General-Purpose Versus Domain-Specific AI for Systematic Confounder Identification in Multiple Sclerosis

MSR212

A Comparative Methodological Study Using IQWiG Assessments as Ground Truth

ISPOR 2026
Philadelphia
May 17–20

Anton O. Wiehe[1] • Florian Woeste, MSc[1] • Pia Ana Cuk, MSc[1]

[1]Pharos Labs, Hamburg, Germany

Background

Systematic confounder identification is mandatory for HTA benefit assessments of non-randomised studies. The only published methodology — the **Pufulete 2022 framework** — combines systematic literature review, clinical expert interviews, and quantitative survey.

IQWiG applied this to RRMS (GA23-02):

- **62 person-days** total effort
- **33 person-days** for literature search alone
- **132** individual variables identified
- Consolidated into **28** confounders (26 measurable)
- 2 confounders found **only** through expert interviews

IQWiG itself called for efficiency improvements: *"The next step should examine which further options exist for reducing effort"* (GA23-02, p.40; translated from German)

Methods

Ground Truth

IQWiG GA23-02 (DOI: 10.60584/GA23-02): 132 individual confounder variables for DMF vs. GA target trial emulation in RRMS.

Systems

System	Type	Citations
Claude Opus 4.5	LLM ± web	Parametric/web
Gemini 3.1 Pro	LLM ± web	Parametric/web
Regulaido	**Domain RAG**	**35-paper SLR**

All systems received the **same exhaustive-extraction prompt**. Claude and Gemini tested ± web-search grounding (Vertex AI).

Evaluation

Two-pass 1-to-1 assignment grading (judge: Claude Opus 4.6):

- **Pass 1:** Each proposal → one of 26 consolidated confounders
- **Pass 2:** Each proposal → one of 132 individual variables
- **Citations:** DOI/PMID verification via Crossref & NCBI APIs

Results

System	Items	R-26	R-132	P-132	Verifiable cites	Cite fabr.
Regulaido	**140**	**88%**	**60%**	**93%**	**100%**	**0%**
Claude (grounded)	124	92%	45%	83%	73%	0%
Claude (no grounding)	131	73%	46%	**97%**	100%	40%
Gemini (grounded)	47	62%	29%	91%	100%	40%
Gemini (no grounding)	85	85%	40%	94%	100%	36%

1 Deepest Coverage of IQWiG Framework

Regulaido identified **83 of 132** individual IQWiG variables (**60%**) — significantly more than Claude (60, 45%) or Gemini (53, 40%). Per-paper agentic extraction from a curated corpus captures variables that LLMs generating from parametric knowledge miss.

2 Citation Fabrication is a Critical Risk

Without web grounding, both Claude and Gemini **fabricated 36–40% of cited DOIs**. Even **with** grounding, 27% of Claude's citations lacked verifiable identifiers (only URLs, no DOIs). Only Regulaido's corpus-bounded architecture guarantees 100% verifiable, 0% fabricated citations.

3 33 Person-Days Replicated in Minutes

IQWiG's systematic literature search consumed **33 of 62 person-days**. All AI systems completed this step in **under 5 minutes**. Two confounders (Patient wish, Tolerability) were identified only through clinical expert interviews — human expertise remains irreplaceable.

4 Volume Matters for Downstream Consolidation

Gemini (grounded) proposed only **47 items** — too few for meaningful human consolidation. Regulaido's **140 items** provide the comprehensive candidate set that the Pufulete 2022 framework requires: filtering is easier than searching for blind spots.

Efficiency Comparison

	IQWiG Manual	AI-Assisted (best)
Lit. search time	33 person-days	**< 5 min**
Total process	62 person-days	< 5 min + expert review
Key confounders (of 26)	**26** (100%)	**23** (88%)
Citation verifiability	100% (manual)	**100%** (architectural)

Conclusions

1. **Domain-specific RAG achieves deepest coverage** of IQWiG's confounder framework (60% of 132 variables, 93% precision) with 100% verifiable and 0% fabricated citations.
2. **Citation provenance is architectural, not behavioral.** LLMs fabricate 36–40% of DOIs without grounding; even with grounding, 27% of citations lack verifiable identifiers. Only corpus-bounded RAG guarantees full verifiability.
3. **AI accelerates the most expensive Pufulete 2022 step** (literature search: 33 person-days → minutes), while clinical expert interviews remain necessary for patient-centric confounders.
4. **Volume matters:** generating more candidates (140 vs. 47) is preferable for human-led consolidation — filtering is easier than searching for blind spots.
5. **The optimal regulatory AI architecture** combines domain RAG (provenance + depth) with human expert input (patient-centric confounders).

Limitations

- Single indication (RRMS); generalisability requires further study
- Single run per system; stochastic variation not assessed
- Regulaido bounded by 35-paper corpus; larger corpus expected to improve coverage
- Judge model (Claude Opus 4.6) not under test but could influence assignment

References

1. IQWiG. GA23-02: Systematische Confounderidentifikation (RRMS). 2025. DOI: 10.60584/GA23-02.
2. Pufulete M, et al. *J Clin Epidemiol.* 2022;148:115-123.
3. Fox RJ, et al. *N Engl J Med.* 2012;367(12):1087-1097.
4. Hernán MA, et al. *JAMA.* 2022;328(24):2446-2447.

Disclosures: AOW, FW, and PAC are employees of Pharos Labs, the developer of Regulaido. Study funded by Pharos Labs.

Presented at ISPOR 2026, Philadelphia, PA, USA