarchy (independent, peer-reviewed > vendor-affiliated, peer-reviewed > vendor-affiliated, non-peer-reviewed > grey literature). Higher-tier evidence was weighted more heavily in the rating.
- Claude (Anthropic) was used for retrieval, extraction, and drafting under human review.

Step 2: Vendor-reported data

- Vendor self-ratings for each of the 41 items were solicited by email; responses were received for 10 of the 14 tools.
- Where possible, vendor self-ratings were verified against web-based research.

Two versions of the results were generated:

1. Ratings based on web-based research and verified vendor self-ratings.
2. All vendor responses accepted, regardless of verification.

* Dimensions informed by: Flemyng et al. 2025 (Cochrane/Campbell/JBI/CEE joint position); Fleurence et al. 2025 (ELEVATE-GenAI); Thomas et al. 2026 (RAISE 1-3).

Commercial Tool Landscape Assessment

Results: Web-based research and verified vendor self-ratings

Domain / Category	ActiveSLR	DistillerSR	EasySLR	Elicit	Laser AI	Loon Lens	MadeAi	Nested Knowledge	otto-SR	Rayyan	scholara.ai	ScholarIQ	Scite	TERA
Functional capabilities														
1. Project setup	Partial	Yes	Yes	Partial	Yes	Partial	Yes	Yes	Partial	Yes	Partial	Partial	Partial	Partial
2. Search	Partial	Yes	Yes	Partial	Yes	Partial	Yes	Yes	Partial	Yes	Partial	Yes	Partial	Yes
3. Screening	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
4. Extraction & appraisal	Partial	Yes	Partial	Partial	Yes	Partial	Partial	Yes	Yes	Partial	Yes	Partial	No	Partial
5. Synthesis & reporting	No	Yes	No	Partial	No	Partial	Partial	Yes	Partial	No	Partial	Partial	No	Partial
6. Data portability	Partial	Yes	Partial	Partial	Partial	Partial	Partial	Partial	Partial	Partial	Partial	Partial	Partial	Partial
Adoption & operations														
7. Industry adoption	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Low	Medium	Low	Medium
8. AI transparency	Low	High	Medium	Low	Medium	Medium	High	High	Medium	Medium	Medium	Medium	Low	Medium
9. Independent validation	Low	Medium	Medium	High	Low	Medium	Medium	Medium	Medium	High	Medium	Medium	Low	Medium
10. User experience	Medium	High	Medium	High	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
11. Total cost of ownership	High	Low	High	High	Medium	Low	Medium	High	Low	Medium	Low	Medium	Medium	Low

Capabilities are rated Yes, Partial, or No for whether a feature is available. Adoption and operational dimensions are rated High, Medium, or Low for maturity and evidence strength.
Umbrella-level ratings roll up from their sub-rows using plurality (the most common rating wins; ties resolve to the middle value).
At time of analysis, vendor responses had not been received for ActiveSLR, Rayyan, Scite, or TERA; For these platforms all ratings are based on web research.

Commercial Tool Landscape Assessment

Results: All vendor responses accepted, regardless of verification

Domain / Category	ActiveSLR	DistillerSR	EasySLR	Elicit	Laser AI	Loon Lens	MadeAi	Nested Knowledge	otto-SR	Rayyan	scholara.ai	ScholarIQ	Scite	TERA
Functional capabilities														
1. Project setup	Partial	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Partial	Partial
2. Search	Partial	Yes	Yes	Partial	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Partial	Yes
3. Screening	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
4. Extraction & appraisal	Partial	Yes	Partial	Partial	Yes	Yes	Yes	Yes	Yes	Partial	Yes	Yes	No	Partial
5. Synthesis & reporting	No	Yes	Partial	Partial	Yes	Yes	Partial	Yes	Partial	No	Yes	Partial	No	Partial
6. Data portability	Partial	Yes	Partial	Partial	Yes	Yes	Yes	Partial	Yes	Partial	Partial	Yes	Partial	Partial
Adoption & operations														
7. Industry adoption	Medium	High	Medium	Medium	High	Medium	Medium	High	Medium	Medium	Medium	Medium	Low	Medium
8. AI transparency	Low	High	Medium	Medium	Medium	Medium	High	High	Medium	Medium	Medium	Medium	Low	Medium
9. Independent validation	Low	High	Medium	High	High	Medium	Medium	High	Medium	High	Medium	High	Low	Medium
10. User experience	Medium	High	High	High	High	Medium	High	High	Medium	Medium	High	High	Medium	Medium
11. Total cost of ownership	High	High	High	High	High	Low	High	High	Low	Medium	High	High	Medium	Low

Capabilities are rated Yes, Partial, or No for whether a feature is available. Adoption and operational dimensions are rated High, Medium, or Low for maturity and evidence strength.
Umbrella-level ratings roll up from their sub-rows using plurality (the most common rating wins; ties resolve to the middle value).
At time of analysis, vendor responses had not been received for ActiveSLR, Rayyan, Scite, or TERA; For these platforms all ratings are based on web research.

