



Journal of Health Economics and Outcomes Research

Systematic Review



Faster but Not Truer, Part 2: The Regulatory Record, the Economics, and the TRACE-HEOR Reporting Standard

Onur Baser^{1,2*}

¹Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA

²Graduate School of Public Health and Health Policy, City University of New York, New York, USA

ARTICLE INFORMATION

This is the second of two articles. Part 1 measured the literature on artificial intelligence in health economics and outcomes research and validated its own classification. Part 2 assembles the regulatory, health technology assessment, and industry record, examines what the economics of automation implies for this field, and proposes a reporting standard.

Accepted September 16, 2026

Keywords: artificial intelligence; large language models; health economics and outcomes research; health technology assessment; regulatory science; reporting standards; real-world evidence

*Corresponding author:

Email address: onur.baser@sph.cuny.edu

➤ [Supplementary Material](#)

ABSTRACT

Background: Part 1 of this series found that measured gains from artificial intelligence (AI) in health economics and outcomes research (HEOR) concentrate in tasks with observable ground truth and disappear where the quantity of interest is a counterfactual. What regulators, health technology assessment (HTA) bodies, and sponsors have said about that division has not been assembled in one place.

Methods: We assembled the primary regulatory, HTA, industry, and journal policy record through September 2026, requiring for each claim a named primary source and, where a quantity was reported, its provenance and whether the reporting party had a commercial interest. We downloaded the Food and Drug Administration (FDA) AI-enabled medical device list in full, tabulated published economic evaluations of AI, and set reported industry efficiency gains against disclosed use in submissions. We then evaluated the field against the task-based productivity framework of Nobel laureate Daron Acemoglu and derived a reporting standard from the documented failure modes.

Results: The FDA has authorized 1614 AI-enabled devices, 1230 (76.2%) in radiology, rising from 18 in 2016 to 335 in 2025, and has received more than 500 drug and biologic submissions containing AI components since 2016; its January 2025 credibility assessment guidance remains a draft and excludes discovery and operational uses. Industry sources report time savings of 40% to 97% in regulatory writing and approximately 80% in payer dossier drafting, but every one of these figures comes from a single-arm comparison without a concurrent control and was reported by the company that deployed or sold the tool. AI use was disclosed in 3 of 1391 HTA documents (0.2%) and 0.6% of one journal's submissions, against surveys reporting use between 25% and 81%. Nineteen economic evaluations of AI have been published against 1614 authorized devices. No regulator or HTA body has been documented accepting or rejecting an AI-derived endpoint as such.

Conclusions: Current disclosure instruments ask whether AI wrote the text. Reviewers need to know whether AI selected the evidence, defined the cohort, derived the endpoint, or generated the code, and whether the recommendation would change without it.

INTRODUCTION

Part 1 of this series reported a systematic scoping review of 2476 indexed studies on artificial intelligence (AI) in health economics and outcomes research (HEOR) published between 2010 and 2025 and measured the error rate of its own classification against a blinded independent classifier. Two findings carry forward. First, the size of this literature is not known to within a factor of 2: the corpus reduces to approximately 1271 genuinely HEOR-relevant studies (95% confidence interval [CI], 1131-1410), and the fastest growing label, economic evaluation, is the one that could least reliably be assigned (Cohen kappa, 0.144).

Second, measured gains concentrate in tasks with an observable answer and are absent where the quantity of interest is a counterfactual.¹⁻³

That division sorts the field into two layers. A clerical layer, comprising screening, extraction, code drafting, and dossier assembly, has measurable ground truth and documented gains. An inferential layer, comprising cohort definition, endpoint derivation, confounder selection, and effect estimation, has no supervised target, because no patient is observed under both the treatment and the comparator. The two layers carry different risks and should be governed differently.

This second article asks what the institutional record says about that distinction. Regulators come first, because a reimbursement



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

dossier is a regulated document and because the FDA has written the most developed framework while becoming one of the largest users of these tools. Health technology assessment (HTA) agencies come second, because they will absorb whatever the automation produces. Industry comes third, because the efficiency claims driving adoption originate there and almost none has been independently replicated. We then turn to how these systems fail, to what productivity economics predicts for a field whose central quantity is unobservable, and to what reporting would make an AI-assisted submission auditable.

METHODS

The review methods, search strategy, classification dictionary, and validation procedure are reported in full in Part 1 and archived in the public replication package.⁴ This article draws on the narrative evidence review conducted alongside that scoping review.

For the regulatory, HTA, industry, and journal policy record, we searched agency, society, and company sources through September 2026 and required a named primary source for every claim. Where a quantity was reported, we recorded its provenance, whether a peer-reviewed study, a conference abstract, a regulatory document, or a company communication, and whether the reporting party had a commercial interest in the result. Conference abstracts and vendor communications are labeled rather than excluded, because in this field they carry information available nowhere else.

We downloaded the FDA AI-enabled medical device list in full and tabulated authorizations by year and specialty in **Supplementary Table S2**. Published economic evaluations of AI in health care appear in **Supplementary Table S1**, industry efficiency claims with their provenance in **Supplementary Table S3**, journal and society policies in **Supplementary Table S4**, and survey estimates of use and proficiency in **Supplementary Table S5**. **Figure 1** places the instruments on a common timeline.

RESULTS

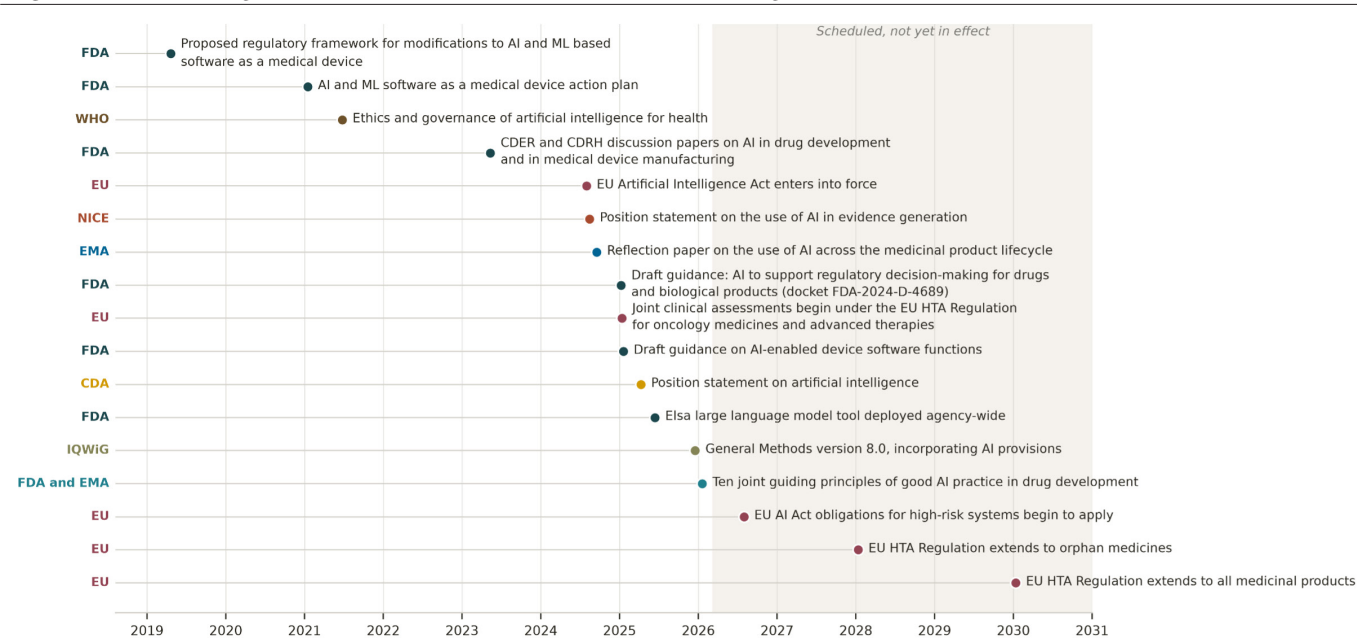
The Regulatory Record

The FDA’s position is the most developed and the least finished. **Table 1** summarizes the instruments now in force or in draft.

For drugs and biologics, the January 2025 draft guidance proposes a 7-step credibility assessment framework in which a sponsor defines the question of interest, the context of use, and the model risk, then establishes credibility proportionate to that risk.⁵ The definition of model risk is the operative sentence: “Model risk is a combination of two factors: (a) model influence... and (b) decision consequence.” Two features are routinely missed. The guidance explicitly excludes AI used in drug discovery and in operational efficiencies, so most publicly announced pharmaceutical AI deployments fall outside it. It remains a draft as of September 2026, with the comment period having closed on April 7, 2025. The agency’s account of the evidence behind it names an FDA-sponsored workshop convened by the Duke Margolis Institute for Health Policy in December 2022, more than 800 comments from 65 entities, and experience with more than 500 submissions containing AI components since 2016.^{6,7} The Pharmaceutical Research and Manufacturers of America (PhRMA) asked the agency to clarify which device provisions apply, how human in the loop mitigation affects the risk assessment, and how models composed of multiple models are treated,⁸ a fair summary of what the framework does not yet answer.

For devices, the record is quantitative. We downloaded the FDA AI-enabled medical device list in full: 1614 authorizations current as of September 4, 2026, of which 1230 (76.2%) are in radiology, 154 (9.5%) in cardiovascular, and 73 (4.5%) in neurology, rising from 18 authorizations in 2016 to 335 in 2025 and 181 in the first 8 months of 2026.⁹ The January 2025 draft guidance on AI-enabled device software functions asks for validation at “at least three geographically diverse US clinical sites” with 2-sided 95% confidence intervals (CIs).^{10,11} The concentration in radiology is the number a health economist should notice: three-quarters of authorized AI in American medicine operates

Figure 1. Timeline of Regulatory, HTA, and Journal Policy Instruments Governing AI in Evidence Generation, 2019-2030



Instruments are grouped by issuing domain, with drug and biologic regulation, device regulation, health technology assessment, payer regulation, and publication policy shown separately. Items in force are distinguished from drafts and from provisions with future application dates, including the staged application of Regulation (EU) 2021/2282 from 2025 through 2030 and the general-purpose AI obligations of the EU AI Act from August 2, 2025.

