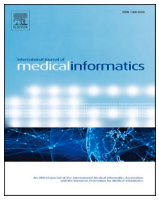
International Journal of Medical Informatics 217 (2026) 106511

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Original Research Article

A multimodal EHR-based phenotyping framework integrating consensus clustering and transformer-based clinical NLP: application to autoimmune gastritis

Daniele Pala [a,b,1,*], Marco Vincenzo Lenti [c,d,1], Giovanni Santacroce [c,d], Laura Bergomi [b], Chiara Curgu [b], Tommaso Buonocore [b], Chiara Sirtoli [a], Enea Parimbelli [b], Giordano Lanzola [b], Antonio Di Sabatino [c,d]

[a] Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy
[b] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
[c] Department of Internal Medicine and Medical Therapeutics, University of Pavia, Pavia, Italy
[d] First Department of Internal Medicine, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

ARTICLE INFO

Keywords:
Atrophy
Consensus clustering
Disease progression
Endoscopy
Multimodal analysis
Natural language processing
Precision medicine

ABSTRACT

Objective: To develop and evaluate a multimodal electronic health record (EHR)-based phenotyping pipeline integrating structured and unstructured clinical data to identify disease subgroups and characterize longitudinal trajectories in a real-world setting.

Materials and methods: We conducted a retrospective multicenter study including 1,598 patients with autoimmune gastritis. Structured demographic and clinical variables were combined with longitudinal endoscopic and histological data extracted from routine care. A consensus clustering strategy integrating partitioning (K-medoids) and hierarchical approaches was applied to identify robust patient subgroups. Free-text endoscopic reports were processed using a fine-tuned transformer-based natural language processing (NLP) model to automatically extract structured phenotypic features. To address irregular follow-up intervals, time-normalized progression indices were developed to capture both severity and temporal dynamics of disease evolution.

Results: After preprocessing, 607 patients were included in the analysis. The consensus clustering approach identified three clinically distinct subgroups. The NLP model demonstrated high performance in extracting endoscopic features (accuracy 90.2%, balanced accuracy 89.3%). Application of the proposed progression indices revealed significant differences in longitudinal patterns of mucosal damage across clusters ($p < 0.01$).

Conclusion: This study demonstrates the feasibility of integrating clustering techniques and transformer-based clinical NLP within a unified EHR phenotyping pipeline. The proposed approach supports scalable secondary use of structured and narrative clinical data for subgroup discovery and trajectory modeling in chronic diseases.

1. Introduction

Autoimmune gastritis (AIG) is a chronic immune-mediated disorder characterized by progressive destruction of the oxyntic mucosa, leading to gastric atrophy and functional impairment [1]. Although its reported prevalence ranges from 0.5% to 4.5% depending on population characteristics [2], AIG is frequently under-recognized due to its heterogeneous and often nonspecific clinical presentation. Over time, disease progression may result in micronutrient deficiencies, anemia, neurological manifestations, and an increased risk of gastric neoplasia [3–7].

Diagnosis relies on a combination of serological, endoscopic, and histological findings, including parietal cell antibody positivity and corpus-predominant atrophy [8–10]. However, patients present at variable stages and with diverse clinical profiles, and tools for systematic phenotypic stratification and longitudinal risk assessment remain limited. As a consequence, management often focuses on treating complications rather than anticipating disease trajectories.

* Corresponding author at: Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy.
E-mail address: daniele.pala@unibg.it (D. Pala).
[1] Co-first authors.

https://doi.org/10.1016/j.ijmedinf.2026.106511
Received 11 March 2026; Received in revised form 4 May 2026; Accepted 26 May 2026
Available online 26 May 2026

From a clinical informatics perspective, AIG represents a prototypical example of a heterogeneous chronic condition documented through both structured variables and narrative clinical reports within electronic health records (EHRs). Leveraging these multimodal data sources for subgroup identification and progression modeling requires methods capable of integrating mixed data types and handling irregular follow-up intervals.

In this study, we developed a multimodal EHR-based phenotyping pipeline combining consensus clustering, transformer-based natural language processing of endoscopic reports, and time-normalized progression indices. Using AIG as a real-world use case, we illustrate a scalable approach for data-driven subgroup discovery and trajectory characterization in routine clinical datasets.

2. Methods

2.1. Description of the datasets

This paper presents a sub-analysis of a previously published study on an Italian cohort comprising two complementary datasets that included information on 1598 patients with AIG diagnosed across nine referral centres. The original study protocol was approved by the local ethics committee (Fondazione IRCCS Policlinico San Matteo, protocol number P3599/2017).

The first dataset contained cross-sectional data collected at the time of diagnosis, including demographic and clinical characteristics. The second dataset consisted of longitudinal data, including endoscopic findings and histological atrophy scores obtained during follow-up.

2.1.1. Cross sectional dataset

The cross-sectional dataset included the following variables:

- Demographic variables: Age, Sex, Marital Status and Socioeconomic Status.
- Risk-related behaviours: Smoking, Alcohol use, use of gastroprotectors.
- Clinical variables: blood count, homocysteine, past Helicobacter Pylori infection, Gastrin and Chromogranin, parietal cell antibodies (PCA).
- Symptoms and Complications: presence and typology of gastrointestinal and neurological symptoms, presence of complications.
- Autoimmune comorbidities: presence of one or more autoimmune comorbidities among vitiligo, thyroid disease, Addison Disease, celiac disease, psoriasis, type 1 diabetes, other.
- Family history and reasons for diagnostic work-up.

