Streaming k-Medoids for Fair and Scalable Patient Phenotyping under Memory Constraints

Syed Ahmar Shah*
University of Edinburgh
Edinburgh, UK, EH16 4UX
ahmar.shah@ed.ac.uk

Fatima Almagharbi
University of Edinburgh
Edinburgh, UK, EH16 4UX
falmaghr@ed.ac.uk

Aziz Sheikh
University of Oxford
Oxford, UK, OX2 6GG
aziz.sheikh@phc.ox.ac.uk

Abstract

Clustering offers a powerful route to identify disease phenotypes, but applying distance-based methods at population scale remains challenging. Standard k-medoids with Gower distance, a natural choice for mixed-type clinical data, has quadratic time and memory complexity that renders it infeasible for modern electronic health record (EHR) datasets with hundreds of thousands of patients. We address this barrier with a streaming+coreset k-medoids framework that scales linearly in runtime and uses bounded memory, enabling clustering under modest hardware limits. Our approach combines chunk-wise distance computation, Hungarian alignment of medoids across chunks, and a coreset-based refinement, with optional feature weighting to incorporate domain knowledge. Experiments on a synthetic 200,000-patient asthma dataset informed by literature show that the method (i) matches the accuracy of full-distance clustering, (ii) scales to population-level datasets under 10 GB RAM, and (iii) recovers minority-dominated phenotypes when ethnicity is appropriately weighted. This work demonstrates a practical and broadly applicable framework for large-scale, mixed-type healthcare clustering, motivated by the needs of precision medicine.

1 Introduction

The promise of precision medicine rests on tailoring care to subgroups, moving beyond a one-size-fits-all model [1]. Chronic conditions account for most of the global healthcare burden, yet many remain umbrella terms for heterogeneous phenotypes [2]. Identifying these subtypes is critical, as patients differ in risk, progression, and treatment response [3]. Clustering offers a data-driven route to such phenotyping, and routinely collected electronic health records (EHRs) now provide unprecedented opportunities at population scale [4]. Analysing EHRs, however, poses unique challenges. Datasets may include hundreds of thousands of patients with mixed data types (numeric, binary, categorical), and are often analysed in privacy-preserving secure environments with strict hardware limits. Standard distance-based clustering methods, with quadratic time and memory requirements, are infeasible in such settings [5, 6].

Asthma provides a motivating exemplar. Affecting over 260 million people worldwide, it is the most common chronic respiratory disease with substantial heterogeneity [7]. Previous cluster analyses have identified eosinophilic, neutrophilic, and obesity-related phenotypes [8–10], but these were

*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Muslims in ML.

based on carefully selected cohorts of limited size. Ethnic minority patients remain under-represented, despite evidence of differential risks and treatment responses [7]. A concurrent national-scale analysis of over 200,000 asthma patients in the Clinical Practice Research Datalink (CPRD)—one of the largest studies of its kind to date—is examining how prescription hormone exposure relates to asthma development and manifestation, and how these associations vary across asthma phenotypes and comorbid metabolic-syndrome–related conditions [11]. In this context, applying standard clustering proved infeasible under the available hardware, motivating the present work. We propose a streaming + coreset clustering framework for large, mixed-type healthcare datasets under constrained memory. Our method extends k-medoids with weighted Gower distances [12], chunk-wise streaming [13], Hungarian alignment of medoids across chunks [14], and a coreset-based refinement step. It scales linearly with dataset size (up to a small quadratic coreset term) and allows feature weighting to incorporate domain knowledge. Using a synthetic 200,000-patient asthma dataset informed by prior literature, we demonstrate: (i) scalability under 16 GB RAM; (ii) performance comparable to full-distance clustering where feasible; and (iii) recovery of minority-dominated clusters through feature reweighting. A concise survey is deferred to Appendix A covering mixed-type clustering and asthma phenotyping.

2 Methods

2.1 Problem setup

We consider clustering a dataset of N patients with mixed feature types—numeric, binary, and categorical. Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ and let K be the target number of clusters. Standard k-medoids (PAM) optimises

$$\min_{M \subseteq \mathcal{X},\, |M|=K} \sum_{i=1}^{N} \min_{m \in M} d(x_i, m),$$

where M are the medoids and $d(\cdot,\cdot)$ is a dissimilarity. PAM (the classical k-medoids algorithm [5]) selects actual data points as representatives and works with arbitrary dissimilarities, which makes it robust to outliers and suitable for mixed-type data. However, standard PAM requires an $N \times N$ distance matrix—$\mathcal{O}(N^2)$ time/memory—motivating our streaming+coreset variant. For mixed clinical variables we use a weighted Gower dissimilarity to encode domain knowledge via feature weights (details next).

