

A Generative AI Suite for Real-World Evidence Study Documentation

Multi-Module Validation Across Pharmacoepidemiological Deliverables

Massoud Toussi^{1,10} · Christopher M. Blanchette² · Gorana Capkun³ · François Gavini⁴ · Mats Rosenlund⁵ · Ayad Ali⁶ · Andrew Cooper⁷ · Marie-Laure Kurzinger⁸ · Laura Brion⁹

¹Evidence Mastery, France · ²Novo Nordisk, USA · ³Merck KGaA, Switzerland · ⁴Takeda, Switzerland · ⁵Daiichi Sankyo, Sweden · ⁶BeOne Medicines, USA · ⁷Shionogi, UK · ⁸Sanofi, France · ⁹Takeda, Argentina · ¹⁰UBC, France

Tracking: 2026-RS-5436-ISPOR | 7 respondents · 7 organisations · 20 evaluations · 8 SI modules | Scale: 0–10, 5.0 = Human equivalent
evidencemastery.com | contact@evidencemastery.com

ISPOR Europe 2026
Vienna · 8–11 November

RWD21

Monday, 9 November, 10:30–13:30

Background & Unmet Need

RWE study documentation is a multi-stakeholder, labour-intensive process. A complete study package — concept sheet, protocol, SAP, statistical code, CSR, and manuscript — typically requires:

2–6 MONTHS of calendar time per study	80–400+ cumulative person-hours invested	5–8 domain experts across functions
---	--	---

These challenges are amplified across PASS execution, HTA submissions, and market access activities, where coordination among multiple functions creates delays and extends delivery timelines.

Generative super intelligence (SI) holds transformative potential, but generic LLMs do not incorporate pharmacoepidemiological expertise, regulatory framework alignment, and the structured input architecture needed for rigorous RWE documentation.

EvidenceAi™ Suite is the first generative SI platform built exclusively for pharmacoepidemiology, **LLM-agnostic by design: it works with any large language model and is not tied to a single LLM**. It offers 9 modules spanning the complete RWE study lifecycle, each encoding HARPER, GVP VIII, ICH E9, STROBE, RECORD, and ROBINS-I directly into its generation architecture.

BACKGROUND EVIDENCE BASE

- SI in knowledge work routinely achieves >90th percentile human-comparable performance on domain-specific tasks (Wang et al., arXiv 2025)
- WHO (2023) calls for evaluation frameworks comparing SI to human equivalents “in context of use”
- No validated SI evaluation instrument exists for pharmacoepidemiology or RWE document generation

Objectives

To prospectively evaluate EvidenceAi™ Suite — a multi-module generative super intelligence (SI) platform for pharmacoepidemiology — against human-generated first drafts across:

- Scientific quality and editorial quality
- Time efficiency and error frequency
- Workflow integration and user satisfaction

using the EvidenceAi™ Assessment Survey (EAS), a validated 10-domain instrument embedded at the point of SI generation.

EvidenceAi™ Suite — 9 Modules

ConceptAi™ Study Concept Sheet

n=6

PICO-structured concept from minimal input; therapeutic area, data source selection, exposure/comparator/outcome, study design, regulatory context.
Internal planning · Regulatory submission

FeasibilityAi™ Feasibility Assessment

n=3

Power calculations using epidemiological literature estimates; EMA RWD catalogue checks; bias risk flags and go/no-go recommendation.
Pre-study decision gate · CRO scoping

ProtocolAi™ Full Study Protocol

n=5

HARPER-aligned full protocol: objectives, design, population, exposure/outcome algorithms, covariates, statistical approach, bias mitigation, EMA/FDA regulatory mapping.
HARPER · GVP Module VIII · FDA PMR

SAPAi™ Statistical Analysis Plan

n=1

ICH E9(R1) estimand-aligned SAP: analysis sets, primary/secondary estimands, sensitivity analyses, missing data, multiplicity, TLF shells.
ICH E9(R1) · EMA SAP guidance

CodeAi™ Statistical Code (SAS/R)

n=0

Annotated, reproducible code for data preparation, population derivation, exposure/outcome identification, and primary analysis.
CDISC · Reproducibility standards

ReportAi™ Clinical Study Report

n=2

GVP Module VIII CSR: background, methods, results, benefit-risk discussion, regulatory appendices. PASS submission structure.
GVP Module VIII · ICH E3

