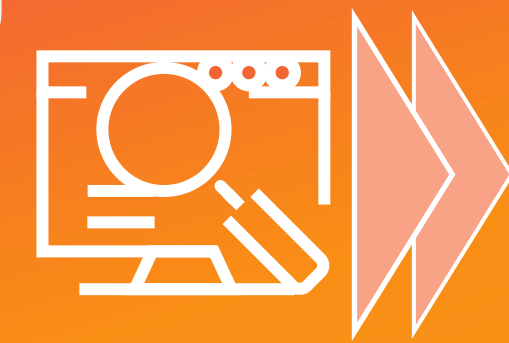


Comparison of data-driven vector representations and ontology-based semantic similarity measures for grouping adverse event terms

François Haguinet,¹ Jeffery L. Painter,² Gregory Powell,² Andrew Bate³

¹GSK, Wavre, Belgium; ²GSK, Durham, NC, USA; ³GSK, London, UK

Disclosures: All authors are employed by GSK and hold financial equities in GSK. The authors declare no other financial or non-financial relationships or activities.

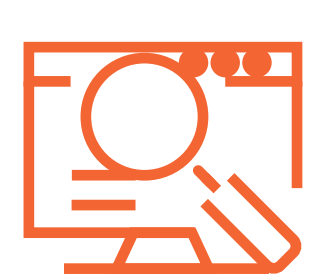


Ontology-based semantic similarity groups MedDRA AE terms into clinically coherent clusters more effectively than vigiVec (FAERS-trained) in this evaluation, suggesting that knowledge-based similarity is a promising baseline for AE term grouping in pharmacovigilance

Background

- Effective grouping of adverse event (AE) terms is crucial for pharmacovigilance (PV).
- Hierarchical terminologies like MedDRA, with their discrete, rigid boundaries, cannot explicitly capture the nuanced clinical relatedness in spontaneous reporting systems.
- Data-driven vector representations derived from reporting patterns are the most common approach for clustering AEs. Previous research evaluated ontology-based SSMs^{1,2} for clustering AEs, we present the first comparison to a data-driven vector representation approach (specifically vigiVec³ which is the most extensively discussed in PV).

Aim



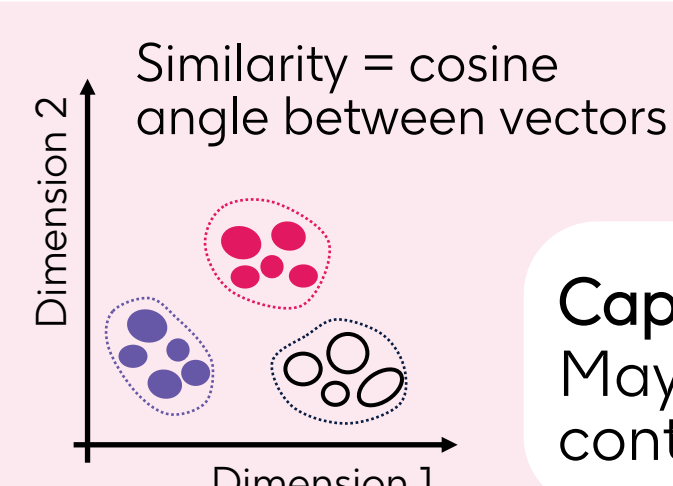
To compare vigiVec's ability to group MedDRA preferred terms (PTs) around medically meaningful centroid PTs to ontology-based semantic similarity measures (SSMs).

Methods

vigiVec (Word2Vec)
Data-driven similarity

FAERS reports
co-reporting patterns

Skip-gram architecture learns AE code embeddings by predicting all other codes on a case report from one 'target' code



Captures quantitative association
May reflect co-reporting, shared context, medication use

Ontology-based SSMs⁴
Knowledge-driven similarity

MedDRA
SNOMED CT

UMLS integration
+ IC

Similarity uses hierarchy and IC (e.g., Lin, Resnik)

$SSM = f(\text{shared ancestors, IC})$

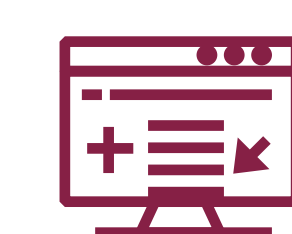
Reflects clinical meaning
Respects curated biomedical relationships

vigiVec

10,986,071 FAERS safety reports (2004–2019)

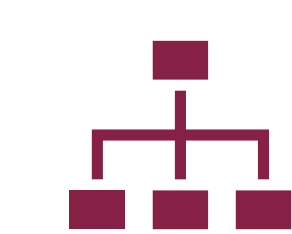
19,751 MedDRA PTs
8,397 product concepts