2.1.2. Longitudinal dataset

The longitudinal dataset comprised data collected during oesophagogastroduodenoscopies (OGDs). At each examination, patients underwent both macroscopic assessment of the gastric mucosa and histological evaluation of the corpus atrophy.

Each patient underwent up to four OGDs, with non-uniform follow-up intervals. For each procedure, the following information was recorded:

- A free-text clinical note in Italian describing the main endoscopic findings.
- A histological atrophy score of the gastric corpus ranging from 0 (no atrophy) to 3 (severe atrophy) was provided.

2.2. Variable selection and missing data imputation

Several variables in the cross-sectional dataset exhibited missing values, exceeding 60% for some features. Socioeconomic indicators and blood count parameters showed the highest rates (>50%) and were excluded from primary analyses to minimize bias while preserving sample size.

For the remaining variables, a complete-case approach was adopted, except for gastrointestinal symptoms and PCA, for which clinically informed imputation was applied:

- Gastrointestinal symptoms: Missing entries were interpreted as absence of symptoms, based on documentation practices and confirmed through consultation with treating physicians, who reported that symptoms are recorded only when present.
- PCA antibodies: PCA values were missing in 9.19% of cases. Given their central diagnostic role, imputing these values allowed retention of a larger cohort. We hypothesized that missing results reflected negative tests. To assess this assumption, patients with missing PCA values were compared with those with documented negative results using age and prevalence of autoimmune comorbidities as reference variables.
- Age distributions were compared using the Mann–Whitney U test and proportions using the χ^2 test. Patients with positive PCA differed significantly from those with negative PCA, whereas no significant differences were observed between missing and negative cases. Missing PCA values were therefore considered compatible with negative results and imputed accordingly.
- Age, Sex, Autoimmune Diseases, Familiarity, Complications and Helicobacter Pylori infection: unfortunately, also these variables presented several missing data, and no information about their cause was available. Therefore, in order to avoid introducing bias in the data, we just removed the lines presenting missing information. This caused the loss of a large number of patients, but ensured the maximum informativeness of the remaining ones.

2.3. General methodology

Since a standard classification framework for AIG is currently lacking, the primary objective of this study was to identify clinically meaningful disease subtypes using a data-driven approach.

The analytical framework consisted of two main steps (see Fig. 1):

1) **Identification of disease subgroups**: Unsupervised learning techniques were applied to o demographic and clinical variables to uncover latent patient subgroups. To enhance robustness and minimise algorithm-specific bias, a consensus clustering strategy combining two complementary clustering methods was implemented.
2) **Clinical validation through longitudinal analysis**: The clinical relevance of the identified clusters was assessed by analysing disease trajectories over time using longitudinal OGD data. This step involved the development of a dedicated NLP pipeline to extract structured information from free-text endoscopic reports, allowing quantitative evaluation of progression patterns across clusters.

Fig. 1 shows the general pipeline of the study.

2.3.1. Consensus clustering

Consensus Clustering was adopted to identify robust patient subgroups while mitigating the variability inherent to single-algorithm approach. Since no prior assumptions regarding the optimal number of clusters required were available, a multi-step unsupervised framework was implemented.

First, a partitional approach was applied using K-Medoids. Given the mixed nature of the dataset (continuous and categorical variables), pairwise dissimilarities were computed using Gower's distance. The optimal number of clusters (K) was determined based on the average silhouette width.

Second, a hierarchical clustering analysis was performed using divisive hierarchical clustering with Gower's distance, allowing exploration of the underlying data structure from a complementary perspective. The cutoff was selected through visual inspection of the

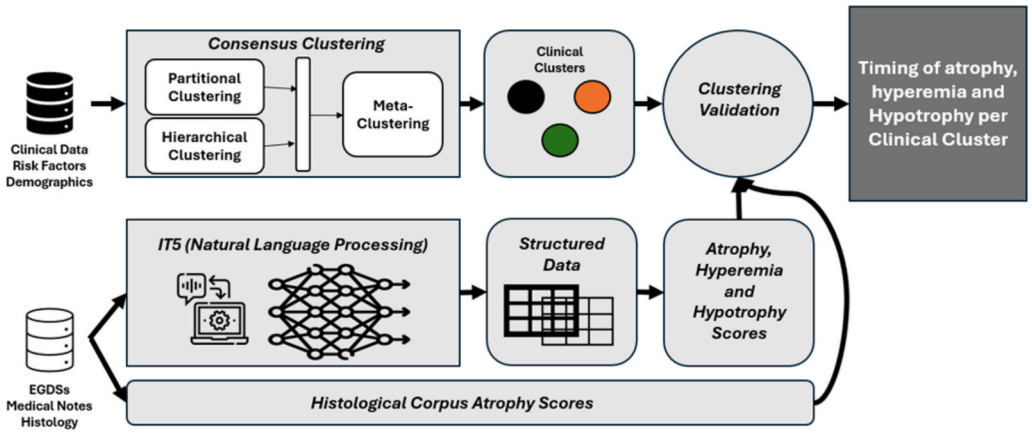


Fig. 1. Pipeline of the study, based on the combination of consensus clustering for the static dataset containing data at the moment of diagnosis, and NLP and atrophy scores accounting for temporal dynamics in the dynamic EGDS dataset..

dendrogram.

Finally, results from the two approaches were integrated through a *meta*-clustering step using the K-Modes algorithm, again with silhouette optimization to choose the best number of clusters, which generated the final consensus clusters.

