

# An experiment fromSeValid: Integration of Artificial Intelligence in the Clinical Validation Pipeline of VAC4EU

Hyeraci Giulia<sup>1\*</sup>, Lippi Marco<sup>2\*</sup>, Maccari Martina<sup>3\*</sup>, Nardoni Valeria<sup>2\*</sup>, Limoncella Giorgio<sup>4\*</sup>, Arana Alejandro<sup>5\*</sup>, Lucenteforte Ersilia<sup>4\*</sup>, Marinai Simone<sup>2\*</sup>, Mohammadi Sima<sup>6\*</sup>, Roberto Giuseppe<sup>1\*</sup>, Virgili Gianni<sup>7\*</sup>, Weibel Daniel<sup>8\*</sup>, Dehghan Tarazjani Amirreza<sup>6\*</sup>, Gini Rosa<sup>1\*</sup>

<sup>1</sup>Pharmacoepidemiology Unit, ARS Toscana, Florence, Italy. <sup>2</sup>Department of Information Engineering, University of Florence, Florence, Italy. <sup>3</sup>SOD Ottica Fisiopatologica – Clinical Trial Center, Florence, Italy. <sup>4</sup>Department of Statistics, University of Florence, Florence, Italy. <sup>5</sup>RTI Health Solutions, Barcelona, Spain. <sup>6</sup>Department of Data Science and Biostatistics, Real World Evidence, University Medical Center Utrecht, Utrecht, the Netherlands. <sup>7</sup>Department NEUROFARBA, University of Florence, Italy. <sup>8</sup>University Medical Center Utrecht, Utrecht, the Netherlands. email: giulia.hyeraci@ars.toscana.it

