CMSS Checklist: Using AI in Clinical Practice Guideline Development

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

EXECUTIVE SUMMARY

The Council of Medical Specialty Societies (CMSS) is a coalition of more than 50 organizations representing nearly 1 million physicians whose mission is to advance the expertise and collective voice of specialty societies in support of physicians and the patients they serve.

Purpose

This implementation checklist is intended to support specialty societies as they incorporate AI methods into their clinical practice guideline (CPG) development processes. It is not meant to be comprehensive; rather, it highlights the most important considerations at each step and provides links to resources for those seeking more detailed information and guidance.

Background and rationale

Many specialty societies are considering how to use AI in various ways, including within CPG development. Implementation of AI tools in the CPG process is challenging and complex, so CMSS is supporting specialty societies by creating a structured document that consolidates important considerations and resources into an easy-to-use checklist.

Approach

CMSS undertook an environmental scan that included: 1) a survey of the members of the CMSS CPG professional peer group; 2) 12 semi-structured interviews with representatives of organizations involved in CPG development and/or systematic review; and 3) a review of the evidence on AI tools for use in guideline development.

Findings

Most of the 17 respondents to the survey used AI weekly or daily for administrative tasks, but only six organizations used it in CPG development itself. Eleven respondents cited barriers such as limited staffing and expertise, cost, validation concerns, copyright restrictions, and hallucination risk. Among adopters, AI use was confined to study selection, data extraction, and search strategy development, with no reported use in risk-of-bias assessment, synthesis, evidence ratings, or drafting of guidelines. Formal AI policies were rare (5 of 16), disclosure practices varied widely, and organizations were roughly split between building tools in-house and partnering with vendors. In open-ended comments, respondents called AI a high priority but still immature area and expressed interest in forthcoming CMSS guidance.

A targeted literature scan found few studies assessing the use of AI for CPG development. Overall, the studies suggest Large Language Models (LLMs) show real promise for accelerating literature screening tasks but perform less reliably when tasked with higher-order judgment such as synthesizing evidence into final recommendations. Adams and colleagues from the Brown Evidence-Based Practice Center are performing a living review of studies assessing the use of artificial intelligence or machine learning to semi- or fully automate aspects of evidence synthesis. This will be a valuable resource for CPG developers as the field advances.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

From 12 semi-structured interviews the following themes emerged: 1) AI use by CPG developers is primarily in literature screening and data extraction; 2) nearly every CPG developer eschewed the use of AI to write guidelines, risk-of-bias assessments, or GRADE judgements; 3) human-in-the-loop is universally considered to be necessary; 4) the use of AI helps with speed, accuracy, and expanding capacity for CPG development; 5) the main concerns are hallucinations, reproducibility, and AI vendors over-promising the maturity of their product; 6) transparency is essential but reporting standards are still emerging; 7) licensing CPG content to AI vendors creates tension; and 8) CPG developers want a cross-organization community.

Key highlights of the Checklist

The results of the survey, interviews, and literature scan informed the development of the Checklist. Several themes recur throughout the Checklist:

- Start small.
- Use AI where the evidence supports it.
- Keep humans involved at every step.
- Validate tools within the specific organizational setting before scaling their use.
- Govern AI use carefully.
- Attend to equity and representativeness in the underlying data and literature.
- Document and disclose AI use transparently.
- Share lessons learned with the broader community.

Limitations

This is a rapidly evolving field; therefore, the checklist should be treated as a living document and revisited periodically as practices, tools, and evidence mature.

Conclusion

The use of AI has real promise in supporting the development of CPGs. Specialty societies should keep abreast of the advances in the field and use a structured checklist when starting to use AI in CPG development.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

STAGE	CONSIDERATIONS	RESOURCES/REFERENCES	KEY RISKS & CAUTIONS

1. Assess organizational readiness and set the strategic approach	Gauge current staff and CPG panel comfort and skill level with AI, recognizing views will range from enthusiastic to skeptical
	• Offer AI training for CPG panel members
	• Identify structural gaps (staffing, technical expertise, budget) that could limit responsible use
	• Decide on a philosophy: move deliberately without waiting too long, and avoid exclusive reliance on a single vendor
	• Discuss financial considerations with organizational leadership including:
	  • Upfront costs with vendors
	  • Licensing costs and other ongoing costs
	  • ROI – tradeoffs: cost of investment vs reduced time to final guideline; reduced staff time; ability to do more guidelines
		Free resources for basic AI training
		• OpenAI Academy
		• Anthropic
		• Google
		• ACP: AI Resource Hub
		• AMA: Artificial Intelligence in Health Care Series
			• No established blueprint yet exists across the field
			• Risk of falling behind peers vs.
			taking on unvalidated risk by moving too fast