This multi-algorithm strategy improves cluster stability and enhances the likelihood of identifying clinically meaningful subphenotypes.

The combination of K-medoids and hierarchical clustering was chosen to leverage complementary properties of partitioning and hierarchical approaches in the context of mixed-type clinical data. K-medoids, using Gower distance, provides robustness to outliers and is well suited for mixed variables, while hierarchical clustering allows exploration of the underlying data structure without pre-specifying the number of clusters. The subsequent K-modes *meta*-clustering step was introduced to integrate the results from both methods and reduce algorithm-dependent variability, in line with established consensus clustering strategies.

2.3.2. NLP approach

Free-text clinical notes from OGDs were analysed using a supervised NLP pipeline to automatically extract structured clinical variables.

Annotation

An experienced physician at the San Matteo Polyclinic Hospital in Pavia (MVL) manually annotated 13 predefined features:

- Twelve binary variables describing mucosal findings (i.e., atrophy, hypotrophy, and hyperemia of the antrum, corpus, and fundus; hiatal hernia; NET; carcinoma)
- One categorical multi-class variable describing polyp burden (categorized as absent, isolated of small size, isolated of large size, multiple, or multiple of small size).

Model architecture

Due to the shortness of the clinical text notes (all under 512 tokens), a lightweight encoder-decoder transformer model trained for Italian language tasks (IT5 model – it5/it5-base-question-answering, Hugging Face) [11] was fine-tuned in a question-answering framework, using predefined questions and answer options to facilitate structured information extraction.

Training, Validation, Testing

Data were split into training (70%), validation (15%), and test (15%) sets. Model fine-tuning was performed using the training and validation sets on 10 of the 13 features. Three rare outcomes —hiatal hernia, NET, and carcinoma—were excluded due to class imbalance. To address class imbalance, polyp categories were collapsed into absent, isolated, and multiple.

Model performance was assessed on the test set. Evaluation metrics included strict accuracy (defined as the exact match between the model's prediction and the reference annotation), as well as precision, recall, specificity, F1-score, and balanced accuracy. Except for strict accuracy, all metrics were computed using a binary classification framework, considering answers equal to present, isolated, and multiple as positive cases, and absent as negative.

2.3.3. Longitudinal analysis

Longitudinal OGD data provided both endoscopic and histological indicators of disease evolution, i.e. histological corpus atrophy score, with severity ranging from 0 to 3, and information about the presence and localization of atrophy, hypotrophy hyperaemia, polyps and cancerous lesions.

Because follow-up intervals and number of examinations varied across patients, conventional time-series modelling was not feasible without significant approximation. Therefore, averaged measures resuming information about both the severity and the timing of the disease progression were designed. In particular, corpus atrophy histological scores were scaled as follows:

$$Score(i) = \frac{MaximumAthrophy(i) - MinimumAtrophy(i)}{YearofMaximumAtr.(i) - YearofMinimumAtr.(i)}$$

Meaning that for each i^{th} patient a score ranging from 0 to 3 is obtained, with higher numbers corresponding to a higher atrophy progressing in a shorter time period. Exceptions were made to this calculation in the following cases:

- If the minimum atrophy value ever reported was 3, the Score value was forced to 3.
- If both the minimum and maximum atrophy values were reported for the same year, the denominator is set to 1 in order to avoid dividing by zero.

Similarly, a set of scores for all the endoscopic features, i.e. atrophy, hypotrophy, hyperaemia and polyps variables obtained from the clinical notes was calculated. In this case, the reported values for each patient/OGDs consisted in only 0 or 1 indicating the absence or presence of atrophy, hypotrophy, hyperaemia and polyps. Therefore, the summarized scores for each metric were calculated as follows:

$$Score(i) = \frac{1}{Yearof(Metric(i,j) = 1) - FirstEGDSyear(i,j)}; ifmetric(i,j) = 1$$

where i indicates the patient and j indicates the specific atrophy,

hypotrophy or hyperaemia metric. Similarly to the histological atrophy score, a number ranging from 0 to 1 is obtained, where 1 indicates that the metric is already present at the first OGD, whereas 0 indicates that the specific metric is never observed in all visits. In short, the presence of the specific OGD prognostic marker is weighted by the time that occurred before its presentation from the first procedure. If two OGDs were performed during the same year and the atrophy developed between the two, the denominator was set to 1.

A Cox proportional hazard model was also performed (Results reported in the Supplementary Material), however the scarceness and high variability of time frames required the application of the Last Observation Carried Forward method to regularize the data.

3. Results

Following preprocessing and handling of missing data, 607 patients from the original cohort were retained for analysis. Their main characteristics are reported in Table 1. Excluded cases lacked sufficient information for inclusion in the clustering and longitudinal analyses. Since the proportion of excluded patients is relatively high, we performed a statistical analysis in order to demonstrate the absence of selection bias. The results are reported in the Supplementary Material (Table S1).

3.1. Clustering results

The consensus clustering pipeline was applied to the selected cohort using nine clinically relevant variables: age, sex, PCA, presence of gastrointestinal symptoms, autoimmune comorbidities, family history, disease complications, and previous *Helicobacter pylori* infection.

Additional variables with high missingness (smoking, alcohol consumption, blood count parameters, NET, and carcinoma) were not included in cluster derivation but were used for clustering validation.

Using K-Medoids clustering with Gower distance, the silhouette coefficient indicated two clusters as the optimal partition (Fig. 2).