Commercial Tool Landscape Assessment

Functional Capabilities: Dimensions and Rating Criteria

Category	
1. Project setup	**Project management features** ● Yes: Multi-user workspace with role-based permissions, progress dashboards, and project-level configuration. Supports team collaboration workflows ● Partial: Basic project workspace exists but limited permissions, no dashboards, or no multi-user collaboration features ● No: No project management functionality; single-user or file-based workflow only
	Protocol development ● Yes: Dedicated protocol module with structured fields, PROSPERO-compatible export or similar, linkage to downstream workflow settings ● Partial: Structured workflow that could support a protocol but no standalone module, no versioning, no registry export ● No: No protocol functionality
	User support functionality ● Yes: Multiple support channels (email + live chat or phone). Documentation/help centre. Training or onboarding programmes available ● Partial: Email support or documentation exists but no live chat/phone. No formal training programmes ● No: No documented support channels or help resources
2. Search	**Direct database searching** ● Yes: Built-in search interface querying at least one major bibliographic database (PubMed, Embase, Cochrane) with Boolean/MeSH support ● Partial: Connects to databases but with limited query syntax (no Boolean/MeSH), or only partial database coverage ● No: No in-platform search; users must search externally and import results
	RIS/file import ● Yes: Supports standard citation file imports (RIS, BIB, NBIB, or CSV) from multiple reference managers and databases. Tracks source database and search date ● Partial: Supports at least one citation format but with limitations (e.g., limited field mapping, no source tracking, only one format accepted) ● No: No file import capability
	Other import methods ● Yes: Supports additional import methods beyond file upload: reference manager integration (Zotero, EndNote), mining from existing reviews, or API-based ingestion ● Partial: One additional import method available but with significant limitations or narrow scope ● No: No import methods beyond basic file upload (if file upload exists)
	Deduplication ● Yes: Automated detection and merging of identical and near-identical citations. Logs all duplicates and allows the user to manually override AI decisions regarding duplicate handling ● Partial: Automated duplicate detection of identical records but not near-identical records, deduplication that requires manual intervention or review, or deduplication that does not allow the user to override AI decisions ● No: No automated or semi-automated deduplication; user must manually identify and deduplicate records

Commercial Tool Landscape Assessment

Functional Capabilities: Dimensions and Rating Criteria

Category	
3. Screening	**Title/abstract human screening** ● Yes: Structured title/abstract screening with inclusion/exclusion decisions. Supports dual screening with conflict resolution. Supports recording exclusion reasons ● Partial: Basic screening functionality exists but missing dual screening, conflict resolution, or recording exclusion reasons ● No: No in-platform screening
	Full-text human screening ● Yes: Full-text screening with inclusion/exclusion decisions, dual-reviewer support, conflict resolution, and full-text retrieval or linking ● Partial: Full-text screening exists but without dual-reviewer support, or no integrated full-text retrieval/linking ● No: No full-text screening functionality
	AI-assisted screening (title/abstract + full-text) ● Yes: ML/AI actively assists screening decisions (prioritisation, recommendation, auto-exclude) with documented accuracy metrics ● Partial: AI provides suggestions but accuracy/recall is not documented, or AI is limited to ranking only ● No: No AI involvement in screening
4. Extraction & appraisal	**Study linking** ● Yes: Can group publications from the same study (e.g., clinical study report + article + abstract from the same trial) and propagate study-level metadata across grouped publications ● Partial: Publications grouped but with significant limitations (e.g., study-level metadata does not propagate across publications, or data extraction occurs entirely at the publication level) ● No: No study linking functionality
	Clinical data extraction templates ● Yes: Pre-built extraction templates for clinical evidence; extraction can handle multi-arm, multi-timepoint (baseline and >1 follow-up timepoint) data structures ● Partial: Customisable extraction fields for clinical data, but no pre-built templates. May not support multi-arm or multi-timepoint structures natively ● No: No capability to build clinical data extraction templates
	Economic data extraction templates ● Yes: Pre-built templates for economic evidence (utility values, costs, resource use); extraction can handle multi-arm, multi-timepoint (baseline and >1 follow-up timepoint) data structures ● Partial: Customisable extraction fields that could be configured for economic data, but no pre-built HEOR templates. May not handle cost/utility data structures natively ● No: No capability to build economic data extraction templates
	AI-assisted data extraction ● Yes: AI extracts data with documented accuracy, supporting citations, and human review/override ● Partial: AI suggests or highlights data but extraction is primarily manual, or accuracy is not documented ● No: No AI involvement in extraction

