GI-Clust: Deep Clustering for Early Gastrointestinal Cancer Detection

Katarina Vukosavljević[*]
Department of Engineering Science
University of Oxford
Oxford, United Kingdom
katarina.vukosavljevic@eng.ox.ac.uk

Emma Bailey
Centre for Cancer Evolution, Barts Cancer Institute
Queen Mary University of London London, United Kingdom
emma.bailey@qmul.ac.uk

Jun Wang
Centre for Cancer Evolution, Barts Cancer Institute
Queen Mary University of London
London, United Kingdom
j.a.wang@qmul.ac.uk

Julia Hippisley-Cox
Wolfson Institute of Population Health
Queen Mary University of London
London, United Kingdom
julia.hippisley-cox@qmul.ac.uk

Tingting Zhu
Department of Engineering Science
University of Oxford
Oxford, United Kingdom
tingting.zhu@ox.ac.uk

Abstract

Early diagnosis of gastrointestinal (GI) cancers remains challenging due to non-specific symptom presentation and the limitations of existing risk stratification tools in primary care. Current models are predominantly static, failing to capture how patient trajectories evolve within electronic health records (EHRs). We present GI-Clust, a predictive deep clustering framework designed for irregular, multivariate EHR time series. GI-Clust employs a dual-encoder architecture: an LSTM-based attention encoder for temporal features with an integrated interpretability framework, and a lightweight MLP encoder for baseline risk factors, fused via a gated mechanism. Latent embeddings are clustered using a Gumbel-Softmax layer, enabling differentiable optimisation. The framework jointly optimises prediction and clustering objectives to uncover clinically interpretable patient subgroups. Evaluated on 210,970 UK primary care patients from the QResearch database, GI-Clust outperforms strong baselines, including XGBoost, LSTM-Encoder, and CAMELOT, achieving AU-ROC 0.870 and F1 0.380, while identifying phenotype-specific feature–time dependencies (e.g., haemoglobin in the six months prior to diagnosis across GI cancer subtypes). Crucially, the model generalises well to geographically distinct test regions, demonstrating robustness. To our knowledge, this is the first predictive clustering approach applied to longitudinal UK primary care data for cancer detection.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time-Series for Health.

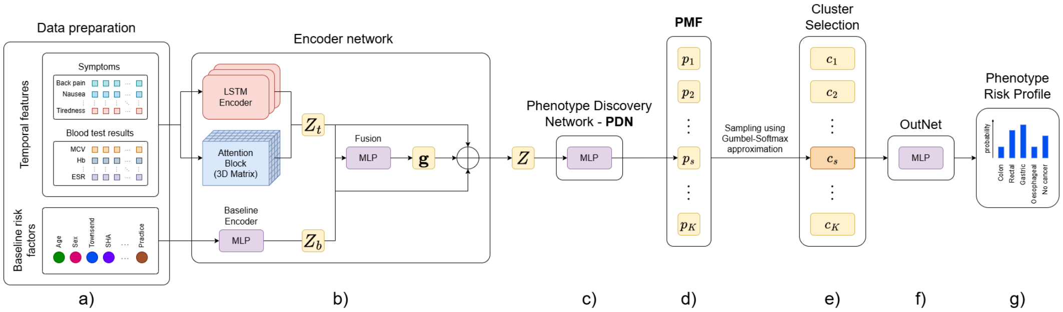


Figure 1: Overview of the GI-Clust architecture.

1 Introduction

Nearly half of all cancer cases in England were diagnosed at stage III or IV in 2018, with rates even higher for gastrointestinal (GI) cancers [1, 2]. These four GI cancers alone represent over 15% of new UK cancer cases, yet their vague symptoms make early detection particularly difficult, contributing to some of the poorest colorectal survival rates in Europe [3, 4]. Existing NHS risk tools, such as QCancer [5, 6], provide only static, cross-sectional estimates and fail to capture how patient symptoms and blood test results evolve over time. Temporal electronic health records (EHRs) offer a opportunity to capture this temporal dynamics and enable earlier diagnosis by providing rich longitudinal information on demographics, symptoms, and laboratory tests. However, they also pose technical challenges due to their heterogeneity, irregular sampling, and missingness.