Divisive hierarchical clustering was then performed to explore alternative structures. Based on dendrogram inspection, a cutoff distance of 0.38 was selected to obtain five different clusters, containing 23, 24, 9, 267 and 294 observations respectively (Fig. 3).

Table 1
Characteristics of the patients selected after preprocessing and removal of missing entries.

Variable	Female (n = 449)	Male (n = 158)	p-value	Test used
Age (years)	56 (45–67)	58.00 (46–69)	0.119	Mann-Whitney U
PCA Ab negative	107 (23.8%)	37 (23.4%)	1.000	Chi-square
PCA Ab positive	342 (76.2%)	121 (76.6%)		
Autoimmunity – No	184 (41.0%)	95 (60.1%)	<0.001*	Chi-square
Autoimmunity – Yes	265 (59.0%)	63 (39.9%)		
Familiarity – no	372 (82.9%)	138 (87.3%)	0.231	Chi-square
Familiarity – yes	77 (17.1%)	20 (12.7%)		
Complications – yes	84 (18.7%)	33 (20.9%)	0.632	Chi-square
Complications – no	365 (81.3%)	125 (79.1%)		
Previous H. Pylori – No	341 (75.9%)	133 (84.2%)	0.041*	Chi-square
Previous H. Pylori – Yes	108 (24.1%)	25 (15.8%)		
NET – negative	397 (88.4%)	141 (89.2%)	0.893	Chi-square
NET – positive	52 (11.6%)	17 (10.8%)		
Carcinoma – negative	447 (99.6%)	158 (100.0%)	1.000	Fisher's exact
Carcinoma – positive	2 (0.4%)	0 (0.0%)		
Smoking – yes/ex	49 (10.9%)	31 (19.6%)	0.008*	Chi-square
Smoking – no	400 (89.1%)	127 (80.4%)		
Alcohol – yes	49 (10.9%)	31 (19.6%)	0.008*	Chi-square
Alcohol – no	400 (89.1%)	127 (80.4%)		

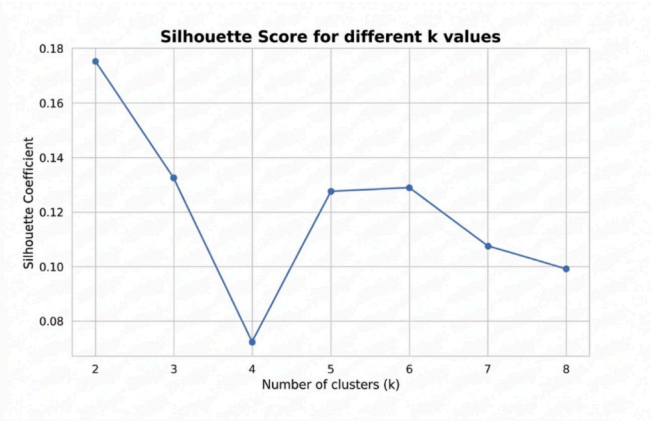


Fig. 2. Silhouette score plot for the K-Medoids clustering.

Finally, K-modes was applied on the clustering results from the two previous algorithms, again using the Silhouette coefficient to select the proper number of clusters, resulting in 3 final clusters.

Table 2 reports the results in terms of cluster dimensions and average age.

A one-way ANOVA combined with a multiple comparison test was applied on the age variable to find whether there were any differences among the three clusters, enlightening a statistically significant difference between the first two clusters ($P < 0.001$).

Fisher's exact test identified PCA positivity, autoimmune comorbidities, sex, and family history as the variables most strongly associated with cluster membership (all $p < 0.001$), whereas complications, H. pylori infection, and gastrointestinal symptoms showed weaker or non-significant associations (Table 3).

The main features of the clusters can be summed up as follows:

- Cluster 1: younger patients, predominantly female, high PCA positivity and high prevalence of autoimmune comorbidities.
- Cluster 2: older patients, higher proportion of males, low prevalence of autoimmune comorbidities.
- Cluster 3: small subgroup characterised by low PCA prevalence and higher family history rates.

The p-value of the test related to the presence of gastrointestinal symptoms was also significant, however the multiple comparisons within groups, applying the correction of Bonferroni, did not return any significant results.

Statistical tests (Chi-squared) were then performed on external variables not used for clustering, in order to check whether the new clusters were associated with them as well. These variables are: smoking, alcohol consumption, anaemia, gastric carcinoma, NET. All p-values were not significant, however it should be noted that some variables, such as NET and carcinoma, had an extremely low number of positive cases.

3.2. NLP performance

The IT5 model was applied to 425 textual clinical notes randomly extracted from the 693 available in the dataset. 297 of these entries were used as training set, 64 as validation set and 64 as test set, following the 70:15:15 proportion. Results of the model performances on the test set are reported in Table 4, and class-wise metrics are reported in Table S3 of the Supplementary Materials. From these results it is possible to notice that the model has generally high performances, especially in terms of accuracy and recall.