Generating evaluation dataset⁴



Step 1: Select centroid PTs

Select 12 centroid PTs based on narrow SMQs and corresponding proprietary groupings



Step 2: Compile base reference sets

For each centroid, create a base list by combining all PTs from its corresponding narrow SMQ, HLGT, and proprietary grouping



Step 3: Augment with semantically similar PTs

For each centroid PT, add the top 200 PTs most semantically similar to the centroid and not yet included in the base reference set



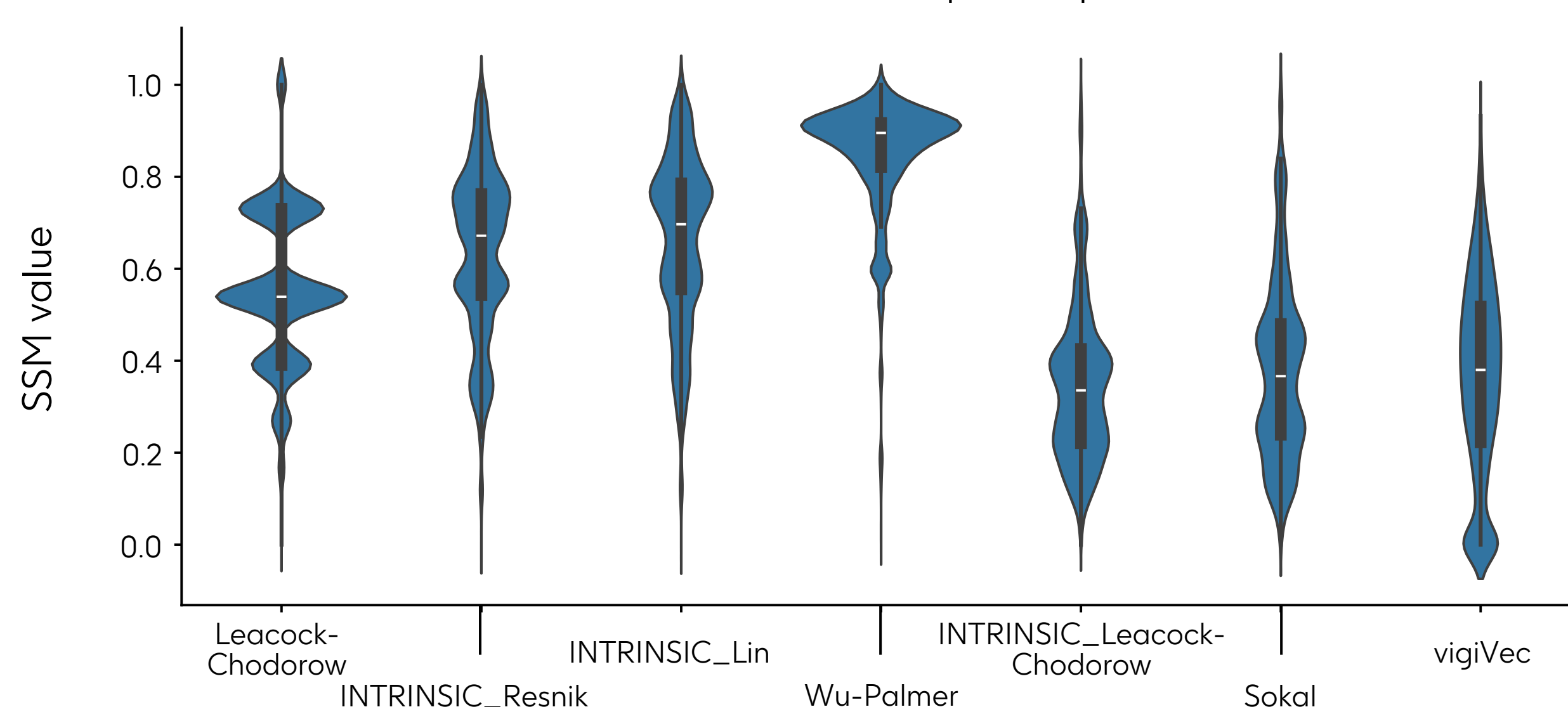
Step 4: Conduct medical review and adjudication