Table 1. Regulatory and HTA Instruments Governing AI in Evidence Generation

Issuing Body	Instrument	Date	Status	Operative Requirement
US FDA	Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products (docket FDA-2024-D-4689)	Jan 2025	Draft as of Sep 2026; comments closed 7 Apr 2025	Seven-step risk-based credibility assessment framework. Model risk is defined as the combination of model influence and decision consequence. Explicitly excludes AI used in drug discovery and in operational efficiencies.
US FDA	Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations (FDA-2024-D-4488)	Jan 2025	Draft	Recommends performance validation at ≥ 3 geographically diverse US clinical sites, with 2-sided 95% confidence intervals reported.
US FDA	AI-Enabled Medical Device List	Current 4 Sep 2026	Published list	1614 authorizations. Radiology 1230 (76.2%), cardiovascular 154 (9.5%), neurology 73 (4.5%). Rose from 18 authorizations in 2016 to 335 in 2025.
US FDA	Considerations for the Use of Real-World Data and Real-World Evidence, electronic health records and medical claims (FDA-2020-D-2307)	2024	Final guidance	States that FDA does not endorse any specific AI technology for real-world data curation.
US FDA and EMA	Guiding Principles of Good Artificial Intelligence Practice in Drug Development	Jan 2026	Joint guiding principles	Ten principles including human-centric by design, risk-based approach, clear context of use, data governance and documentation, risk-based performance assessment, and life cycle management.
US FDA	Digital Health Advisory Committee on generative AI-enabled devices	20-21 Nov 2024	Advisory committee meeting	First dedicated advisory committee consideration of generative AI in regulated devices.
EMA	Reflection paper on the use of artificial intelligence in the medicinal product lifecycle (EMA/CHMP/CVMP/83833/2023)	9 Sep 2024	Final reflection paper	States that the use of transparent models is preferred. Applies a risk-based approach across the product lifecycle.
EMA CHMP	Qualification opinion on prognostic covariate adjustment (PROCOVA)	2022	Qualification opinion	Qualified for use, while stating that excluding dropout data would not be acceptable for regulatory purposes.
European Union	Regulation (EU) 2024/1689, the AI Act	In force 1 Aug 2024	In force; general-purpose AI obligations apply from 2 Aug 2025	Risk-tiered obligations. EFPIA has argued that AI used solely for medicines research and development is exempt under Recital 25 and Articles 2.6 and 2.8.
European Union	Regulation (EU) 2021/2282 on health technology assessment	In force 11 Jan 2022	Applies from 12 Jan 2025 for oncology and ATMPs; 13 Jan 2028 for orphans; 13 Jan 2030 for all others in scope	Joint clinical assessment. Approximately 50 assessments of medicines and 5 of devices estimated for 2026, with 12 ongoing and 14 countries involved.
NICE	Use of AI in evidence generation: NICE position statement (ECD11)	15 Aug 2024	Final position statement	AI should only be used when there is demonstrable value from doing so. Where more explainable methods are potentially robust, those should be presented in the first instance. Directs submitters to the PALISADE checklist.
Canada's Drug Agency (CDA-AMC)	Guidance on Reporting the Use of AI in Evidence Generation	Apr 2025	Final guidance	Built on the principle of augmentation, not replacement. Permits AI generated search strategies, screening, extraction, meta-analysis code, and synthetic or external controls subject to transparent reporting.

Table 1. Regulatory and HTA Instruments Governing AI in Evidence Generation

Issuing Body	Instrument	Date	Status	Operative Requirement
IQWiG	General Methods, version 8.0	2025	Final methods manual	German reference methods manual, incorporating requirements applicable to AI derived evidence.
ICER and Peterson Health Technology Institute	Assessment framework for digital health technologies	2024	Published framework	Extends value assessment methods to digital and algorithm-based technologies.
CMS	HPMS memorandum on Medicare Advantage utilization management (CMS-4201-F)	6 Feb 2024	Final rule guidance	Algorithms or AI alone cannot be used as the basis to deny admission.
State of California	Senate Bill 1120, Chapter 879, Statutes of 2024	2024	Enacted statute	Requires licensed physician review of medical necessity determinations rather than reliance on an algorithm alone.
WHO	Ethics and governance of AI for health: guidance on large multimodal models	2024	Guidance	Notes hallucination rates of ≥3% and as high as 27%.
ISPOR	Generative AI taxonomy, ELEVATE-GenAI reporting guideline, and Working Group report for HTA	2024-2025	Published society guidance	First HEOR-specific reporting instruments for generative AI. AI ranked first among HEOR trends for 2026-2027.

Status is as of September 9, 2026. Quoted text is verbatim from the cited instrument. Full citations with uniform resource locators appear in the reference list.

in one specialty, so the cost-effectiveness literature is being asked to generalize from a narrow deployment base.

The agency is also now a user. Elsa was deployed agency-wide in June 2025 on a FedRAMP High environment, with the explicit statement that it does not train on data submitted by regulated industry.¹² A regulator that uses these tools internally will develop views about their failure modes faster than one that only reviews them.

The European position is more prescriptive on one dimension. The European Medicines Agency reflection paper states that “the use of transparent models is preferred,”¹³ a preference the National Institute for Health and Care Excellence (NICE) echoes in requiring that where more explainable methods are potentially robust “those should be presented in the first instance, with supplementary use of less transparent approaches.”¹⁴ The EU AI Act entered into force on August 1, 2024 with general purpose AI obligations applying from August 2, 2025,¹⁵ and the European Federation of Pharmaceutical Industries and Associations (EFPIA) has argued that AI used solely for medicines research and development is exempt under Recital 25 and Articles 2.6 and 2.8.¹⁶ Industry association frameworks converge on similar content.¹⁷⁻¹⁹

Payer-side regulation has moved faster than either. The Centers for Medicare and Medicaid Services stated in February 2024 that “algorithms or artificial intelligence alone cannot be used as the basis to deny admission” in Medicare Advantage,²⁰ and California enacted Senate Bill 1120 requiring physician review of medical necessity determinations.²¹ World Health Organization guidance on large multimodal models notes hallucination rates of “at least 3 percent” and “as high as 27 percent.”^{22,23} The asymmetry is worth stating plainly: using AI to deny care is now regulated in US statute, while using AI to generate the evidence that determines whether care is covered at all is governed by voluntary position statements.

The HTA Record

Table 2 sets the claims against the record.

NICE published its position statement on August 15, 2024, requiring demonstrable value, a preference for explainable methods, justification of assumptions using instruments such as PALISADE, and naming algorithmic bias, cybersecurity, and reduced human oversight as the principal risks.¹⁴ Canada’s Drug Agency followed in April 2025, drawing explicitly on NICE, permitting model generated search strategies, automated screening and extraction, meta-analysis code, and synthetic data and external control arms subject to transparent reporting, and framing the whole underaugmentation rather than replacement.²⁴⁻²⁷

The workload argument for automation is real. Regulation (EU) 2021/2282 applies from January 12, 2025, for new oncology active substances and advanced therapy medicinal products, from January 13, 2028, for orphan products, and from January 13, 2030, for all others in scope.²⁸⁻³⁰ The HTA Coordination Group estimated approximately 50 joint clinical assessments of medicines and approximately 5 of devices for 2026, with 12 ongoing and 14 countries involved as of December 2025.^{31,32} A doubling of dossier volume with no increase in HEOR headcount is the strongest business case for automation available.

Yet, disclosed use remains 3 documents out of 1391, at NICE technology appraisals TA1154 for AI-assisted study screening, TA1071 for AI-assisted searches to identify treatment effect modifiers, and TA962 for a natural-language processing approach, with no explicit reporting in any Canada’s Drug Agency or IQWiG dossier.³³ A separate audit reported that 7 agencies are adopting AI internally, with NICE recording 49 AI initiatives across 25 appraisals.³⁴ Agencies are adopting faster than submitters are disclosing. Conference activity confirms both volume and concentration: of 28 184 abstracts submitted to ISPOR and ISPE between January 2020 and December 2024, 1153 addressed

AI, with 89% of applications in literature review automation and 9% in economic modeling.³⁵⁻³⁷

The Industry Record

For a pharmaceutical reader, the question is not whether these tools work but whether the reported gains are the kind a budget can be built on. **Table 2** and **Supplementary Table S3** present the claims with provenance labeled.

Reported reductions in regulatory writing time span Pfizer at 40% faster first draft clinical study reports³⁸; Merck at 180 hours to 80 hours per fully human-reviewed first draft, with errors reduced by 50%³⁹; Novo Nordisk at more than 10 weeks to 10 minutes for clinical study documentation⁴⁰; Takeda at a projected halving of a 20-week tabulation process⁴¹; and a Takeda and Weave Bio preprint at a 97% reduction, from about 100 hours to 3.7 hours for 18 870 pages across 61 source documents in one investigational new drug application.⁴² For comparison, Certara, which sells regulatory writing software, describes the usual baseline: a clinical study report takes about 120 hours of writing over 4 weeks, and a fully reviewed and approved report takes an average of 83 days from database lock. Certara does not, however, report a measured time saving for its own product.⁴³

Each of the reductions listed above is reported by the party deploying or selling the tool, and none has been independently replicated. The most methodologically explicit, the Weave and Takeda preprint, compares against a historical estimate with no concurrent control and sample size of 2. Elsewhere, Novartis reported site selection falling from 4 to 6 weeks to a 2-hour meeting for a 14 000-patient trial, and GSK reported approximately £8 million saved in late-stage asthma studies.⁴⁴⁻⁴⁹

On the market-access side, the quantified evidence reduces to a single conference abstract. A retrieval augmented generation multi-agent pipeline ingesting 110 documents and generating sections per AMCP Format version 5.0 produced a draft reported as approximately 90% complete with approximately 80% less development time, while the authors noted repetition in all sections and subject-matter experts who only somewhat agreed that responses were largely complete.^{50,51} Vendor claims for payer landscape monitoring are unaudited.^{52,53} For pricing, gross to net, contracting analytics, and international reference pricing, we located no primary source with a measured, attributable efficiency or accuracy figure. That absence is itself a finding.

Against the efficiency claims stand the failures. DSP-1181, presented as the first AI-designed drug when it entered the clinic in January 2020, was abandoned in January 2022 for failing to meet study criteria, and EXS-21546 was discontinued after modeling suggested a suitable therapeutic index would be difficult to reach, with Exscientia, BenevolentAI, and Recursion having accumulated a combined deficit above US \$1.5 billion.^{54,55} Independent analyses of AI-discovered molecules report phase 1 success of 80% to 90% and phase 2 success of approximately 40%,⁵⁶ against historical benchmarks of 13.8% overall probability of success from phase 1 to approval and 3.4% in oncology across 406,038 trial entries.^{57,58} Counts of AI discovered candidates in trials range from 31 across 8 companies⁵⁹ to 117 assets across 63 companies,⁶⁰ against a total pipeline of several thousand. Phase 2 is where efficacy is tested, and phase 2 success for AI-discovered molecules sits at the historical average, judged against a cost per approved drug of US \$2558 million capitalized preapproval in 2013 dollars⁶¹ or a median of US \$985.3 million (95% CI, 683.6-1228.9) in 2018 dollars.⁶²⁻⁶⁴

Failure Modes

Four failure modes are documented well enough to be expected rather than exceptional.