Fig. 4 shows the confusion matrices for the model's predictions of the various categories, where it can be noticed that the model predicts most of the categories correctly, even in the presence of a strong data

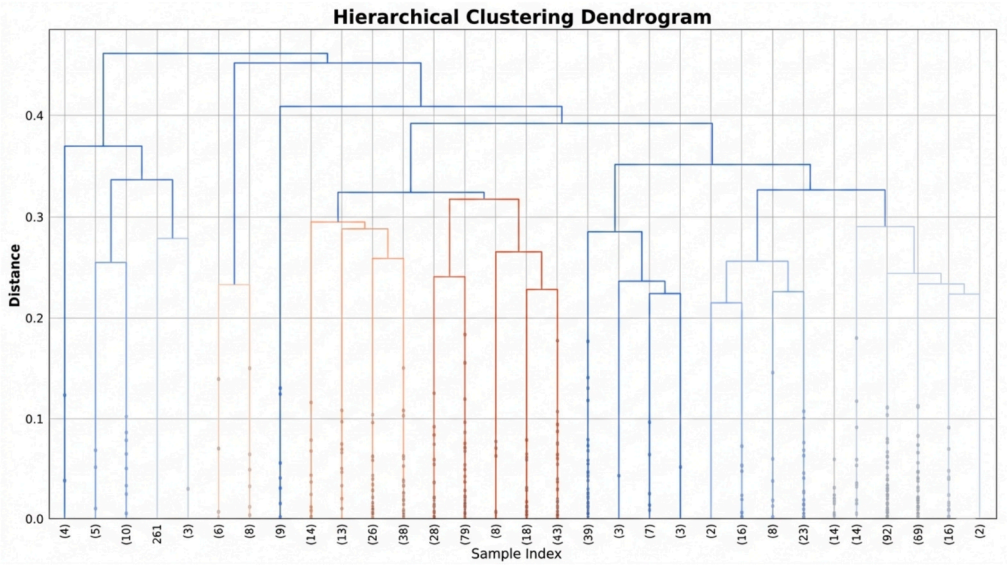


Fig. 3. Dendrogram resulting from the divisive hierarchical clustering.

Table 2
Dimension and average age of the three clusters found by the consensus clustering procedure.

Cluster	Numerosity	Average age
Cluster 1	342	53.6
Cluster 2	228	58.7
Cluster 3	37	57.7

Table 3
Fisher's exact test results on the variables influencing the clusters.

Variable	Clus.1 (yes/no)	Clus.2 (yes/no)	Clus.3 (yes/no)	P-Value	Notes
APCA	278/64	180/48	5/32	≪**0.001*****	
Autoimmunity	314/28	5/223	9/28	≪**0.001*****	
Female Sex	289/53	141/87	19/18	≪**0.001*****	*Not significant clusters 2–3*
Familiarity	59/283	24/204	14/23	**0.00005**	*Not significant clusters 1–2*
Complications	66/276	47/181	4/33	0.37	
H.Pylori	82/260	45/183	6/31	0.33	
GI Symptoms	153/189	76/152	14/23	**0.0236**	*Not significant pairwise comparisons*

Table 4
Results of IT5's performances.

Strict Accuracy	**Format Accuracy**	**F1**	**Precision**	**Recall**	**Balanced Accuracy**
90.2 (95% CI: 82.9%–97.5%)	100	78.45	71.15	87.4	89.3

imbalance for several of them, such as polyps, hyperaemia, atrophy and antrum hypotrophy. With a more detailed error analysis, however, it was possible to notice that most of the errors made by the model were due to the occasional use of infrequent synonyms or conditions related to the categories (e.g., "inflammation" or "hyperplasia" were used in a couple of notes to indicate the presence of hyperaemia), a few were due to mistakes made by the clinicians who made the reference annotations, and some were due to ambiguous notes (e.g. "some polyps" without specifying whether they were numerous or isolated).

3.3. Histological and endoscopic progression scores

Fig. 5 shows the distribution of the atrophy scores for the first two clusters. All patients in Cluster 3 had a score equal to zero, denoting the absence of progressive atrophy despite the presence of other markers, suggesting a phenotype of potential AIG.

A Kruskal-Wallis test was performed on the three distributions, obtaining a p-value equal to 0.025. However, pairwise comparisons applying the Bonferroni correction did not return any significant results.

Concerning the endoscopic scores, among 359 patients with available endoscopy follow-up, only 4 patients belonged to Cluster 3; therefore this cluster was removed from the subsequent analyses.

Concerning Clusters 1 and 2, a series of Mann-Whitney tests were performed, with results visible in Table 5. All were higher in Cluster 2, suggesting earlier or more rapid manifestation of these features.

A cumulative endoscopic damage score was also computed for each patient, summing the number of values equal to 1 by the last endoscopy, leading to a distribution of values ranging from zero to ten. Mann-Whitney test results showed that there is a significant difference also in the distribution of these values across the two clusters, with a p-value of 0.0007. Also in this case, the highest ranks belong to Cluster 2. Cox Regression and Kaplan-Meier analysis (Table S2 and Figs. S1 and S2 in the Supplementary Material) also show a significant higher risk of progression for Fundus Atrophy, diffuse Hypotrophy and Polyps in Cluster 2, confirming the possible identification of a patient group characterized by a faster disease progression.

4. Discussion

The increasing availability of electronic health records (EHRs) offers new opportunities for data-driven phenotyping of complex diseases. However, real-world clinical datasets are characterized by mixed structured and unstructured data [12,13], heterogeneous documentation practices, and irregular follow-up intervals [14], which complicate subgroup identification and longitudinal modeling. In this study, we implemented a multimodal framework integrating structured clinical variables, longitudinal endoscopic–histological data, and transformer-based natural language processing (NLP) to identify disease subphenotypes and characterize their progression trajectories in autoimmune gastritis (AIG).

