Comparative Evaluation of Sequence Encoding Strategies for Clustering Inhaled Therapy Pathways in COPD Patients Using Real-World Data

MSR56
ISPOR Europe 2025

Authors: Romane Péan, MSc[1], **Marie Génin, MSc[1]**, Nina Temam, PharmD[1], Diane Vincent, MSc[1], Rachel Nadif, PhD[2], Sofiane Kab, PharmD, PhD[2], Nicolas Roche, MD, PhD[3], Pauline Guilmin, MSc[1]
Affiliation: [1]Quinten Health, Paris, France ; [2]INSERM, Villejuif, France ; [3]Assistance Publique–Hôpitaux de Paris (AP-HP), Paris, France

INTRODUCTION

- Identifying and comparing treatment pathways is essential for informed healthcare decision-making. Machine learning techniques can uncover insights from real-world therapeutic trajectories, offering new perspectives on patient management.
- However, these trajectories are highly variable and complex. Most machine learning algorithms, such as clustering, cannot inherently capture the sequential or temporal structure of treatment data.
- To overcome this limitation, treatment sequences must be transformed into numerical representations through a process known as encoding. Yet, despite the growing adoption of such methods, there is no consensus on the most appropriate encoding strategy.
- This study compares three sequence encoding approaches applied to real-world therapeutic sequences in COPD, aiming to evaluate their capacity to produce clinically meaningful patient clusters.

OBJECTIVE

To evaluate and compare three sequence encoding strategies for clustering COPD therapeutic sequences and to assess the clinical interpretability of the resulting clusters:

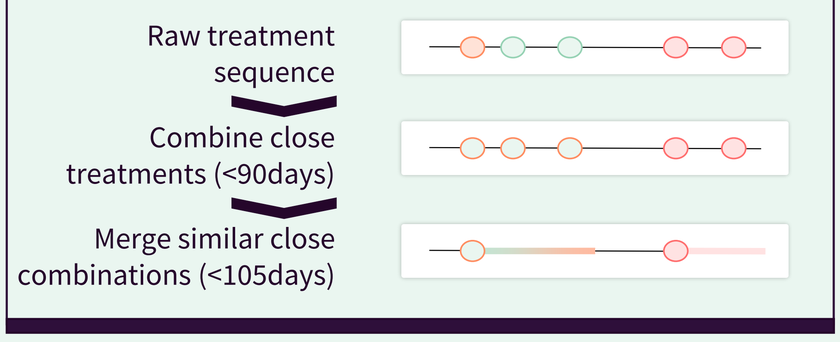
- **SeqMining**
- **SeqToChar**
- **Autoencoder**

METHODS

Data were obtained from the French CONSTANCES[1] cohort, linked to the national health claims database (SNDS). Participants were identified as COPD patients based on spirometry results or validated questionnaires. Five-year sequences of inhaled maintenance therapies (mono-, bi- or triple therapy)[2], mapped using ATC7 codes, were extracted.

Prescription data were processed into standardised treatment sequences by applying temporal rules that accounted for overlaps and treatment durations (see Figure 1).

Figure 1: Sequence cleaning and creation pipeline



Three encoding approaches were applied:

- **SeqMining**: Frequent subsequences were extracted using the SPADE[3] algorithm, and binary vectors were generated to indicate the presence or absence of specific patterns.
- **SeqToChar**: Sequences were represented as character strings, and pairwise similarities were calculated using the Jaro[4] distance.
- **Autoencoder**: A deep learning model was trained to learn continuous, abstract representations of sequences in a reduced latent space[5], taking into account treatment durations (see Figure 2 for more details).

- Each encoding strategy was subsequently fed into a k-means clustering algorithm.
- The optimal number of clusters was determined through silhouette score analysis.
- Cluster validity was further evaluated using both silhouette scores[6] and UMAP[7] projections.
- The best-performing method was explored in greater detail through trajectory visualizations[8] to assess its clinical interpretability.