2a. Select the process step(s) and AI tool(s) to use	• Consider steps for which there is benefit to reducing time/costs
	• Favor steps that are easily automated and repeatable
	• Prioritize steps where AI tools have been validated as accurate across several independent studies
	• Focus on one step in the process when getting started.
	Literature search, screening, and triage:
	• This is the most studied and piloted step for AI use in guideline development so consider starting with this step
	  • Use active-learning screening tools with a defined stopping rule (e.g., a probability/run-length threshold) OR processes using LLMs as reviewer (s)
	  • Retain human review rather than fully "handing over" screening, especially for reviews likely to have very few eligible studies
	  • Document the stopping rule and screening approach in the review protocol/methods section
		• Living review of AI/ML use in evidence synthesis (Adams and colleagues) https://effectivehealthcare.ahrq.gov/products/machine-learning-tools/white-paper
		• Table 1. (below) of readiness of tools for specific steps in evidence synthesis
		• Evidence on AI use in literature search screening (Tables 2, 3, 4below)
		Other resources:
		• Tools with AI extensions for assessing individual studies
		  • PROBAST-AI
		  • TRIPOD+AI
			• Screening sensitivity/specificity vary considerably by model and tool; random chance could cause a tool to miss an entire sub-topic's literature base when eligible studies are very few
			• Human extraction error rates are often higher than assumed, which should inform benchmarking expectations; complex/quantitative outcomes remain a weaker use case
			• Even human raters show high variability on evidence-synthesis/GRADE judgments, complicating validation; no organization reported routine AI use for this step as of this report

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

STAGE	CONSIDERATIONS	RESOURCES/REFERENCES	KEY RISKS & CAUTIONS
2a. Select the process step(s) and AI tool(s) to use (continued)	• Pilot AI-assisted extraction paired with human review, particularly for straightforward background/demographic data • Recognize that complex outcomes data (e.g., results tables) are extracted less reliably than background characteristics **Evidence synthesis, risk-of-bias, and GRADE/certainty judgments:** • Recognize this is currently the least mature application area; avoid routine reliance on AI for these judgments • If piloting, maintain full human adjudication given the high degree of subjective, nuanced judgment required	• AGA example of a 2-LLM pipeline that consistently identified studies with 100% sensitivity (Chung et al)	• No established blueprint yet exists across the field • Risk of falling behind peers vs. taking on unvalidated risk by moving too fast
2b. Select topic for review	• When getting started, pilot a topic that you are confident has an RCT evidence base • Start with a clinical question that has focused and well-defined PICOTS • Avoid piloting on topics known to be harder for AI tools: behavioral interventions, observational/real-world evidence, or reviews with very high or very low inclusion rates		• Choosing a poorly suited pilot topic (nuanced, judgment-heavy, or low/high-inclusion-rate) can produce misleadingly poor early results and undermine staff trust in AI tools before they've had a fair test
3. Establish governance and policy	• Develop (or extend) an organization-wide AI-use policy into a guideline-development-specific policy • Define what data may/may not be uploaded to AI tools, including copyright and licensing constraints on source literature • Build in mandatory human-in-the-loop oversight at every step where AI is used • Establish an audit-trail/logging process capturing AI inputs, decisions, and outputs, since models change over time	• CMSS resources on AI and specialty societies • RAISE framework • GIN responsible-AI principles • Anticipated National Academy of Medicine update to systematic review standards	• Guideline-specific AI policy often lags far behind general organizational AI policy • Policies can become outdated quickly given the pace of change