2.2 Weighted Gower distance

Each patient record x is split into numeric ($x^{(\mathrm{num})}$), binary ($x^{(\mathrm{bin})}$), and categorical ($x^{(\mathrm{cat})}$) features. For two patients x_i and x_j, the weighted Gower dissimilarity is

$$d(x_i, x_j) = \frac{1}{W}\Bigg(\sum_{u \in \mathrm{num}} w_u \frac{|x_{iu} - x_{ju}|}{r_u} + \sum_{v \in \mathrm{bin}} w_v \, \mathbf{1}[x_{iv} \neq x_{jv}] + \sum_{c \in \mathrm{cat}} w_c \, \mathbf{1}[x_{ic} \neq x_{jc}] \Bigg),$$

where $w_\bullet$ are feature weights, r_u are numeric ranges, and

$$W = \sum_{u \in \mathrm{num}} w_u \; + \sum_{v \in \mathrm{bin}} w_v \; + \sum_{c \in \mathrm{cat}} w_c.$$

This formulation enables domain experts to re-weight features such as ethnicity, ensuring minority subgroups are not overlooked.

2.3 Streaming + Coreset framework

To formalise our approach, we outline the procedure in Algorithm 1, which summarises the streaming k-medoids framework with weighted Gower distance. The pseudocode highlights how chunking, medoid alignment, and coreset refinement enable efficient clustering of large mixed-type datasets. All code, configuration files, and scripts for dataset generation, clustering, and evaluation are openly available at <https://github.com/syedahmar/Memory-Constrained-Clustering>. Default settings and software environment details are documented in Appendix B, which also includes scripts to reproduce all figures reported in this paper.

This algorithm reduces memory usage from quadratic in N to bounded by the coreset size M, while runtime grows linearly with N, all while preserving fidelity to full k-medoids clustering.

Algorithm 1 Streaming k-Medoids with Weighted Gower Distance

Require: Dataset $\mathcal{X}$ of N points; target clusters K; chunk size C; coreset size $M \ll N$
Ensure: Refined global medoids M_g and labels for all N points
for each chunk of size C in $\mathcal{X}$ **do**
    Compute weighted Gower distances within the chunk
    Run local k-medoids $\rightarrow$ local medoids
end for
Align local medoids to global medoids using greedy/Hungarian matching
for each cluster **do**
    Retain candidate pool (closest points + summaries)
    Maintain hard-point reservoir for outliers
end for
Construct coreset S of size M from candidate pools + reservoirs
Run k-medoids on $S \rightarrow$ refined global medoids M_g
Final assignment: Assign all N points to nearest medoid in $O(NK)$ time with $O(N)$ memory by streaming distances one medoid at a time

2.4 Synthetic dataset

To evaluate scalability and accuracy under controlled conditions, we generated a literature-informed synthetic dataset of 200,000 patients with asthma, with proportional subsamples from 10k–200k in 10k increments. Ten phenotypes were defined, reflecting patterns reported in prior asthma clustering studies (eosinophilic, neutrophilic, obesity-related, etc.), with two minority-dominated clusters to test sensitivity to under-represented groups. Features included demographics (age, Body Mass Index (BMI)), binary indicators (e.g., smoking, allergy, inflammatory markers, metabolic-syndrome drugs), and a categorical variable for ethnicity. Ground-truth cluster labels were retained for evaluation. See Appendix C for further details.

2.5 Complexity analysis

With chunk size C, coreset size M, and K clusters, our algorithm runs in

$$T(N) = \mathcal{O}(NC + M^2 + NK), \qquad \text{Memory} = \mathcal{O}(C^2 + M^2 + N).$$

Here, $\mathcal{O}(NC)$ arises from per-chunk distance builds, $\mathcal{O}(M^2)$ from coreset refinement, and $\mathcal{O}(NK)$ from final assignment. Memory is dominated by the $\mathcal{O}(C^2 + M^2)$ distance buffers, while the $\mathcal{O}(N)$ label array is small. Unlike full k-medoids ($\mathcal{O}(N^2)$ time/memory), runtime scales linearly and memory remains bounded for fixed C and M (see Appendix D for the full derivation).

3 Results

Figure C1 summarises ten literature-informed asthma phenotypes (feature probabilities, age/BMI ranges, ethnicity), including two minority-dominated phenotypes used as a fairness probe. In Figure 1, streaming+coreset matches full PAM on small N and remains stable up to $N = 200{,}000$ based on various performance metrics (ARI, NMI, Silhouette, Purity; definitions in Appendix E); runtime grows linearly while standard Gower PAM becomes infeasible beyond $\sim$20k (middle); peak memory stays < 10 GB at $N = 200{,}000$, whereas full-distance would require > 160 GB (lower), consistent with our complexity analysis. Figure 2 shows that increasing the ethnicity weight recovers minority-dominated clusters (8: Black/mixed; 9: Asian) that are otherwise absorbed under uniform weights, improving subgroup recoverability without degrading overall quality. Extended bootstrap analyses with confidence intervals for the weighting modes are provided in Appendix F.