True label / Predicted label		ANTRUM ATROPHY		CORPUS ATROPHY		FUNDUS ATROPHY		ANTRUM HYPOTROPHY		CORPUS HYPOTROPHY		FUNDUS HYPOTROPHY		ANTRUM HYPEREMIA		CORPUS HYPEREMIA		FUNDUS HYPEREMIA		POLIPS			
		no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	isolated	multiple	no	
Test set (64 records)	no	55	3	53	3	54	1	45	9	41	7	40	5	35	3	44	6	47	6	3	1	2	isolated
	yes	0	6	1	7	0	9	1	9	1	15	1	18	2	24	5	9	2	9	1	2	1	multiple
																				2	0	52	no

Fig. 4. IT5 Confusion Matrices for each one of the tested variables. The rows represent the true labels, while the columns represent the labels predicted by the IT5 model.

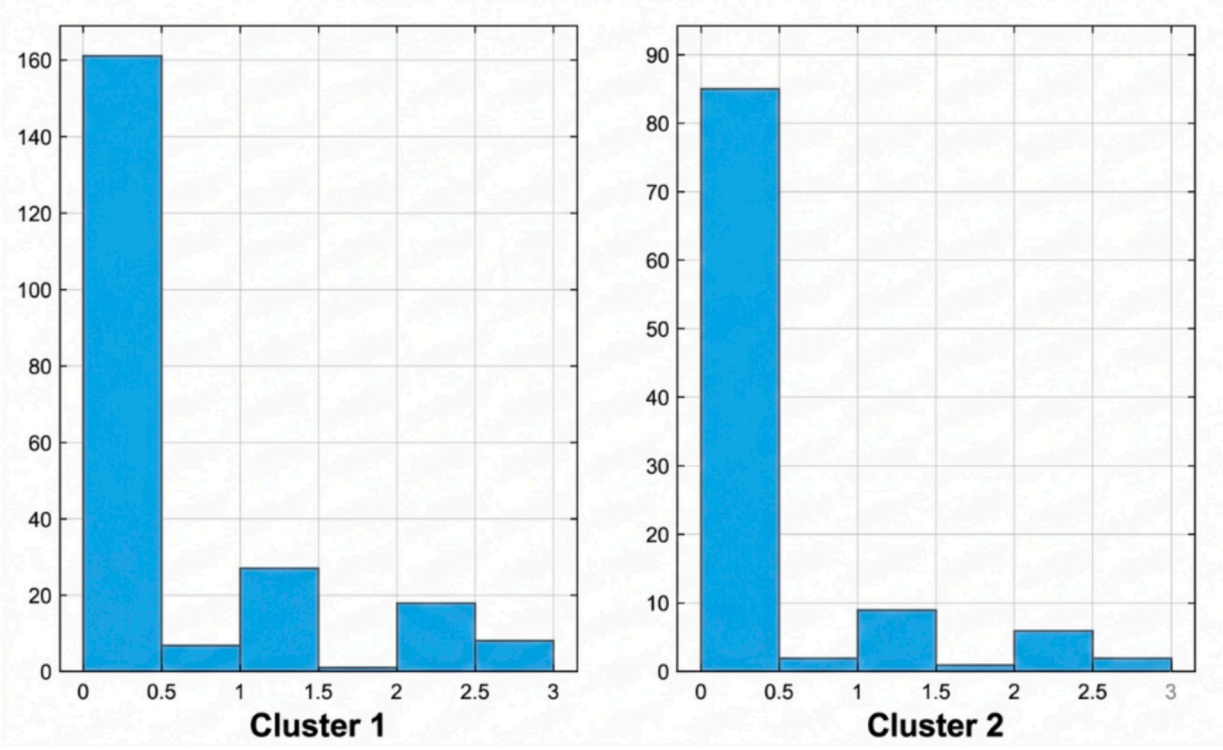


Fig. 5. Distribution of the scores for Cluster 1 and Cluster 2.

Table 5
P-Values of the Mann-Whitney test on the endoscopic scores in Cluster 1 and 2.

Variable	p-value
Atrophy antrum	0.4325
Atrophy corpus	0.0966
Atrophy fundus	**0.0450**
Hypotrophy antrum	**0.0198**
Hypotrophy corpus	0.5097
Hypotrophy fundus	0.5781
Hyperemia antrum	**0.0028**
Hyperemia corpus	0.8063
Hyperemia fundus	0.2378
Polyps	0.3288

Using a large multicenter cohort, consensus clustering identified three phenotypes, partly consistent with previous observations [8]. A predominantly autoimmune cluster with female predominance and high PCA positivity contrasted with an older, male-predominant subgroup with fewer autoimmune comorbidities and more rapid progression of endoscopic and histological damage. A smaller subgroup exhibited minimal progression, suggesting a potential early or incomplete phenotype [15]. These findings align with prior reports describing heterogeneity in AIG presentation and evolution [7,9,16], but provide a structured framework linking baseline characteristics with longitudinal patterns.

The identification of a subgroup with fewer autoimmune features but faster progression raises possible explanations, including differences in immune regulation, environmental exposures, or diagnostic timing. Previous studies have reported variability in disease course and neoplastic risk without a clear phenotypic framework [7,9,16]. Our results suggest that data-driven clustering may help formalize this heterogeneity and support risk-adapted monitoring strategies, in line with current efforts toward personalized management of chronic gastrointestinal diseases [17–22].

Our findings suggest that differences in underlying pathobiology, — potentially reflecting immune response intensity, environmental modifiers, or diagnostic timing, — may translate into distinct trajectories of mucosal damage. Recognizing these trajectories could help refine endoscopic surveillance intervals, earlier monitoring of micronutrient deficiencies, and careful assessment of neoplastic risk. Currently, clinical management of AIG relies on monitoring complications rather than anticipating disease evolution, therefore these results may lead to the definition of more personalized treatment strategies.