Commercial Tool Landscape Assessment

Functional Capabilities: Dimensions and Rating Criteria

Category	
4. Extraction & appraisal (continued)	**Data validation** ● Yes: Automated validation rules (range checks, consistency checks, duplicate detection); built-in differencing/comparison between extraction versions; platform supports dual extraction with adjudication; auto-adjudication is available (not all conflicts be resolved manually) ● Partial: Dual extraction or manual review supported, but no automated validation rules ● No: No validation infrastructure
	Critical appraisal ● Yes: AI-assisted appraisal against at least one standard framework (RoB, JBI, etc.) with reviewable/editable outputs ● Partial: Manual template-based appraisal (platform provides structured forms but no AI assistance) ● No: No built-in appraisal functionality
5. Synthesis & reporting	**PRISMA diagram** ● Yes: Auto-generated PRISMA flow diagram (2020 or 2009 format) from workflow data. Configurable, editable, and exportable at publication quality ● Partial: Simplified flow summary or mini-PRISMA included in reports but not a full PRISMA 2020 standard diagram. May require manual reformatting ● No: No PRISMA diagram generation
	Report tables ● Yes: Tables for study characteristics, patient characteristics, and outcomes are automatically generated from extracted data. Configurable, editable, and exportable ● Partial: Data can be exported to build tables externally, or basic tabular views exist but are not configurable or editable ● No: No report table generation
	Report figures ● Yes: Auto-generated statistical visualisations (forest plots, funnel plots, network diagrams) ● Partial: Conceptual/qualitative figures, PRISMA diagrams only, or integration with external BI tools ● No: No figure generation
	Quantitative synthesis ● Yes: Fully automated in-platform meta-analysis (pairwise or NMA) that is directly linked to extracted data (e.g., forest plots, pooled effect sizes generated from extraction fields) ● Partial: Meta-analytic capability available in-platform but not automatically linked to extracted data; requires manual configuration or data re-entry ● No: No in-platform meta-analytic capability; quantitative synthesis requires export to external tools (R, Stata, RevMan)
	Report writing ● Yes: Integrated manuscript editor or AI-assisted narrative report generation with structured sections (methods, results, discussion). In-app editing ● Partial: AI can draft summaries or partial report sections, but no integrated manuscript editor. Or report generation exists but is limited to specific sections only ● No: No report writing functionality; users must write reports externally using exported data

Commercial Tool Landscape Assessment

Functional Capabilities: Dimensions and Rating Criteria

Category	
5. Synthesis & reporting (continued)	**GRADE assessment** ● Yes: Automated workflow covering all five GRADE domains, with support for per-outcome ratings. Exportable Summary of Findings or Evidence Profile tables ● Partial: Some functionality (e.g., a template, partial domain coverage, or manual-only assessment) ● No: No structured GRADE tool
	Living review / update support ● Yes: Built-in support for living reviews: automated search monitoring, incremental screening (new records only), and version tracking between updates ● Partial: Some update functionality (search re-run, new records identified) but no incremental screening, no automated monitoring, or no version tracking ● No: No living review or update functionality; updating requires starting a new project or re-running all steps
6. Data portability	**Data export** ● Yes: All data (screening decisions, extracted data, appraisal) exportable as CSV/Excel. Exports configurable at study-level, arm-level, or intervention-level. Format usable in R/Stata/Excel ● Partial: Some data exportable but not all stages, or limited format options (e.g., only CSV, no arm-level granularity) ● No: No structured data export; data is locked in the platform
	API & integration ● Yes: Documented REST API or equivalent for programmatic data retrieval. Integrations with reference managers, NMA tools, or economic modelling software ● Partial: API exists but poorly documented, or integrations are limited to basic reference managers only. No HEOR-specific integration examples ● No: No API or programmatic integration capability