4 Discussion

We introduced a streaming+coreset k-medoids framework for clustering large, mixed-type healthcare datasets under memory constraints. Identifying distinct disease phenotypes is a key step toward precision medicine: patients with the same diagnosis can differ markedly in risk, progression, and

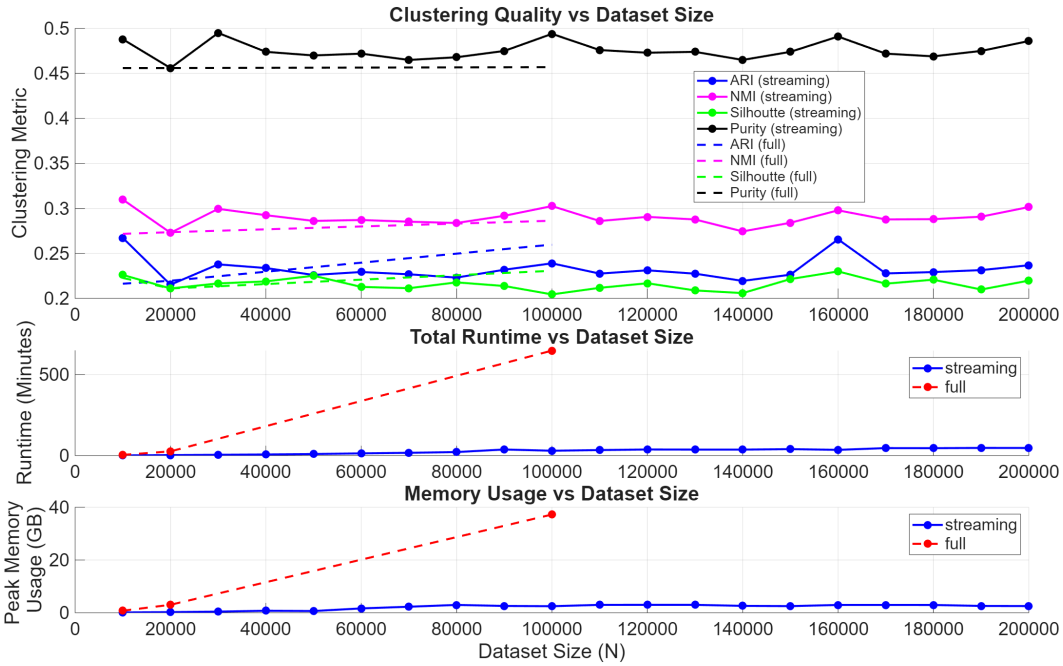

Figure 1: Scalability and performance of streaming + coreset k-medoids compared with full PAM. (Upper) ARI/NMI/silhouette: streaming matches full PAM at small N and remains stable to $N = 200{,}000$. (Middle) Runtime grows linearly; full PAM becomes infeasible beyond ∼20k. (Lower) Peak memory fits in <10 GB at $N = 200{,}000$; full-distance would need > 160 GB.

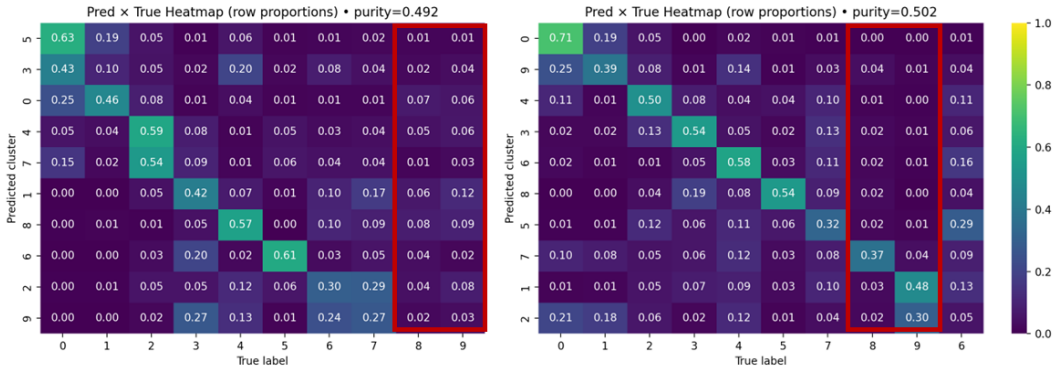

Figure 2: Feature weighting effect (100k). Uniform weights merge minority phenotypes; increasing ethnicity weight recovers clusters 8 (Black/mixed) and 9 (Asian) as distinct patterns.

treatment response. In asthma, for example, discovering clinically meaningful subtypes from national-scale electronic health records could enable more targeted prevention and therapy, but has been infeasible with existing distance-based clustering methods due to their quadratic time and memory requirements. Motivated by this medical challenge, our method overcomes the scalability barriers of standard PAM, achieving linear runtime scaling and bounded memory usage. In practice, clustering 200,000 patients stayed below 10 GB peak RAM in our runs where full-distance clustering would require more than 160 GB. While the final assignment maintains an $\mathcal{O}(N)$ label array, its absolute size is tiny compared to the bounded ($\mathcal{O}(C^2 + M^2)$) distance buffers.