In this paper, we present GI-Clust, Figure 1, a predictive clustering framework tailored for irregular, multivariate EHR time-series. Our model includes a novel encoder that incorporates an interpretability framework, combining an LSTM-based attention encoder for temporal features with a lightweight MLP encoder for baseline risk factors, fused through a gated mechanism. Cluster assignments are modelled using a Gumbel-Softmax layer to enable differentiable optimisation of discrete subgroups. We evaluate GI-Clust on 210,970 UK primary care patients from the QResearch database and outperform strong baselines while providing phenotype-specific risk profiles for GI cancer subtypes. Our phenotypes highlight clinically relevant feature–time pairs linked to early GI cancer detection.

2 Related Work

GI cancer detection. Most existing approaches rely on imaging modalities (endoscopy, CT, pathology) with CNN-based models achieving strong performance [4, 7, 8]. Risk models such as QCancer [5, 6, 9] and CRISP [10] use logistic regression or decision trees with static covariates, but rarely incorporate temporal dynamics or routine primary care data.

Mixed temporal EHR data modelling. Deep learning models using LSTMs and GRUs [11, 12, 13, 14] capture sequential dependencies but model static and temporal variables together, which can be particularly detrimental in sparse primary care data, as static features may dominate and obscure subtle temporal features. Recent methods that encode data types separately [15, 16, 17, 18] improve prediction but remain limited in interpretability and robustness to irregular sampling.

Phenotype discovery. Clustering methods like SOM-VAE [19], AC-TPC [12], and CAMELOT [11] uncover patient subgroups, but are mainly applied to hospital datasets and often overlook noise, missingness, and heterogeneity in primary care records.

3 Method

We propose **GI-Clust**, a predictive clustering framework for irregular EHR time series that jointly models baseline risk factors and temporal trajectories to discover clinically interpretable subgroups. Our models learns clusters that are predictive of the clinical outcomes of interest while remaining interpretable by revealing feature-time pairs contributing to phenotype assignment. The framework, Figure 1, consists of three main sub-networks: an *attention-based dual encoder* with an interpretability framework, a *Phenotype Discovery Network (PDN)*, and an *OutNet*.

Encoder Network. Baseline features (e.g. age and sex) are compressed into a low-dimensional representation that captures overall background risk, $\mathbf{Z}_b$. Temporal features (e.g symptoms, blood test results) are encoded using an LSTM with attention, $\mathbf{Z}_t$. This separation is designed to prevent static risk factors from dominating the representation and to encourage the model to exploit subtle clinical signals in the temporal features. The LSTM-based encoder, inspired by [11], uses a stacked LSTM not to directly produce the final temporal embedding, but to approximate the feature–time importance captured by the attention block. The custom attention block, a dense-layer block, disentangles feature contributions by projecting each input into time-independent latent representations. A more detailed technical description of the encoder architecture is provided in the Appendix A.

The two representations are fused via a gated mechanism that adaptively balances baseline and temporal embeddings:

$$\mathbf{Z} = \mathbf{g} \odot \mathbf{Z}_t + (1 - \mathbf{g}) \odot \mathbf{Z}'_b, \tag{1}$$

where $\mathbf{Z}_t$ and $\mathbf{Z}'_b$ are the temporal and projected baseline embeddings, $\mathbf{g} \in [0, 1]^L$ is a learnt gate, and $\odot$ denotes element-wise multiplication.

