Evaluating an Automated AI-Driven Pipeline for Literature Surveillance and Synthesis: A Proof-of-Concept for Health Economics and Outcomes Research (HEOR) Communications

Kiessling J[1], O'Donovan P[1], Heron L[1], Yakob L[1]
[1]Adelphi Values PROVE, Bollington, Cheshire SK10 5JB, United Kingdom

Adelphi
ADELPHI VALUES
PROVE
Expertise in Access and Value Evidence Outcomes

MSR90

Introduction

Health Economics and Outcomes Research (HEOR) relies on timely access to high-quality and up-to-date evidence to inform policy and payer decision-making, and stakeholder communications.[1] However, recent bibliometric analyses show that the publication rate in healthcare research has markedly increased since the mid-2010s, creating a moving evidence base that can quickly outdate HEOR analyses that synthesise literature identified and collated through manual processes.[2-4] A responsibility to synthesise the most up-to-date evidence for accurate decision making drives the need for scalable methods that keep pace with this increased volume of new evidence.[3,4]

The emergence of artificial intelligence (AI) and machine learning (ML), and their potential to markedly reduce the time required for evidence identification, screening, and synthesis, can enable more continuously up-to-date research dissemination, with ML classification alone proving to reduce the screening burden by 60-80% in a previous application.[5] With rising volumes of primary research, these capabilities underscore the necessity of integrating AI into evidence-generation pipelines, while maintaining the methodological rigor and transparency required for decision-making contexts.[1,6] However, robust, scalable solutions for fully automated, end-to-end evidence surveillance and synthesis remain largely untested in the HEOR context.

- Primary Objective: To evaluate the feasibility and performance of a fully automated, AI-driven pipeline for identification, summarization, and dissemination of newly published literature relevant to HEOR.
- Secondary Objective: To propose a scalable framework for future AI-assisted evidence communications that can be adapted to the evolving needs of HEOR professionals.

Methods

AI Workflow
A modular, proof-of-concept pipeline was developed, comprising of the following elements:

1. Automated Retrieval:
Regular harvesting of newly indexed publications from major bibliographic databases (PubMed, Embase, etc.) via application programming interface (API) integration.

2. Custom ML Classifier Logic:
Due to the limited volume of evidence on HEOR outcomes in RSV vaccination, two supervised machine learning (ML) classifiers were trained on dual screened articles using the DistillerSR software. These were applied in combination through Boolean logic to identify articles relevant to the defined PICO inclusion criteria: RSV vaccination (P and I), and HEOR outcomes (O). The comparator component of the criteria was not restricted. Training involved feature engineering on titles, abstracts, and metadata.

3. Generative AI (GenAI) Agent: Topic and Evidence Identification
An AI agent utilizing multi-shot prompting and instructions defining relevance and impact of HEOR research, based on expert input, was created to:
- Identify 3-5 themes from structured data prioritizing the impact of the theme on the RSV landscape, as well as the quality and volume of evidence.
- Identify evidence for each theme, prioritizing relevance to the theme, and the potential impact of reported findings according to guidelines developed through expert input.

4. GenAI Synthesis Module:
A retrieval-augmented generation (RAG) agent was developed to generate:
- Structured summaries of the evidence identified as relevant by the ML classifier pipeline (tabular format),
- Narrative newsletter-style syntheses of the evidence deemed impactful and relevant by topic, adhering to strict guidelines which included instructions on referencing and evidence prioritization.