Bias in deployed algorithms: Obermeyer et al showed that a commercial risk algorithm applied to approximately 200 million

Table 2. Reported Performance and Efficiency of AI Across HEOR Workflows, with the Provenance of Every Claim

HEOR Workflow	Task	Reported Result	Comparator	Provenance
Evidence synthesis	Title and abstract screening	Sensitivity 0.962, specificity 0.996; ~110 records/min at US \$59	Human reviewer decisions in 1381 and 3146 record sets	Peer-reviewed
Evidence synthesis	Title and abstract screening	Sensitivity 81.1%-96.5%, specificity 25.8%-80.4%; 10,279 false positives in the worst configuration	Human reviewers across 22 665 citations	Peer-reviewed
Evidence synthesis	Title and abstract screening	Sensitivity 0.85, specificity 0.98, with a median specificity gap of 0.48 across reviews	Human reviewers across 16 669 citations	Peer-reviewed
Evidence synthesis	Title and abstract screening	Sensitivity 0.42, kappa 0.34 against human reviewer kappa 0.77	2 human reviewers	Peer-reviewed
Evidence synthesis	Screening prioritization	Work saved over sampling 0.55; ~75% of screening effort saved at recall above 95%	Exhaustive human screening of 11 984 records	Peer-reviewed
Evidence synthesis	Agentic screening and extraction, human in the loop	Screening F1 0.917-0.977, kappa 0.844-0.906; extraction F 0.96-0.998	Human reviewers	Peer-reviewed
Evidence synthesis	Data extraction	96.3% accuracy across 160 extracted items	Human extraction	Peer-reviewed
Evidence synthesis	Data extraction	94.8% accuracy at 88 seconds per randomized trial	Human extraction	Peer-reviewed
Evidence synthesis	Data extraction	Model 96.2%, human 95.3%, human-assisted by model 97.9%; time 86.9-14.7 minutes	Human extraction across 12 814 extractions in 107 trials	Peer-reviewed
Evidence synthesis	Time and cost	17/25 studies reported time reductions above 50%, range 36.0% to more than 99.0%	Unaided human workflow	Peer-reviewed
Evidence synthesis	Time and motion, hybrid vs autonomous	Hybrid 97.0% accuracy, 80.6% recall, 51.4% time saved; autonomous 81.1% time saved but 48.4% recall; unaided humans 90.1% accuracy, 88.7% recall	Unaided human workflow across 234 publications	Conference abstract
Economic modeling	Automated model coding in R	14/15 scripts error-free in the lung cancer set; every error-free script reproduced the published ICER to within 1%	Published cost effectiveness results	Peer-reviewed
Economic modeling	Network meta-analysis input extraction	>99% extraction accuracy	Human extraction	Peer-reviewed
Economic modeling	Expected value of partial perfect information by metamodeling	12.5 million model runs reduced from more than 1000 days to ~2 days, a roughly 500-fold reduction	Brute force nested simulation	Peer-reviewed
Economic modeling	Generative AI in health economic evaluation, evidence base composition	25 studies included from 2322 screened; 92% conference abstracts, 92% published 2024, 96% company authored	Not applicable	Preprint
RWD and causal inference	High-dimensional propensity score	Estimated relative risk moved from 1.09 crude to 0.94 with 15 investigator specified covariates to 0.88 with 500 algorithmically selected covariates	Investigator specified propensity score	Peer-reviewed
RWD and causal inference	High dimensional propensity score weighting, simulation	Bias 0.094 with 23.6% CI coverage, against 0.018 bias and 91.4% coverage for a modified procedure	Known simulated truth	Peer-reviewed
RWD and causal inference	Trial emulation, PARADIGM-HF	Logistic propensity model AUC 0.79 gave HR 0.93; gradient boosted model AUC 0.95 gave HR 0.61	Randomized trial HR 0.81 (95% CI 0.61-1.06)	Preprint
Unstructured data	Social and behavioral determinants from notes	Prevalence recorded rose from 0.03% to 8.3% in claims to 1.40% to 36.3% with notes; lack of social support text only in 99.8% of cases	Structured claims and Z codes	Peer-reviewed
Unstructured data	ECOG performance status extraction	AUROC 0.95 across 79 698 patients and 495 862 notes; imputed poor performance status HR 11.9	Documented performance status	Peer-reviewed

Table 2. Reported Performance and Efficiency of AI Across HEOR Workflows, with the Provenance of Every Claim

HEOR Workflow	Task	Reported Result	Comparator	Provenance
Unstructured data	Oncology variable extraction	Stage value precision 98.8% and recall 92%, but histology linked to neoplasm recall 79%	Human abstraction of 200 patients	Peer-reviewed
Unstructured data	Cancer registry staging extraction	T stage F1 0.89, M stage F1 0.24	Registry abstraction	Peer-reviewed
Unstructured data	Real-world progression-free survival derivation	Sentence-level accuracy 98.2%; patient-level within 30 days 88%; median rwPFS 20 months (95% CI 18-25) computationally vs 25 months (15-35) manually; 30 months from radiology reports vs 23 months from clinical notes	Manual curation in a 100 patient subset	Peer-reviewed
Unstructured data	Human-abstracted real-world response, the reference standard	Inter-abtractor agreement 0.94 (0.92-0.96) on assessment presence and 0.82 (0.78-0.86) on best response; 15.0 minutes per patient line; Spearman 0.99 vs trial ORR	Second human abtractor	Peer-reviewed
Unstructured data	Scaled machine learning real-world response	r = 0.86 against human abstraction in a test subset of 4047	Human abstraction	Conference abstract
Unstructured data	AI-augmented vs traditional real-world data	Recall 49.9%-93.2%, precision 73.6%-93.5%, completeness 46.1% (38.2-54.0) to 96.6% (85.8-107.4)	Traditional structured approach	Peer-reviewed
Unstructured data	Vendor accuracy claim	>92% accuracy	Not stated	Company communication
Synthetic data	Control-arm simulation, generator comparison	CART-based generation recovered the real median PFS inside the 95% CI in 88.8%-98.1% of replicates, against 6.5%-37.8% for a conditional tabular GAN	Real trial control arm	Peer-reviewed
Synthetic data	Synthetic control arm in myelodysplastic syndromes	Clinical fidelity 93%, genomic fidelity 90%, 7133 patients, 187 matched synthetic controls	Real patient records	Peer-reviewed
Synthetic data	Reporting completeness	70/73 studies (95%) evaluated utility; 31 of 67 (46%) claiming privacy purpose evaluated privacy; 8 of 73 (11%) used a differentially private generator	Not applicable	Peer-reviewed
Regulatory writing	Clinical study report first draft	40% faster first draft; manuscript submission timeline 15% shorter	Historical internal process	Company communication
Regulatory writing	Clinical study report first draft	180 hours to 80 hours per fully human-reviewed draft; errors reduced 50%; first draft in as little as 5 minutes	Historical internal process	Company communication
Regulatory writing	Clinical study documentation	>10 weeks to 10 minutes; 90% reduction in CSR writing time; review cycles halved; 95% reduction for device verification protocols	Historical internal process, baseline 2.3 CSRs per writer per year	Company communication
Regulatory writing	Analysis and tabulation process	20-week process projected to be cut approximately in half	Historical internal process	Company communication
Regulatory writing	IND document generation	~100 hours to 3.7 hours for 18 870 pages across 61 documents; to 2.6 hours for 11 425 pages across 58 documents	Historical estimate, no concurrent control, n = 2 INDs	Preprint
Regulatory writing	Industry baseline	~120 hours of regulatory writing over 4 weeks; 83 days from database lock to approved CSR	Not applicable, baseline only	Company communication
Clinical development	Site selection for a 14 000 patient cardiovascular outcomes trial	4-6 weeks reduced to a 2-hour meeting; enrollment closed 13 patients above target	Historical internal process	Company communication
Clinical development	Late-stage study cost	Approximately GBP 8 million saved; all trials targeted 15% faster	Historical internal process	Company communication
Market access	Payer dossier drafting per AMCP Format v5.0	Draft ~90% complete; ~80% less development time; 70%-75% lower cost; authors noted repetition in all sections	Single team, no comparator, 110 source documents	Conference abstract
Market access	Payer policy monitoring	Policy and formulary alerts within 24-48 hours; analysis in minutes rather than days or weeks	Not stated	Company communication

Market access	Pricing, gross to net, contracting, and international reference pricing analytics	No primary source with a measured, attributable efficiency or accuracy figure was located	Not applicable	Not available
Drug discovery	AI-discovered molecule phase success	Phase 1 success 80%-90%; phase 2 success approximately 40%	Historical benchmarks of 13.8% phase 1 to approval overall and 3.4% in oncology	Peer-reviewed
Drug discovery	AI-discovered assets in clinical development	31 assets across 8 companies; 117 assets across 63 companies, with 60 of 117 (51.3%) completing phase 1 and 8 of 117 (6.8%) completing phase 2	Total clinical pipeline of several thousand	Peer-reviewed

Abbreviations: AUC, area under the curve; AUROC, area under the receiver operating characteristic curve; GAN, generative adversarial network; HR, hazard ratio; ICER, incremental cost-effectiveness ratio; ORR, objective response rate; rwPFS, real-world progression-free survival.

Provenance is classified as peer reviewed, preprint, conference abstract, or company communication. Company communication denotes a figure reported by the organization deploying or selling the tool, without independent replication or a concurrent control. Every company reported efficiency figure in this table falls into that category. Where no primary source with a measured figure could be located, the cell records that absence rather than an estimate.

Americans, by using cost as a proxy for need, identified 17.7% of Black patients for extra help where correcting the target would have identified 46.5%.⁶⁵ The mechanism was label choice, not model class, which is why it is a HEOR problem: the outcome variable was a health economics variable.

Weak external validity: The Epic sepsis model, evaluated externally across 38 455 hospitalizations, achieved area under the curve of 0.63 (95% CI, 0.62-0.64) against a vendor reported range of 0.76 to 0.83.⁶⁶ High risk of bias is the norm rather than the exception here.^{67,68}

Post deployment decay: Rahmani et al documented data drift across 112 972 encounters,⁶⁹ and Kopanitsa found across 4 deployed clinical AI systems that calibration degraded earlier and more consistently than discrimination.⁷⁰ For HEOR, this is the crux, because a cost-effectiveness result depends on calibration, not on ranking. A model that still discriminates well and no longer calibrates will keep passing the metric that is monitored and stop supporting the estimate that is used.