Predictive clustering. The fused representation $\mathbf{Z}$ is passed to the PDN, an MLP that outputs cluster assignment probabilities over K trainable cluster representations $(\mathbf{c_1}, \ldots, \mathbf{c_K})$. To make cluster assignment differentiable, we apply the Gumbel-Softmax trick [20, 21], which relaxes categorical sampling as:

$$\tilde{p}_k = \frac{\exp\left(\frac{\log p_k + g_k}{\tau}\right)}{\sum_{j=1}^{K} \exp\left(\frac{\log p_j + g_j}{\tau}\right)}, \tag{2}$$

where p_k is the cluster probability, $g_k \sim \text{Gumbel}(0, 1)$, and τ is the temperature hyperparameter. Here, we set $\tau = 0.5$. At inference, patients are assigned to their most likely cluster.

Outcome prediction. The selected cluster representation $\mathbf{c_k}$ is passed to the OutNet, which gives a probability distribution over the five clinical outcomes (Colon, Rectal, Gastric, Oesophageal, No Cancer). A detailed training and optimisation procedure is provided in the Appendix A.

4 Results

We evaluated GI-Clust on the QResearch primary care database, comprising 210,970 patients (8,118 with GI cancer across colon, rectal, gastric, and oesophageal subtypes, matched to controls). Records include up to five years of 43 clinically relevant features [5, 6, 9] including demographics, diagnoses, symptoms, and blood tests. We provide the cohort selection and preprocessing strategies, full list of variables, and summary statistics in the Appendix B. Evaluation is performed on a geographically distinct region (South East and South West of England).

We implemented GI-Clust in TensorFlow 2.10 and trained on a single Tesla V100 PCIe 32GB GPU, using fixed random seeds for reproducibility and stratified cross-validation for robustness. We provide the full model and training hyperparameters for GI-Clust and our benchmarks in the Appendix D.

We compare against four benchmarks: CAMELOT [11], the state-of-the-art for interpretable clustering of time-series EHR data; XGBoost, widely used in clinical settings for its interpretability and strong performance on tabular data; a simple LSTM Encoder, included to directly test the usefulness of our clustering framework beyond sequence modelling; and TSKM [22], the state of the art in time-series clustering, used here to confirm that the raw data does not contain trivial clustering patterns.

Two tasks were considered: (1) Clinical outcome prediction, a five-class classification problem (Colon, Rectal, Gastric, Oesophageal, No Cancer); and (2) Phenotype discovery, evaluating whether cluster assignments reveal clinically interpretable patient subgroups. Performance was assessed using AU-ROC, AU-PRC, F1, Recall, and Precision, computed via macro-averaging across classes to target imbalance, the results are shown in Table 1. Results are reported as the mean and standard deviation over five random seeds.

Table 1: Classification performance across five clinical outcome categories: Colon, Rectal, Oesophageal, Gastric, and No Cancer. The best value for each metric is shown in bold. For all reported metrics, higher values indicate better performance.

Method	AUROC	AUPRC	F1	Recall	Precision
XGB	0.817(±0.003)	0.314(±0.002)	0.253(±0.002)	**0.467(±0.006)**	0.245(±0.001)
LSTMClassifier	0.769(±0.126)	0.247(±0.028)	0.243(±0.031)	0.334(±0.087)	0.235(±0.026)
TSKM	0.553(±0.002)	0.248(±0.010)	0.196(±0.0)	0.2(±0.0)	0.192(±0.0)
CAMELOT	0.858(±0.011)	0.322(±0.009)	0.311(±0.038)	0.298(±0.041)	**0.448(±0.033)**
GI-Clust (ours)	**0.870(±0.009)**	**0.337(±0.009)**	**0.380(±0.026)**	0.425(±0.039)	0.382(±0.019)

Beyond classification, GI-Clust discovers clinically meaningful subgroups, shown in Figure 2 in the Appendix. The learnt clusters correspond to distinct GI cancer profiles (e.g. cancer vs. no cancer, rectal- vs. gastric/oesophageal-dominant), while the attention maps, Figure 6, highlight early risk signals such as haemoglobin in the months before diagnosis.