Beyond scalability, the framework allows feature weighting to integrate domain knowledge. This was critical for recovering minority-dominated phenotypes in our synthetic asthma dataset: under uniform weights, these groups were absorbed into majority clusters, but re-weighting ethnicity enabled their detection. Such flexibility highlights the role of algorithm design not only in efficiency but also in promoting fairness and inclusivity in healthcare analytics.

Although we demonstrated the approach in asthma, the framework is broadly applicable to other chronic diseases and routinely collected EHRs, particularly in contexts where compute resources are

constrained. Our framework makes population-scale, mixed-type clustering feasible under realistic hardware limits, while allowing domain knowledge to guide discovery of clinically meaningful — and often overlooked — patient subgroups (see Appendix G for contributions, limitations, and future work).

5 Acknowledgements

This work was supported by Asthma+Lung UK, grant number: WAPG22\00019

References

[1] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[2] Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258): 1204–1222, 2020.

[3] J Larry Jameson and Dan L Longo. Precision medicine—personalized, problematic, and promising. *Obstetrical & gynecological survey*, 70(10):612–614, 2015.

[4] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[5] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[6] Olivier Bachem, Mario Lucić, and Andreas Krause. Scalable k-means clustering via lightweight coresets. *Journal of Machine Learning Research*, 20(104):1–49, 2018.

[7] Syed Ahmar Shah, Aziz Sheikh, and Colin Simpson. Understanding asthma, a global epidemiological perspective. *Clinical Asthma: Theory and Practice*, pages 1–13, 2025.

[8] Sally E. Wenzel. Asthma phenotypes: The evolution from clinical to molecular approaches. *Nature Medicine*, 18(5):716–725, 2012.

[9] Pranabashis Haldar, Ian D. Pavord, and et al. Cluster analysis and clinical asthma phenotypes. *American Journal of Respiratory and Critical Care Medicine*, 178(3):218–224, 2008.

[10] W. C. Moore, D. A. Meyers, S. E. Wenzel, and et al. Identification of asthma phenotypes using cluster analysis in the sarp cohort. *American Journal of Respiratory and Critical Care Medicine*, 181(4):315–323, 2010.

[11] Fatima Almaghrabi, Bright I Nwaru, Aziz Sheikh, Athanasios Tsanas, Holly Tibble, Hilary Critchley, Tracy Jackson, Azhar Ali, and Syed Ahmar Shah. Exogenous sex steroid hormones and asthma phenotypes: a study protocol for a prospective cohort analysis with uk-wide primary care data. *BMJ open*, 15(3):e097126, 2025.

[12] John C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4): 857–871, 1971.

[13] Sudipto Guha, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan. Clustering data streams. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 359–366, 2003.

[14] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[15] Erick Forno and Juan C. Celedón. Health disparities in asthma. *American Journal of Respiratory and Critical Care Medicine*, 185(10):1033–1045, 2012.

[16] Precision medicine needs an equity agenda. *Nature Medicine*, 27(5):737, 2021. doi: 10.1038/s41591-021-01373-y. URL https://doi.org/10.1038/s41591-021-01373-y.

[17] Leonard Kaufman and Peter J. Rousseeuw. Clustering by means of medoids. In *Statistical Data Analysis Based on the L1-Norm and Related Methods*, 1987. Original PAM write-up; details vary by proceedings/tech report.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2009.

[19] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 52nd ACM Symposium on Theory of Computing (STOC)*, 2020.

[20] Zhexue Huang. Extensions to the k-means algorithm for clustering large datasets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[21] Dharmendra S. Modha and William S. Spangler. Feature weighting in k-means clustering. *Machine Learning*, 52(3):217–237, 2003.

A Appendix A: Extended Related Work

A.1 Asthma Phenotyping via Clustering

Asthma is a heterogeneous condition encompassing a range of symptom profiles, inflammatory patterns, and treatment responses. Early work highlighted that asthma should not be treated as a single disease entity but as a collection of phenotypes with distinct underlying mechanisms [8]. Cluster analysis has been instrumental in uncovering these phenotypes.

Haldar et al. [9] used unsupervised clustering on clinical and biomarker data from patients in the UK and demonstrated the existence of distinct groups such as early-onset atopic asthma and late-onset eosinophilic asthma. Moore et al. [10] extended this work in the Severe Asthma Research Program (SARP) cohort in the US, again identifying clinically meaningful subgroups that aligned with different prognoses and treatment responses.

Despite their importance, these studies had notable limitations. They typically analysed highly selected populations (specialist clinics, severe asthma registries), with sample sizes in the hundreds or low thousands. As a result, they may not generalise to the wider asthma population captured in routine care. Furthermore, ethnic minority patients are often under-represented, even though disparities in asthma outcomes across racial and ethnic groups are well documented [15]. Recent calls in precision medicine stress the importance of studying large, diverse, real-world populations to ensure equitable advances in treatment [16].

This motivates the use of routinely collected electronic health records (EHRs) for large-scale phenotyping. EHR data can capture hundreds of thousands of patients across demographics, comorbidities, and prescribing patterns, providing a more representative basis for discovering subtypes. However, the sheer scale and complexity of EHR data (mixed numeric, binary, and categorical variables; missingness; longitudinality) pose methodological challenges that existing clustering methods were not designed to handle.