Commercial Tool Landscape Assessment

Adoption & Operations: Dimensions and Rating Criteria

Category	
7. Industry adoption	**HEOR relevance** ● High: Documented use in HEOR-specific SLRs (pharmacoeconomic, cost-effectiveness, or budget impact reviews). Named pharma/HEOR clients or ISPOR case studies ● Medium: Platform is positioned for HEOR or pharma but no documented HEOR-specific SLR use cases. General life sciences adoption only ● Low: No evidence of HEOR or pharma use. Platform focused on general research, clinical, or academic use only
	HTA acceptance ● High: Platform methods positively referenced in HTA guidance or used in an accepted HTA submission. Or HTA body explicitly names the platform as acceptable ● Medium: Methods align with HTA guidelines AND active engagement with HTA processes (consultations, HTAi presentations, HTA-adjacent guidance citations) ● Low: No alignment with HTA methods and no evidence of engagement with HTA bodies or processes
8. AI transparency	**Regulatory audit trail** ● High: Tamper-proof audit logs of all user actions, exportable, GxP/21 CFR Part 11/Annex 11 compatible ● Medium: Some audit logging exists but not exportable, not tamper-proof, or not documented as GxP-compatible ● Low: No audit trail functionality, or limited to login history only
	AI disclosure & documentation ● High: Clearly discloses AI model types, versions, and developer. Provides rationale or provenance for AI outputs. Exportable audit logs of AI decisions. Documentation enables users to complete RAISE reporting template and ELEVATE-GenAI checklist (model characteristics, accuracy, factuality, reproducibility, calibration/uncertainty domains) ● Medium: Discloses some AI information (e.g., model type or a general description of AI functions) but lacks version visibility, detailed provenance, or auditability of outputs ● Low: Does not disclose model type, model version, or rationale for outputs. Does not provide a mechanism for users to audit AI decisions
	Human oversight capability ● High: Human review mechanisms exist at every AI-assisted step. Confidence scores, uncertainty flags, or stopping rules provided ● Medium: Human review available at some AI-assisted steps but not all. Limited confidence scoring ● Low: No documented human review mechanisms. AI outputs cannot be reviewed or overridden
	Human oversight enforcement (default-on) ● High: Platform MANDATES human review at every AI step. Cannot bypass. Audit logged ● Medium: Default workflow but can be bypassed. Encourages but does not enforce review ● Low: No enforcement. AI outputs automatically accepted

Commercial Tool Landscape Assessment

Adoption & Operations: Dimensions and Rating Criteria

Category	
9. Independent validation	**Independent accuracy validation** ● High: At least one peer-reviewed study by authors independent of the vendor validates the platform's accuracy claims (screening recall, extraction accuracy) ● Medium: Peer-reviewed validation exists but all authors are affiliated with the vendor (founders, employees, consultants). Or independent evaluation is planned/in progress ● Low: No peer-reviewed validation of any kind; all accuracy claims are from vendor marketing materials, website, or conference posters only
	Peer-reviewed evidence base ● High: Multiple peer-reviewed publications about the platform in recognised journals (methods papers, validation studies, or comparative evaluations) ● Medium: One peer-reviewed publication exists, or publications are in non-indexed or low-impact venues only ● Low: No peer-reviewed publications about the platform
	Time efficiency evidence ● High: Independent peer-reviewed time-savings study with documented methodology ● Medium: Vendor-authored peer-reviewed study or white paper with methodology ● Low: Marketing claims only or no time-savings claims
10. User experience	**Reference customers & case studies** ● High: ≥3 named reference customers OR ≥2 published case studies with outcomes ● Medium: 1–2 named customers or brief testimonials ● Low: No named customers or case studies
	Onboarding & learning curve ● High: Self-service onboarding (guided tutorials, templates, in-app help) enabling productive use within 1–2 days. Formal training available but not required ● Medium: Moderate learning curve requiring formal training or significant documentation review. Platform is usable but not intuitive for new users ● Low: Steep learning curve with no onboarding resources. Requires extensive training or technical expertise to use effectively
	Known UX limitations ● High: No significant UX pain points reported in reviews or user forums. Interface is modern and responsive ● Medium: Some UX issues reported (e.g., slow performance, unintuitive navigation, limited mobile support) but core functionality is usable ● Low: Significant UX problems widely reported; platform usability is a documented barrier to adoption