STAGE	CONSIDERATIONS	RESOURCES/REFERENCES	KEY RISKS & CAUTIONS
4. Vet and select AI tools/ vendors	• Compare multiple tools/vendors rather than adopting the first one presented • Test tools directly against organizational data rather than relying solely on vendor demonstrations or claims • Screen out tools that do not adhere to recognized standards (e.g., RAISE) • Consider locked-down/secure or enterprise instances where copyright or data-privacy protection is needed • Avoid exclusivity agreements with a single vendor; seek multiple opinions	• Peer organizations and informal cross-society networks • Considerations when vetting vendors (Figure 1 below) • AI Vendor Intake Checklist (Figure 2 below)	• Rapid proliferation of unproven vendor tools ("gold rush"); difficulty distinguishing credible tools from unproven ones • Vendor claims may not reflect independently validated performance • Wasted staff time on pilot partnerships that fail to deliver a working product • Unsolicited vendor outreach can complicate objective evaluation
5. Approach drafting of guideline text and recommendations with caution	• Reserve AI for assisted editing/polishing of a human-drafted content rather than original drafting of recommendations • Require human authorship and verification of all recommendation statements • Screen AI-assisted drafts carefully for fabricated or non-existent citations before circulation		• Documented cases of AI-fabricated citations entering panel-submitted drafts • Wide variability in expert agreement with AI-drafted clinical algorithms (most nodes under 60% agreement in one study)
6. Validate rigorously before scaling use	• Formally test AI tools against existing human-generated results (for example, run AI-driven processes in parallel with usual non-AI processes and compare results) before scaling • Repeat prompts/questions over time and across models to check reliability and reproducibility • Build a reference standard that accounts for the limitations of human performance as an imperfect gold standard • Publish or share validation results so the field is not "reinventing the wheel"		• Identical prompts can yield different results across time or model versions • Model updates may complicate reproducing decisions made years earlier

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

STAGE	CONSIDERATIONS	RESOURCES/REFERENCES	KEY RISKS & CAUTIONS
7. Document, disclose, and report AI use	• Include a dedicated section in protocols/methods documenting which steps used AI, which tools, and how outputs were verified • Default toward oversharing rather than undersharing when documenting AI use • Confirm journal-specific AI disclosure requirements before submission, as these vary by publisher • Communicate clearly that piloting or partnering with a vendor does not constitute an organizational endorsement	• RAISE reporting guidance • PRISMA-AI reporting extension	• Journal-level AI disclosure policies vary and are still evolving • Inconsistent internal documentation practices
8. When organizationally ready, develop a plan for living guidelines and ongoing surveillance	• Define how AI will support continuous literature monitoring, identification of new evidence, and thresholds to trigger guideline update • Keep judgments about whether/how to change recommendations with the human panel, not the AI tool • Set a realistic update cadence based on organizational capacity	• Cochrane statement on AI for rapid reviews • Study in 2025 by Cochrane, et al, reported 15% of published living reviews used machine learning	• Version drift: models and their outputs change over time, complicating long-term consistency
9. Plan dissemination and downstream/ patient-facing uses	• Consider AI-assisted plain-language summaries, translation, guideline chatbots/assistants, and clinical decision support (CDS) integration • Ensure downstream tools draw only from vetted, evidence-based sources with clear sourcing and version-currency safeguards • Evaluate licensing arrangements with AI vendors for guideline content, weighing revenue opportunities and wider scientific dissemination against content-protection concerns.	Example: ASCO Guideline Assistant	• End-user (clinician/public) trust in AI-mediated guideline content is less examined than internal staff trust • Risk that commercial tools extract value from guideline content without crediting the originating organization

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

STAGE	CONSIDERATIONS	RESOURCES/REFERENCES	KEY RISKS & CAUTIONS
10. Invest in workforce training and change management	• Ensure staff and CPG panel members involved in AI-assisted work have strong evidence-synthesis methodology expertise • Build consensus across a spectrum of comfort levels, from enthusiastic adopters to skeptics, without moving faster than the most cautious members are comfortable with	• Internal training materials on AI and guideline-methodology language • GIN North America AI community • Cochrane AI Methods Group • ICASR: International Collaboration for the Automation of Systematic Reviews	• Lack of in-house expertise can be a meaningful barrier to entry and to evaluating vendor claims
11. Attend to equity and representativeness	• Assess whether AI training data and underlying literature are adequately representative • Distinguish higher bias risk in interpretive AI tasks (e.g., drafting discussion sections) from comparatively lower risk in mechanical tasks (e.g., screening abstracts)	• Encoding Equity AI tool kit	• AI-enabled tools may widen rather than narrow existing care disparities if equity is not explicitly addressed
12. Engage the broader community and shared infrastructure	• Participate in cross-organizational forums, working groups, or convenings to share lessons learned • Contribute to or draw on shared resources	• GIN North America AI community • Cochrane AI Methods Group • ICASR: International Collaboration for the Automation of Systematic Reviews • CMSS -facilitated connections between specialty societies and evidence-synthesis/ methodology experts	• Working in isolation risks duplicating effort already validated elsewhere in the field