A.2 Clustering Methods for Large-Scale, Mixed-Type Data

Distance-based clustering. Partitioning Around Medoids (PAM) [17] is a natural choice for clinical datasets. Unlike k-means, which requires Euclidean space, PAM operates directly on dissimilarity matrices and is more robust to outliers. For mixed-type data, Gower's dissimilarity [12] is widely used because it can handle numeric, binary, and categorical variables simultaneously, while also permitting feature weighting. These properties have made PAM with Gower dissimilarity a common choice in healthcare phenotyping [18]. The main drawback is scalability: computing and storing the full $N \times N$ distance matrix requires $O(N^2)$ time and memory. For datasets with tens or hundreds of thousands of patients, as in EHRs, this becomes computationally infeasible.

Streaming and coreset approaches. To address scalability, researchers have explored streaming and coreset-based clustering methods. Guha et al. [13] introduced scalable k-medoids approximations in data stream settings. More recently, Bachem et al. [6] and Feldman & Langberg [19] studied coreset constructions that approximate clustering objectives by summarising subsets of the data, reducing both time and memory requirements. These approaches have shown strong theoretical and empirical performance on large numeric datasets. However, most work has focused on k-means in Euclidean space, not k-medoids with general dissimilarities, and almost none address the additional challenges of mixed-type clinical data. Moreover, healthcare-specific constraints such as operating in Trusted Research Environments (TREs) with strict memory and compute limits (often < 32 GB RAM) make these scalability concerns especially pressing.

Feature weighting and domain knowledge. Another line of work has considered incorporating domain knowledge into clustering. Weighted dissimilarities allow certain features to have greater influence, which is particularly important in healthcare where minority-defining attributes (e.g., ethnicity) may otherwise be swamped by majority features. Previous studies proposed weighting schemes for mixed data [20, 21], though many require supervision or are limited to small datasets. Our contribution builds on this tradition, showing how explicit feature weighting within a scalable clustering framework can help recover minority-dominated phenotypes that are otherwise overlooked.

A.3 Summary

In summary, while clustering has played a central role in asthma phenotyping, prior studies have been limited to small, selective cohorts. At the same time, advances in scalable clustering have focused on numeric data in machine-learning benchmarks, with little translation to mixed-type healthcare datasets. Our work bridges this gap: motivated by the clinical need for large-scale, equitable asthma phenotyping, we extend streaming and coreset methods to *weighted* Gower dissimilarities, enabling population-scale clustering under modest hardware constraints.

B Appendix B: Code and Experimental Settings

All code used in this study is openly available at: `https://github.com/syedahmar/Memory-Constrained-Clustering`.

B.1 Repository contents

The repository includes:

- **Core algorithm implementation:** `stream_kmedoids_pipeline_v5.py` — streaming+coreset k-medoids with weighted Gower dissimilarity.
- **Config files:** YAML files in `configs/` specifying dataset path, feature groups, chunk size C, coreset size M, weighting mode, and other options.
- **Synthetic dataset generator:** script to reproduce the asthma phenotypes described in Appendix C.
- **Experiment runner:** pipeline to execute runs across dataset sizes (10k–200k) and save outputs under `runs/`.
- **Evaluation scripts:** functions for ARI, NMI, silhouette, purity, and cluster-level metrics (precision/recall for minority clusters).
- **Plotting utilities:** scripts to recreate the figures shown in the main paper.

B.2 Default settings

Unless stated otherwise in the main text, experiments were run with:

- **Chunk size (C):** 2000
- **Coreset size (M):** up to 20,000
- **Target clusters (K):** 10

- **Weighting modes:** *none* (uniform), *manual* ($w_{\text{ethnicity}}$=3), *supervised* (using true labels; benchmarking only)
- **Hardware constraint:** runs limited to 32 GB RAM to reflect typical TRE/resource-constrained environments

B.3 Reproducibility notes

- **Random seeds:** fixed seeds for initialisation and sampling across all experiments.
- **Software environment:** Python 3.10; key dependencies listed in `requirements.txt` (e.g., `scikit-learn`, `pyclustering`, `numpy`, `pandas`, `tqdm`).
- **Runtime tracking:** each run logs runtime, peak memory, and clustering metrics to `runs/`, along with learned feature weights (when applicable).

B.4 Usage example

Example command:

```
python stream_kmedoids_pipeline_v5.py --config configs/baseline_uniform.yml
```

Example configuration files for the baseline (uniform weights), manual weighting, and supervised weighting are provided under `configs/`.