Experts independently review each final list

Results

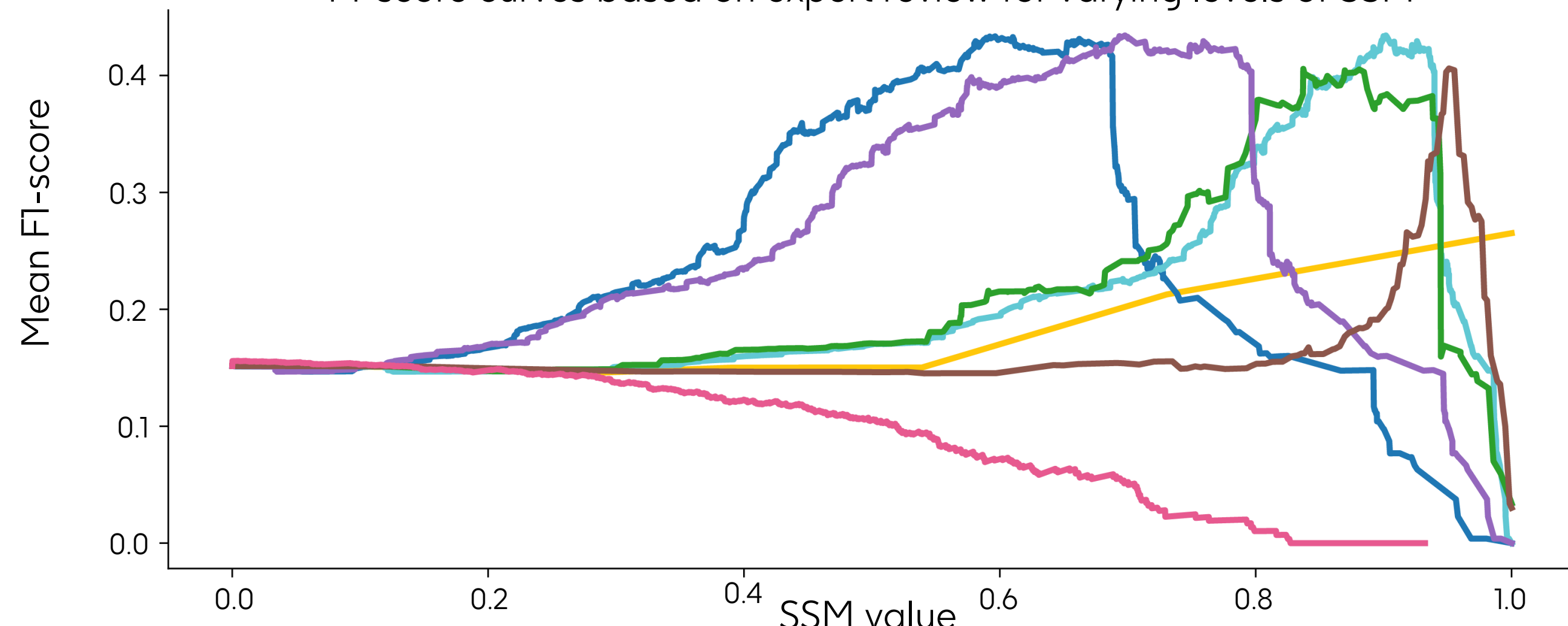
- Low values were more represented in the vigiVec similarity distribution, with many PTs exhibiting zero similarity.

Distribution of SSM values across all MedDRA PT pairs represented in the reference set