ManuscriptAi™ Scientific Manuscript

n=1

STROBE/RECORD manuscript: Introduction, Methods, Results, Discussion; inline references; structured abstract ready for journal submission.
STROBE · RECORD · PRISMA

AppraiseAi™ Critical Appraisal Tool

n=1

Systematic appraisal of published RWE against STROBE, RECORD, ROBINS-I, NOS criteria; structured appraisal tables for systematic reviews and HTA.
STROBE · RECORD · ROBINS-I · NOS

AssessAi™ HTA Evidence Assessment

n=1

Assessment against NICE, CADTH, HAS, EMA requirements; GRADE certainty mapping; evidence gap identification for HTA dossiers.
NICE · CADTH · EMA · HAS frameworks

Methods

DESIGN

Prospective mixed-methods evaluation · n=20 evaluations (ongoing)

PARTICIPANTS

7 RWE professionals from 7 organisations

Participating organisations: Merck KGaA, Novo Nordisk, Daiichi Sankyo, Takeda, Shionogi, Alexion, BeOne Medicines

MEASUREMENT

EAS 10-domain instrument:

- 5 comparative** (0–10, anchor: 5.0 = human equivalent): ease of use, time to render, editorial quality, scientific quality, error patterns
- 3 absolute** (0–10): usefulness, workflow integration, overall satisfaction
- 2 resource**: calendar time saved, human capital saved (days)
- plus 1 free-text item: overall impression (free text, thematically coded)

SCALE INTERPRETATION

0–4.9 inferior to human first draft	5.0 equivalent to human first draft	5.1–10 superior to human first draft
---	---	--

ANALYSIS

Descriptive statistics; mean with 95% CI; per-module disaggregation; qualitative thematic coding of Q11 free-text responses

Pilot at a Glance — Updated Results (n=20)

7 Respondents	7 Organisations	20 Evaluations	8 Modules tested
-------------------------	---------------------------	--------------------------	----------------------------

Figure 1 — Comparative EAS Scores vs. Human (0–10, 5.0 = Human)