Commercial Tool Landscape Assessment

Adoption & Operations: Dimensions and Rating Criteria

Category	
11. Total cost of ownership	**Estimated annual cost (15–20 users)** ● High: Publicly estimable AND competitive. Published pricing tiers ● Medium: Partially public. Rough estimate possible ● Low: No pricing info. Fully contact-us
	Pricing model & transparency ● High: Clear and fully documented pricing model ● Medium: Partially documented. Enterprise pricing requires sales contact ● Low: Completely opaque pricing model
	Total cost predictability ● High: Predictable total cost. No significant hidden costs ● Medium: Some additional costs but estimable ● Low: Significant unpredictable costs
	Implementation & migration effort ● High: Documented onboarding programme, data migration tools/services, clear implementation timeline. Migration from existing tools straightforward ● Medium: Some onboarding support but no formal migration pathway. Significant internal effort required ● Low: No migration support, no onboarding programme, no documentation of implementation requirements

Commercial Tool Landscape Assessment

Functional Capabilities: Key Takeaways

Category	
1. Project setup	• Multi-user collaboration with role-based permissions and project-level configuration is widespread. Differentiation sits in progress dashboards and enterprise governance, where the enterprise platforms pull ahead of the research-assistant tools.
2. Search	• Built-in Boolean/MeSH-capable search against PubMed/Embase/Cochrane is a clear fault line: enterprise platforms (DistillerSR, Nested Knowledge) offer it; several AI-screening tools (Loon Lens, EasySLR) explicitly do not. • Reference-manager integration (Zotero, EndNote) and API-based ingestion separate the platforms with engineering depth from those anchored to file upload. • Automated identical-and-near-identical deduplication with audit logs is effectively universal; near-identical/fuzzy matching and user-override of AI decisions are where some tools are differentiated.
3. Screening	• Structured title/abstract screening with dual-reviewer support, conflict resolution, and exclusion-reason capture is universal. • Full-text screening with dual-reviewer workflows is universal; integrated full-text retrieval/linking varies.
4. Extraction & appraisal	• Every vendor cites AI involvement, but only a subset backs it with published sensitivity/specificity numbers against independent benchmarks. • Pre-built clinical templates are common; handling multi-arm, multi-timepoint structures natively is the stricter bar that separates LLM extractor tools from the purpose-built SLR platforms. • Pre-built economic extraction templates (utility values, costs, resource use with multi-arm, multi-timepoint structure) are the weakest capability.
5. Synthesis & reporting	• Auto-generated PRISMA 2020 diagrams from workflow data are common across the enterprise platforms; the distinction is whether the diagram is configurable/editable/ exportable at publication quality. • Fully automated in-platform meta-analysis directly linked to extracted data is rare; several vendors state that it is on the roadmap. • Integrated manuscript editors with structured section support are rare. • No vendor demonstrates a fully automated five-domain GRADE workflow with per-outcome ratings and exportable Summary-of-Findings tables. • Automated search monitoring + incremental screening + version tracking is patchy; some vendors state that it is forthcoming.
6. Data portability	• Structured exports (CSV/Excel) of screening decisions, extracted data, and appraisal are universal; granularity (study-level, arm-level, intervention-level) differentiates the enterprise platforms. • Documented REST APIs with HEOR-specific integrations (NMA tools, economic modelling) are rare.

Commercial Tool Landscape Assessment

Adoption & Operations: Key Takeaways

Category	
7. Industry adoption	• Named pharma customers and documented HEOR-specific SLR use cases are rare; Partnerships and case studies by CROs and consultancies are more common. • Evidence of tools referenced in HTA guidance or used in accepted HTA submissions is rare.
8. AI transparency	• Tamper-proof, exportable audit logs with GxP / 21 CFR Part 11 / Annex 11 compatibility are claimed by a narrow subset of the enterprise vendors; most others offer login-history-level logging. • Disclosing AI model types, versions, developer, and providing rationale/provenance sufficient for a RAISE / ELEVATE-GenAI reporting pass is a bar the market is only beginning to reach. • Human review mechanisms at every AI-assisted step with confidence scores or uncertainty flags are common. However, mandatory, non-bypassable human review at every AI step with audit logging is rare; most vendors encourage but do not enforce review.
9. Independent validation	• Peer-reviewed validation by authors independent of the vendor is exceedingly rare. • Model version drift (model or version changes) creates challenges for accuracy validations.
10. User experience	• High-volume vendors uniformly report no significant UX pain points.
11. Total cost of ownership	• Publicly posted pricing tiers are the exception; most vendors are partially public or sales-contact only. At the enterprise end, annualised cost is opaque. • Cost predictability is hampered by usage-based components (API usage, compute tiers) which only emerge post-contract. • Documented onboarding programmes, data migration tools/services, and clear implementation timelines are claimed by the enterprise platforms; self-service tools rely on self-onboarding and do not offer migration pathways per se.

