Verification and validation framework for data generated with artificial intelligence in the context of the OMOP CDM.

Please list all authors and their affiliations in the correct order
Gabriel Maeztu[1,2], Paula Chocron[1], Alejandro Castrelo[1], Flavius Nicu[1], Marc Asenjo[1], María Quijada[1], Sandra Pulido[1],Mónica Arrúe[1], Marc Oliver[1], Oriol Moles[1], Mariona Forcada[1], Irene López[1], Luis Leon[1], Alvaro Abella[1]
IOMED, Universitat de Barcelona

Background

The effective implementation of Artificial Intelligence (AI) to infer new data points presents significant opportunities for enhancing clinical databases and healthcare research. This paper delves into the critical process of verifying and validating the output of AI before integrating the output to the OMOP CDM, specifically within the context of data extracted from non-structured data sources such as text or images from production systems. The study’s methodology encompasses a two-tiered approach: first, a rigorous verification process conducted by a group of physicians, assessing the accuracy of AI inferences such as symptom identification, correct ID assignments or contextual information inference among others. Second, a comprehensive validation phase executed by clinical data scientists evaluates the data quality.

Methods

The study's methodology involves the use of a comprehensive framework for the verification and validation of AI outputs in the form of concept_ids, applied across 11 different cohorts at 11 distinct hospitals in different European regions. This approach is designed to demonstrate the generalisability of the evaluation framework across multiple therapeutic areas and various instances of the OMOP CDM in different geographic regions.

The verification process is physician-led, involving detailed assessments of AI inferences to ensure the accurate identification of concepts, correct ID assignments, and appropriate contextual information. Physicians conduct remote source data verification (rSDV)[1] to critically evaluate the AI-generated data points, confirming their clinical relevance and correctness.In parallel, the validation phase, carried out by clinical data scientists, assesses the AI outputs against six key dimensions of data quality: accuracy, completeness, consistency, reliability, timeliness, and relevance. Finally, Data Quality Dashboards were used for further checks[2].

The evaluation process includes cross-validation techniques that compare AI-identified concept IDs with data from structured sources, enhancing the robustness of the validation. A Bayesian statistical framework is employed to calculate sample sizes, ensuring rigorous performance evaluation with high confidence levels. Additionally, an algorithm is used to monitor annotator agreement[3] in real-time, thereby maintaining a high standard of data accuracy and reliability.

This multi-faceted approach aims to ensure the integrity and utility of AI-generated clinical data across diverse healthcare settings, illustrating the framework's scalability and applicability in various clinical contexts.

Results

The evaluation framework lets the analyst of the study set a threshold for each of the concept_ids generated through the AI process to be included or not in the OMOP CDM.

In this 11 studies and 11 cohort, a mean of 0.89 f1-score and a stddev 0.10 in 523 concept_ids were evaluated, and 92.3% of these concepts were included in the OMOP CDM. The results demonstrate the potential of AI in production environments to enhance the amount of available data in the OMOP CDM, while also highlighting the importance of systematic verification and validation to ensure the transparency and reliability of AI outputs to be included in the OMOP Databases.