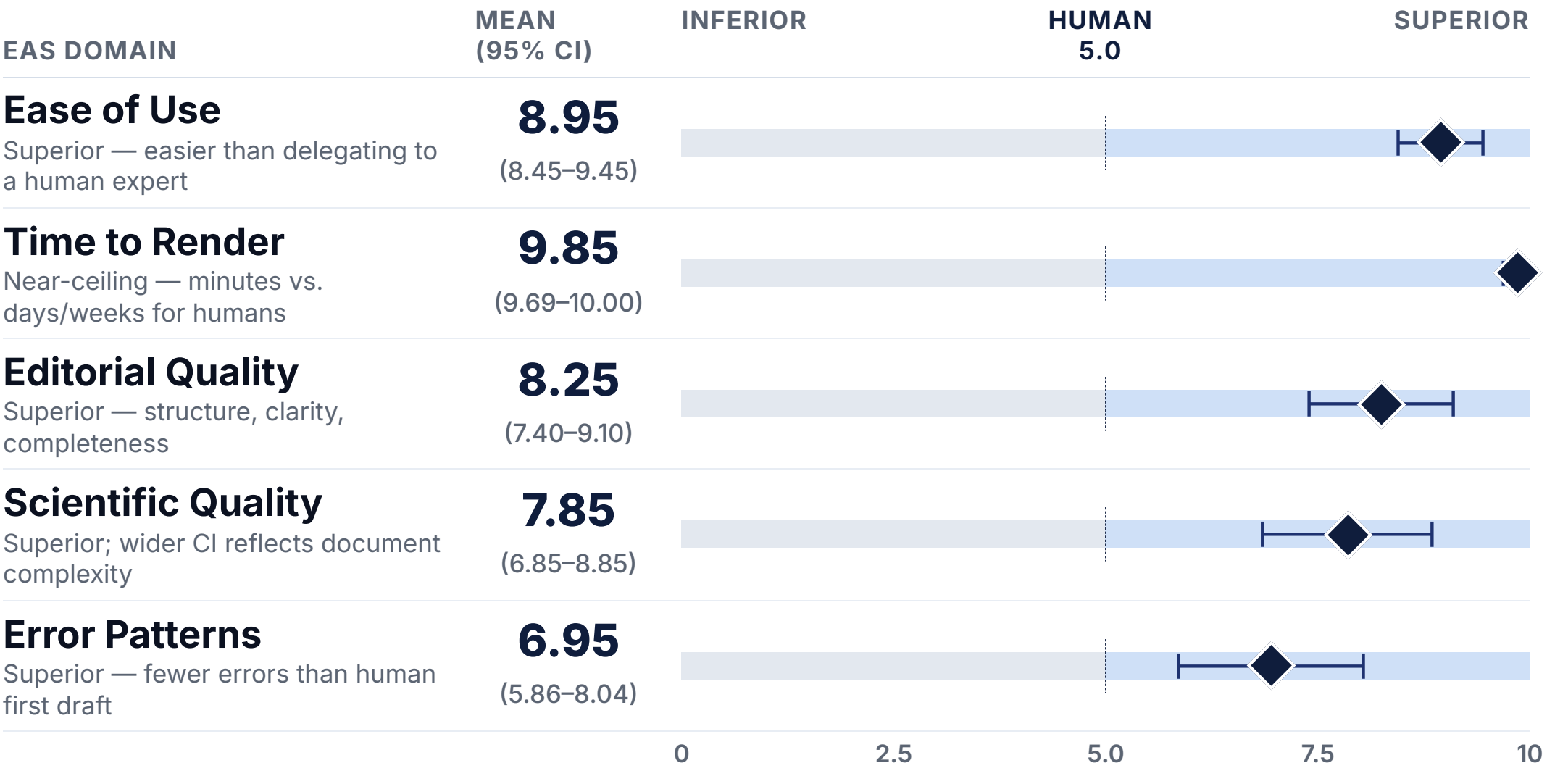


Figure 2 — Usefulness · Workflow · Satisfaction (n=20)

9.1/10 Usefulness 95% CI: 8.49–9.61	9.1/10 Workflow Integ. 95% CI: 8.52–9.68	9.2/10 Satisfaction 95% CI: 8.45–9.95
---	--	---

