Leveraging AI for automated extraction of drug–event details from complex ICSR narratives

LLMs offer a promising approach to automating drug–event attribute extraction from complex ICSR narratives, strengthening their role in pharmacovigilance workflows and automated causality tools

Click for details

Olivia Mahaux,[1] Naomi Oladele Aduroja,[2] Louise-Emma Ngwi Sama,[2] Vijay Kara,[3] Gregory Powell,[4] Jeffery L Painter,[4] Andrew Bate[3,5]

[1]GSK, Wavre, Belgium; [2]UNC Eshelman School of Pharmacy, Chapell Hill, NC, USA; [3]GSK, London, UK; [4]GSK, Durham, NC, USA; [5]London School of Hygiene and Tropical Medicine, London, UK

Disclosures: OM, VK, GP, JLP, and AB are employed by GSK and hold financial equities in GSK. NOA and LENS have nothing to disclose. The authors declare no other financial or non-financial relationships and activities.

Background

- Reviewing ICSRs for causality assessment is time-consuming, especially for complex cases with multiple drugs and AEs,[1,2] because reviewers must manually retrieve key drug–event details relevant to causality assessment from the narrative. In our experience, manual narrative review typically requires 5–20 minutes per case.
- To improve efficiency, we evaluated LLMs for automatically extracting drug–event pair details from narratives into a structured format to support automated causality assessment for regulatory reporting.

Aim

To conduct a feasibility study assessing the ability of LLMs to reliably extract key attributes from complex ICSR narratives.

Conclusions

LLMs show promise in automating drug–event attribute extraction from complex ICSR narratives but face limitations in distinguishing "Unknown" and "Not applicable" categories for dechallenge and rechallenge, and in handling attributes requiring clinical judgment.

The additional 30-case validation set provided mixed support for the initial 20-case findings.

Improved prompts, refined attribute definitions, and model fine-tuning may address these limitations and enhance their utility in pharmacovigilance workflows and automated causality tools.

Methods

Data acquisition

- ✓ 50 complex post-marketing ICSRs, defined as cases with:
  - ≥2 suspect products
  - ≥2 AEs

Gold standard dataset

- ✓ extracted by a pharmacovigilance expert
- ✓ quality-controlled by a second reviewer

Gemini 2.5 Pro & GPT 5.1

- ✓ Exploratory set: 20 cases - **prompt v1**
- ✓ Validation set: 30 cases - **prompt v1**
- ✓ Same validation set re-tested: 30 cases - **prompt v2***

Extracted variables

- ✓ dechallenge
- ✓ rechallenge
- ✓ time to onset (TTO)**
- ✓ medical history (concurrent conditions)
- ✓ concomitant medications

Evaluation approach

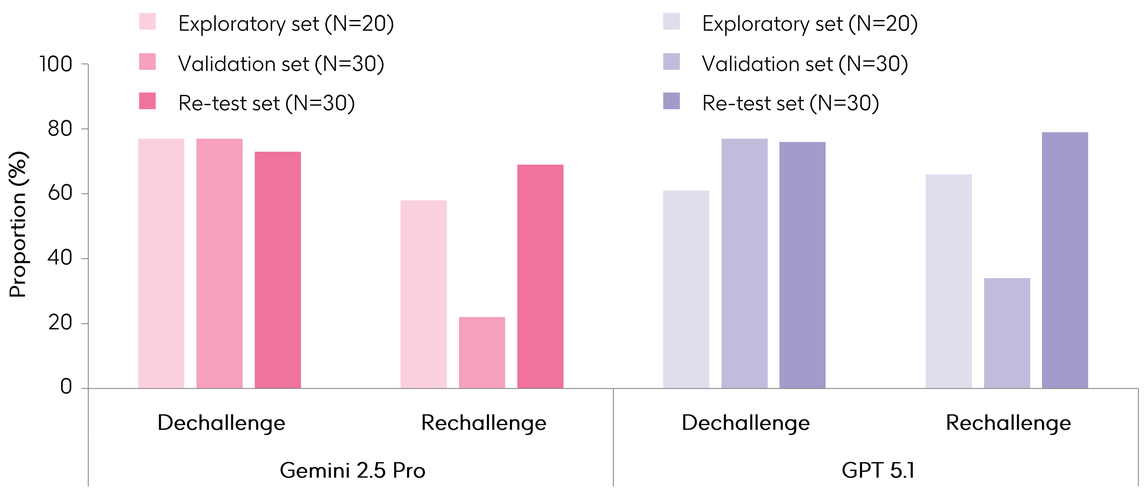
- ✓ LLM-extracted variables compared with the gold-standard dataset
- ✓ Agreement assessed for each extracted variable