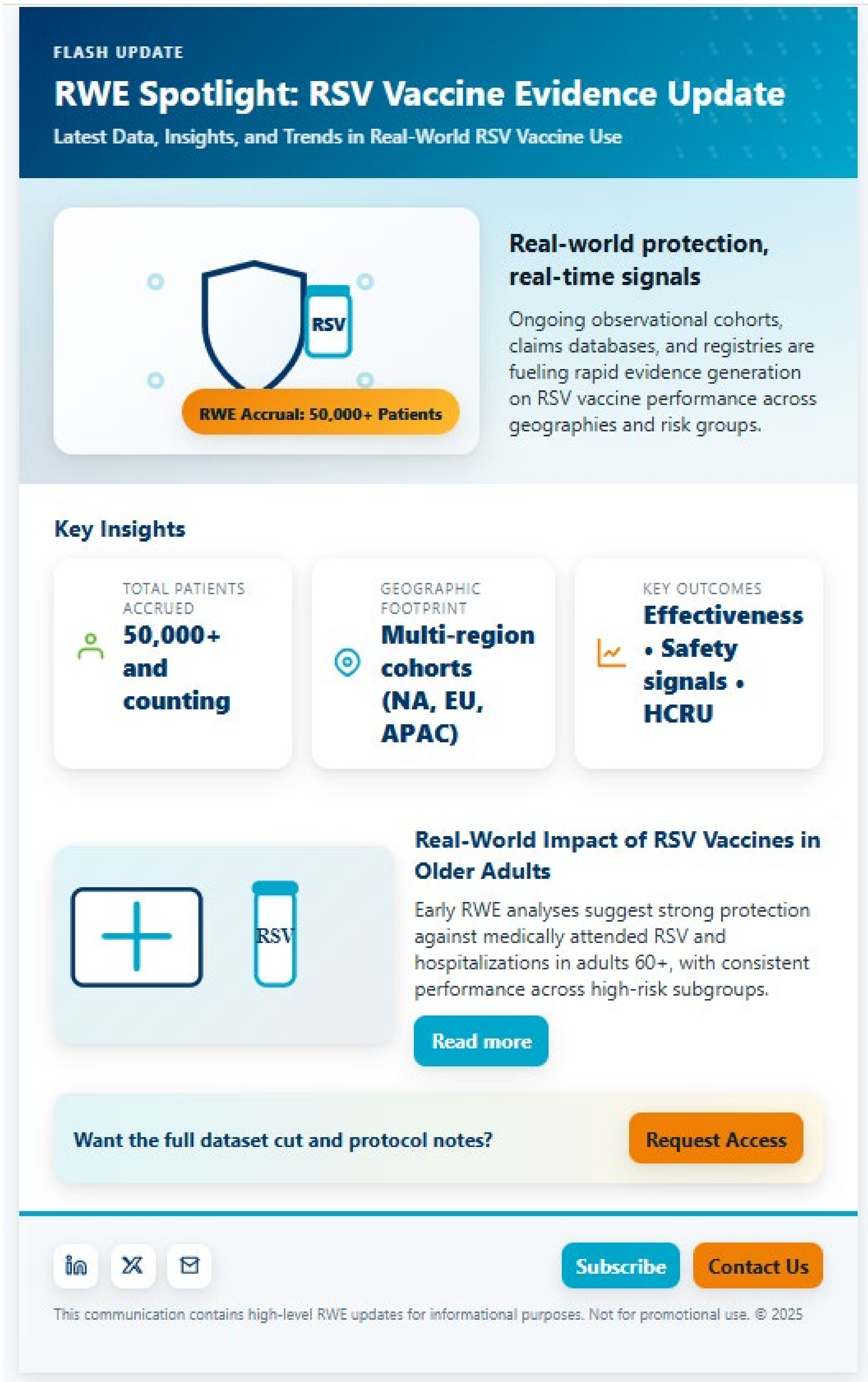
Evaluation Framework:
- Classifiers: Sensitivity (precision), specificity (recall), and the balanced accuracy score were compared to a gold-standard set (n=300) of manually dual screened articles.
- Generative Outputs: Evaluated by a blinded panel of three HEOR experts for:
  - Accuracy (agreement with the referenced source's data and sentiment),
  - Consistency (internal coherence between topic sections),
  - Readability (Flesch-Kincaid score, qualitative assessment),
  - Value (qualitative assessment of the value of the disseminated findings based on their depth and breadth of impact, implications for clinical practice and future research, and the reporting of actionable insights).
- Adaptability: Tested framework's capacity to integrate new HEOR topics and evidence, and integrate additional agent rules with minimal re-training.

Results

Classifier Performance
- Mean Specificity: 99%
- Mean Sensitivity: 87%
- Precision/Recall: Maintained high positive predictive value, with low false positive rate.

Generative AI Synthesis Module
- Compilation of written content into a HTML template, with consistent formatting and live web-links (**Figure 1**).
- Accuracy: 93% of generated summaries were rated as accurate and aligned with manual extractions.
- Consistency: Narrative and structured outputs maintained logical flow, completeness, and adhered to formatting guidelines.
- Readability: All outputs scored within the "easy" to "plain English" range on Flesch-Kincaid; minimal post-editing (<10% of summaries required correction).
- Value: Panel highlighted strengths in rapid turnaround and consistent formatting, with minor issues related to nuanced study design elements.

Figure 1. Comparison of themes identified by manual analysis and AI



Response to Adaptation and Refinement
- The modular design allowed rapid updating for new topics and inclusion criteria changes.
- Minimal technical intervention required for classifier retraining or synthesis prompt revision.

Feasibility and Scalability
- End-to-end automation led to a significant reduction in manual screening and summarization time.
- The pipeline was able to process and synthesize new literature updates weekly, compared to traditional review cycles spanning several weeks.
- The timeline reduction demonstrated by this workflow could substantially improve the timeliness of evidence delivery for HEOR stakeholders.

Conclusions

This pilot demonstrates that a fully automated, AI-driven pipeline can effectively support HEOR communications by rapidly identifying and synthesizing relevant literature. This application of AI, and the associated reduction in manual effort, can be used to identify and communicate high impact research, with the potential to significantly accelerate access to meaningful insights from the latest published evidence. Ongoing development will focus on broader topic coverage, more granular evaluation metrics, and integration into real-world HEOR and HTA workflows.

References

1. Hennessy S, Atsuta Y, Hill S, Rago L, Juhaeri J, Council for International Organizations of Medical Sciences Working Group X. Real-World Data and Real-World Evidence in Regulatory Decision Making: Report Summary From the Council for International Organizations of Medical Sciences (CIOMS) Working Group XIII. Pharmacoepidemiol Drug Saf. Mar 2025;34(3):e70117. doi:10.1002/pds.70117; **2.** Dalky A, Altawalbih M, Alshanik F, et al. Global Research Trends, Hotspots, Impacts, and Emergence of Artificial Intelligence and Machine Learning in Health and Medicine: A 25-Year Bibliometric Analysis. Healthcare (Basel). Apr 13 2025;13(8)doi:10.3390/healthcare13080892; **3.** Zwack CC, Haghani M, de Bekker-Grob EW. Research trends in contemporary health economics: a scientometric analysis on collective content of specialty journals. Health Econ Rev. Jan 25 2024;14(1):6. doi:10.1186/s13561-023-00471-6; **4.** Okoroiwu KL, Okoroiwu HU, Ude LO, Ezuma CO, Omeje EI. Global bibliometric analysis of cost effectiveness analysis in healthcare research from 2013 to 2023. Cost Eff Resour Alloc. Sep 16 2024;22(1):68. doi:10.1186/s12962-024-00576-7; **5.** Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol. Nov 2017;91:31-37. doi:10.1016/j.jclinepi.2017.08.011; **6.** Zisis K, Pavi E, Geitona M, Athanasakis K. Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. J Pharm Pharm Sci. 2024;27:12302. doi:10.3389/jpps.2024.12302