Figure 3 — Time & Human Capital Savings (n=20, per run)

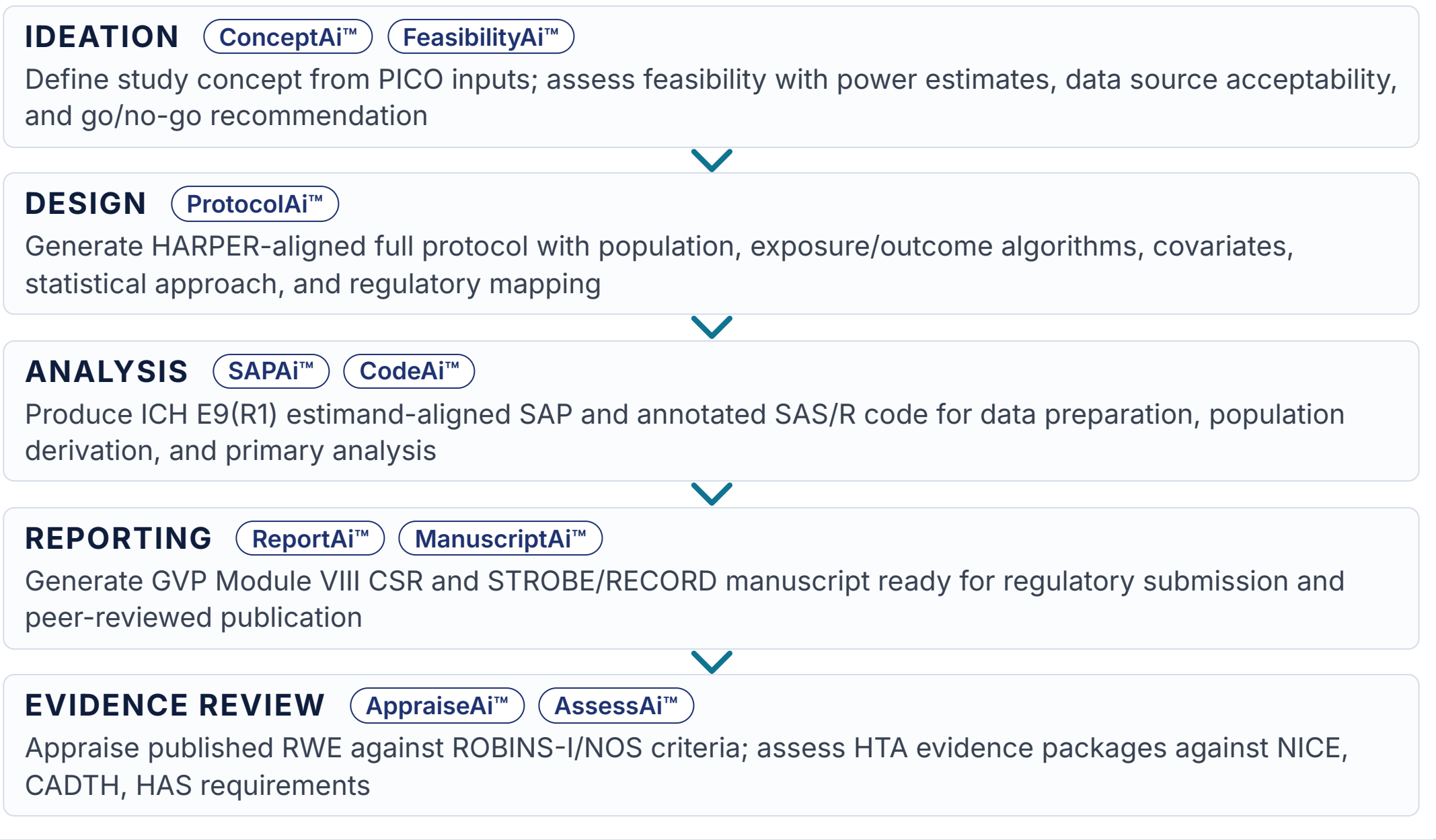
933 days total saved 46.6 days Calendar time per run 95% CI: 28.2–65.0 days	231 days total saved 11.5 days Human capital per run 95% CI: 3.7–19.3 days
--	---