Note: This checklist reflects a rapidly evolving field. Organizations are encouraged to treat it as a living document, revisit it periodically, and validate any AI tool against their own data before relying on it for guideline-relevant work.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Table 1. Readiness of AI and active learning tools for systematic reviews by task

SYSTEMATIC REVIEW TASK	STATUS (MEDIAN RECALL/ ACCURACY)	RECOMMENDED USE	KEY SAFEGUARDS / CONDITIONS
Study design identification (e.g., RCT tagging)	Ready for use with oversight (98% recall)	First-pass filter to flag likely RCTs for human review; reduces manual screening of study designs	Retain human adjudication of borderline cases; document recall threshold used
Abstract screening: semiautomated (active learning)	Ready for use with oversight (95% recall)	Screening prioritization tool to substantially reduce burden while maintaining high recall; use Table 2 targets as benchmarks	Use conservative stopping rules; monitor recall against a hold-out sample; restart algorithm at each update cycle; performance varies by topic
Abstract screening: semiautomated with stopping criteria	Not yet ready for routine use; some stopping rules show consistently sufficient recall for use but may not dramatically reduce screening burden.	Using screening prioritization with specific stopping rules that have been validated as having high recall to either stop screening change from double to single screening	Use conservative stopping rules; monitor recall against a hold-out sample
Abstract screening: fully automated (LLM)	Not yet ready for routine use; inconsistent performance (87% recall); better performance with pretraining or few shot prompting (97%)	Supplementary triage aid or first pass to create an enriched set for active-learning screening; do not replace human screening without verification	Performance not yet sufficient for standalone exclusion decisions without human verification; few shot and fine-tuned models outperform zero-shot
Full-text screening (LLM)	Not yet ready for routine use; limited data (97% recall)	Second-reviewer check on full text exclusions	Use conservative stopping rules; monitor recall against a hold-out sample
Data extraction of descriptive data	Not yet ready for routine use; inconsistent performance (88% accuracy)	Structured suggestion layer or second-reviewer check, especially for descriptive data; semi-automated workflows preferred over fully automated	Pilot on a representative sample before production adoption; accuracy highly sensitive to prompting strategy
Risk of bias assessment for Cochrane 1 RoB tool	Not yet ready for routine use; inconsistent performance (75% accuracy)	Structured suggestion layer or second reviewer for human extraction; human adjudication of all domains required	Performance not yet sufficient for standalone RoB decisions without human verification.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Table 1. Readiness of AI and active learning tools for systematic reviews by task (continued)

SYSTEMATIC REVIEW TASK	STATUS (MEDIAN RECALL/ ACCURACY)	RECOMMENDED USE	KEY SAFEGUARDS / CONDITIONS
Search strategy development	Not yet ready for use; poor performance for zero-shot LLMs and limited evidence for fine-tuned or few shot LLMs (zero-shot 25% recall)	Ancillary term generation and query expansion; human-in-the-loop editing of AI generated queries before execution	Do not use as primary search method; AI-generated queries perform substantially worse than librarian-developed strategies; human-in-the-loop editing improves performance considerably
Data extraction of results	Not yet ready for use; poor performance (54% accuracy)	Second-reviewer check; semiautomated workflows preferred over fully automated	Performance not yet sufficient for standalone use; accuracy highly sensitive to prompting strategy and data source
Risk of bias assessment for Cochrane 2 RoB tool	Not yet ready for use; performance (53% accuracy)	Starting-point structured prompt or second check; human adjudication of all domains required	Performance much lower for RoB 2.0.; do not use without full human review
Risk of bias assessment for other RoB tools	Not yet ready for use; sparse data and inconsistent results	Starting-point structured prompt or second check; human adjudication of all domains required	Few studies and variable results; do not use without full human review
Meta-analysis	Not yet ready for use; sparse data and inconsistent results	May be used for exploratory analysis; all code should be verified by a statistician	Few studies and variable results; do not use without full human review
Meta-analysis	Exploratory	Drafting assistance and summarization aid under close human supervision	Evidence base sparse; inconsistent quality criteria; documented limitations in accuracy and comprehensiveness