Beyond the clinical findings, the primary contribution of this study lies in the proposed informatics framework. First, the integration of

partitioning and hierarchical approaches within a consensus clustering strategy enhances subgroup robustness and mitigates algorithm-dependent bias. Second, transformer-based NLP enabled automated extraction of structured features from routine narrative endoscopic reports, achieving high balanced accuracy despite variability in clinical language. This demonstrates the feasibility of leveraging unstructured documentation to enhance phenotypic resolution without requiring changes to clinical workflows. Third, the introduction of time-normalized progression indices provides a pragmatic approach for modeling disease trajectories in the presence of irregular follow-up intervals, a common limitation in retrospective EHR-based research. Together, these components illustrate how combining structured and unstructured data sources can generate insights not accessible through single-modality analyses.

Several limitations should be acknowledged. The retrospective design and extended inclusion period may have introduced heterogeneity in documentation and diagnostic practices. A substantial proportion of patients were excluded due to missing data, limiting the power of the results and leading to the risk of selection bias. Irregular follow-up intervals precluded standard time-to-event modeling; therefore, the proposed progression indices represent a simplified summary of temporal dynamics. Cox regression could be applied also after a substantial pre-processing operation that lead to an oversimplification of the time progression of the disease through the LOCF method, which is known to have some important limitations [23], therefore more frequent follow-ups and advanced longitudinal analysis could improve the results. Although the NLP pipeline demonstrated strong performance, some misclassification—particularly for rare findings—remains possible. Even if the classification metrics show good performances overall, addressing data imbalance directly in the model architecture could improve the identification of minority classes even more [24–26]. Furthermore, clustering validation could refine the clusters even more in order to have better defined clinically significant groups. From a clinical point of view, the exclusion of rare outcomes such as NET and carcinoma is an important limitation, as prediction of these outcomes is important in AIG. Unfortunately the extremely low number of observations (2 observations with carcinoma and 13 with NET in total) made it impossible to perform a reliable analysis, signaling the necessity to gather more complete datasets for future analyses. Finally, the identified phenotypes were derived within a single national cohort and require external validation to assess reproducibility and clinical utility. Although computational cost was not a limiting factor in the present study, parallelization strategies may support future applications to larger EHR datasets.

Despite these possible improvements, this study demonstrates the feasibility of integrating consensus clustering, clinical NLP, and longitudinal normalization within a unified EHR-based phenotyping pipeline. Applied to AIG, this approach identified subgroups with distinct progression patterns and provides a scalable strategy for data-driven subgroup discovery and trajectory modeling in heterogeneous chronic diseases.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 5.2 in order to check grammar/spelling and improve the style of some limited parts of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Daniele Pala: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marco Vincenzo Lenti:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Data curation, Conceptualization. **Giovanni Santacroce:** Writing – original draft, Validation, Formal analysis. **Laura Bergomi:** Writing – original draft, Software, Resources, Methodology, Formal analysis, Data curation. **Chiara Curgu:** Resources, Formal analysis, Data curation. **Tommaso Buonocore:** Supervision, Software, Resources, Methodology. **Chiara Sirtoli:** Data curation, Formal analysis. **Enea Parimbelli:** Supervision, Methodology, Investigation. **Giordano Lanzola:** Supervision, Investigation, Funding acquisition. **Antonio Di Sabatino:** Supervision, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) – project n. PNC0000003 – AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them - CUP B53C22006700001. The Authors wish to acknowledge all the clinicians that contribute to the data collection in the nine Italian centers mentioned in the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2026.106511.

References

[1] M.V. Lenti, et al., Autoimmune gastritis, Nat. Rev. Dis. Primers 6 (1) (2020) 56, https://doi.org/10.1038/s41572-020-0187-8.
[2] M.V. Lenti, et al., Determinants of diagnostic delay in autoimmune atrophic gastritis, Aliment. Pharmacol. Ther. 50 (2) (2019) 167–175, https://doi.org/10.1111/apt.15317.
[3] E. Miceli, et al., Pregnancy-related complications in autoimmune atrophic gastritis: a monocentric experience, Dig. Liver Dis. 55 (1) (2023) 146–148, https://doi.org/10.1016/j.dld.2022.10.003.
[4] M.V. Lenti, et al., Infertility and recurrent miscarriage in a patient with autoimmune atrophic gastritis, Intern. Emerg. Med. 13 (5) (2018) 815–816, https://doi.org/10.1007/s11739-018-1823-0.
[5] E. Lahner, et al., Clinical and endoscopic-histological features of multifocal and corpus-restricted atrophic gastritis patients with non-cardia gastric cancer or dysplasia: a multicenter, cross-sectional study, Clin. Transl. Gastroenterol. 16 (8) (2025) e00862, https://doi.org/10.14309/ctg.0000000000000862.
[6] S. Massironi, et al., Occurrence and characteristics of endoscopic gastric polyps in patients with autoimmune gastritis (AGAPE study): a multicentric cross-sectional study, Dig. Liver Dis. 57 (1) (2025) 198–205, https://doi.org/10.1016/j.dld.2024.07.024.
[7] M.V. Lenti, G. Broglio, A. Di Sabatino, Unravelling the risk of developing gastric cancer in autoimmune gastritis, Gut 72 (7) (2023) 1429–1430, https://doi.org/10.1136/gutjnl-2022-328345.
[8] M.V. Lenti, et al., Distinguishing features of autoimmune gastritis depending on previous *Helicobacter pylori* infection or positivity to anti-parietal cell antibodies: results from the autoimmune gastRitis Italian network study group (ARIOSO), Am. J. Gastroenterol. 119 (12) (2024) 2408–2417, https://doi.org/10.14309/ajg.0000000000002948.
[9] E. Lahner, et al., Autoimmune gastritis: diagnosis, clinical management and natural history. A position paper by the Autoimmune gastRitis Italian netwOrk Study grOup (ARIOSO), Dig. Liver Dis. 58 (1) (2026) 38–50, https://doi.org/10.1016/j.dld.2025.10.015.
[10] M.V. Lenti, E. Miceli, M.C. Camargo, A. Di Sabatino, Letter to the editor, Am. J. Gastroenterol. 120 (12) (2025) 2964, https://doi.org/10.14309/ajg.0000000000003522.