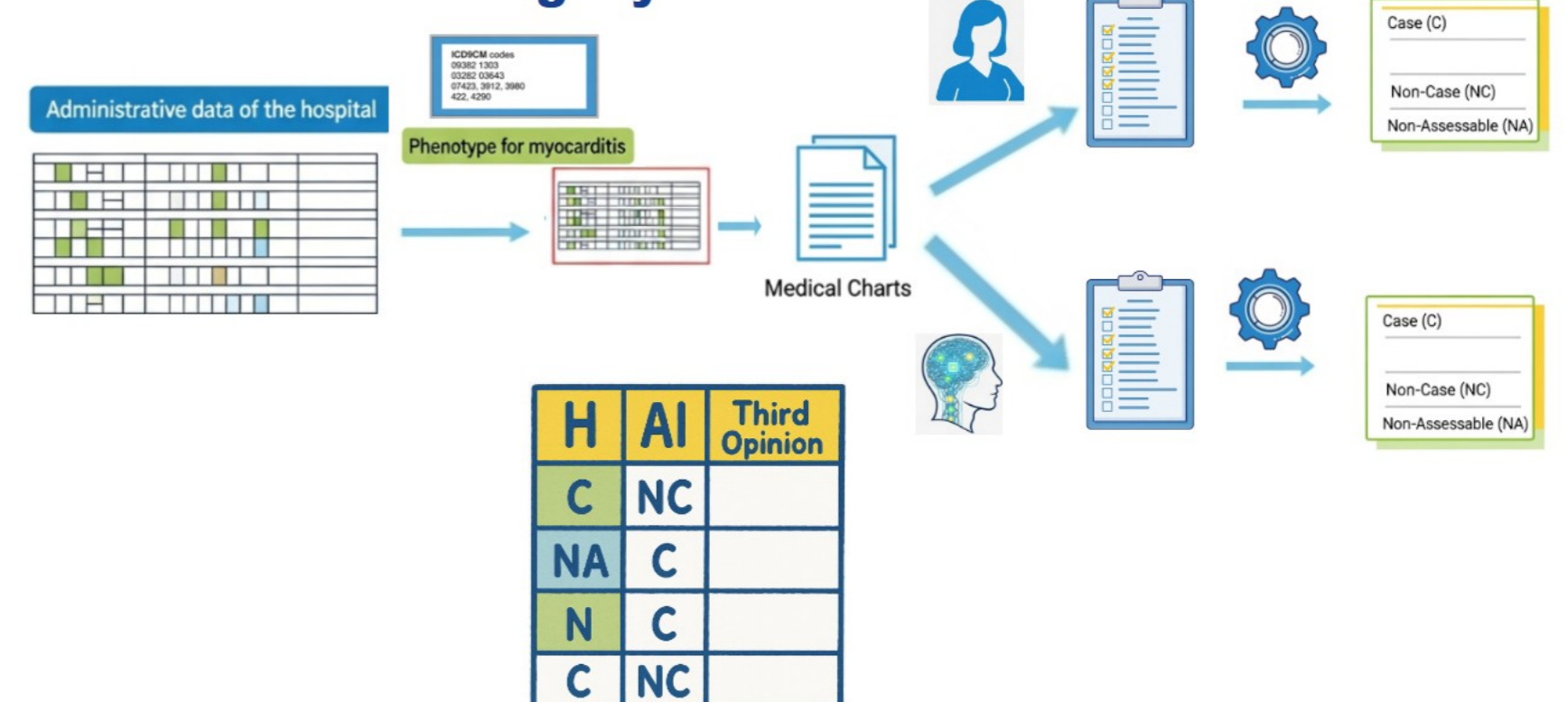
## BACKGROUND

Validation of study outcomes is crucial to assess the accuracy and reliability of results generated by real-world studies, and can provide valuable input for quantitative bias analysis. VAC4EU has established a pipeline to estimate outcome positive predictive value (PPV) through human review of medical data. However, this process is resource-intensive and may limit scalability. The SeValid project, co-funded by ARS Toscana and VAC4EU, aims to develop a comprehensive strategy to facilitate validation studies and reduce misclassification bias. Large language models (LLMs) offer the potential to complement human assessment and extend validation to larger samples. Therefore, the objective of this experiment is to integrate the VAC4EU validation pipeline, based on human assessment, with an LLM, using an algorithm to identify myocarditis events in hospital discharge records as a case study.

## METHODS

- A 16-item clinical questionnaire, designed by VAC4EU based on the Brighton Collaboration definition, was used to classify events as:
  - Case (C)
  - Non-case (NC)
  - Non-assessable (NA)
- Dummy medical charts (MC) from VAC4EU were used to:
  - Train a human assessor (H)
  - Design and refine prompts for an open-source LLM (Gemma2 9B)
- Real events were extracted from hospital discharge records in Florence, Italy (July 2022 – June 2024) using an algorithm based on ICD9CM codes: 09382, 1303, 03282, 03643, 07423, 3912, 3980, 422, 4290
- For each event, H and LLM independently completed the questionnaire.
- Concordance between H and LLM was assessed; a medical expert provided a third opinion for discordant cases.
- The reference standard (RS) was defined as:
  - Concordant H + LLM assignments, or
  - The expert's third opinion
- Positive predictive value (PPV) of the algorithm was calculated using H, LLM, and RS, and RS was used to assess validity of H and LLM.

### Use case: validating myocarditis



## RESULTS

Events were 38. Out of 38 pairs of assessments (H,LLM):

- Concordant pairs were **30** (79%), all (C,C).
- Discordant pairs were: (C,NA), N = **5**, 13%, (C,NC), N = **2**, 5% and (NC,C), N = **1**, 3%.
- Algorithm's **PPV** were (H) **97%** and (LLM) **97%**, respectively.
- Reference standard evaluated had 37 as C, 1 as NC, 0 as NA (PPV = 97%).
- **Sensitivity** of H and LLM were **97%** and **81%** respectively, while the one true NC were misclassified by both H and LLM (specificity = 0).

Table 1. Comparison between Human and LLM

Human	LLM	N	%
C	C	30	79%
C	NA	5	13%
C	NC	2	5%
NC	C	1	3%
Total		38	

C= case; NC= Non-case; NA= Non-assessable

Table 2. Comparison between Human, LLM and Reference standard

	TP	FP	NA	PPV
Human	37	1	0	97%
LLM	30	1	5	97%
Reference standard	37	1	0	97%

TP=True positives; FP: False positives; NA= Not assessable; PPV= Positive Predictive Value

## CONCLUSION

### LLM behavior

The LLM showed an overly cautious attitude, sometimes leading to misclassified NA/NC results.

### Complex cases

Two NC cases were correctly captured by RS, reflecting complex classifications missed by both H and LLM.

### Next steps

Iterative prompt optimization, ideally guided by real MC data, is essential to improve LLM performance.

### Lessons learned

This experiment provides valuable insights to refine and strengthen the analytical pipeline.

LLMs bring both risks and opportunities to:

- support outcome validation, and
- foster high-quality real-world evidence (RWE) generation from real-world data (RWD).