Adam GP, Kuhn E, Davies M, Holmer H, Htun J, Coppola EL, Wethington H, Ivlev I, Balk EM, Trikalinos TA. Machine Learning Tools To (Semi-)Automate Evidence Synthesis, 2026 Review Update. (Prepared by the Brown Evidence-based Practice Center [contract no. 75Q80120D00001], Kaiser Permanente Research Affiliates Evidence-based Practice Center [contract no. 75Q80120D00004], Pacific Northwest Evidence-based Practice Center [contract no. 75Q80120D00006], and the Scientific Resource Center [contract no. 75Q80122C00002].) AHRQ Publication No. 26-EHC011. Rockville, MD: Agency for Healthcare Research and Quality. August 2026. Posted final reports are located on the Effective Health Care Program page. https://effectivehealthcare.ahrq.gov/products/machine-learning-tools/white-paper

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Table 2. Active learning tools for abstract screening

TOOL NAME	N STUDIES	YEAR RANGE	ACCESS	NOTES	RECALL, MEDIAN (RANGE)	BURDEN, MEDIAN (RANGE)
Abstrackr	4	2010-2020	free; open source	Defunct, incorporated into VADRR	99.5 (95-100)	47 (40-57)
ASReview	3	2023-2026	free; open source	Stopping rule studies excluded; see Riepke	90.5 (86-95)	27 (15-30)
Claude-3 Haiku	2	2025	tiered subscription	LLM used as an active learning tool	94 (90-97)	50 (50-50)
DistillerSR	2	2019-2025	subscription		96.5 (94-98)	44 (39-53)
DoCTER	1	2025	free		98	23
EPPI Reviewer	2	2020-2023	subscription		94 (88-100)	50 (50-50)
Gemini-1.5 Flash	2	2025	tiered subscription	LLM used as an active learning tool	93.5 (92-97)	50 (50-50)
Gemini-1.5 Flash+GPT-4o mini+-Claude-3 Haiku	2	2025	tiered subscription	LLM used as an active learning tool	97 (95-98)	50 (50-50)
GPT-4o mini	2	2025	tiered subscription	LLM used as an active learning tool	91 (90-98)	50 (50-50)
Rayyan	2	2022-2023	subscription		48 (30-66)	55 (50-60)
Research Screener	1	2021	subscription		95	11
RobotAnalyst	1	2019	free; open source		98	35
SWIFT-Active Screener	2	2016-2020	free		95 (95-95)	43 (33-54)

Recall = The number of relevant studies or data items correctly identified (also referred to as sensitivity); Burden = the percent of studies evaluated by a human for a given recall for screening tools.

Table created from data in: Adam GP, Kuhn E, Davies M, Holmer H, Htun J, Coppola EL, Wethington H, Ivlev I, Balk EM, Trikalinos TA. Machine Learning Tools To (Semi-) Automate Evidence Synthesis, 2026 Review Update. (Prepared by the Brown Evidence-based Practice Center [contract no. 75Q80120D00001], Kaiser Permanente Research Affiliates Evidence-based Practice Center [contract no. 75Q80120D00004], Pacific Northwest Evidence-based Practice Center [contract no. 75Q80120D00006], and the Scientific Resource Center [contract no. 75Q80122C00002].) AHRQ Publication No. 26-EHC011. Rockville, MD: Agency for Healthcare Research and Quality. August 2026. Posted final reports are located on the Effective Health Care Program page. https://effectivehealthcare.ahrq.gov/products/machine-learning-tools/white-paper

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Table 3. LLMs for fully automated abstract screening

TOOL NAME	N STUDIES	YEAR RANGE	RECALL, MEDIAN (RANGE)	PRECISION, MEDIAN (RANGE)
Claude-3	4	2024-2026	90 (87-97)	NR
Claude-3.5	3	2025	82 (82-94)	92.5 (10-99)
Claude-3.7	1	2025	93	92
DeepSeek-chat 2.5	1	2025	98	98
DeepSeek-R1	2	2025-2026	82.5 (74-91)	54.5 (15-98)
DeepSeek-V3	2	2025-2026	82 (81-83)	NR
Gemini-1.0	1	2024	75	96
Gemini-1.5	3	2025-2026	92 (73-94)	87 (82-98)
Gemini-2.0	2	2025	81 (79-83)	Specificity 93 (86-100)
GPT-4	25	2024-2026	77 (12-100)	82 (2-98)
GPT-5	2	2025-2026	85 (72-98)	NR
Llama-3	8	2024-2026	84 (4-100)	59 (1-100)
Combinations of 2 models (citation included if either model screened in)	4	2025-2026	92 (80-100)	Specificity 97 (20-100)
Combinations of 3 or more models (included if either majority voted in or 3rd model adjudicated as in)	6	2025-2026	94 (75-100)	Specificity 88 (56-96)
Model given a set of relevant and irrelevant citations as part of the prompt (few-shot)	8	2023-2026	96 (71-100)	Specificity 63 (49-98)