Figure 2: Examples of numerical representation of treatment sequences through different encoding methods

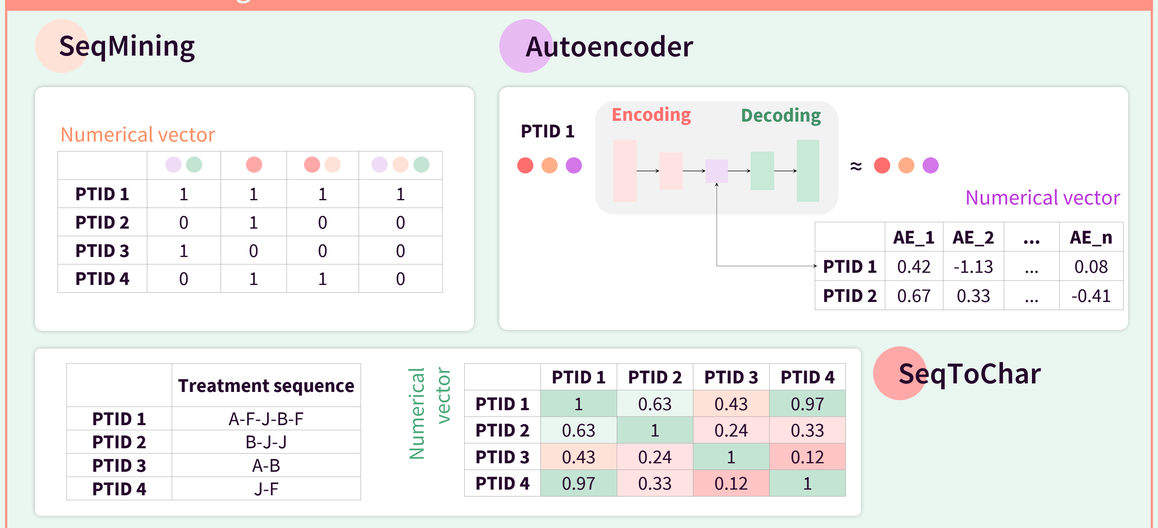
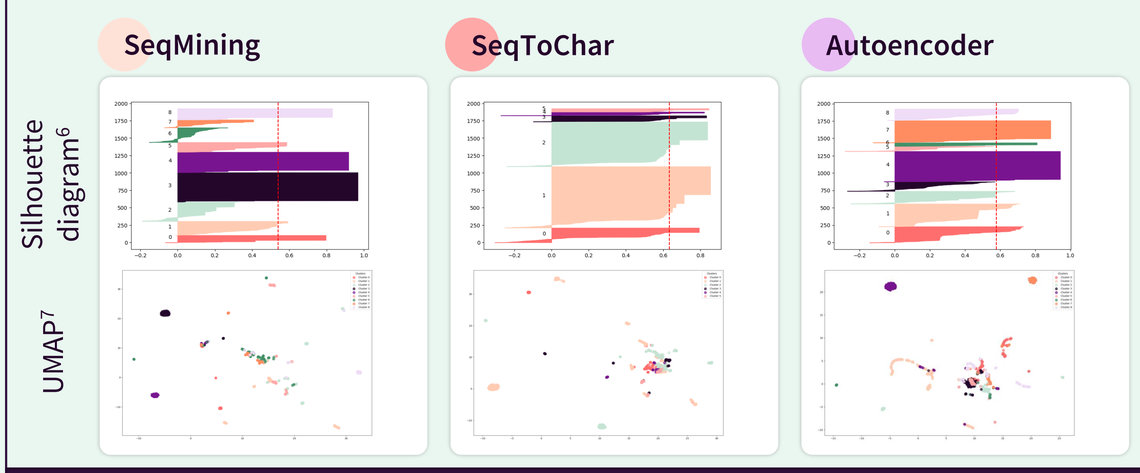


Illustration of how COPD treatment pathways are transformed into numerical vectors using SeqMining, SeqToChar and Autoencoder

RESULTS

- Among 4,982 participants identified with COPD, 1,926 met the criteria for five-year follow-up and treatment inclusion.
- On average, participants were exposed to two therapeutic combinations, and 90% received inhaled corticosteroids.
- When evaluating the homogeneity of clusters using silhouette diagrams[6] and UMAP[7] projections, we observed that all three encoding methods produced relatively coherent and homogeneous groupings. SeqToChar achieved the highest silhouette score[9] (0.63), compared with SeqMining (0.54) and the Autoencoder (0.58) (see Figure 3).