[11] G. Sarti, M. Nissim, "IT5: Text-to-text Pretraining for Italian Language Understanding and Generation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 9422–9433. Accessed: Mar. 02, 2026. [Online]. Available: https://aclanthology.org/2024.lrec-main.823/.

[12] D. Pala, B. Lee, X. Ning, D. Kim, L. Shen, "Mediation Analysis and Mixed-Effects Models for the Identification of Stage-specific Imaging Genetics Patterns in Alzheimer's Disease," in: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2022, pp. 2667–2673. doi: 10.1109/BIBM55620.2022.9995405.

[13] L. Bergomi, E. Parimbelli, D. Pala, T.M. Buonocore, "BAT: A Toolkit for Biomedical Text Augmentation," in: *Artificial Intelligence in Medicine*, R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, and B. Zupan, Eds., Cham: Springer Nature Switzerland, 2025, pp. 35–39. doi: 10.1007/978-3-031-95841-0_7.

[14] M. Salzler, A. Saithna, M.P. Cote, E. Matzkin, M.J. Rossi, Lost but not forgotten: how to manage follow-up loss in clinical research, Arthroscopy 41 (11) (2025) 4381–4388, https://doi.org/10.1016/j.arthro.2025.07.038.

[15] M.V. Lenti, E. Miceli, A. Vanoli, C. Klersy, G.R. Corazza, A. Di Sabatino, Time course and risk factors of evolution from potential to overt autoimmune gastritis, Dig. Liver Dis. 54 (5) (2022) 642–644, https://doi.org/10.1016/j.dld.2021.10.001.

[16] E. Miceli, et al., Long-term natural history of autoimmune gastritis: results from a prospective monocentric series, Am. J. Gastroenterol. 119 (5) (2024) 837–845, https://doi.org/10.14309/ajg.0000000000002619.

[17] M. Iacucci, G. Santacroce, M. Yasuharu, S. Ghosh, Artificial intelligence-driven personalized medicine: transforming clinical practice in inflammatory bowel disease, Gastroenterology 169 (3) (2025) 416–431, https://doi.org/10.1053/j.gastro.2025.03.005.

[18] M.V. Lenti, et al., Personalize, participate, predict, and prevent: 4Ps in inflammatory bowel disease, Front. Med. 10 (2023), https://doi.org/10.3389/fmed.2023.1031998.

[19] J.L. Jameson, D.L. Longo, Precision medicine–personalized, problematic, and promising, N. Engl. J. Med. 372 (23) (2015) 2229–2234, https://doi.org/10.1056/NEJMsb1503104.

[20] M.V. Lenti, et al., Latent class analysis identifies novel coeliac disease subgroups with distinctive clinical features: a multicentric study, Eur. J. Intern. Med. 139 (2025) 106355, https://doi.org/10.1016/j.ejim.2025.05.020.

[21] S.C. Shah, M.B. Piazuelo, E.J. Kuipers, D. Li, AGA clinical practice update on the diagnosis and management of atrophic gastritis: expert review, Gastroenterology 161 (4) (2021) 1325–1332.e7, https://doi.org/10.1053/j.gastro.2021.06.078.

[22] M. Dinis-Ribeiro, et al., Management of epithelial precancerous conditions and early neoplasia of the stomach (MAPS III): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG) and European Society of Pathology (ESP) Guideline update 2025, Endoscopy 57 (5) (2025) 504–554, https://doi.org/10.1055/a-2529-5025.

[23] D. Mavridis, G. Salanti, T.A. Furukawa, A. Cipriani, A. Chaimani, I.R. White, Allowing for uncertainty due to missing and LOCF imputed outcomes in meta-analysis, Stat. Med. 38 (5) (2019) 720–737, https://doi.org/10.1002/sim.8009.

[24] I. A. C. M *et al.*, "Development and validation of a class imbalance-resilient cardiac arrest prediction framework incorporating multiscale aggregation, ICA and explainability," IEEE Trans. Biomed. Eng. 72(5) 2025 1674–1687, doi: 10.1109/TBME.2024.3517635.

[25] M.Y. Ansari, et al., A survey of transformers and large language models for ECG diagnosis: advances, challenges, and future directions, Artif. Intell. Rev. 58 (9) (2025) 261, https://doi.org/10.1007/s10462-025-11259-x.

[26] V. Chandrasekar, et al., Integrated approaches for immunotoxicity risk assessment: challenges and future directions, Discov. Toxicol. 1 (1) (2024) 9, https://doi.org/10.1007/s44339-024-00010-w.