Recall = number of relevant studies or data items correctly identified (i.e., sensitivity); Precision (unless noted to be specificity) = number of relevant studies or data items among the total identified; Specificity = The number of irrelevant studies or data items correctly identified. NR = not reported

Table created from data in: Adam GP, Kuhn E, Davies M, Holmer H, Htun J, Coppola EL, Wethington H, Ivlev I, Balk EM, Trikalinos TA. Machine Learning Tools To (Semi-) Automate Evidence Synthesis, 2026 Review Update. (Prepared by the Brown Evidence-based Practice Center [contract no. 75Q80120D00001], Kaiser Permanente Research Affiliates Evidence-based Practice Center [contract no. 75Q80120D00004], Pacific Northwest Evidence-based Practice Center [contract no. 75Q80120D00006], and the Scientific Resource Center [contract no. 75Q80122C00002].) AHRQ Publication No. 26-EHC011. Rockville, MD: Agency for Healthcare Research and Quality. August 2026. https://effectivehealthcare.ahrq.gov/products/machine-learning-tools/white-paper

Table 4. LLMs for fully automated full text screening

TOOL NAME	N STUDIES	YEAR RANGE	ACCESS	RECALL, MEDIAN (RANGE)	PRECISION, MEDIAN (RANGE)
Covidence	1	2025	subscription	100	44
DeepSeek	2	2025	tiered subscription	80.5 (61-100)	NR
Elicit	1	2025	subscription	87	87
Gemini-1.5	1	2025	tiered subscription	75.5	NR
GPT-4	7	2024-2025	tiered subscription	82.5 (38-100)	specificity 83 (48-99)
Kimi (moonshot-v1-128k)	1	2025	tiered subscription	100	100
Llama-3.1	1	2025	free; open source	38	100
LoonLens	1	2025	subscription	95	58
Nested Knowledge	1	2025	subscription	53	NR
SWIFT-Active Screener	1	2025	free	100	51
Combinations of models	1	2025		94.5 (89-100)	specificity 76.5 (38-97)

Recall = number of relevant studies or data items correctly identified (i.e., sensitivity); Precision (unless noted to be specificity) = number of relevant studies or data items among the total identified; Specificity = The number of irrelevant studies or data items correctly identified. NR = not reported

Table created from data in: Adam GP, Kuhn E, Davies M, Holmer H, Htun J, Coppola EL, Wethington H, Ivlev I, Balk EM, Trikalinos TA. Machine Learning Tools To (Semi-) Automate Evidence Synthesis, 2026 Review Update. (Prepared by the Brown Evidence-based Practice Center [contract no. 75Q80120D00001], Kaiser Permanente Research Affiliates Evidence-based Practice Center [contract no. 75Q80120D00004], Pacific Northwest Evidence-based Practice Center [contract no. 75Q80120D00006], and the Scientific Resource Center [contract no. 75Q80122C00002].) AHRQ Publication No. 26-EHC011. Rockville, MD: Agency for Healthcare Research and Quality. August 2026. https://effectivehealthcare.ahrq.gov/products/machine-learning-tools/white-paper

Figure 1: Specialty society considerations when vetting guideline development vendors

1. DOES THE SOCIETY UNDERSTAND ITSELF FIRST?
Before talking to any vendor:
• Do we have a defined, documented guideline development process?
• Do we understand our own "data generation process" (e.g., PICOs, evidence tables, expert narratives) well enough to state it explicitly?
• Which discrete step(s) are we trying to improve (title/abstract screening, data abstraction, meta-analysis, risk-of-bias, GRADE application)?
2. TECHNICAL SUBSTANCE
• Ask vendor to break down their process into discrete steps.
• Is the vendor's pipeline generic/one-size-fits-all across societies, or built to reflect your specific process and standards?
• Probe claims about using "agentic" AI to ensure they are offering something beyond an LLM with document extraction.
• Ask specifically how they extract data (PDF/XML/scraping), what model(s) underlie it, and how outputs are validated.tion)?
• Ask about their depth of understanding of guidelines and methodology (GRADE, PICO, evidence tables).
• During discussions with vendors, ensure you have a guideline content expert in addition to a technical expert.
3. VERIFICATION AND VALIDATION
• What is their plan and budget for verifying outputs? Will they pay for the expert time needed to validate their product, or expect the society's volunteer experts to do it for free?
• Are they transparent about the need for iterative human curation/annotation to become useful or do they market it as fully autonomous, "click of a button," end-to-end?
• Can the society independently replicate/validate a subset of their output before committing (e.g., "run these numbers, we'll run them ourselves and compare")?
• For which components has the vendor published or externally validated performance (e.g., screening, data abstraction) versus components that remain largely untested (e.g., automated GRADE application)?