5 Discussion

Classification results show that GI-Clust outperforms the benchmarks, demonstrating that our novel encoder and predictive clustering framework are well-suited for capturing temporal dynamics in sparse primary care records and improving early GI cancer detection. Importantly, the model not only improves prediction but also reveals clinically meaningful patient phenotypes, as the discovered clusters map onto distinct GI cancer subtypes, Figure 2. Each subplot shows the outcome-normalised proportion of patients per cancer subtype. Distinct patterns emerge: Cluster 1 concentrates gastric (22%) and oesophageal (42%) cancers, Cluster 3 contains the majority of "No Cancer" patients (37%), Cluster 4 mixes Colon and Gastric, and Cluster 5 is dominated by Rectal cancer. Important to note, assignment to a cluster does not imply diagnosis, but indicates that a patient's trajectory resembles those of patients later diagnosed with that outcome, supporting risk stratification for screening or monitoring.

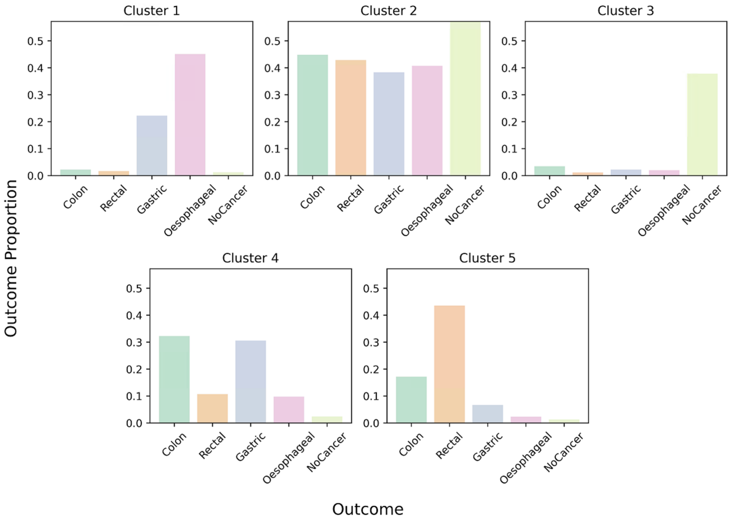


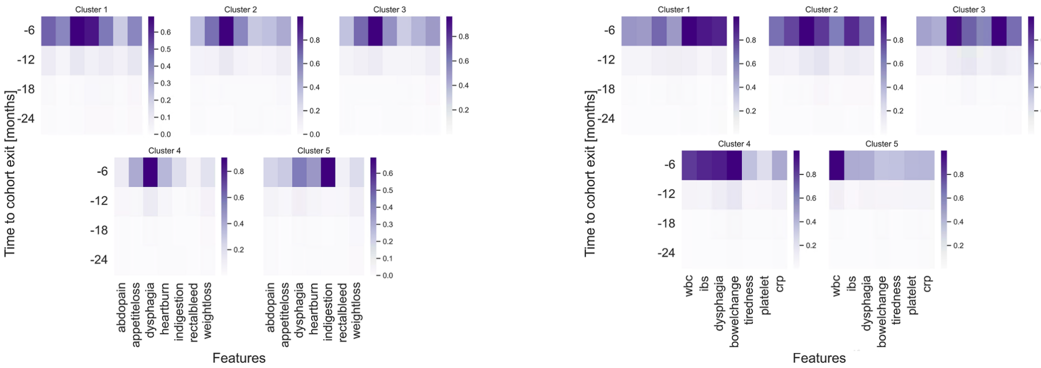
Figure 2: Distribution of clinical outcomes across identified clusters in GI-Clust. For each cluster, the bar plot represents the percentage of patients per each of the five outcome classes: Colon, Rectal, Gastric, Oesophageal, and No Cancer. Each subplot is labelled Cluster 1 through Cluster 5, corresponding to the cluster numbers assigned by the model.