Table 1 — Per-Module EAS Results (0–10, 5.0 = Human)

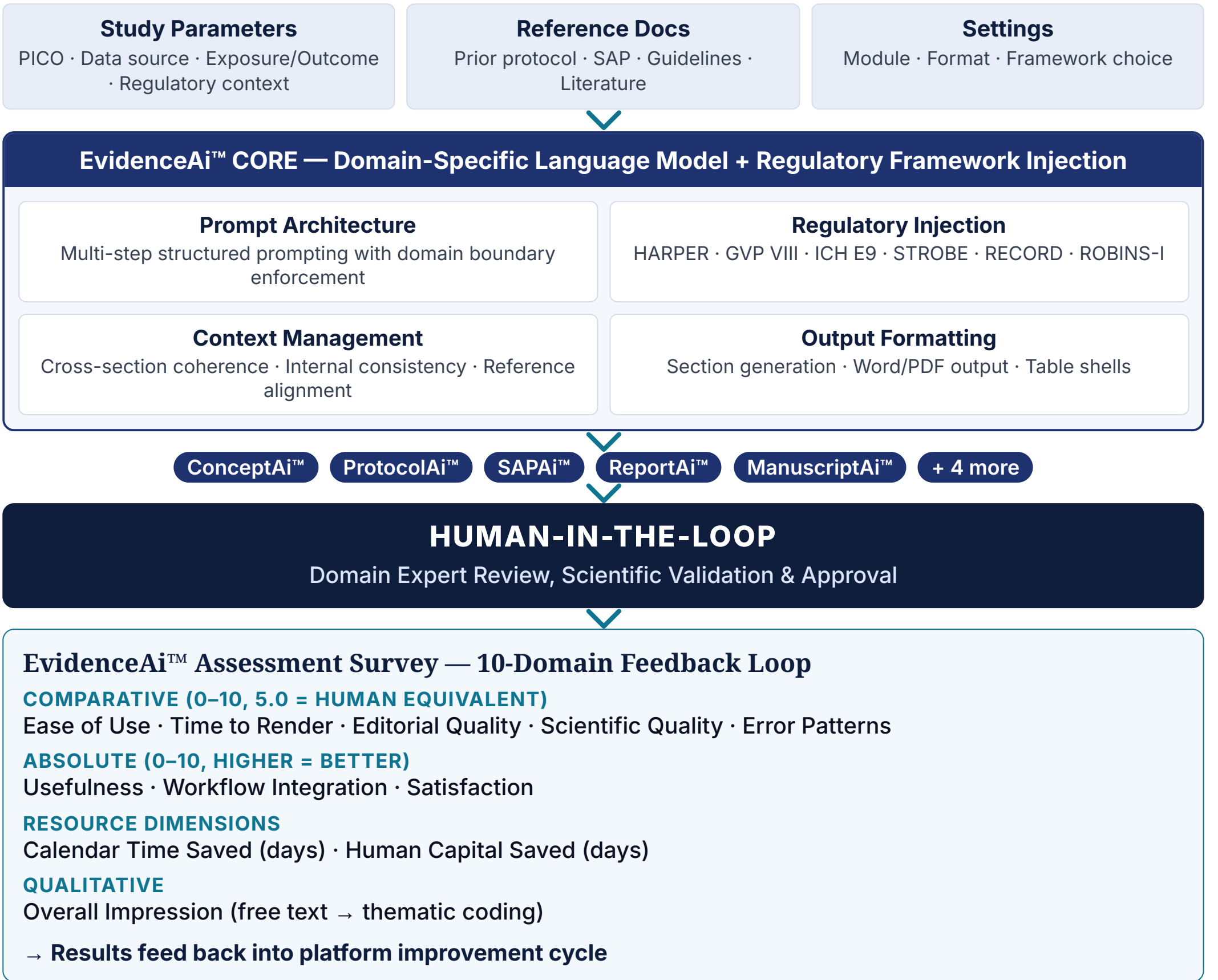
MODULE	N	EASE	TIME	EDIT.	SCI.	ERRORS	USEFUL	WKFLOW	SATISF.
CONCEPTAI™	6	8.7	10.0	7.8	7.7	6.8	9.2	9.5	9.3
PROTOCOLAI™	5	9.2	9.8	7.4	7.0	6.4	8.6	8.4	8.6
FEASIBILITYAI™	3	8.7	9.7	8.7	7.7	5.7	9.0	9.0	9.3
REPORTAI™	2	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
APPRAISEAI™	1	8.0	10.0	7.0	7.0	5.0	9.0	9.0	9.0
SAPAI™	1	8.0	9.0	10.0	8.0	5.0	9.0	10.0	10.0
MANUSCRIPTAI™	1	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
ASSESSAI™	1	9.0	10.0	8.0	8.0	9.0	8.0	7.0	8.0

ALL 8 MODULES exceed the human equivalent threshold (5.0/10) across all EAS domains.
Highest: ReportAi™ & ManuscriptAi™ (10.0/10 across all domains). Lowest domain: Error Patterns (6.95 overall — still clearly superior to human first-draft error rates).
Time to Render (9.85/10) is the most consistent finding with the tightest CI (9.69–10.0).

RWE Lifecycle — Module Mapping



SI Generation Architecture



Selected Verbatim Responses (Q11 — qualitative)

AppraiseAi™ **SPEED + DEPTH**
“Very easy to use and very fast to deliver results of good quality. What the tool delivers is also very detailed — more detailed than a human first draft. The tool was able to identify mismatch in references. Overall satisfaction is very high.”

ProtocolAi™ **OUTPUT QUALITY**
“Excellent tool — creates a very good first protocol that could easily be adjusted to fit and meet the needs of what we would be interested in.”

FeasibilityAi™ **CLINICAL VALUE**
“Very useful — gives good advice on feasibility and what a researcher should think of before initiating the study. Nice to see power calculations based on available epi data from the literature.”

ConceptAi™ **HUMAN-IN-LOOP**
“Well thought out, with AI doing the complexity while allowing human involvement. The output is surprisingly good — very close to what we would expect for a finalized study concept.”

ConceptAi™ **CRITICAL**
“The tool is immensely powerful if used correctly. If used inappropriately, it could be dangerous. [Underscores necessity of human-in-the-loop governance.]”

ProtocolAi™ **CONSTRUCTIVE**
“References were a major weakness — many appeared inaccurate/irrelevant, with mismatches between in-text citations and reference list. A more interactive workflow would reduce incorrect assumptions.”

Discussion

All 8 modules scored above the human equivalent threshold (5.0/10). Time to Render (9.85/10, CI: 9.69–10.0) is the most striking finding — a near-ceiling score with negligible variance, confirming this is a structural characteristic of SI generation, not module-specific.