C Appendix C: Synthetic Dataset Generation

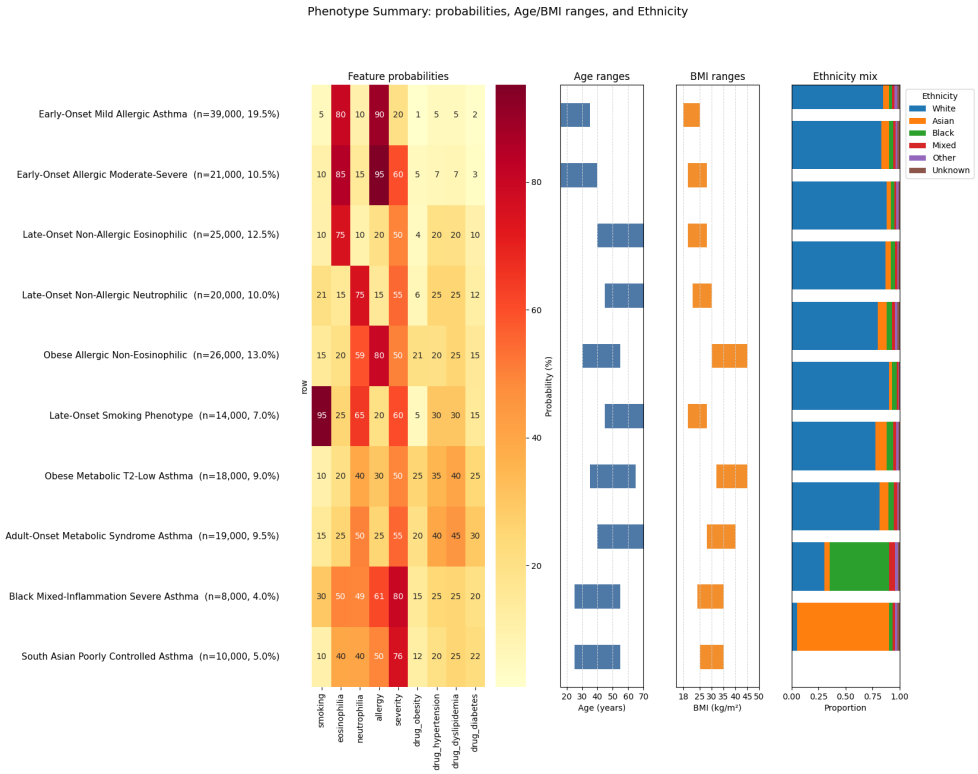

Figure C1: Synthetic asthma phenotypes. Ten clusters were defined with distributions across demographic, inflammatory, and treatment features. Age and BMI ranges are shown for each phenotype, alongside ethnicity proportions. Minority-dominated phenotypes (clusters 8–9) were included explicitly to test the ability of clustering methods to recover under-represented subgroups.

To evaluate the proposed framework at scale, we generated a synthetic dataset of 200,000 asthma patients, informed by prior literature on asthma phenotypes. The aim was to build a testbed that captures key heterogeneity seen in real-world populations while enabling controlled experiments with ground-truth cluster labels.

C.1 Phenotype design

Based on seminal cluster analyses and subsequent reviews [8–10], we defined ten distinct phenotypes spanning demographic, clinical, and treatment-related characteristics. These include early-onset allergic asthma (mild and moderate–severe), late-onset eosinophilic and neutrophilic asthma, obesity-related phenotypes, smoking-related asthma, and metabolic-syndrome–associated phenotypes. Two minority-dominated phenotypes were specified as fairness probes: cluster 8 (primarily Black and mixed-ethnicity patients) and cluster 9 (primarily South Asian patients).

Figure C1 (Appendix) provides an overview of phenotype definitions, including feature probabilities, age/BMI distributions, and ethnicity proportions. The ten clusters range from 4%–20% prevalence in the synthetic cohort, approximating heterogeneity reported in clinical studies.

C.2 Feature set

Each patient record includes:

- **Demographics:** age, BMI, smoking status.
- **Clinical indicators:** eosinophilia, neutrophilia, allergy history, overall severity.
- **Treatment exposures:** indicators of prescribing for hypertension, diabetes, dyslipidaemia, and obesity (proxying metabolic syndrome).
- **Ethnicity:** categorical with six levels (White, Asian, Black, Mixed, Other, Unknown).

Values are assigned probabilistically according to per-phenotype distributions (see Figure C1, left panel).

C.3 Scaling

To examine scalability, we evaluated subsets of increasing size from the full dataset: 10k, 20k, $\ldots$, 200k patients. This enabled benchmarking of runtime and memory as a function of N, and comparison against standard k-medoids where feasible.

C.4 Fairness probe

We intentionally set clusters 8 and 9 at 4% and 5% prevalence, respectively. Under uniform feature weights these minority phenotypes are readily absorbed into majority clusters. As shown in the main results (Figure 2), re-weighting ethnicity recovers both as distinct groups, illustrating how domain knowledge can improve equitable phenotyping.

D Appendix D: Detailed Complexity Analysis

We analyse the runtime and memory complexity of the proposed streaming+coreset k-medoids algorithm, and contrast it with standard full-distance k-medoids.

D.1 Preliminaries

Let

- N: total number of patients,
- K: number of clusters (fixed and small),
- C: chunk size,
- M: coreset size with $M \ll N$.