Note: *Prompt v2 was refined post hoc based on observations from the exploratory set only; the initial 20 cases were not re-tested with prompt v2. **Extracted as numeric + unit, text, or Unknown with detail.

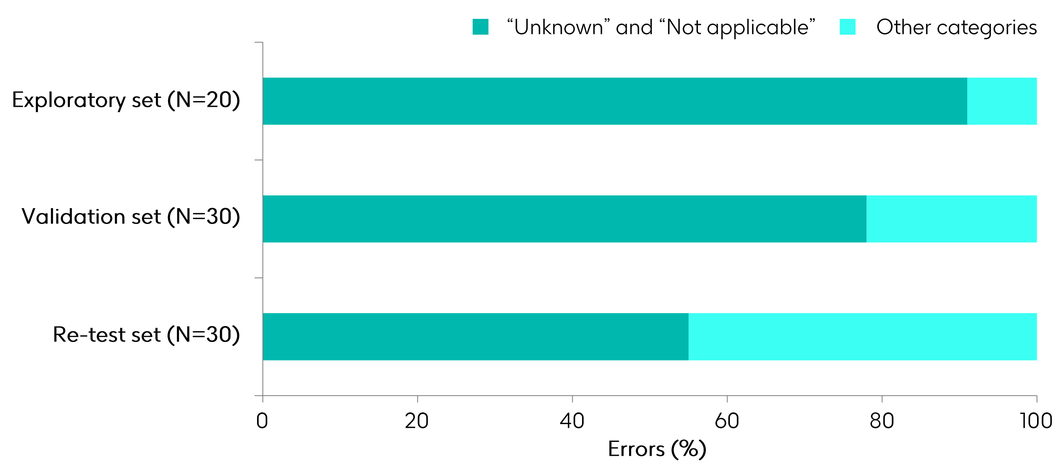
Results

Dechallenge/Rechallenge

- **Dechallenge/rechallenge matching with gold standard**: Dechallenge agreement was broadly stable and similar between models except in the exploratory set, whereas rechallenge was more variable, higher for GPT 5.1, and improved in the re-test.