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Figure 1: Specialty society considerations when vetting guideline development vendors

4. DATA, IP, AND VALUE EXCHANGE

- What are we giving them, and what do we get in return? Recognize that guideline content, evidence, and especially expert labeling/annotation are valuable and cost to produce.
- What are the terms around our data/guidelines — are we granting broad rights, and does that limit our leverage later (e.g., strong contractual footing vs. giving content away as a "public good")?
- Is the vendor also trying to identify and directly recruit our expert network (to label data or provide review) in a way that could eventually bypass the society entirely?
- What is the carryover? If our experts' time improves their product, how do they acknowledge or compensate that when they take the resulting product to the next organization? Get this in writing.

5. REFERENCE-CHECKING AND PEER NETWORK

- Ask other societies: what was their experience, and how did they verify the vendor's claims?
- Recognize that processes differ meaningfully across societies; another society's good experience with a vendor doesn't guarantee fit if your process is different.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Figure 2: AI Vendor Intake Checklist

For Guideline Development & Evidence Synthesis Vendors

Please review and complete this checklist prior to our introductory call.

1. Technical Substance

QUESTION / ITEM	VENDOR RESPONSE / NOTES
☐ Please describe your process broken down into discrete steps.	
☐ Is your product a one-size-fits-all across societies, or built to reflect a specific society's process and standards? Please explain.	
☐ You describe your system as "agentic." Please explain specifically what that means in your architecture, beyond an LLM performing document extraction.	
☐ How do you extract data (PDF, XML, web scraping)? What underlying model(s) do you use, and how are outputs validated?	
☐ Describe your team's depth of experience with guideline methodology (GRADE, PICO, evidence tables).	

2. Verification and Validation

QUESTION / ITEM	VENDOR RESPONSE / NOTES
☐ What is your plan and budget for verifying outputs? Will you cover the cost of the expert time needed to validate your product, or is that expected from our volunteer experts?	
☐ How does iterative human curation/annotation fit into your process? Please describe realistically, rather than as fully autonomous or "one-click."	
☐ Can we independently replicate or validate a subset of your output before committing (e.g., we run a sample ourselves and compare results)?	
☐ For which components has your performance been published or externally validated, and which remain largely untested?	

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Figure 2: AI Vendor Intake Checklist

For Guideline Development & Evidence Synthesis Vendors

Please review and complete this checklist prior to our introductory call.

3. Data, IP, and Value Exchange

QUESTION / ITEM	VENDOR RESPONSE / NOTES
☐ What data/content are we providing to you, and what do we receive in return?	
☐ What are the proposed terms around use of our guidelines/data?	
☐ Will you be identifying or directly recruiting members of our expert network (for labeling or review)?	
☐ If our experts' time and input improve your product, how will that be acknowledged or compensated when the resulting product is offered to other organizations? Please provide this in writing.	

4. Reference-Checking and Peer Network

QUESTION / ITEM	VENDOR RESPONSE / NOTES
☐ Please provide references from other specialty societies you have worked with, including a contact name and email/phone.	
☐ For each reference, please describe the scope of work performed and the outcome.	
☐ Have you worked with societies whose guideline development process differs from ours? If so, please describe how your approach was adapted.	

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Acronyms

ACP – American College of Physicians.

AGA – American Gastroenterological Association.

AHRQ – Agency for Healthcare Research and Quality, the federal agency that funds the Evidence-based Practice Center Program.

AI – Artificial intelligence.

AMA – American Medical Association.

ASCO – American Society of Clinical Oncology.

CDS – Clinical decision support. Tools that bring guideline recommendations into clinical workflows, such as electronic health record alerts or order sets.

CMSS – Council of Medical Specialty Societies.

CPG – Clinical practice guideline.

EPC – Evidence-based Practice Center. AHRQ-funded centers, such as the Brown EPC, that conduct systematic reviews and methods research.

GIN – Guidelines International Network, a global network of guideline developers that has issued principles for the responsible use of AI.

GRADE – Grading of Recommendations Assessment, Development and Evaluation. A widely used framework for rating the certainty of evidence and the strength of recommendations.