The algorithm proceeds in three phases:

1. **Streaming over chunks:** build within-chunk distance matrices and run local k-medoids;
2. **Coreset refinement:** run k-medoids on a reduced set of size M;
3. **Final assignment:** assign all N points to their nearest refined medoid.

D.2 Time complexity

1) Chunked distance builds. For chunk t of size $B_t \approx C$, constructing the weighted Gower distance matrix costs

$$\mathcal{O}(B_t^2) \text{ per chunk.}$$

With $\frac{N}{C}$ chunks in total,

$$\sum_t \mathcal{O}(B_t^2) \;\approx\; \frac{N}{C} \cdot \mathcal{O}(C^2) \;=\; \mathcal{O}(NC).$$

2) Coreset refinement. After streaming, we construct a coreset of size M (via candidate pools/reservoirs). Running k-medoids on this subset requires an $M \times M$ distance matrix:

$$\mathcal{O}(M^2).$$

3) Final assignment. Each of the N points is assigned to the nearest of the K refined medoids by computing weighted Gower distances:

$$\mathcal{O}(NK).$$

Total time:

$$T(N) \;=\; \mathcal{O}\big(NC \;+\; M^2 \;+\; NK\big)\,.$$

D.3 Memory complexity

1) Chunk storage. At any time only one chunk is resident; its distance matrix requires

$$\mathcal{O}(C^2).$$

2) Coreset storage. The coreset distance matrix contributes

$$\mathcal{O}(M^2).$$

3) Label storage. We maintain cluster labels/indices for all points:

$$\mathcal{O}(N).$$

Total memory:

$$\text{Memory} \;=\; \mathcal{O}\big(C^2 \;+\; M^2 \;+\; N\big)\,.$$

D.4 Comparison with standard k-medoids

Standard full-distance k-medoids computes and stores the entire $N \times N$ distance matrix:

$$T(N) \;=\; \mathcal{O}(N^2), \qquad \text{Memory} \;=\; \mathcal{O}(N^2).$$

This quadratic growth is typically infeasible for datasets beyond ∼50k–100k patients on 32 GB machines. By contrast, our streaming+coreset method scales linearly in N, with only modest quadratic terms in C and M.

D.5 Intuition

- The quadratic bottleneck is shifted from the full dataset (N^2) to manageable subsets (C^2 and M^2).
- The final assignment $\mathcal{O}(NK)$ is linear in N and cheap in practice since K is small (e.g., K=10).
- Memory is dominated by the chunk and coreset distance matrices and thus remains bounded as N grows (for fixed C and M).

Summary. Standard k-medoids has prohibitive $\mathcal{O}(N^2)$ time and memory. Our method achieves $T(N) = \mathcal{O}(NC + M^2 + NK)$ with memory $\mathcal{O}(C^2 + M^2 + N)$, enabling clustering of hundreds of thousands of patients under 16 GB RAM.

Compared to standard full-distance k-medoids ($\mathcal{O}(N^2)$ time and memory), our method achieves $T(N) = \mathcal{O}(NC + M^2 + NK)$ and $\text{Memory} = \mathcal{O}(C^2 + M^2) + \underbrace{\mathcal{O}(N)}_{\text{labels}}$. For typical settings (e.g. $C \sim 2{,}000$, $M \leq 20{,}000$, $K \leq 10$), the quadratic terms are bounded and dominate peak RAM, while the $\mathcal{O}(N)$ labels are negligible in practice.

E Appendix E: Evaluation Metrics

We report clustering performance using four widely used metrics:

Adjusted Rand Index (ARI). Measures agreement between predicted and true labels, corrected for chance. Values range from -1 (worse than random) to 1 (perfect match), with 0 indicating random assignment.

Normalised Mutual Information (NMI). Quantifies mutual dependence between predicted and true clusters, normalised between 0 and 1; 1 indicates identical clusterings.

Silhouette Score. An internal metric based on pairwise distances. For each point, it compares cohesion (similarity to its own cluster) and separation (dissimilarity to the nearest other cluster). Scores range from -1 (poor fit) to 1 (well-separated clusters).

Purity. An external measure of cluster homogeneity. Each predicted cluster is assigned the majority ground-truth label; purity is the fraction of correctly assigned samples and ranges from 0 to 1.

Together, these metrics capture external alignment with ground truth (ARI, NMI, Purity) and internal separation (Silhouette).

Formal definitions (optional)

Let $U = \{U_i\}$ be ground-truth classes and $V = \{V_j\}$ predicted clusters over n samples. Let $n_{ij} = |U_i \cap V_j|$, $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$.