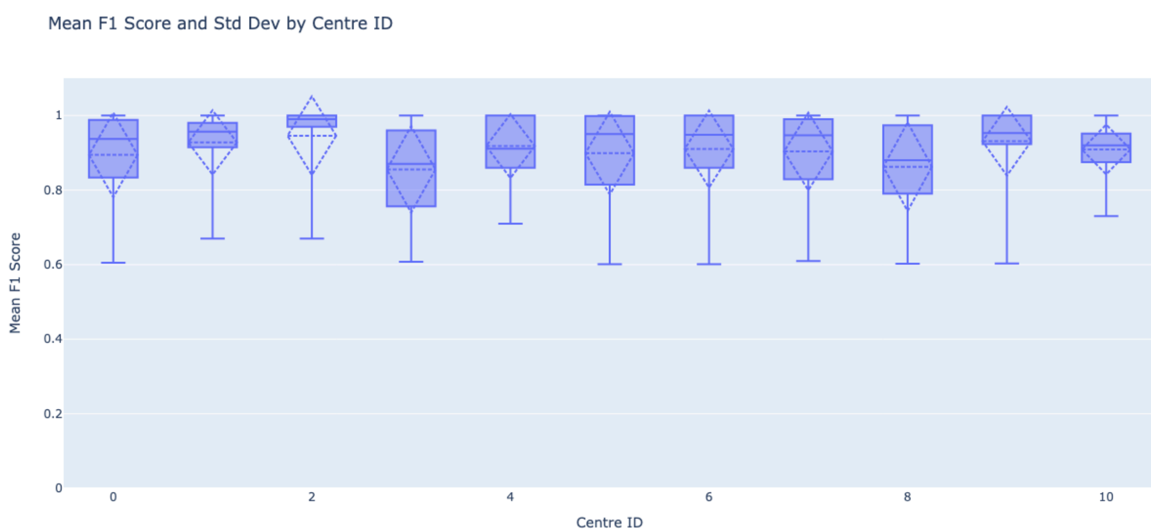

Figure 1. Mean f1-score of the models by Centre

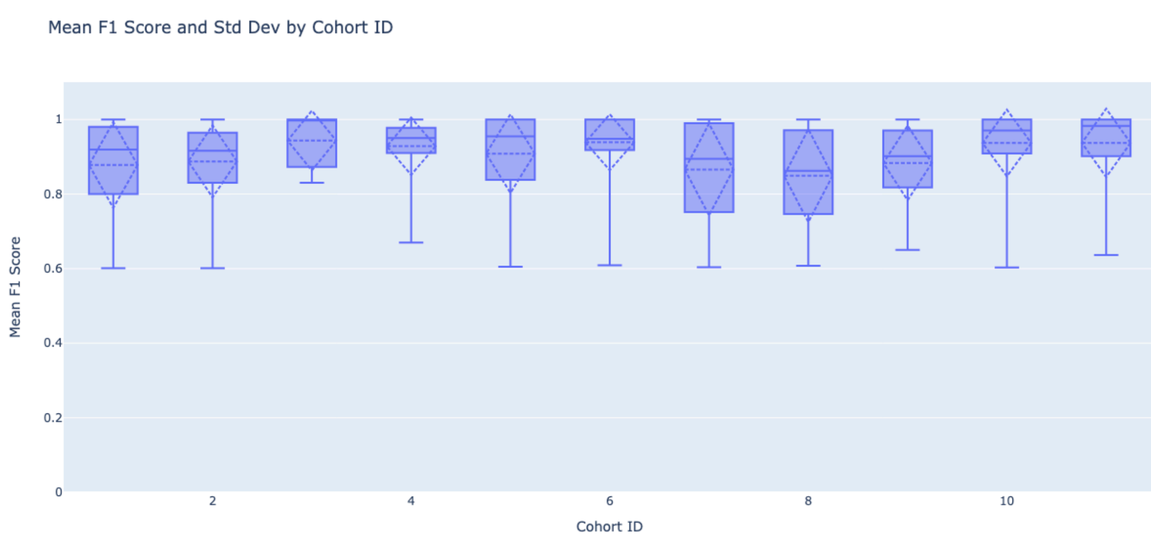

Figure 2. Mean f1-score of the models by Cohort

	Mean F1-score										
Centre ID	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Cohort ID											
1	0.863906	0.961547	0.905	0.852662	0.909953	0.851091	0.872655	0.918655	0.843112	0.80817	0.876314
2	0.881087						0.892431				
3							0.943333				
4		0.917222					0.939801				
5	0.907632					0.899925	0.953897	0.902561			
6			1.0	0.94011	0.955429	0.917944		0.975128	0.897959	0.81156	0.961526
7	0.86299			0.829598		0.902367	0.889918	0.861992	0.848953		
8				0.83742					0.862491		
9				0.840678	0.9			0.937292	0.873563	0.855031	0.907641
10	0.922133					0.951668	0.923109			0.952495	
11	0.897327			0.947899					0.954934		

Table 1. Mean f1-score of by Cohort and Centre

Conclusion

In conclusion, the positive outcomes of this study highlight the significant potential of AI in augmenting the data available in the OMOP CDM. The demonstrated results in data accuracy, quality, and standardization underscore the value of integrating advanced AI methodologies in healthcare data systems.

References

1. Pettus JA, Pajk AL, Glatz AC, Petit CJ, Goldstein BH, Qureshi AM, Nicholson GT, Meadows JJ, Zampi JD, Law MA, Shahanavaz S, Kelleman MS, McCracken CM; Congenital Cardiac Research Collaborative. Data quality methods through remote source data verification auditing: results from the Congenital Cardiac Research Collaborative. Cardiol Young. 2021 Nov;31(11):1829-1834. doi: 10.1017/S1047951121000974. Epub 2021 Mar 17. PMID: 33726868.
2. Blacketer C, Schuemie FJ, Ryan PB, Rijnbeek P (2021). "Increasing trust in real-world evidence through evaluation of observational data quality." *Journal of the American Medical Informatics Association*, **28**(10), 2251-2257.
3. Colosimo ME, Morgan AA, Yeh AS, Colombe JB, Hirschman L. Data preparation and interannotator agreement: BioCreAtIvE task 1B. BMC Bioinformatics. 2005;6 Suppl 1(Suppl 1):S12. doi: 10.1186/1471-2105-6-S1-S12. Epub 2005 May 24. PMID: 15960824; PMCID: PMC1869005.