The Guidance Landscape — and the Validation Gap

Available guidance and frameworks:

- PRISMA-AI — reporting extension for AI use in systematic reviews
- Cochrane AI guidance — framework for responsible AI use in evidence synthesis
- RAISE / ELEVATE-GenAI (ISPOR) — AI disclosure checklists for evidence synthesis
- Individual HTA body guidance — Varies significantly by jurisdiction; some bodies have begun issuing specific expectations; many have not yet done so.
- NICE AI position statement — first HTA body to publish principles for AI in evidence generation; referenced in our Frontiers in Pharmacology paper as a landmark development
- Vienna Principles (ICASR) — foundational principles for automation in systematic reviews

What the tool landscape reveals about the validation gap:

- Most accuracy claims in the field are vendor-reported — not independently validated
- Where independent validation exists, it is often conducted on older model versions (e.g. Elicit validated against GPT-3 when current product uses Claude)
- No tool in the assessed landscape has documented, broad HTA body acceptance
- GRADE assessment — a core HTA requirement — is absent in every commercial tool assessed
- Economic data extraction templates — essential for HEOR submissions — are only available in two tools (Nested Knowledge, Laser AI)

Where gaps remain:

- No agreed minimum reporting standard specifically for AI use in HTA submission dossiers
- No validated, publicly available benchmark dataset for HTA SLR performance evaluation
- Limited guidance on agentic AI in evidence synthesis
- Significant variation in HTA body expectations across jurisdictions
- Formal HTA body acceptance of any specific AI-assisted review approach
- GRADE assessment capability in any commercial tool assessed

The field-level implication:

"The absence of a public benchmark for HTA SLR tasks means every organisation must validate for itself. This is expensive, inconsistent, and does not produce shared knowledge. A public leaderboard — similar to those used in LLM development — using the same datasets could change this. But it requires collective commitment from industry, academia, and HTA bodies."

How to Prepare for a Successful HTA Submission

Preparing for Submission

Starting Earlier Than You Think

The most common mistake organisations could make when using AI in HTA submissions is treating AI governance as something to address after the review is complete. It needs to be addressed before the first tool is switched on

Three questions to answer before you begin:

01

Is your AI tooling approved, documented, and validated for this type of review?

02

Is your team clear on roles, responsibilities, and escalation pathways?

03

Do you know what the receiving HTA body expects – or have you asked?

Tool Selection and Validation

What to Have in Place Before You Start

Before selecting an AI tool for an HTA SLR, establish:

Fitness for purpose

- Has the tool been validated in a context comparable to your review type – HTA-specific documents, PICO structures, therapeutic area?
- What performance metrics are available – sensitivity, specificity, recall – and against what reference standard?

Version control and stability

- What is the current version of the tool, and how frequently is it updated?
- What is the vendor's policy on notifying users of updates that may affect outputs?

Data classification and security

- What data classification does the tool support – is it approved for the sensitivity level of your submission data?
- Where are your data processed and stored?

Validation approach

- How will you validate the tool's performance for your specific review before deploying it in a submission?
- What is your reference standard – and is it sufficiently representative of your review population?

Documentation readiness

- Can the tool generate a complete, exportable log of its decisions and outputs?
- Is that log in a format that can be included in or appended to a submission dossier?

Engaging HTA Bodies

Earlier and More Directly Than You Might Expect

What HTA bodies are increasingly asking about AI use in submissions:

- Was AI used at any stage of the SLR or evidence synthesis process?
- Which tools were used, at which stages, and with what human oversight?
- How was the AI tool validated for this specific review type?
- What is the estimated impact of AI use on the completeness and accuracy of the evidence base?
- Can the AI-assisted process be replicated – and if not, how is variability characterised?

What you should be prepared to answer:

- A clear, stage-by-stage account of AI use – not a blanket disclosure
- Validation evidence specific to your review – not just vendor-supplied metrics
- Documentation of human oversight at each consequential decision point
- An honest characterisation of any limitations introduced by AI use

Several HTA bodies are willing to engage in pre-submission dialogue about AI use in dossiers. In our experience, those conversations are more productive – and less uncomfortable – than discovering a body's expectations at the point of submission.