Human factors: Jabbour et al showed in a randomized vignette study of 457 clinicians that a standard AI raised diagnostic accuracy from a baseline of 73.0% by 2.9 percentage points, while a systematically biased AI lowered it by 11.3 points, with explanations providing no significant protection.⁷¹ Dratsch et al found experienced radiologists falling from 82.3% to 45.5% accuracy when shown an incorrect AI suggestion,⁷²⁻⁷⁴ and Budzyń et al reported adenoma detection in non-AI-assisted colonoscopies falling by 6.0 percentage points after AI exposure, from 28.4% to 22.4% (95% CI, 10.5 to -1.6 lower; $P=.0089$).^{75,76} Goh et al randomized 50 physicians and found diagnostic reasoning scores of 76% with a large language model against 74% without, an adjusted difference of 2 percentage points (95% CI, -4 to +8; $P=.60$), while the model alone scored 16 points higher than physicians (95% CI, 2-30; $P=.03$).⁷⁷ The model beat the physician and beat the physician using the model. Any workflow that assumes a human reviewer will catch the model's errors is assuming away the best measured finding in this literature.

Fabrication is the failure mode most specific to evidence work. Chelli et al assessed 471 references and found hallucination rates of 39.6% for GPT-3.5, 28.6% for GPT-4, and 91.4% for Bard,⁷⁸ and Walters and Wilder examined 636 citations and found 55% fabricated by GPT-3.5 and 18% by GPT-4.⁷⁹ This has already reached consequential documents: a US federal health policy report cited studies that do not appear to exist, with a listed first author confirming she had never written the paper attributed to her,⁸⁰ a filed expert declaration in federal court contained a citation to a nonexistent statistics article,^{81,82} and one journal retracted 129 papers in the first 6 weeks of 2025.⁸³ No

equivalent incident has yet been documented inside a drug regulatory or HTA submission. That is the state of the record, not a reassurance.

The Economic Evaluation of AI Itself

HEOR is being asked to evaluate AI and has barely done so. El Arab et al found 19 economic evaluations of AI in health care published from 2020 through 2024, with CHEERS scores between 21 and 27 of 28.⁸⁴ Nineteen evaluations against 1614 authorized AI-enabled devices⁹ is the field's most direct self-indictment (**Supplementary Table S1**).

The evaluations that exist are instructive. Mital and Nguyen estimated an incremental cost effectiveness ratio of US \$23 755 per quality-adjusted life-year for AI vs polygenic risk scores in breast cancer risk stratification, with an AI cost threshold of US \$318 per mammogram.^{85,86} Against this, Vargas-Palacios et al modeled AI in the UK breast screening program and found a scenario in which no price per scan made the technology cost effective.⁸⁷ That is the result to hold onto: for a diagnostic AI, cost effectiveness is frequently determined by the downstream care pathway rather than by the price of the algorithm, so a procurement negotiated on price per scan is negotiating the wrong parameter.⁸⁸⁻⁹⁰

The regulatory record on external controls is more informative than the synthetic data literature acknowledges. Jahanshahi et al identified 45 FDA approvals between 2000 and 2019 that accepted external control data, 36 (80%) in rare disease,^{91,92} and in its qualification opinion on prognostic covariate adjustment, the EMA Committee for Medicinal Products for Human Use stated that excluding dropout data "would not be acceptable for regulatory purposes."⁹³ The model may be qualified, and the analysis population still has to be complete.

Transparency, Disclosure, and Workforce

The baseline was poor before AI was introduced. In a random sample of 200 health-economic evaluations, 10 (5%) mentioned access to raw data and analytic code,⁹⁴ and in healthcare discrete event simulation 47 of 564 studies (8.3%) cited an openly available model.⁹⁵⁻⁹⁸ Replication attempts are sobering: of 5 economic analyses reimplemented, 2 were fully replicated and 2 were unequivocally not replicable⁹⁹; 4 obesity models rebuilt from 87 publications showed cost-utility ratio deviations from 3.0% lower to 17.9% higher.¹⁰⁰

Disclosure of AI use runs far below reported use. Journals and societies have converged on rules that prohibit AI authorship and require disclosure of model name, version, and dates.¹⁰¹⁻¹¹⁰ These are tabulated in **Supplementary Table S4**. Against those rules, 0.6% of submissions to *The Lancet Infectious Diseases* in February 2024 declared ChatGPT use and 1.8% declared AI as a research tool,¹¹¹ while surveys report use between 25% and 81%, including more than half of about 1600

academics across 111 countries having used AI in peer review and 81% of 1692 physicians aware of or using AI in 2026 against 38% in 2023.¹¹²

¹¹⁷ The survey estimates are shown in **Supplementary Table S5**.

Within HEOR, the workforce is not yet equipped. Of 133 HEOR, real-world data, and medical affairs professionals surveyed in April 2026, 61% rated their AI proficiency as basic or very limited and 4% as expert.¹¹⁸ Among 100 HEOR leaders at 72 companies, 69% personally use AI, 29% expect AI to replace some HEOR staff within 2 years, and 59% report that their team does not use AI for any listed HEOR workflow.¹¹⁹ Medical affairs reports higher confidence than capability, with 49% still using full manual expert review of every AI generated output.^{120,121} Competency frameworks are appearing, but none is HEOR specific,¹²²⁻¹²⁷ while capital has already moved: AI-enabled startups took 54% of US digital health funding in 2025 against 37% in 2024.¹²⁸⁻¹³²

DISCUSSION

The governance finding is the sharpest contrast in this series. Adoption claims run at 40% to 97% time savings in pharmaceutical regulatory writing and approximately 80% in payer dossier drafting, while disclosed use in HTA submissions is 3 of 1391 documents (0.2%) and declared use in journal submissions is 0.6%. There are only three explanations. Either the tools are not actually used in submission grade work, which the industry evidence makes implausible, or they are used and not disclosed, which the survey evidence makes likely. Alternatively, the disclosure requirement is understood to apply to authorship rather than to method, so a systematic review screened by a large language model is not thought to require declaration in the way a manuscript drafted by one would be. The third explanation carries most of the weight and is fixable, because it is a definitional failure rather than a compliance failure. Current instruments ask whether AI wrote the text. Reviewers need to know whether AI selected the evidence, defined the cohort, extracted the parameter, or generated the code.

What Acemoglu's Critique Means for This Field

Daron Acemoglu, awarded the 2024 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel with Simon Johnson and James Robinson,¹³³ estimates total factor productivity gains no greater than 0.66% over 10 years, roughly 0.064% annually, with gross domestic product effects near 1%. The estimate derives from a task-based decomposition rather than from pessimism about model capability.^{134,135} His central mechanism is the distinction between easy and hard learning tasks. A task is easy to learn when outcomes are observed, labels are abundant, and feedback is fast, and hard to learn when the correct answer is not observed. His illustration is clinical: diagnosing a persistent cough is hard because "there is no large, well-curated data set of successful diagnoses and cures."

Health economics is the limiting case of a hard learning task, for structural rather than technological reasons. The quantity of interest is a counterfactual, so no data set of correct incremental cost effectiveness ratios exists, will exist, or could exist. This is why the empirical pattern reported in Part 1 is exactly what his framework predicts: the gains disappear precisely where the label disappears.

Acemoglu's second mechanism is task share. Even large improvements on a small share of total task time yield small aggregate gains.^{136,137} Applied to HEOR, the clerical layer is where the 40% to 97% figures live, and it is a minority of the total work and the least remunerated portion of it. The judgment layer, choosing the comparator, specifying the model structure, defending the extrapolation, negotiating the price, is unaffected. Under this arithmetic an 80% reduction in a task that is

20% of the total yields a 16% aggregate gain in the best case, before accounting for the review burden the automation creates.

The counterargument deserves fair statement, because it is empirically supported. Noy and Zhang found writing time down 40% and quality up 18% across 453 professionals with the largest gains for weaker performers,¹³⁸ Brynjolfsson et al found 14% to 15% more resolutions per hour across 5179 support agents with 34% gains for novices,^{139,140} and Peng et al found programmers 55.8% faster.¹⁴¹ Three features are uncomfortable for HEOR. The gains concentrate in the least experienced, which compresses the junior training pathway through which HEOR expertise has been built. They are largest where output is verifiable, which HEOR output is not. And the direction reverses outside the frontier of capability: in a randomized study of 758 consultants, those using AI inside that frontier completed 12.2% more tasks at a 25.1% faster rate, while on a task outside it, accuracy fell from 84.5% to 70.6% with AI and to 60.0% with AI plus an overview of its limits.¹⁴² The overview made it worse. HEOR's hardest tasks sit outside the frontier, and its practitioners cannot yet reliably tell which side a given task is on, since 61% rate their proficiency as basic or very limited.¹¹⁸

What follows is not that AI will not matter. It is that the mechanism by which it matters will be cost structure rather than evidence quality, and that the two are separable. The realistic 10-year expectation is a HEOR function producing more dossiers per analyst at similar quality, with a compressed junior tier, a widened tail of low-quality submissions from parties who could not previously afford to submit at all, and a corresponding increase in the burden on reviewers. Reviewers, not producers, absorb the externality. That is the case for the reporting standard below, and it is an economic case rather than an ethical one.

TRACE-HEOR: A Reporting Standard for AI-Assisted HEOR Evidence

Existing instruments cover the parts of this problem that resemble clinical prediction or clinical trials: TRIPOD+AI and PROBAST+AI for prediction model reporting and bias,^{143,144} CONSORT-AI and SPIRIT-AI for trials of AI interventions,^{145,146} DECIDE-AI for early live clinical evaluation,¹⁴⁷ STARD-AI for diagnostic accuracy,¹⁴⁸ TRIPOD-LLM for large language model studies,¹⁴⁹ CANGARU for disclosure in scholarly writing,¹⁵⁰ CHEERS 2022 for economic evaluation reporting,¹⁵¹ PALISADE for real-world data study credibility,¹⁵² and ELEVATE-GenAI with the ISPOR taxonomy for the HEOR-specific work.¹⁵³⁻¹⁵⁵ None answers the question an HTA reviewer actually has, which is whether the reimbursement recommendation would change if the AI component were removed.

We propose TRACE-HEOR, for Transparent Reporting of Artificial intelligence Components in Health Economics and Outcomes Research. The following 10 items are designed as an extension to CHEERS 2022 and PALISADE rather than as a replacement, and stated so that each is verifiable by a reviewer with access only to the submission. **Table 3** presents them in full, with the evidence supporting each.