To further examine the learnt clusters, we analyse their corresponding attention heatmaps in Figure 3. Figure 3a presents the attention scores for alarm or red-flag symptoms across clusters, where similar patterns appear in Clusters 2, 3, and 4. This suggests that red-flag symptoms alone are insufficient to distinguish between the trajectories of the majority-mixed (Cluster 2), No-Cancer (Cluster 3), and

Cancer (Cluster 4) subgroups. Regardless, abdominal pain, dysphagia and heartburn, risk factors for gastric and oesophageal cancer, are identified as significant features in Cluster 1, which predominantly contains patients with gastric and oesophageal cancers. Similar patterns appear in other clusters as well, confirming that GI-Clust is capable of identifying clinically meaningful features within each cluster.

In contrast, Figure 3b displays the heatmaps for the seven most informative features identified by GI-Clust. These are clinical risk factors for GI cancer that are not red-flag symptoms. The distinct temporal patterns observed in the six months preceding cohort exit highlight the importance of incorporating broader clinical risk factors when phenotyping patients. They also help explain why Clusters 2, 3, and 4 correspond to different patient sub-cohorts in Figure 2. The complete attention heatmaps, including all features, are provided in the Appendix, Figure 6.

These findings indicate that GI-Clust works well with highly imbalanced datasets and is learning temporal dependencies that align with established clinical knowledge, Table 6. The features highlighted by our model correspond well with known clinical risk patterns. Importantly, their emergence predominantly within the six months before cohort exit is consistent with clinical understanding while also emphasising the inherent difficulty of the diagnostic task.



(a) Heatmap of attention weights across clusters for GI Cancer alarm symptoms.

(b) Heatmap of attention weights across clusters for the 7 most important features identified by GI-Clust.

Figure 3: Attention heatmaps for each cluster learnt by GI-Clust for selected features in the 2 years before diagnosis or cohort exit. For each cluster, the attention maps show the class-weighted averaged attention scores across time-varying features and time windows. The weights are normalised to a 0.0 to 1.0 range. Higher attention indicates that changes in those feature–time pairs played a greater role in the model's decision to assign patients to that cluster. The x-axis shows clinical features, the y-axis shows time to cohort exit in 6-month intervals (top = most recent).

6 Conclusion

In this paper, we introduced GI-Clust, a semi-supervised clustering framework for temporal EHR data applied to early GI cancer detection. Methodologically, the use of separate baseline and temporal encoders with gated fusion, combined with Gumbel-Softmax clustering, enables learning from sparse and irregular primary care records and improves classification accuracy over strong benchmarks. Clinically, GI-Clust uncovers interpretable patient subgroups that align with established cancer phenotypes and highlights subtype-specific temporal signals, offering a pathway toward earlier screening and stratified monitoring. Future work will extend validation to external datasets and explore multimodal integration to enrich phenotype discovery.

Key takeaway: GI-Clust uses a novel dual-encoder with gated fusion and predictive clustering to learn from sparse multivariate EHR time series, boosting early cancer detection while uncovering meaningful, interpretable patient subgroups.

References

[1] Cancer Research UK. Cancer statistics for the uk, 2017. URL `https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk`.

[2] Cancer Research UK. Cruk: Early diagnosis hub, 2017. URL `https://crukcancerintelligence.shinyapps.io/EarlyDiagnosis/`.

[3] Marzieh Araghi and et al. Colon and rectal cancer survival in seven high-income countries 2010–2014: variation by age and stage at diagnosis (the icbp survmark-2 project). *Gut*, 70(1): 114–126, 2021. doi: https://doi.org/10.1136/gutjnl-2020-320625.

[4] Ganji Purnachandra Nagaraju and et al. Artificial intelligence in gastrointestinal cancers: diagnostic, prognostic, and surgical strategies. *Cancer Letters*, 612:217461, 2025. doi: https://doi.org/10.1016/j.canlet.2025.217461.