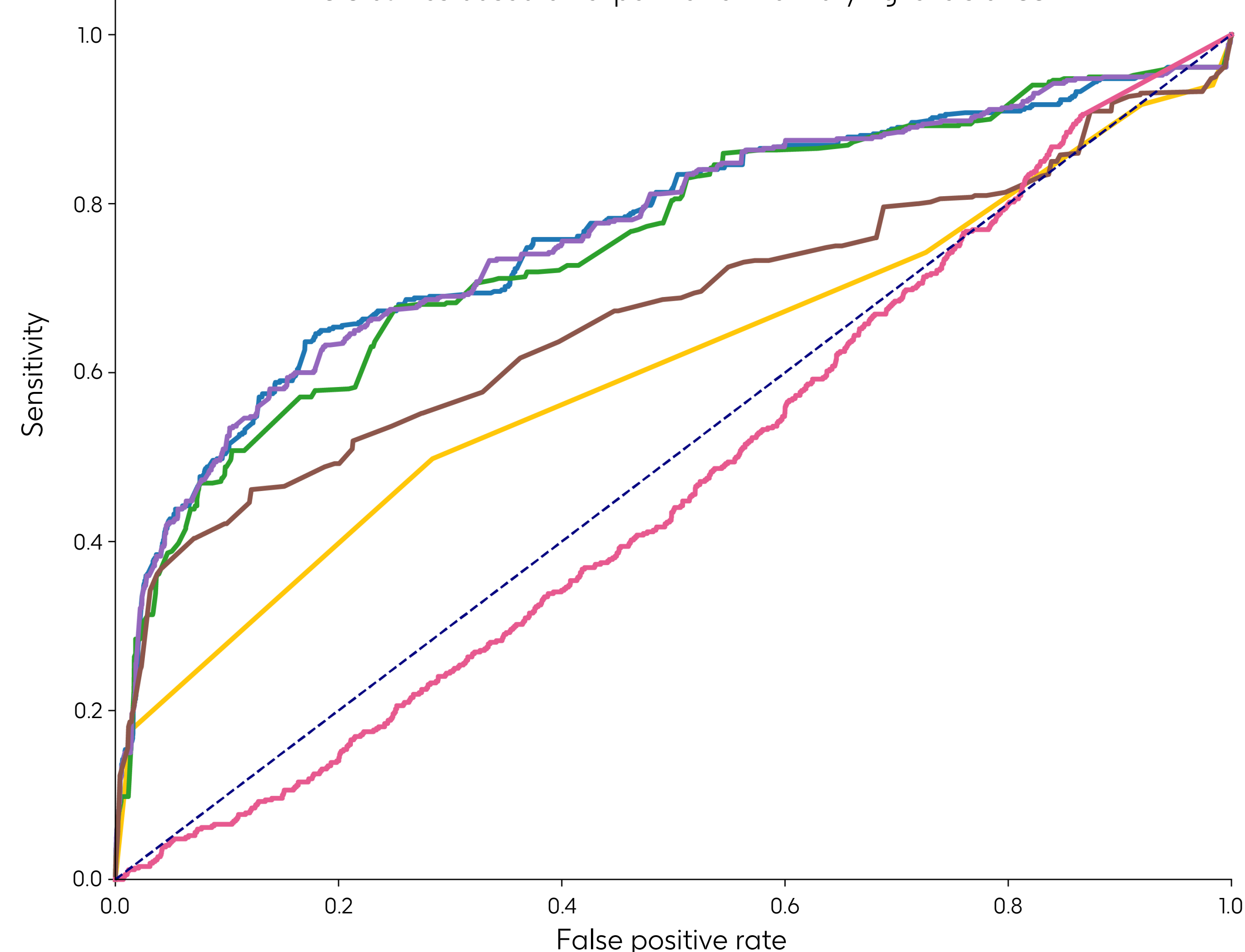
- F1-score analysis showed that increasing the similarity threshold for vigiVec generally decreased its performance as the loss in sensitivity (recall) consistently outweighed gains in precision. In contrast, the other SSMs, initially showed improved performance as precision gains exceeded sensitivity losses, and at a certain point, when the threshold became too strict, a sharp drop in sensitivity was seen.

F1-score curves based on expert review for varying levels of SSM



- ROC analysis showed that vigiVec approximated random classification across much of the sensitivity range (0.0–0.6), with modest improvement thereafter. AUC was 0.47, while ontology-based SSMs AUC ranged from 0.60 to 0.77.

ROC curves based on expert review for varying levels of SSM



The curves of Sokal and INTRINSIC_Lin are exactly the same due to their strictly monotonic relationship.

Legend depicting the curves in the graph on the bottom left side and the graph above:

INTRINSIC_Leacock-Chodorow, INTRINSIC_Lin, INTRINSIC_Resnik, Leacock-Chodorow, Sokal, Wu-Palmer, vigiVec

- Other implementations of vigiVec were tested with similar results.

Abbreviations

AUC, area under the curve; FAERS, US Food and Drug Administration Adverse Event Reporting System; HLGT, high-level group term; IC, information content; MedDRA, Medical Dictionary for Regulatory Activities; ROC, receiver operating characteristic; SMQ, Standardised MedDRA Query; SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms; UMLS, Unified Medical Language System.

References

- Pesquita C, et al. PLoS Comput Biol. 2009; 5:e1000443.
- Amar F, et al. J Med Internet Res. 2024; 26:e45209.
- Erlanson N, et al. Drug Saf. 2025; 48:401–413.
- Painter JL, et al. AMIA Annu Symp Proc. 2025; 2024:979–988.

Data sharing statement

The data that support the findings of this study are available from the corresponding author (FH), upon reasonable request.

Authors' contributions

Study concept and design: FH, JLP, GP, AB. Data acquisition: FH, JLP, GP. Data analysis: FH, JLP. Data interpretation: FH, JLP, GP, AB.

Acknowledgements

Medical writing (Maria C. Maior), design, and coordination support were provided by Akkodis Belgium c/o GSK.

Funding

GSK