1. Declare the function, not the tool. State, for each AI component, which of the following it performed: evidence identification or screening, data extraction, cohort or exposure definition, outcome or endpoint derivation, covariate selection, effect estimation, parameter derivation, model code generation, synthetic data or external control generation, or narrative drafting. Name the model, version, vendor, and dates of use, as *JAMA* already requires.¹⁰⁵
2. Specify the context of use and the model risk, using the FDA's 2-factor definition of model influence and decision consequence,⁵ and state which decision the model influences.

Table 3. TRACE-HEOR: Proposed 10-Item Standard for Transparent Reporting of AI Components in Health Economics and Outcomes Research

Item	Short Name	Requirement	Empirical Basis	Reviewer Verifiable
1	Function, not tool	Declare, for each AI component, which function it performed: evidence identification or screening, data extraction, cohort or exposure definition, outcome or endpoint derivation, covariate or confounder selection, effect estimation, parameter derivation, model code generation, synthetic data or external control generation, or narrative drafting. Name the model, version, vendor, and dates of use.	Disclosure instruments currently ask whether AI wrote the text. Disclosed AI use was 3/1391 HTA documents (0.2%) and 0.6% of <i>Lancet</i> submissions, against self-reported use of 25%-69%.	Yes
2	Context of use and model risk	State the question of interest, the context of use, and the model risk using the 2-factor definition of model influence and decision consequence; name the decision the model influences.	Adopts the FDA credibility assessment framework verbatim so that a single characterization serves both regulatory and HTA review.	Yes
3	Calibration, not only discrimination	Report calibration in the deployment population and time period, not only discrimination. Where the estimate depends on absolute risk or absolute cost, discrimination alone is insufficient.	Across 4 deployed clinical AI systems, calibration degraded earlier and more consistently than discrimination. Cost effectiveness results depend on calibration.	Yes
4	Reliability of the human reference	Where AI output is validated against human abstraction, report the reliability of that human reference, including inter-abstractor agreement.	The reference standard for scaled machine learning response abstraction itself had agreement of 0.94 on assessment presence and 0.82 on best response category.	Yes
5	Disaggregated, variable level performance	Report performance separately for each variable the analysis depends on. Aggregate F1 or accuracy is not sufficient.	Stage value reached 98.8% precision while histology linkage reached 79% recall in the same model; T stage reached F1 0.89 while M stage reached 0.24.	Yes
6	Decision reversal analysis	Re-estimate the primary result with the AI component replaced by the pre-AI alternative or by a plausible worst case, and state whether the incremental cost effectiveness ratio crosses the relevant threshold or whether the recommendation changes.	No existing instrument requires this. Deriving real-world progression from radiology reports rather than clinical notes moved median survival by 7 months, which is decision relevant.	Yes
7	Bias and subgroup performance	Report performance for the subgroups the decision affects and report the demographic composition of any training or synthetic population.	A commercial risk algorithm applied to approximately 200 million people identified 17.7% of Black patients for extra help where a corrected target identified 46.5%.	Yes
8	Synthetic data and external controls	Report generator family, fidelity against the real target quantity rather than against summary statistics, and a privacy evaluation.	Generator choice alone moved recovery of the true median progression free survival from 6.5% to 98.1% of replicates. Only 11% of studies used a differentially private generator.	Yes
9	Human oversight actually performed	Report quantitatively what fraction of AI output was reviewed, by whom, with what qualifications, and what the disagreement rate was.	A biased AI lowered clinician diagnostic accuracy by 11.3 percentage points and explanations did not protect; experienced radiologists fell from 82.3% to 45.5% with an incorrect suggestion.	Yes
10	Archive prompts, code, and versions	Deposit prompts, code, model versions, random seeds, and intermediate outputs in a citable archive at the time of submission.	5% of health-economic evaluations mention code availability. An AI assisted analysis is not auditable without the prompts.	Yes

TRACE-HEOR is designed as an extension to CHEERS 2022 and PALISADE rather than as a replacement and is aligned to the FDA credibility assessment framework and to ELEVATE-GenAI. Items 1, 2, 6, 9, and 10 constitute the minimum set. Item 6 is the item no current instrument requires and the item a reimbursement committee needs most. Every item is stated so that it can be verified by a reviewer with access only to the submission.

3. Report calibration, not only discrimination, in the deployment population and time period. Where an estimate depends on absolute risk or absolute cost, discrimination alone is insufficient.⁷⁰
4. Report the human reference standard and its own reliability. Where AI output is validated against human abstraction, report inter-abstractor agreement for the reference.^{156,157}
5. Report performance for the variables the analysis depends on, disaggregated. Aggregate F1 conceals the pattern in which stage extraction reaches 0.988 precision, while histology linkage reaches 0.79 recall¹⁵⁸ and T stage reaches F1 0.89 while M stage reaches 0.24.¹⁵⁹
6. Report a decision reversal analysis. Re-estimate the primary result with the AI component replaced by the pre-AI alternative, or by a plausible worst case, and state whether the incremental cost-effectiveness ratio crosses the relevant threshold or whether the recommendation changes. This item exists in no current instrument and is the one reviewers need most.
7. Report bias and subgroup performance for the populations the decision affects and state the demographic composition of any training or synthetic population.^{65,160}
8. For synthetic data or model based external controls, report generator family, fidelity against the real target quantity rather than against summary statistics, and a privacy evaluation. Generator choice alone moved recovery of the true median progression free survival from 6.5% to 98.1% of replicates.^{161,162}
9. Report the human oversight actually performed, quantitatively: what fraction of AI output was reviewed, by whom, with what qualifications, and the disagreement rate.^{71,72,77,120}
10. Deposit prompts, code, model versions, seeds, and intermediate outputs in a citable archive at submission. Five percent of health economic evaluations currently mention code availability⁹⁴, and an AI-assisted analysis is not auditable without the prompts.

Items 1, 2, 6, 9, and 10 are the TRACE-HEOR minimum set, and a submission reporting only those five is materially more auditable than any submission we located. If an agency were to adopt only one, it should be item 6, because decision reversal is the only item that converts a technical property of a model into the currency a reimbursement committee reasons in.

A Research Agenda

Five priorities follow, ordered by the ratio of decision relevance to current investment.

First, benchmark AI-derived endpoints against regulatory grade human abstraction in a multisponsor, multivendor study with a preregistered protocol, and publish the discordant cases. Current validation is vendor-authored,^{157,163,164} which is a conflict rather than a failing, and the remedy is independent replication.

Second, obtain a regulatory precedent. Sponsors should seek scientific advice or a qualification opinion on an AI-derived endpoint, and agencies should publish the reasoning when they give it, as the EMA did on prognostic covariate adjustment.⁹³ One published opinion would be worth more to this field than 50 conference abstracts.

Third, establish an empirical decision reversal literature. Take completed cost-effectiveness analyses and HTA submissions, substitute AI-assisted components for the original ones, and measure how often the recommendation changes. The replication literature indicates the base rate of nonreplication is already high before AI is introduced,^{99,100} which makes the marginal question both answerable and urgent.

Fourth, run randomized trials of AI-assisted HEOR workflows with quality, not speed, as the primary endpoint. Every existing industry

efficiency figure is single-arm and self-reported, and the designs already exist in adjacent fields.^{2,138,142}

Fifth, evaluate AI itself economically at a scale proportionate to its deployment, and build HEOR-specific competency and audit capacity. Nineteen published economic evaluations against 1614 authorized devices is a ratio of about 1:85^{9,84}; the finding that no price per scan made an AI cost-effective in a national screening program should be reproduced or refuted,^{87,88} and reviewer side capacity is the binding constraint because joint clinical assessment volume rises on a fixed schedule.^{28,32,118}

Limitations

The regulatory, HTA, and industry record assembled here is a narrative review rather than a systematic one, and negative results are less visible in it: a sponsor that abandoned an AI-assisted approach has no reason to publish that fact. Several central quantities rest on conference abstracts and company communications, which we label rather than exclude because they are the only available source. The disclosure denominators are audits of published documents, so they measure disclosure and not use. TRACE-HEOR is a proposal derived from documented failure modes rather than a consensus instrument; it has not been through a Delphi process, and it should be treated as a starting draft for one.

CONCLUSIONS

Regulators have written more about AI in evidence generation than sponsors have disclosed using it. The FDA has authorized 1614 AI-enabled devices and received more than 500 drug and biologic submissions containing AI components, while its credibility framework remains a draft that excludes discovery and operational uses. Industry reports time savings of 40% to 97% in regulatory writing, none independently replicated. Three of 1391 health technology assessment documents disclose an AI component.

The gap is not primarily a compliance problem. Disclosure instruments were built to establish who wrote a document, and the operations that matter in this field are the ones that select the evidence, define the cohort, derive the endpoint, and generate the code. Under the task-based productivity arithmetic that Acemoglu applies to the wider economy, the tasks AI does well in health economics are the minority of the work and the least remunerated part of it, while the burden of checking the output falls on reviewers who did not choose the tool. TRACE-HEOR states ten items that would make an AI-assisted submission auditable, and one of them, a decision reversal analysis, would tell a reimbursement committee the only thing it needs to know.

Disclosures: The author reports no conflicts of interest related to this work. No funding was received for this study. The author is affiliated with an academic institution and a health economics consultancy that conducts real-world evidence studies for pharmaceutical and life sciences clients; no client had any role in the conception, conduct, analysis, or reporting of this review, and no client reviewed the manuscript before submission. The large language model classifier described in Part 1 is a measured object of study in this series rather than an assistive writing tool. Its prompts, inputs, and outputs are archived in the replication package. No AI tool is listed as an author.

Data Availability: All search strings, retrieved records, the classification dictionary, the blinded classifier prompts and outputs, the analysis scripts, the figure generation code, the regulatory and industry source tables, and the complete reference library are available in the public replication package 4. No individual patient data were used.