Adjusted Rand Index (ARI).

$$\mathrm{ARI} \;=\; \frac{\displaystyle\sum_{i,j}\binom{n_{ij}}{2} \;-\; \frac{\displaystyle\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\displaystyle\tfrac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] \;-\; \frac{\displaystyle\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}.$$

Normalised Mutual Information (NMI). Let $p_{ij} = n_{ij}/n$, $p_i = a_i/n$, $p_j = b_j/n$. Define

$$I(U;V) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i p_j}, \qquad H(U) = \sum_i p_i \log \frac{1}{p_i}, \quad H(V) = \sum_j p_j \log \frac{1}{p_j}.$$

We use the geometric normalisation:

$$\mathrm{NMI} \;=\; \frac{I(U;V)}{\sqrt{H(U)\,H(V)}} \;\in\; [0,1].$$

Silhouette. For point x, let $a(x)$ be its mean distance to points in its own cluster, and $b(x)$ the minimal mean distance to any other cluster. Then

$$s(x) \;=\; \frac{b(x) - a(x)}{\max\{a(x),\, b(x)\}}, \qquad \mathrm{Silhouette} \;=\; \frac{1}{n}\sum_x s(x).$$

Purity. With predicted clusters $\{V_j\}$ and true classes $\{U_i\}$,

$$\mathrm{Purity} \;=\; \frac{1}{n}\sum_j \max_i |U_i \cap V_j|.$$

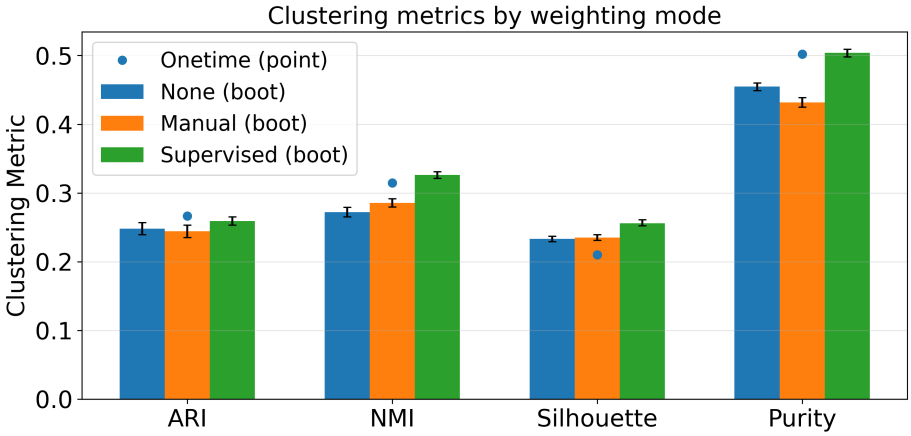

Figure F1: Bootstrap mean ± SE for ARI, NMI, silhouette, and purity across the three weighting modes (uniform, manual, supervised).

F Appendix F: Additional Results on Weighting

To further evaluate the impact of feature weighting, we conducted bootstrap experiments across three modes:

- **None (uniform):** all features weighted equally;
- **Manual:** ethnicity weight set to 3 (others fixed at 1);
- **Supervised:** weights derived from ground-truth labels (upper bound).

Interpretation. The supervised setting—unrealistic in deployment but informative—provides an upper bound and consistently yields the highest performance. Manual weighting offers modest gains in global metrics relative to uniform weighting, but its key benefit is qualitative: it enables recovery of minority-dominated clusters (clusters 8–9; see Appendix Fig. C1 and main Fig. 2) that are otherwise absorbed into majority groups. This shows that simple domain-informed adjustments can improve inclusivity without sacrificing global performance.

Practical significance. While global improvements in ARI/NMI are modest, the ability to recover under-represented subgroups has outsized importance in healthcare contexts where equity is central. These findings suggest that flexible feature weighting should be a core component of scalable clustering frameworks for clinical data.

G Appendix G: Key Contributions, Limitations, and Future Work

G.1 Key Contributions

- **Scalability:** a streaming+coreset k-medoids algorithm with (empirically) linear runtime and bounded memory, enabling clustering of 200k+ patients under modest (16 GB) hardware.
- **Fairness:** support for domain-informed feature weighting, allowing recovery of minority-dominated phenotypes that uniform clustering overlooks.
- **Generalisability:** a broadly applicable framework for large, mixed-type healthcare datasets—motivated by asthma but extendable to other chronic diseases and EHR contexts.

G.2 Limitations

- **Synthetic evaluation:** experiments use a literature-informed synthetic dataset to ensure controlled ground truth and stress-test scalability
- **Single-machine implementation:** current results are from a single-node Python pipeline under RAM constraints; distributed/GPU variants remain future work.

- **Coreset sensitivity:** clustering quality depends on coreset size M; adaptive or theoretically guaranteed coreset selection is an open problem.

G.3 Future Directions

- **Real-world deployment:** apply the method to national EHR datasets (e.g., CPRD, DataLoch) under TRE conditions for asthma and other chronic diseases.
- **Temporal extension:** incorporate longitudinal signals (e.g., repeated prescriptions, symptom trajectories) into the streaming framework.
- **Automatic feature weighting:** learn weights adaptively to balance domain knowledge with statistical signal.
- **Scalable infrastructure:** develop distributed implementations for larger-scale or near–real-time applications (e.g., patient monitoring systems).