ICASR – International Collaboration for the Automation of Systematic Reviews.

IP – Intellectual property.

LLM – Large language model. A type of AI model, such as GPT, Claude, Gemini, Llama, or DeepSeek, trained on large volumes of text to generate and analyze language.

ML – Machine learning.

NR – Not reported.

PICO / PICOTS – Population, Intervention, Comparator, Outcome (plus Timing and Setting). The standard framework for structuring a clinical question.

PRISMA-AI – An in-development AI extension of PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), the standard reporting guideline for systematic reviews.

PROBAST-AI – The AI extension of the Prediction model Risk Of Bias Assessment Tool, used to assess risk of bias in prediction model studies that use AI.

RAISE – Responsible use of AI in evidence synthesis. A framework and set of recommendations for the responsible, transparent use of AI in evidence synthesis.

RCT – Randomized controlled trial.

RoB (RoB 1 / RoB 2) – Risk of bias. Refers here to Cochrane's original and revised tools for assessing risk of bias in randomized trials.

ROI – Return on investment.

TRIPOD+AI – The AI update to TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis), a reporting guideline for prediction model studies.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

Glossary

Active learning – A machine learning approach used in screening. The tool learns from a reviewer's include/exclude decisions and continually re-ranks the remaining citations so the most likely relevant ones appear first.

Agentic AI – AI systems that plan and carry out multistep tasks with some autonomy, rather than simply responding to a single prompt. The checklist advises probing vendor claims that a product is "agentic."

Audit trail – A log of AI inputs, decisions, and outputs. It allows work to be reviewed or reproduced later, which matters because models change over time.

Burden – In screening studies, the percentage of citations a human must still review to reach a given level of recall.

Data extraction (abstraction) – Pulling specific information from included studies into structured tables. Examples are descriptive data such as participant characteristics and results data such as outcomes.

Enterprise / locked-down instance – An organization-specific, access-controlled deployment of an AI tool. It is used to protect confidential data and comply with copyright and licensing restrictions.

Evidence synthesis – The process of identifying, appraising, and combining research findings, as in a systematic review or meta-analysis.

Few-shot prompting – Giving an LLM a small number of worked examples (e.g., relevant and irrelevant citations) within the prompt to improve its performance.

Fine-tuning – Further training an existing AI model on task-specific data to improve its performance on that task.

Fully automated – A process in which AI makes decisions without human review of each output. Compare with semiautomated.

Gold standard / reference standard – The benchmark against which a tool's performance is measured. The checklist notes that human performance is an imperfect gold standard.

Hallucination – AI-generated content that is false or fabricated but presented as fact, such as citations to studies that do not exist.

Hold-out sample – A subset of citations set aside and screened by humans. It is used to check an AI tool's recall.

Human-in-the-loop – A workflow in which humans review, verify, and retain final decision-making authority over AI outputs.

Living guideline / living review – A guideline or systematic review that is continually updated as new evidence emerges, rather than revised only at long intervals.

Meta-analysis – A statistical method for combining quantitative results from multiple studies into a pooled estimate.

Precision – The proportion of items identified by a tool that are actually relevant.

Recall (sensitivity) – The proportion of all relevant studies or data items that a tool correctly identifies. High recall is critical in screening so that eligible studies are not missed.

Reproducibility – The ability to obtain consistent results when a process is repeated. It is a challenge with LLMs, whose outputs can vary between runs.

Risk-of-bias assessment – Evaluation of the methodological quality of individual studies to judge how likely their results are to be biased.

Screening (title/abstract and full-text) – The stages of a systematic review in which citations are assessed against eligibility criteria. Screening is done first on titles and abstracts, then on full articles.

Semiautomated – A process in which AI assists or prioritizes the work while humans make or verify final decisions.

Specificity – The proportion of irrelevant studies or data items that a tool correctly excludes.

Stopping rule – A predefined criterion, such as a probability or run-length threshold, for when reviewers can stop screening in an active-learning tool.

Surveillance (evidence surveillance) – Ongoing monitoring of newly published literature to determine when a guideline may need updating.

Systematic review – A rigorous, protocol-driven review that identifies, appraises, and synthesizes all relevant studies on a defined question.

Version drift – Changes in an AI model's behavior or outputs as the model is updated over time. It complicates long-term consistency and reproducibility.

Zero-shot prompting – Asking an LLM to perform a task with no examples provided in the prompt.

CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES


CMSS
COUNCIL OF MEDICAL
SPECIALTY SOCIETIES