REFERENCES

- Lai H, Liu J, Bai C, et al. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digit Med*. 2025;8(1):74. doi:10.1038/s41746-025-01457-w
- Lee K, Paek H, Ofoegbu N, et al. A4SLR: an agentic artificial intelligence-assisted systematic literature review framework to augment evidence synthesis for health economics and outcomes research and health technology assessment. *Value Health*. 2025;28(11):1655-1664. doi:10.1016/j.jval.2025.08.002
- Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial intelligence to automate health economic modeling: a case study to evaluate the potential application of large language models. *Pharmacoeconomics Open*. 2024;8(2):191-203. doi:10.1007/s41669-024-00477-8
- Baser O. Replication package for artificial intelligence in health economics and outcomes research: a systematic scoping review. GitHub repository. 2026. Accessed October 6, 2026. <https://github.com/obaser3791/ai-heor-scoping-review>
- US Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products. Draft guidance for industry. Docket FDA-2024-D-4689. Silver Spring, MD; January 2025. Accessed October 6, 2026. <https://www.fda.gov/media/184830/download>
- US Food and Drug Administration. FDA proposes framework to advance credibility of AI models used for drug and biological product submissions. Press announcement. January 6, 2025. Accessed October 6, 2026. <https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions>
- Poddar A, Samson M, Innes GK, et al. Leveraging artificial intelligence in drug and biological product development: an FDA and Clinical Trial Transformation Initiative Workshop Report. *NEJM AI*. 2025;2(12). doi:10.1056/aipc2500801
- Pharmaceutical Research and Manufacturers of America. *PhRMA Comments on Docket FDA-2024-D-4689*. Accessed October 6, 2026. https://cdn.aglty.io/phrma/global/resources/import/pdfs/FDA-2024-D-4689_PhRMA%20Comments.pdf
- US Food and Drug Administration. Artificial intelligence-enabled medical devices. Device list current as of September 4, 2026. Accessed October 6, 2026. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>
- US Food and Drug Administration. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. Draft guidance. Docket FDA-2024-D-4488. January 2025. Accessed October 6, 2026. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing>
- US Food and Drug Administration. Digital health advisory committee meeting: total product life cycle considerations for generative AI-enabled devices. November 20-21, 2024. Accessed October 6, 2026. <https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-112024>
- US Food and Drug Administration. FDA launches agency-wide AI tool to optimize performance for the American people. News release. June 2, 2025. Accessed October 6, 2026. <https://www.fda.gov/news-events/press-announcements/fda-launches-agency-wide-ai-tool-optimize-performance-american-people>
- European Medicines Agency. Reflection paper on the use of artificial intelligence in the lifecycle of medicines. EMA/CHMP/CVMP/83833/2023. September 9, 2024. Accessed October 6, 2026. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medical-product-lifecycle_en.pdf
- National Institute for Health and Care Excellence. Use of AI in evidence generation: NICE position statement. ECD11. August 15, 2024. Accessed October 6, 2026. <https://www.nice.org.uk/corporate/ecd11>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. Accessed October 6, 2026. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- European Federation of Pharmaceutical Industries and Associations. *EFPIA Position on the Use of Artificial Intelligence in the Medicinal Product Lifecycle*. October 2024. Accessed October 6, 2026. <https://efpia.eu/media/tzeavw1t/efpia-position-on-the-use-of-artificial-intelligence-in-the-medicinal-product-lifecycle.pdf>
- International Federation of Pharmaceutical Manufacturers and Associations. *IFPMA Artificial Intelligence Ethics Principles*. July 2022. Accessed October 6, 2026. https://www.ifpma.org/wp-content/uploads/2023/01/i2023_IFPMA-Artificial-Intelligence-AI-Ethics-Principles.pdf
- Novartis. Novartis commitment to the ethical and responsible use of AI systems. Accessed October 6, 2026. https://www.novartis.com/sites/novartis_com/files/novartis-responsible-use-of-ai-systems.pdf
- Roche. Roche AI ethics principles. Updated April 1, 2025. Accessed October 6, 2026. <https://assets.roche.com/f/176343/x-/401c28049f/roche-ai-ethics-principles.pdf>
- Centers for Medicare and Medicaid Services. Coverage Criteria and Utilization Management Requirements in CMS-4201-F. Health Plan Management System memorandum. February 6, 2024. Accessed October 6, 2026. <https://www.cms.gov/files/document/faq-contract-year-2024-policy-and-technical-changes-medicare-advantage-program-final-rule-utilization.pdf>
- California Senate Bill 1120, Chapter 879, Statutes of 2024 (Physicians Make Decisions Act). Accessed October 6, 2026. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1120
- World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*. Geneva: WHO; 2024. ISBN 978-92-4-008475-9.
- World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva: WHO; 2021. ISBN 978-92-4-002920-0.
- Canada's Drug Agency. Position statement on the use of artificial intelligence in the generation and reporting of evidence. April 2025. Accessed October 6, 2026. https://www.cda-amc.ca/sites/default/files/MG%20Methods/Position_Statement_AI_Renumbered.pdf
- CDA-AMC. Development of an evaluation instrument on artificial intelligence search tools for evidence synthesis. *CJHT*. 2024;4(10). doi:10.51731/cjht.2024.1004
- Institute for Quality and Efficiency in Health Care. General methods; version 8.0. IQWiG; 2026. doi:10.60584/general-methods_v8.0
- Pearson SD, Singh P, Beaudoin F, et al. Institute for Clinical and Economic Review: Peterson Health Technology Institute value

- assessment framework for digital health technologies. *J Comp Eff Res*. 2023;12(12):e230154. doi:10.57264/cer-2023-0154
28. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment. *Official Journal of the European Union*. Accessed October 6, 2026. <https://eur-lex.europa.eu/eli/reg/2021/2282/oj/eng>
 29. European Commission. Implementation of the regulation on health technology assessment. Accessed October 6, 2026. https://health.ec.europa.eu/health-technology-assessment/implementation-regulation-health-technology-assessment_en
 30. European Commission. Joint clinical assessments. Accessed October 6, 2026. https://health.ec.europa.eu/health-technology-assessment/implementation-regulation-health-technology-assessment/joint-clinical-assessments_en
 31. Barbuto V, European Commission DG SANTE. The HTA regulation. Presentation to the 15th EMA Industry Standing Group. December 11, 2025. Accessed October 6, 2026. https://www.ema.europa.eu/en/documents/presentation/presentation-hta-regulation-v-barbuto-ec_en.pdf
 32. Health Technology Assessment Coordination Group. *Annual Work Programme 2026*. November 28, 2025. Accessed October 6, 2026. https://health.ec.europa.eu/document/download/81db097f-a9cb-4404-837c-115b4289321d_en
 33. Use of artificial intelligence and machine learning in evidence synthesis for health technology assessment submissions: a review of NICE, CDA, and IQWiG documents. Poster presented at: ISPOR 2026; Philadelphia, PA; May 17-20, 2026. Accessed October 6, 2026. <https://www.ispor.org/docs/default-source/cti-meeting-21463-documents/9774d614-3a98-4fba-bcf4-c1ab426db452.pdf>
 34. Use of artificial intelligence in health technology assessment submissions and agency practice. Poster MSR96 presented at: ISPOR 2026; Philadelphia, PA. Accessed October 6, 2026. <https://www.ispor.org/docs/default-source/cti-meeting-21463-documents/edf7be4a-ac97-45e7-82f1-ed3fc505aa61.pdf>
 35. Emerging trends in AI applications shaping pharmacoepidemiology and health technology assessments. Poster presented at: ISPOR 2025. Accessed October 6, 2026. <https://www.ispor.org/heor-resources/presentations-database/presentation-cti/ispor-2025/poster-session-3/emerging-trends-in-ai-applications-shaping-pharmacoepidemiology-and-health-technology-assessments>
 36. Assessing the use of artificial intelligence in publication-grade research: a literature review of applications in HEOR. Poster SA14 presented at: ISPOR Europe 2025. Accessed October 6, 2026. <https://www.ispor.org/heor-resources/presentations-database/presentation-cti/ispor-europe-2025/poster-session-1-2/assessing-the-use-of-artificial-intelligence-in-publication-grade-research-a-literature-review-of-applications-in-heor>
 37. Artificial intelligence tools in health technology assessment: a scoping review. Conference proceeding OP07. Accessed October 6, 2026. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11718724/>
 38. Pfizer. How AI is already reshaping cancer care. Accessed October 6, 2026. <https://www.pfizer.com/news/articles/how-ai-is-already-reshaping-cancer-care>
 39. Merck & Co. Merck expands innovative internal generative AI solutions, helping to deliver medicines to patients faster. News release. June 25, 2025. Accessed October 6, 2026. <https://www.merck.com/news/merck-expands-innovative-internal-generative-ai-solutions-helping-to-deliver-medicines-to-patients-faster/>
 40. Anthropic. Novo Nordisk transforms clinical documentation with Claude. Customer story. Accessed October 6, 2026. <https://claude.com/customers/novo-nordisk>
 41. Takeda. Aiming to enhance regulatory submissions with GenAI. December 12, 2024. Accessed October 6, 2026. <https://www.takeda.com/our-impact/our-stories/aiming-to-enhance-regulatory-submissions-with-genai/>
 42. Automated generation of investigational new drug applications with large language models. arXiv preprint 2509.09738. Accessed October 6, 2026. <https://arxiv.org/pdf/2509.09738>
 43. Certara. A tailored generative AI solution for regulatory writing. White paper. November 2024. Accessed October 6, 2026. <https://www.certara.com.cn/app/uploads/2024/11/2024-11-08-White-Paper-CoAuthor-and-GPTs.pdf>
 44. Reuters. Drugmakers turn to AI to speed trials, regulatory submissions. January 26, 2026. Accessed October 6, 2026. <https://www.reuters.com/legal/litigation/drugmakers-turn-ai-speed-trials-regulatory-submissions-2026-01-26/>
 45. Merck & Co. Merck and Google Cloud partner to accelerate agentic AI enterprise transformation. News release. April 22, 2026. Accessed October 6, 2026. <https://www.merck.com/news/merck-and-google-cloud-partner-to-accelerate-agentic-ai-enterprise-transformation/>
 46. Eli Lilly and Company. Lilly partners with NVIDIA to build industry's most powerful AI supercomputer owned and operated by a pharmaceutical company. October 28, 2025. Accessed October 6, 2026. <https://investor.lilly.com/news-releases/news-release-details/lilly-partners-nvidia-build-industrys-most-powerful-ai>
 47. Tempus AI. Tempus signs expanded strategic agreements with AstraZeneca and Pathos AI. April 23, 2025. Accessed October 6, 2026. <https://investors.tempus.com/news-releases/news-release-details/tempus-signs-expanded-strategic-agreements-astrazeneca-and/>
 48. OpenAI. Moderna: advancing mRNA medicine with generative AI. Customer story. Accessed October 6, 2026. <https://openai.com/index/moderna/>
 49. Sanofi. Sanofi to become the first pharma company powered by artificial intelligence at scale. News release. June 13, 2023. Accessed October 6, 2026. <https://www.sanofi.com/en/media-room/press-releases/2023/2023-06-13-12-00-00-2687072>
 50. Automated AMCP dossier development using a retrieval-augmented generation (RAG) based multi-agent approach. Abstract MSR207. *Value Health*. 2026;29(S6). Accessed October 6, 2026. <https://www.ispor.org/heor-resources/presentations-database/presentation-cti/ispor-2026/poster-session-5-3/automated-amcp-dossier-development-using-a-retrieval-augmented-generation-rag-based-multi-agent-approach>
 51. Academy of Managed Care Pharmacy. AMCP format for formulary submissions. Accessed October 6, 2026. <https://www.amcp.org/resource/amcp-format-formulary-submissions-guidance>
 52. Norstella. Norstella AI: principles and product capabilities. Accessed October 6, 2026. <https://ai.norstella.com/>
 53. IQVIA. Market access insights. Fact sheet. November 7, 2023. Accessed October 6, 2026. <https://www.iqvia.com/library/fact-sheets/market-access-insights>
 54. Endpoints News. The first AI-designed drugs fall short in the clinic following years of hype. Accessed October 6, 2026. <https://endpoints.news/first-ai-designed-drugs-fall-short-in-the-clinic-following-years-of-hype/>
 55. Fierce Biotech. Several months after Exscientia merger, AI outfit recursion reworks pipeline. Accessed October 6, 2026. <https://www.fiercebiotech.com/biotech/several-months-after-exscientia-merge-ai-outfit-recursion-reworks-pipeline>
 56. Jayatunga MK, Ayers M, Bruens L, Jayanth D, Meier C. How successful are AI-discovered drugs in clinical trials? A first analysis