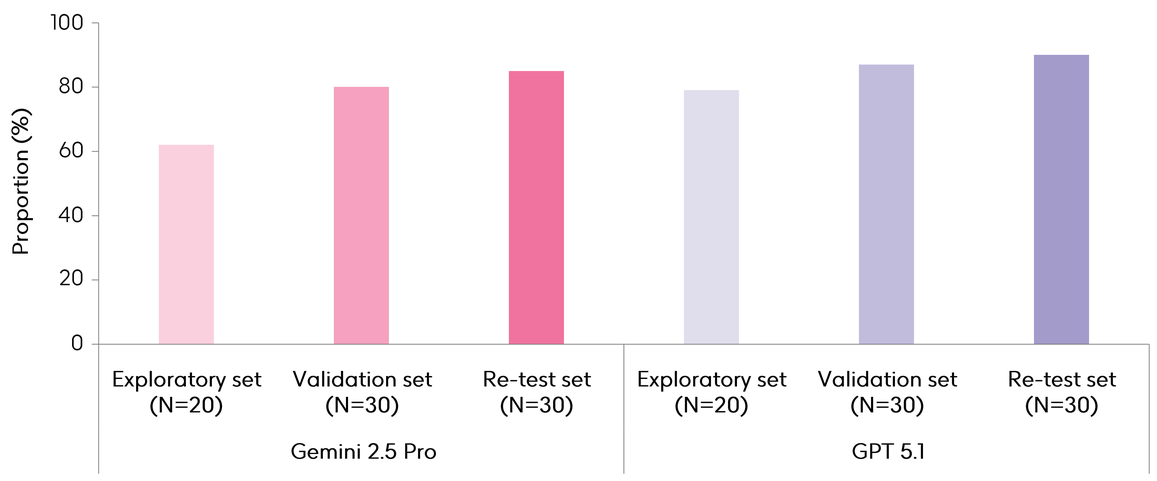


- **Errors breakdown**: Most errors were in the "Unknown" and "Not applicable" categories, although their proportion decreased in the re-test.



TTO

- **TTO matching with gold standard**: TTO matching improved across sets for both models and was consistently higher for GPT 5.1.

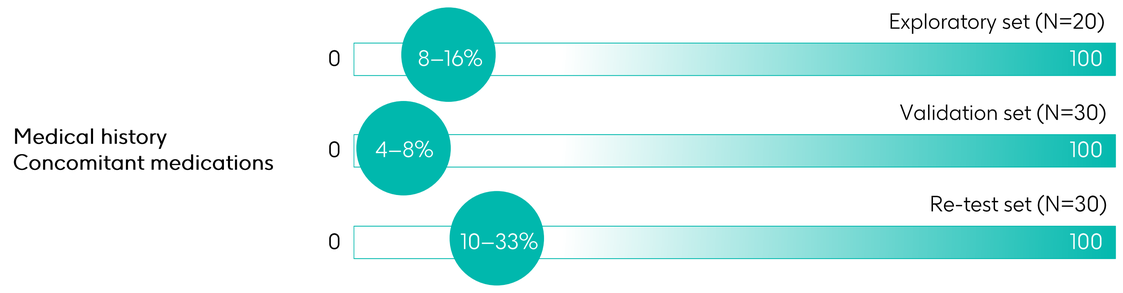


- **TTO numeric error**: Numeric errors showed a different pattern from TTO matching, particularly in the validation set for GPT 5.1.

		Exploratory set (N=20)	Validation set (N=30)	Re-test set (N=30)
Mean absolute TTO error (days)	Gemini 2.5 Pro	0.1	3.8	4.4
	GPT 5.1	1.1	199.3	0.0

Medical history/Concomitant medication

- **Medical history and concomitant medication matching**: accuracy remained low across sets, consistent with strict matching criteria and the need for clinical judgment when identifying relevant confounders.



Abbreviations

AE, adverse event; **ICSR**, individual case safety report; **LLM**, large language model; **N**, number of cases.

References

1. Ralph Edwards I. Drug Saf. 2017;40:365–72.
2. Rodrigues PP, et al. Artif Intell Med. 2018;91:12–22.

Data sharing statement

The data that support the findings of this study are available from the presenting author upon reasonable request.

Authors' contributions

OM, VK, GP, JLP, and AB contributed to the study concept or design. OM, NOA, and LENS contributed to data acquisition. OM, NOA, LENS, VK, GP, and JLP contributed to data analysis. All authors contributed to data interpretation. The authors have reviewed the content and approved the final version of the poster.

Acknowledgements

Medical writing (Bianca Ciocotisan), design, and coordination support were provided by Akkodis Belgium c/o GSK.

Funding

GSK

Presenting author: Gregory Powell, gregory.e.powell@gsk.com

Leveraging AI for automated extraction of drug–event details from complex ICSR narratives

LLMs offer a promising approach to automating drug–event attribute extraction from complex ICSR narratives, strengthening their role in pharmacovigilance workflows and automated causality tools

Olivia Mahaux,[1] Naomi Oladele Aduroja,[2] Louise-Emma Ngwi Sama,[2] Vijay Kara,[3] Gregory Powell,[4] Jeffery L Painter,[4] Andrew Bate[3,5]

[1]GSK, Wavre, Belgium; [2]UNC Eshelman School of Pharmacy, Chapell Hill, NC, USA; [3]GSK, London, UK; [4]GSK, Durham, NC, USA; [5]London School of Hygiene and Tropical Medicine, London, UK

Disclosures: OM, VK, GP, JLP, and AB are employed by GSK and hold financial equities in GSK. NOA and LENS have nothing to disclose. The authors declare no other financial or non-financial relationships and activities.

