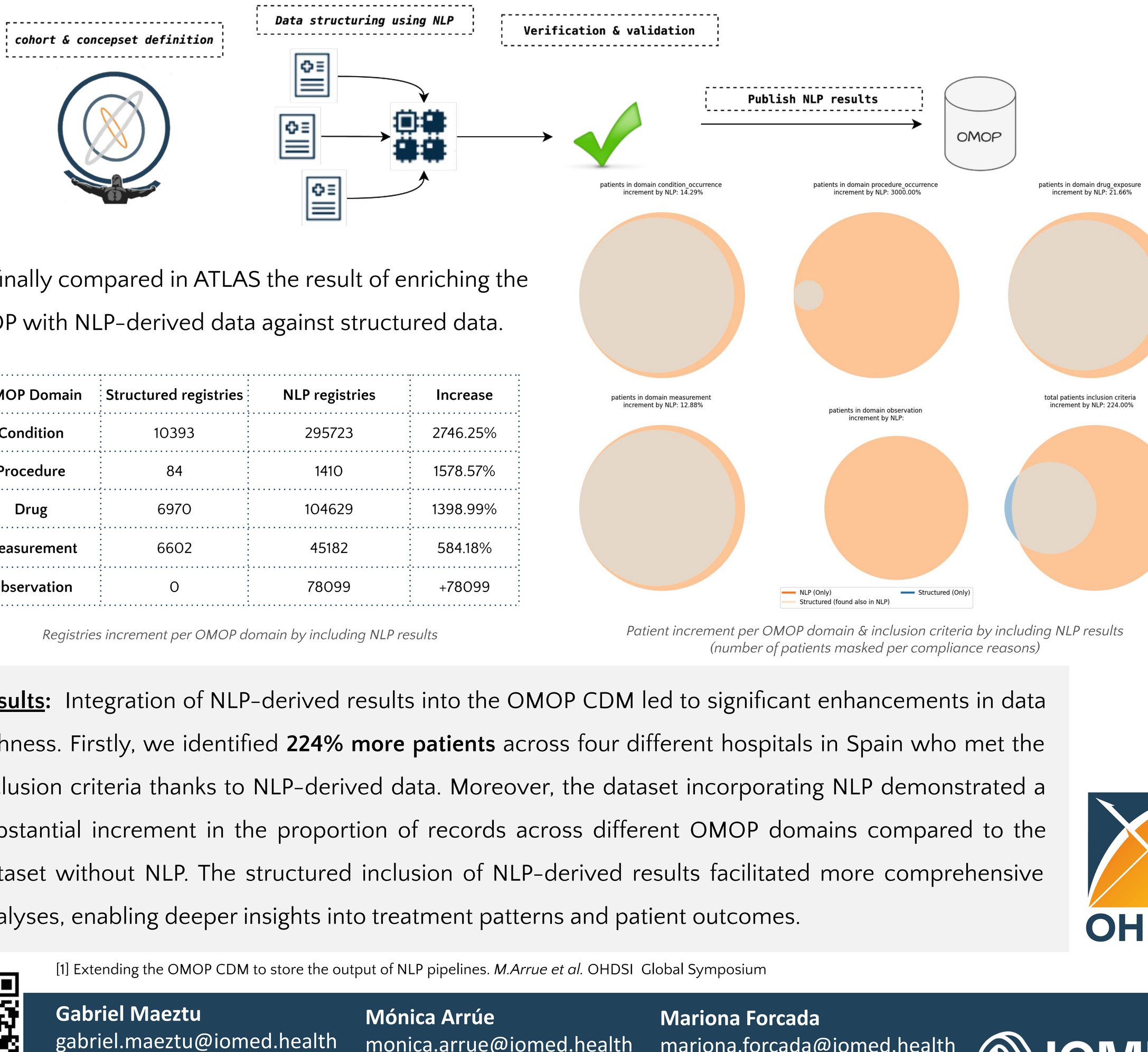
Integrating NLP-derived results in the OMOP CDM

Enhancing Real-World Clinical Data Analysis

Hospitals generate vast quantities of data daily, with **up to 80% of this data being collected in an unstructured** format such as free text. Converting this data into a structured format is essential to unlock its value. In this context, **integrating the output of NLP algorithms with the OMOP CDM** offers a compelling solution for enhancing the analysis of real-world clinical data using standard tools such as ATLAS or HADES.

Methods

A study in the field of dermatology was conducted across four hospitals in Spain. We employed multiple NLP techniques to standardize clinical data from clinical notes into OMOP concept_ids. In this study, 132 relevant concepts were defined and extracted using NLP. The results of the NLP were stored in an expanded NLP schema[1] specifically tailored to efficiently store the NLP outputs. This schema was developed to address the complexities of NLP outputs and facilitate their transformation into the standard OMOP CDM v5.4 structures. Notably, OMOP CDM v5.4 includes a *'note_nlp'* table intended for storing NLP results; however, querying this table can often be complex and inefficient.



Patient increment per OMOP domain & inclusion criteria by including NLP results (number of patients masked per compliance reasons)

We finally compared in ATLAS the result of enriching the OMOP with NLP-derived data against structured data.

OMOP Domain	Structured registries	NLP registries	Increase
Condition	10393	295723	2746.25%
Procedure	84	1410	1578.57%
Drug	6970	104629	1398.99%
Measurement	6602	45182	584.18%
Observation	0	78099	+78099

Registries increment per OMOP domain by including NLP results

Results: Integration of NLP-derived results into the OMOP CDM led to significant enhancements in data richness. Firstly, we identified **224% more patients** across four different hospitals in Spain who met the inclusion criteria thanks to NLP-derived data. Moreover, the dataset incorporating NLP demonstrated a substantial increment in the proportion of records across different OMOP domains compared to the dataset without NLP. The structured inclusion of NLP-derived results facilitated more comprehensive analyses, enabling deeper insights into treatment patterns and patient outcomes.

[1] Extending the OMOP CDM to store the output of NLP pipelines. *M.Arrue et al.* OHDSI Global Symposium

Gabriel Maeztu
gabriel.maeztu@iomed.health
María Quijada
maria.quijada@iomed.health
Mónica Arrúe
monica.arrue@iomed.health
Paula Chocrón
paula.chocron@iomed.health
Mariona Forcada
mariona.forcada@iomed.health
Miguel Ángel Mayer
miguelangel.mayer@upf.edu