- and emerging lessons. *Drug Discov Today*. 2024;29(6):104009. doi:10.1016/j.drudis.2024.104009
57. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. 2018;20(2):273-286. doi:10.1093/biostatistics/kxx069
 58. Wong CH, Siah KW, Lo AW. Corrigendum: estimation of clinical trial success rates and related parameters. *Biostatistics*. 2018;20(2):366-366. doi:10.1093/biostatistics/kxy072
 59. Wilczok D, Zhavoronkov A. Progress, pitfalls, and impact of AI-driven clinical trials. *Clin Pharmacol Ther*. 2024;117(4):887-890. doi:10.1002/cpt.3542
 60. Haque W, Sai-Palm K, Kasparcova MZ, et al. Clinical-stage AI-enabled therapeutic assets: an industry-wide analysis of scale, clinical progression, and company characteristics. *J Clin Oncol*. 2026;44(16_suppl):11072-11072. doi:10.1200/jco.2026.44.16_suppl.11072
 61. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Econ*. 2016;47:20-33. doi:10.1016/j.jhealeco.2016.01.012
 62. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*. 2020;323(9):844. doi:10.1001/jama.2020.1166
 63. Wouters OJ, McKee M, Luyten J. Errors in source data for study of drug development costs. *JAMA*. 2022;328(11):1110. doi:10.1001/jama.2022.14317
 64. Error in data in study of estimated cost of bringing a drug to market. *JAMA*. 2022;328(11):1111. doi:10.1001/jama.2022.14733
 65. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
 66. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065. doi:10.1001/jamainternmed.2021.2626
 67. Wynants L, Van CB, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*. 2020;m1328. doi:10.1136/bmj.m1328
 68. Andaur NCL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;n2281. doi:10.1136/bmj.n2281
 69. Rahmani K, Thapa R, Tsou P, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. 2022. doi:10.1101/2022.06.06.22276062
 70. Kopanitsa G. Validation is not enough: longitudinal evidence of post-deployment fragility in clinical AI systems. *PLOS Digit Health*. 2026;5(7):e0001534. doi:10.1371/journal.pdig.0001534
 71. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients. *JAMA*. 2023;330(23):2275. doi:10.1001/jama.2023.22295
 72. Dratsch T, Chen X, Rezazade MM, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4):e222176. doi:10.1148/radiol.222176
 73. Wang D, Ding J, Sun A, et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *J Am Med Inform Assoc*. 2023;30(10):1684-1692. doi:10.1093/jamia/ocad118
 74. Sakamoto T, Harada Y, Shimizu T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: quasi-experimental study. *JMIR Form Res*. 2024;8:e58666. doi:10.2196/58666
 75. Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol*. 2025;10(10):896-903. doi:10.1016/s2468-1253(25)00133-5
 76. Budzyń K, Romańczyk M, Mori Y. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: authors' reply. *Lancet Gastroenterol Hepatol*. 2025;10(12):1062. doi:10.1016/s2468-1253(25)00324-3
 77. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969
 78. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164. doi:10.2196/53164
 79. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045. doi:10.1038/s41598-023-41032-5
 80. Science. Trump officials downplay fake citations in high-profile report on children's health. Accessed October 6, 2026. <https://www.science.org/content/article/trump-officials-downplay-fake-citations-high-profile-report-children-s-health>
 81. Reuters. Anthropic expert accused of using AI-fabricated source in copyright case. May 13, 2025. Accessed October 6, 2026. <https://www.reuters.com/legal/litigation/anthropic-expert-accused-using-ai-fabricated-source-copyright-case-2025-05-13/>
 82. Washington State Bar Association. Federal court excludes expert for AI-hallucinated citations. NWSidebar. August 31, 2026. Accessed October 6, 2026. <https://nwsidebar.wsba.org/2026/08/31/federal-court-excludes-expert-for-ai-hallucinated-citations/>
 83. Retraction Watch. As Springer Nature journal clears AI papers, one university's retractions rise drastically. February 10, 2025. Accessed October 6, 2026. <https://retractionwatch.com/2025/02/10/as-springer-nature-journal-clears-ai-papers-one-universitys-retractions-rise-drastically/>
 84. El ARA, Al MOA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. *npj Digit Med*. 2025;8(1):548. doi:10.1038/s41746-025-01722-y
 85. Mital S, Nguyen HV. Cost-effectiveness of using artificial intelligence vs polygenic risk score to guide breast cancer screening. *BMC Cancer*. 2022;22(1):501. doi:10.1186/s12885-022-09613-1
 86. Yala A, Mikhael PG, Lehman C, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat Med*. 2022;28(1):136-143. doi:10.1038/s41591-021-01599-w
 87. Vargas-Palacios A, Sharma N, Sagoo GS. Cost-effectiveness requirements for implementing artificial intelligence technology in the Women's UK Breast Cancer Screening service. *Nat Commun*. 2023;14(1):6110. doi:10.1038/s41467-023-41754-0
 88. Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the economic value of clinical artificial intelligence: challenges and opportunities. *Value Health*. 2022;25(3):331-339. doi:10.1016/j.jval.2021.08.015
 89. Huang X, Yang B, Zheng W, et al. Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural China. *BMC Health Serv Res*. 2022;22(1):260. doi:10.1186/s12913-022-07655-6
 90. Kalavadekar PN, Khandarkar KR, Dhotre DR, Chincholkar A, Swarnkar SK, Joshi HJ. Reinforcement learning to optimize tuberculosis screening strategies in resource-limited settings. *Indian J Tuberc*. 2025;72:S159-S167. doi:10.1016/j.ijtb.2025.11.019

91. Jahanshahi M, Gregg K, Davis G, et al. The use of external controls in FDA regulatory decision making. *Ther Innov Regul Sci*. 2021;55(5):1019-1035. doi:10.1007/s43441-021-00302-y
92. Yin X, Mishra-Kalyan PS, Sridhara R, Stewart MD, Stuart EA, Davi RC. Exploring the potential of external control arms created from patient-level data: a case study in non-small cell lung cancer. *J Biopharm Stat*. 2022;32(1):204-218. doi:10.1080/10543406.2021.2011901
93. European Medicines Agency. Qualification Opinion for Prognostic Covariate Adjustment (PROCOVA). Committee for Medicinal Products for Human Use. Accessed October 6, 2026. https://www.ema.europa.eu/en/documents/regulatory-procedural-guide-line/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf
94. Catalá-López F, Ridao M, Tejedor-Romero L, et al. Transparency, openness, and reproducible research practices are frequently underused in health economic evaluations. *J Clin Epidemiol*. 2024;165:111208. doi:10.1016/j.jclinepi.2023.10.024
95. Monks T, Harper A. Computer model and code sharing practices in healthcare discrete-event simulation: a systematic scoping review. *J Simul*. 2023;19(1):108-123. doi:10.1080/17477778.2023.2260772
96. Henderson RH, Sampson C, Pouwels XG, et al. Mapping the landscape of open-source health economic models: a systematic database review and analysis: an ISPOR Special Interest Group Report. *Value Health*. 2025;28(6):813-820. doi:10.1016/j.jval.2025.01.019
97. Zavalis EA, Ioannidis JPA. A meta-epidemiological assessment of transparency indicators of infectious disease models. *PLoS ONE*. 2022;17(10):e0275380. doi:10.1371/journal.pone.0275380
98. Naudet F, Sakarovitch C, Janiaud P, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *The BMJ* and *PLOS Medicine*. *BMJ*. 2018;k400. doi:10.1136/bmj.k400
99. Bermejo I, Tappenden P, Youn J. Replicating health economic models: firm foundations or a house of cards? *Pharmacoeconomics*. 2017;35(11):1113-1121. doi:10.1007/s40273-017-0553-x
100. Schwander B, Nuijten M, Evers S, Hilgsmann M. Replication of published health economic obesity models: assessment of facilitators, hurdles and reproduction success. *Pharmacoeconomics*. 2021;39(4):433-446. doi:10.1007/s40273-021-01008-7
101. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: artificial intelligence use by authors. Accessed October 6, 2026. <https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html>
102. Committee on Publication Ethics. Authorship and AI tools. COPE position statement. Accessed October 6, 2026. <https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools>
103. NEJM AI. Editorial policies. Accessed October 6, 2026. <https://ai.nejm.org/about/editorial-policies>
104. NEJM AI. About NEJM AI. Accessed October 6, 2026. <https://ai.nejm.org/about>
105. JAMA. Instructions for authors. Accessed October 6, 2026. <https://jamanetwork.com/journals/jama/pages/instructions-for-authors>
106. Health Affairs. New article requirements. Accessed October 6, 2026. <https://www.healthaffairs.org/help-for-authors/journal/new-article-requirements>
107. Elsevier. Generative AI policies for journals. Accessed October 6, 2026. <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals>
108. Springer Nature. Artificial intelligence: editorial policies. Accessed October 6, 2026. <https://www.nature.com/nature-portfolio/editorial-policies/ai>
109. International Society for Pharmacoeconomics and Outcomes Research. ISPOR artificial intelligence policy. Accessed October 6, 2026. <https://www.ispor.org/ai-policy>
110. *Value in Health. Guide for Authors*. Accessed October 6, 2026. <https://www.sciencedirect.com/journal/value-in-health/publish/guide-for-authors>
111. Writing and reviewing for us in AI times. Editorial. *Lancet Infect Dis*. April 2024. Accessed October 6, 2026. [https://www.thelancet.com/journals/laninf/article/PIIS1473-3099\(24\)00160-9/fulltext](https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(24)00160-9/fulltext)
112. Kwon D. Is it OK for AI to write science papers? Nature survey shows researchers are split. *Nature*. 2025;641(8063):574-578. doi:10.1038/d41586-025-01463-8
113. Naddaf M. More than half of researchers now use AI for peer review, often against guidance. *Nature*. 2025;649(8096):273-274. doi:10.1038/d41586-025-04066-5
114. American Medical Association. Augmented intelligence research: physician sentiments around the use of AI in health care. Chicago, IL: AMA; 2026. Accessed October 6, 2026. <https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf>
115. American Medical Association. AI usage among doctors doubles, confidence in technology grows. News release. 2026. Accessed October 6, 2026. <https://www.ama-assn.org/press-center/ama-press-releases/ama-ai-usage-among-doctors-doubles-confidence-technology-grows>
116. McKinsey & Company. Generative AI in healthcare: Adoption matures as agentic AI emerges. 2026. Accessed October 6, 2026. <https://www.mckinsey.com/industries/healthcare/our-insights/generative-ai-in-healthcare-current-trends-and-future-outlook>
117. HIMSS and Medscape. AI Adoption by Health Systems Report 2024. Accessed October 6, 2026. <https://www.himss.org/news-center/himss-and-medscape-unveil-groundbreaking-report-ai-adoption-health-systems/>
118. Polygon Health Analytics. AI in HEOR, RWD and medical affairs: what 133 professionals told us. April 2026. Accessed October 6, 2026. <https://polygonhealthanalytics.com/ai-in-heor-rwd-medical-affairs-what-133-professionals-told-us-and-what-it-means-for-the-industry/>
119. Center for the Evaluation of Value and Risk in Health, Tufts Medical Center. HEOR trends: results of a survey of HEOR leaders. Poster presented at: ISPOR 2026; April 7, 2026. Accessed October 6, 2026. https://cevr.tuftsmedicalcenter.org/assets/images/2026_04_07_ISPOR_Poster_HEOR_Trends.pdf
120. Medical Affairs Professional Society. Workflows, not chatbots: findings from the 2026 AI in Medical Affairs Industry Benchmark Survey. Accessed October 6, 2026. <https://medicalaffairs.org/solution-provider-perspective/workflows-not-chatbots/>
121. ZS Associates. Medical affairs outlook report 2025. Accessed October 6, 2026. <https://www.zs.com/content/dam/global/master/pdfs/zs-medical-affairs-outlook-report-2025.pdf>
122. Cai C, Zhu G, Zhou S, et al. A competency framework for medical AI education: mixed methods study. *JMIR Med Educ*. 2026;12:e91116. doi:10.2196/91116
123. Acosta JA. Perspective: advancing public health education by embedding AI literacy. *Front Digit Health*. 2025;7:1584883. doi:10.3389/fdgrh.2025.1584883