Designing the AI-Assisted Review Process

What to Specify Before You Start

At each stage of the review (preferably at the protocol-writing stage), specify in advance:

Search

- Which databases will be searched and by what means – manual, API, AI-assisted query generation?
- How will the AI-assisted search be validated against a manual equivalent?

Screening

- At what stage does AI assist – title/abstract, full-text, or both, and how?
- What is the configured sensitivity threshold and what is the evidence base for that threshold?
- How are borderline cases identified and escalated to human review?

Data extraction

- Which data fields are extracted by AI and which by humans?
- What is the validation process for AI-extracted data – full human review, sampled review, or output-level QC?

Synthesis

- Is AI used in any aspect of evidence synthesis or meta-analysis?
- If so, how are AI-generated outputs reviewed, verified, and approved?

Reporting

- Is AI used in drafting any part of the submission document?
- What is the human review and approval process for AI-drafted content?

A Pre-Submission Checklist for AI-Assisted HTA SLRs

01.
Governance and approval

- [ ] AI tools used in this review are organisationally approved for HTA submission use
- [ ] Roles and responsibilities for AI-assisted steps are documented and assigned
- [ ] An escalation pathway exists for contested or uncertain AI outputs
- [ ] Tool version control is documented for the full review period

02.
Validation and performance

- [ ] Each AI tool has been validated against a reference standard relevant to this review type
- [ ] Performance metrics – sensitivity, recall, error rate – are documented for each automated step
- [ ] Variability in AI outputs across runs has been characterised where relevant

A Pre-Submission Checklist for AI-Assisted HTA SLRs

03.
Documentation and transparency

- [ ] A complete AI use log exists covering all tools, steps, configurations, and dates
- [ ] Human review and override decisions are documented at each AI-assisted step
- [ ] The methods section of the submission discloses AI use at each stage — not as a blanket statement but step by step

04.
HTA body engagement

- [ ] The receiving HTA body's expectations on AI use have been reviewed or confirmed
- [ ] Pre-submission dialogue on AI methodology has been considered and pursued where appropriate
- [ ] The submission is prepared to respond to specific methodological questions about AI use

Looking Further Ahead

Does HTA SLR Methodology Need to Evolve?

What AI is enabling today:

- Reduction in time-to-delivery for evidence synthesis
- Increased capacity to handle larger evidence bases
- More consistent application of eligibility criteria at scale
- Earlier identification of relevant evidence in rapidly evolving therapeutic areas

What AI may enable in the near future:

- Continuous, living evidence synthesis updated in near real-time
- AI-driven network meta-analysis and indirect treatment comparisons
- Agentic systems coordinating entire HTA dossier preparation workflows
- Automated benchmarking against publicly available HTA decision datasets

The methodological question the field needs to answer:

Are the methods we use to synthesise evidence – including our statistical approaches, our PICO frameworks, our reporting standards – designed for a world where AI can do in hours what previously took months? And if not, which methods need to evolve, and how?

A concrete example:

Will we still rely on MCMC-based methods for indirect treatment comparisons – or will AI-driven equivalents eventually offer greater speed without methodological compromise? The answer is not yet clear. But the question is live.

Audience Discussion and Survey Questions

Let's Start With a Snapshot of This Room

Current AI use

Are you currently using AI in any part of your systematic review or evidence synthesis process?

A	Yes – actively using in submissions
B	Yes – piloting or evaluating
C	No – considering it
D	No – not currently on our agenda

Let's Start With a Snapshot of This Room

Stage of automation

Which best describes your current approach?

A Stage 1: AI assists, humans review all outputs	**C** Stage 3: Exploring or piloting agentic workflows
B Stage 2: Some steps fully automated with validation	**D** Not yet at any stage

Let's Start With a Snapshot of This Room

Primary barrier

What is the single biggest barrier to AI adoption in your evidence synthesis work?

A	Absence of clear methodological guidance
B	Lack of validated tools for HTA-specific tasks
C	Organisational governance and approval processes
D	Skills and capability gaps within our team
E	Uncertainty about HTA body expectations
F	No significant barriers – we are progressing well

Let's Start With a Snapshot of This Room

Confidence in submission

How confident would you be submitting an AI-assisted SLR to a major HTA body today?