[5] Julia Hippisley-Cox and Carol Coupland. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *The British Journal of General Practice*, 63(606):e1, 2012. doi: https://doi.org/10.3399/bjgp13X660724.

[6] Julia Hippisley-Cox and Carol Coupland. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *The British Journal of General Practice*, 63(606):e11, 2012. doi: https://doi.org/10.3399/bjgp13X660733.

[7] Runnan Cao and et al. Artificial intelligence in gastric cancer: applications and challenges. *Gastroenterology Report*, 10:goac064, 2022. doi: https://doi.org/10.1093/gastro/goac064.

[8] Sreetama Mukherjee, Sunita Vagha, and Pravin Gadkari. Navigating the future: a comprehensive review of artificial intelligence applications in gastrointestinal cancer. *Cureus*, 16(2):e54467, 2024. doi: https://doi.org/10.7759/cureus.54467.

[9] Julia Hippisley-Cox and Carol Coupland. Development and external validation of prediction algorithms to improve early diagnosis of cancer. *Nature Communications*, 16(1):1–11, 2025. doi: https://doi.org/10.1038/s41467-025-57990-5.

[10] Jon D. Emery and et al. The colorectal cancer risk prediction (crisp) trial: a randomised controlled trial of a decision support tool for risk-stratified colorectal cancer screening. *British Journal of General Practice*, 73(733):e556–e565, 2023. doi: https://doi.org/10.3399/BJGP.2022.0480.

[11] Henrique Aguiar, Mauro Santos, Peter Watkinson, and Tingting Zhu. Learning of cluster-based feature importance for electronic health record time-series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 161–179, Baltimore, MD, USA, july 2022. PMLR.

[12] Changhee Lee and Mihaela Van Der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, Virtual, 2020. JMLR.

[13] Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. T-phenotype: Discovering phenotypes of predictive temporal patterns in disease progression. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 3466–3492, Valencia, Spain, 2023. PMLR.

[14] Changhee Lee, Jem Rashbass, and Mihaela Van der Schaar. Outcome-oriented deep temporal phenotyping of disease progression. *IEEE Transactions on Biomedical Engineering*, 68(8): 2423–2434, 2021. doi: https://doi.org/10.1109/TBME.2020.3041815.

[15] Anna Leontjeva and Ilya Kuzovkin. Combining static and dynamic features for multivariate sequence classification. In *2016 IEEE International Conference on Data Science and Advanced Analytics*, DSAA, pages 21–30, Montreal, QC, Canada, 2016. IEEE.

[16] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *2016 IEEE International Conference on Healthcare Informatics*, ICHI, pages 93–101, Chicago, IL, USA, 2016. IEEE.

[17] Pu Zhang and et al. Simulation model of vegetation dynamics by combining static and dynamic data using the gated recurrent unit neural network-based method. *International Journal of Applied Earth Observation and Geoinformation*, 112:102901, 2022. doi: https://doi.org/10.1016/j.jag.2022.102901.

[18] Molla Hafizur Rahman, Shuhan Yuan, Charles Xie, and Zhenghui Sha. Predicting human design decisions with deep recurrent neural network combining static and dynamic data. *Design Science*, 6:e15, 2020. doi: https://doi.org/10.1017/dsj.2020.12.

[19] Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. Som-vae: Interpretable discrete representation learning on time series, 2018. URL `https://doi.org/10.48550/arXiv.1806.02199`.

[20] Chris J. Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, volume 27 of *NIPS'14*, page 3086–3094, Cambridge, MA, USA, 2014. MIT Press.

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, Nov 2016. URL `https://doi.org/10.48550/arXiv.1611.01144`.

[22] Tavenard Romain and et al. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. URL `http://jmlr.org/papers/v21/20-091.html`.

[23] Julia Hippisley-Cox, David Stables, and Mike Pringle. Qresearch: a new general practice database for research. *Informatics in Primary Care*, 12(1):49–50, 2004. doi: https://doi.org/10.14236/jhi.v12i1.108.