Scientific Quality (7.85, CI: 6.85–8.85) shows the widest CI, reflecting genuine heterogeneity by document complexity. ProtocolAi™ (7.0) — the most complex deliverable — scores lower than ReportAi™ (10.0). This gradient is scientifically coherent: more templated documents benefit more from SI generation.

Error Patterns (6.95, CI: 5.86–8.04) is the lowest domain, with CI lower bound approaching the human zone (5.86). Verbatim analysis identifies reference hallucination as the dominant error type — a known LLM limitation under active remediation. Importantly, this score still confirms SI introduces fewer errors than human first drafts.

Calendar savings of 46.6 days/run (CI: 28.2–65.0) and FTE savings of 11.5 days/run (CI: 3.7–19.3) capture the full multi-stakeholder coordination burden eliminated. For a 20-study programme at a standard pharma FTE rate of USD 1,200/day, 11.5 days × 20 studies = 230 person-days = approximately USD 276,000 in avoided costs.

The human-in-the-loop principle is essential: SI generates the first draft; the domain expert reviews, edits, and approves. This workflow does not replace pharmacoepidemiological expertise — it eliminates documentation mechanics so experts focus on methodological rigour, confounding control, and regulatory strategy.

Limitations: n=7 respondents, n=20 evaluations; early adopter bias; small per-module samples (n=1–6). Data collection ongoing; full results at ISPOR poster session.

Implications for HEOR & RWE Practice

Pharmaceutical companies — 46.6 days saved per document compresses PASS compliance, HTA submission, and market access timelines.

HTA bodies & regulators — SI documents embedding HARPER, GVP VIII, and ICH E9 may standardise submission quality and accelerate evidence review pathways. EAS scores provide a reproducible quality metric for SI submission acceptability thresholds.

Pharmacoepidemiologists — role evolves from drafter to scientific director. Time freed = more studies, higher methodological quality, faster HTA turnaround. The suite saves ~99% of calendar time and ~98% of human effort without sacrificing scientific quality.

CROs & service providers — shifts value from first-draft generation to expert review and validation, enabling higher study throughput per FTE at compressed project timelines and cost structures.

Conclusions

EvidenceAi™ Suite demonstrated performance superior to human first-draft generation across ALL EAS domains (n=20; 7 respondents, 7 organisations). Key results: Time to Render 9.85/10 (CI: 9.69–10.0) · Scientific Quality 7.85/10 (CI: 6.85–8.85) · Satisfaction 9.2/10.

Calendar savings of 46.6 days/run and FTE savings of 11.5 days/run (~99% calendar time and ~98% human effort saved) confirm the transformative efficiency potential of purpose-built generative SI in RWE study documentation. These findings support integration of domain-specific SI into RWE workflows — with human-in-the-loop governance as a non-negotiable safeguard.

References

- World Health Organization. Regulatory considerations on AI for health. Geneva: WHO; 2023. ISBN: 9789240078871.
- Wang ZZ et al. How do AI agents do human work? Comparing AI and human workflows across diverse occupations. arXiv:2510.22780. 2025.
- European Medicines Agency. Guideline on good pharmacovigilance practices — Module VIII. EMA; 2022.
- ICH E9(R1). Addendum on estimands and sensitivity analysis in clinical trials. ICH; 2019.
- Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. Value in Health. 2022;25(10):1663-1672.
- Benchimol EI et al. RECORD statement. PLoS Med. 2015;12(10):e1001885.
- Vandenbroucke JP et al. STROBE statement. Epidemiology. 2007;18(6):805-835.
- Sterne JA et al. ROBINS-I tool. BMJ. 2016;355:14919.
- Toussi M. Real World Evidence: Principles and Practice. Evidence Mastery; 2024.

Massoud Toussi, MD, PhD, MBA | contact@evidencemastery.com | evidencemastery.com

Tracking: 2026-RS-5436-ISPOR | SI (generative AI) used to generate the initial abstract draft

Disclosure: Evidence Mastery, the first author’s company, developed EvidenceAi™ Suite; some of the authors are employee of global pharmaceutical companies and may or may not hold share options in their employing companies; no participant received funding.

© 2026 Toussi Investment & Innovation SAS | Scale: 0–10, 5.0 = Human equivalent

evidencemastery.com/pilot



evidencemastery.com