- **A** Very confident – we have done it
- **B** Somewhat confident – with appropriate documentation
- **C** Not yet confident – too many unresolved questions
- **D** Not confident at all – the guidance is not there yet

Discussion 1 — The Adoption Question

What is the most significant barrier your organisation has faced in adopting AI for evidence synthesis — and how have you navigated it, or not?

Prompts to keep the conversation moving if needed:

- Is the barrier primarily technical, methodological, organisational, or legal?
- Has the barrier shifted over time — or has it remained constant?
- Have you found any approaches that partially addressed it — even if not fully resolved?
- Is there something the field could do collectively that would help?

Discussion 2 — The Standards Question

What would a minimum reporting standard for AI use in HTA SLR submissions look like — and who should be responsible for developing it

Prompts to keep the conversation moving if needed:

- Should reporting requirements vary by stage of automation — or be consistent across all AI use?
- Who has the authority and legitimacy to set this standard — HTA bodies, methodological organisations, industry, or a combination?
- What would make you trust an AI-assisted submission from another organisation?
- Is there a risk that premature standardisation locks in approaches that will quickly become outdated?

Discussion 3 — The Future Question

What does an HTA-ready evidence synthesis ecosystem look like in five years — and what needs to happen in the next twelve months to get us there?

Prompts to keep the conversation moving if needed:

- Should HTA bodies develop public benchmark datasets against which AI tools can be evaluated — similar to the approach used in the LLM development community?
- What should a leaderboard for AI SLR tools look like — and who should maintain it?
- How should the field prepare for agentic AI in evidence synthesis — and what governance frameworks need to be in place before agents are deployed in submission workflows?
- What should startups entering this space be required to demonstrate before their tools are used in HTA submissions?
- If you could change one thing about how the field is approaching this today, what would it be?

What We Hope You Take Home

Five things worth taking back to your team:

01 Locate yourself on the automation spectrum; know whether you are at Stage 1, 2, or 3 – and be honest about whether the boundary between stages is as clear in practice as it is in principle.

02 Build your documentation infrastructure before you need it; an AI use log, a validation record, and a governance trail are significantly easier to build from the start than to reconstruct after the fact.

03 Engage your HTA body before you submit; pre-submission dialogue on AI methodology is more productive – and less risky – than discovering expectations at the point of submission.

04 Contribute to the field, not just your organisation; the benchmark problem, the reporting gap, the guidance vacuum – these are shared problems; organisations that publish their barriers and their methodological approaches accelerate progress for everyone.

05 Start asking the agentic questions now; the governance, accountability, and methodological frameworks for agentic AI in evidence synthesis do not yet exist. The time to start building them is before the technology makes the question urgent.

What We Hope You Take Home

The goal of today was not to give you answers
*— **it was to give you better questions**.*

The field is moving quickly. The organisations that navigate it well will be the ones that move thoughtfully, document carefully, engage openly, and contribute generously.

Thank you.

Continue the Conversation

Speaker:

Dr. Sèye Abogunrin

Global Access Evidence Lead,
Roche

Email: seye.abogunrin@roche.com

What is the one thing your organisation could do in the next 90 days to be better prepared for AI-assisted SLRs for HTA submissions?

Selected resources for further reading:

- PRISMA-AI reporting extension
- Cochrane Collaboration guidance on AI in evidence synthesis
- ISPOR task force outputs on AI in HTA evidence generation

Selected resources for further reading

- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.
- Abogunrin S, Liu Y, Zerbini CH. A systematic literature review (SLR) on the adoption of artificial intelligence-assisted SLRS: implications for health technology assessments. International Journal of Technology Assessment in Health Care. 2026 Feb 16:1-28.
- Hanegraaf P, Wondimu A, Mosselman JJ, De Jong R, Abogunrin S, Queiros L, Lane M, Postma MJ, Boersma C, Van Der Schans J. Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: a mixed-methods review. BMJ open. 2024 Mar 1;14(3):e076912.
- Abogunrin S, Slob BP, Lane M, Emamipour S, Twardowski P, Boersma C, Van der Schans J. The introduction and adoption of artificial intelligence in systematic literature reviews: a discrete choice experiment. BMJ open. 2025 Oct 1;15(10):e099921.
- Abogunrin S, Muir JM, Zerbini C, Sarri G. How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. Frontiers in Pharmacology. 2025 Jan 31;16:1454245.

Doing now what patients need next