A GI-Clust Architecture

Encoder Network. Baseline covariates (e.g., age, sex, ethnicity) are processed by a shallow MLP, while temporal features are encoded using an LSTM with attention. The LSTM Encoder consists of a stacked LSTM block and a custom attention layer. The LSTM-based encoder, inspired by [11], combines a stacked LSTM with a custom attention block. The LSTM captures temporal dependencies, while the attention block, a dense layer block, disentangles feature contributions by projecting each input into time-independent latent representations. Following the notation of [11], the adapted output from the attention block for temporal features can be described as follows:

$$\mathbf{R}_t = \sigma(\mathbf{D} \odot \mathbf{x}_{dyn,t} + \mathbf{B}). \tag{3}$$

where $\mathbf{R}_t$ is the collection of the temporal feature representations at time t, $\mathbf{D} \in \mathbb{R}^{l \times F}$ the learnable attention block kernel, $\mathbf{B} \in \mathbb{R}^{l \times F}$ the learnable attention block bias, where l is the latent dimension and F the number of temporal features, and $x_{dyn,t} \in \mathbb{R}^F$ the temporal feature input at time t. The LSTM encoder output can be written as: $[o_1, \ldots, o_T]$ with the state at time i, $o_i \in \mathbb{R}^l$, where T is the time observation window. Crucially, to ensure that the attention block approximates the LSTM output, the two outputs are combined through the following approximation:

$$\mathbf{o_t} \approx \mathbf{R}_t \alpha_t = \hat{\mathbf{o}}_\mathbf{t}, \tag{4}$$

with α_t computed as the least-squares solution. Finally, the temporal embedding $\mathbf{Z_t}$ is derived as:

$$\mathbf{Z_t} = \sum_t \beta_t \hat{o}_t, \tag{5}$$

where $\hat{o}_t$ is the optimal solution from 4 and attention weights β_t are learnt to emphasise informative time steps. To ensure that we don't leak future inputs from the temporal data in the custom attention block, we additionally apply a mask to the input data directly. The mask ensures that unobserved (future) feature-time entries are zeroed out and prevents the attention mechanism from treating missing values as informative. Finally, the two representations are fused via a gated mechanism that adaptively balances static and dynamic embeddings:

$$\mathbf{Z} = \mathbf{g} \odot \mathbf{Z}_t + (1 - \mathbf{g}) \odot \mathbf{Z}'_b, \tag{6}$$

where $\mathbf{Z}_t$ and $\mathbf{Z}'_b$ are the temporal and projected baseline embeddings, $\mathbf{g} \in [0, 1]^L$ is a learnt gate, and $\odot$ denotes element-wise multiplication.

Model training and optimisation. GI-Clust is trained in two stages: *pre-clustering* and *fine-tuning*. In the pre-clustering stage, the Encoder and Outcome Network are first trained in an encoder–predictor setup [12, 11]. Latent embeddings $\mathbf{Z}$ are then clustered with k-means to initialise K cluster representations, which in turn supervise the PDN. During fine-tuning, the learning rate is reduced and the full model is trained end-to-end. Unlike Aguiar et al. [11], which relies on staged initialisation, we train all sub-networks until convergence, a strategy better suited to sparse primary care data. Different from [12], GI-Clust jointly optimises all sub-networks with a unified objective. Optimisation combines classification and clustering losses, including the loss targeting cluster collapse in [11], with tunable weights to balance tasks.