124. Semi MMA, Das S, Utsho MR, et al. Artificial intelligence in public health education: a scoping review of workforce competency development. *Health Sci Rep.* 2026;9(3):e72066. doi:10.1002/hsr2.72066
125. Wolthusen RPF, El-tohami M, Hendler RA, Riordan PA, Stoklosa JB. A framework for artificial intelligence in medical education: could I, would I, should I? *J Grad Med Educ.* 2025;17(4):430-433. doi:10.4300/jgme-d-25-00009.1
126. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach.* 2024;46(4):446-470. doi:10.1080/0142159x.2024.2314198
127. Pizzi LT, Onukwugha E, Corey R, Albarmawi H, Murray J. Competencies for professionals in health economics and outcomes research: the ISPOR Health Economics and Outcomes Research Competencies Framework. *Value Health.* 2020;23(9):1120-1127. doi:10.1016/j.jval.2020.04.1834
128. Rock Health. 2025 year-end digital health funding overview: a tale of two markets. Accessed October 6, 2026. <https://rockhealth.com/insights/2025-year-end-digital-health-funding-overview-a-tale-of-two-markets/>
129. Rock Health. 2024 year-end market overview: Davids and Goliaths. Accessed October 6, 2026. <https://rockhealth.com/insights/2024-year-end-market-overview-davids-and-goliaths/>
130. White & Case and Mergermarket. Life sciences and AI report. January 2026. Accessed October 6, 2026. https://www.whitecase.com/sites/default/files/2026-01/W%26C_Life%20Sciences%20%26%20AI%20Report_January%202026.pdf
131. McKinsey & Company. AI budgets grow in life sciences. Week in Charts. 2025. Accessed October 6, 2026. <https://www.mckinsey.com/featured-insights/week-in-charts/ai-budgets-grow-in-life-sciences>
132. Grand View Research. Artificial intelligence in healthcare market size, share and trends analysis report. Accessed October 6, 2026. <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market>
133. The Royal Swedish Academy of Sciences. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2024: Daron Acemoglu, Simon Johnson and James A. Robinson, for studies of how institutions are formed and affect prosperity. Stockholm: NobelPrize.org; 2024. Accessed October 6, 2026. <https://www.nobelprize.org/prizes/economic-sciences/2024/summary/>
134. Acemoglu D. The simple macroeconomics of AI. *Econ Policy.* 2024;40(121):13-58. doi:10.1093/epolic/eiae042
135. Acemoglu D. The simple macroeconomics of AI. 2024. doi:10.3386/w32487
136. Acemoglu D, Restrepo P. Automation and new tasks: how technology displaces and reinstates labor. *J Econ Perspect.* 2019;33(2):3-30. doi:10.1257/jep.33.2.3
137. Acemoglu D, Restrepo P. The race between man and machine: implications of technology for growth, factor shares, and employment. *Am Econ Rev.* 2018;108(6):1488-1542. doi:10.1257/aer.20160696
138. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science.* 2023;381(6654):187-192. doi:10.1126/science.adh2586
139. Brynjolfsson E, Li D, Raymond L. Generative AI at work. *Q J Econ.* 2025;140(2):889-942. doi:10.1093/qje/qjae044
140. Brynjolfsson E, Li D, Raymond L. Generative AI at work. 2023. doi:10.3386/w31161
141. Peng S, Kalliamvakou E, Cihon P, Demirel M. The impact of AI on developer productivity: evidence from GitHub Copilot. *arXiv.* 2023. doi:10.48550/arxiv.2302.06590
142. Dell'Acqua F, McFowland E, Mollick E, et al. Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organ Sci.* 2026;37(2):403-423. doi:10.1287/orsc.2025.21838
143. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378
144. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ.* 2025;388:e082505. doi:10.1136/bmj-2024-082505
145. Liu X, Cruz RS, Moher D, Calvert MJ, The SACWG, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
146. Cruz RS, Liu X, Chan A, The SACWG, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med.* 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7
147. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med.* 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9
148. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open.* 2021;11(6):e047709. doi:10.1136/bmjopen-2020-047709
149. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med.* 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5
150. Cacciamani GE, Eppler MB, Ganjavi C, et al. Development of the ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for Accountable Reporting and Use (CANGARU) Guidelines. *arXiv.* 2023. doi:10.48550/arxiv.2307.08974
151. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) Statement: updated reporting guidance for health economic evaluations. *Value Health.* 2022;25(1):3-9. doi:10.1016/j.jval.2021.11.1351
152. Padula WV, Kreif N, Vanness DJ, et al. Machine learning methods in health economics and outcomes research: the PALISADE Checklist: a good practices report of an ISPOR Task Force. *Value Health.* 2022;25(7):1063-1080. doi:10.1016/j.jval.2022.03.022
153. Fleurence RL, Dawoud D, Bian J, et al. ELEVATE-GenAI: reporting guidelines for the use of large language models in health economics and outcomes research: an ISPOR Working Group Report. *Value Health.* 2025;28(11):1611-1625. doi:10.1016/j.jval.2025.06.018
154. Fleurence RL, Wang X, Bian J, et al. A taxonomy of generative artificial intelligence in health economics and outcomes research: an ISPOR Working Group Report. *Value Health.* 2025;28(11):1601-1610. doi:10.1016/j.jval.2025.04.2167
155. Fleurence RL, Bian J, Wang X, et al. Generative artificial intelligence for health technology assessment: opportunities, challenges, and policy considerations: an ISPOR Working Group Report. *Value Health.* 2025;28(2):175-183. doi:10.1016/j.jval.2024.10.3846

156. Ma X, Bellomo L, Magee K, et al. Characterization of a real-world response variable and comparison with RECIST-based response rates from clinical trials in advanced NSCLC. *Adv Ther.* 2021;38(4):1843-1859. doi:10.1007/s12325-021-01659-0
157. Estevez M, Singh N, Dyson L, et al. Ensuring reliability of curated electronic health record-derived data: the validation of accuracy for large language model-/machine learning-extracted information and data (VALID) framework. *JCO Clin Cancer Inform.* 2026;10(1):e2500215. doi:10.1200/cci-25-00215
158. Kenney RC, Chen X, Shintani K, et al. Validation of non-small cell lung cancer clinical insights using a generalized oncology natural language processing model. *JCO Clin Cancer Inform.* 2024:e2300099. doi:10.1200/cci.23.00099
159. Abedian S, Sholle ET, Adekkanattu PM, et al. Automated extraction of tumor staging and diagnosis information from surgical pathology reports. *JCO Clin Cancer Inform.* 2021:1054-1061. doi:10.1200/cci.21.00065
160. Wang D, Florian H, Lynch S, et al. Using AI-generated digital twins to boost clinical trial efficiency in Alzheimer's disease. *A&D Transl Res Clin Interv.* 2025;11(4):e70181. doi:10.1002/trc2.70181
161. Akiya I, Ishihara T, Yamamoto K. Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study. *JMIR Med Inform.* 2024;12:e55118. doi:10.2196/55118
162. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digit Med.* 2025;8(1):60. doi:10.1038/s41746-024-01359-3
163. Riskin DJ, Monda KL, Gagne JJ, et al. Implementing accuracy, completeness, and traceability for data reliability. *JAMA Netw Open.* 2025;8(3):e250128. doi:10.1001/jamanetworkopen.2025.0128
164. Varma G, Yenukoti RK, Kumar MP, et al. A deep learning-enabled workflow to estimate real-world progression-free survival in patients with metastatic breast cancer: study using deidentified electronic health records. *JMIR Cancer.* 2025;11:e64697. doi:10.2196/64697