Supplemental data

Main page

Dechallenge metrics (stratified by model)

		Exploratory set (N=20)				Validation set (N=30)				Re-test set (N=30)			
	Class	Total	Precision	Recall	F1-score	Total	Precision	Recall	F1-Score	Total	Precision	Recall	F1-score
Gemini 2.5 Pro	Positive	231	0.71	0.91	0.80	409	0.75	0.69	0.72	363	0.91	0.64	0.75
	Negative	231	1.00	1.00	1.00	409	0.00	0.00	0.00	363	0.00	0.00	0.00
	Unknown	231	0.45	0.85	0.58	409	0.89	0.80	0.84	363	0.84	0.78	0.81
	Not applicable	231	0.78	0.56	0.65	409	0.85	0.70	0.77	363	0.59	0.54	0.56
GPT 5.1	Positive	210	0.85	1.00	0.92	414	0.79	0.75	0.77	384	0.85	0.60	0.70
	Negative	210	0.60	1.00	0.75	414	0.00	0.00	0.00	384	0.00	0.00	0.00
	Unknown	210	0.43	0.66	0.52	414	0.86	0.82	0.84	384	0.80	0.87	0.83
	Not applicable	210	0.66	0.40	0.49	414	0.53	0.51	0.52	384	1.00	0.50	0.67

Rechallenge metrics (stratified by model)

		Exploratory set (N=20)				Validation set (N=30)				Re-test set (N=30)			
	Class	Total	Precision	Recall	F1-score	Total	Precision	Recall	F1-score	Total	Precision	Recall	F1-score
Gemini 2.5 Pro	Positive	231	0.50	1.00	0.67	409	0.31	0.83	0.45	363	0.10	0.67	0.17
	Negative	231	0.00	0.00	0.00	409	0.71	0.24	0.36	363	0.00	0.00	0.00
	Unknown	231	1.00	0.13	0.23	409	0.62	0.09	0.15	363	0.90	0.69	0.78
	Not applicable	231	0.55	0.99	0.71	409	0.16	0.95	0.27	363	0.41	0.98	0.58
GPT 5.1	Positive	210	1.00	1.00	1.00	414	0.50	0.20	0.29	384	1.00	1.00	1.00
	Negative	210	0.00	0.00	0.00	414	0.00	0.00	0.00	384	0.00	0.00	0.00
	Unknown	210	0.48	0.17	0.25	414	0.72	0.26	0.38	384	0.87	0.86	0.87
	Not applicable	210	0.68	0.91	0.78	414	0.19	0.88	0.31	384	0.53	0.73	0.62

Definition of metrics

- **Precision:** It is the proportion of correctly identified positive classes out of all instances that the model labeled as positive. It shows how much the model can be trusted when it indicates a product-event pair is causal in a case.

$$\text{Precision} = \frac{True\ positives}{True\ positives + False\ positives}$$

- **Recall (sensitivity):** It is the proportion of actual positive classes that the model correctly identifies as positive. It measures how well the model can detect causal product-event pairs in cases.

$$\text{Recall} = \frac{True\ positives}{True\ positives + False\ negatives}$$

- **F1-score:** It is the harmonic average of precision and recall. It provides a single metric that balances both precision and recall, useful for evaluating the model's accuracy in scenarios where there is an uneven class distribution.

$$\text{F1-score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Abbreviations

ICSR, individual case safety report; LLM, large language model; N, number of cases.

Presenting author: Gregory Powell, gregory.e.powell@gsk.com