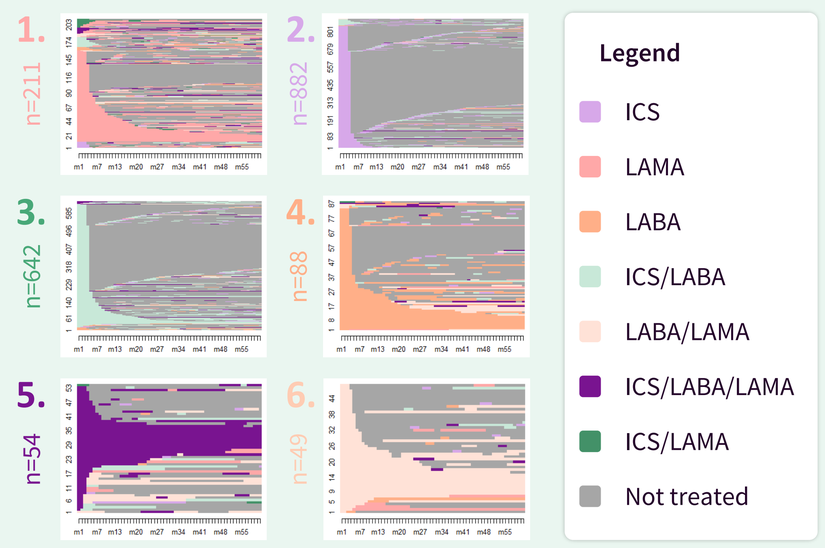
Figure 3: Homogeneity of clusters for encoding methods



Silhouette diagram and UMAP graph for the three encoding methods. SeqToChar demonstrated the highest silhouette score (see red line on silhouette diagram).

Clinical examination of the clusters obtained with SeqToChar method highlighted heterogeneous treatment trajectories, with numerous early treatment discontinuations without therapeutic replacement, reflecting the well-known issue of poor adherence in COPD. Most clusters were primarily grouped according to the initial treatment received (see Figure 4).

Figure 4: Visual representation of SeqToChar clusters using TraMineR[10]



Sequence index plots displaying the treatment trajectories within each cluster.

- Cluster **2**: initiated on ICS monotherapy (not recommended in isolation) and appeared to represent occasional use.
- Clusters **5** and **6**: started with either triple therapy (ICS/LABA/LAMA) or dual therapy (LABA/LAMA) and showed the most durable treatment trajectories over time.
- Cluster **3**: initiated mainly with ICS/LABA, demonstrated lower persistence but partial transitions to other treatment categories.
- Clusters **1** and **4**: started with LABA or LAMA monotherapies, displayed low long-term continuity but frequent transitions toward other therapeutic categories.

CONCLUSION

- **The three encoding methods all produced satisfying results, with SeqToChar achieving the highest silhouette score and allowing the grouping of COPD treatment pathways into clinically interpretable clusters.**
- **Clustering appeared to be driven primarily by the first treatment in the sequence, likely due to the string-based distance computation and the short length of the sequences.**
- **These findings suggest that relying exclusively on technical performance metrics may not be sufficient to guide the choice of an encoding strategy.**
- **Introducing predefined clinical relevance criteria would provide a more balanced assessment of clustering quality, combining methodological rigor with real-world interpretability.**
- **Such an approach could offer a more comprehensive basis for evaluating patient trajectories in health technology assessment (HTA) contexts.**

REFERENCES:
1. Zins, M. et al. The French CONSTANCES Population-Based Cohort: Design,Inclusion and Follow-Up. https://doi.org/10.1007/s10654-015-0096-4
2. Mirza, S. et al. COPD Guidelines: A Review of the 2018 GOLD Report. https://doi.org/10.1016/j.mayocp.2018.05.026
3. Zaki, M. J. SPADE: An Efficient Algorithm for Mining Frequent Sequences.
4. Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. https://doi.org/10.1080/01621459.1989.10478785
5. Baytas, I. M. et al. Patient Subtyping via Time-Aware LSTM Networks. https://doi.org/10.1145/3097983.3097997.
6. Rousseeuw, P.J. et al. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. 20. 53-65. 10.1016/0377-0427(87)90125-7
7. McInnes, L. et al. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426
8. Guo, Y. et al. Survey on Visual Analysis of Event Sequence Data. http://arxiv.org/abs/2006.14291
9. Kaufman & Rousseeuw (1990), Finding Groups in Data, pp. 79–81
10. TraMineR : https://traminer.unige.ch/doc/seqdef.html

ABBREVIATIONS:
COPD: Chronic Obstructive Pulmonary Disease; SNDS: Système National des Données de Santé; ATC: Anatomical Therapeutic Chemical Classification; UMAP: Uniform Manifold Approximation and Projection; ICS: Inhaled corticosteroids; LABA: Long-acting beta-agonist; LAMA: Long-acting muscarinic antagonists; HTA: Health technology assessment

CONFLICT OF INTEREST: NA

CONTACT INFO: m.genin@quinten-health.com

quinten health