B QResearch Dataset

QResearch [23] is a UK primary care database containing anonymised EHRs from over 35 million patients, linked at the individual level to the National Cancer Registry, Death Registry, and Hospital Episode Statistics. For this work, we used a random 25% subset covering 13 million patients in England (2010–2019), providing demographics, prescriptions, diagnoses, laboratory tests, and cancer registry information (diagnosis, grade, stage). This large, diverse dataset is well suited for early cancer detection, while also reflecting the typical challenges of EHRs: heterogeneity, missingness, and irregular sampling. We constructed a case–control cohort from QResearch comprising four GI cancer types (colon, rectal, gastric, oesophageal), defined using ICD-10 codes and excluding non-melanoma skin cancers. Eligible patients were 20–100 years old with at least two years of GP registration prior to entry (2015–2019). Cancer cases were restricted to first diagnoses within this period, excluding prior cancers, multiple synchronous cancers, or diagnoses recorded only at death. Controls were cancer-free up to 2019 and matched to cases 1:25 (to reflect the national cancer prevalence) by sex and general practice, with a maximum age difference of 10 years. The final cohort included 8,118 GI cancer cases and 210,970 patients overall. Table 2 shows the breakdown of the number of patients per cancer type and the diagnostic codes used for extraction, revealing a highly imbalanced dataset. Figure 4 shows the cohort extraction flowchart.

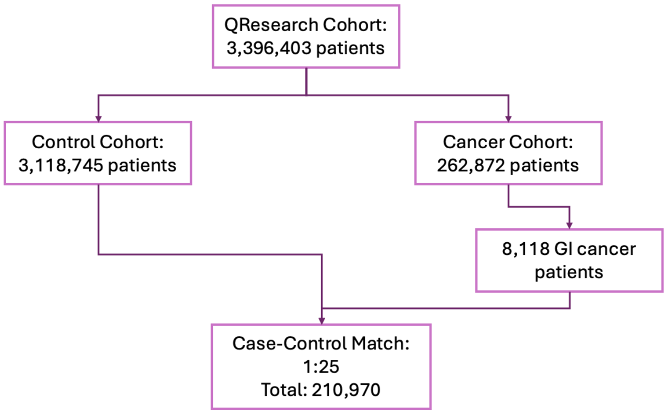


Figure 4: Cohort selection from the QResearch database: The flowchart shows the extraction process used to define the case–control cohort. A total of 8,118 GI cancer cases were identified and matched to controls in a 1:25 ratio, based on sex and general practice, with a maximum age difference of 10 years. When more than 25 matches were available, controls were ranked by age proximity. Controls could be matched to multiple cases, but differing index dates limited exposure window overlap.

Table 2: Cancer coding and case distribution: The table lists the ICD-10 codes used to identify cases of gastrointestinal cancer in the study cohort, along with the number of patients per cancer type.

Cancer Type	Colon	Rectal	Gastric	Oesophageal
ICD-10 Code	C18	C19, C20	C16	C15
Cases	3,926	1,966	893	1,333

Preprocessing involved standardising units across blood tests, removing extreme outliers (outside the 0.1–99.9th percentiles) and clipping values beyond three standard deviations. Age was fixed at cohort entry, while other baseline covariates remained static. Time-varying features (risk factors, symptoms, blood tests) were aggregated into 6-month intervals (maximum 10 per patient). Continuous measures (e.g., blood tests, BMI, alcohol intake) were averaged per interval, while symptoms are binary. Missing data were imputed using last observation carried forward, population means (for entirely missing features), or an "Unknown" category (for categorical variables). Finally, features were min–max and batch normalised. The full list of used clinical variables is provided in Table 3.

Table 3: Overview of clinical variables from the QResearch dataset: The table presents the full list of clinical variables extracted from the QResearch database, grouped into four clinically relevant categories: baseline characteristics, risk factors, symptoms, and blood test results. Where applicable, we report the percentage of missing values in the study population. For binary variables (excluding sex), a value of 1 indicates the presence of the feature and 0 its absence. As such, symptoms do not have missingness. Sex is encoded as 1 for male and 0 for female. All symptoms are time-varying binary variables, and all blood test results are continuous and time-varying. For baseline characteristics and risk factors, we include additional columns describing each feature's nature (static or time-varying) and type (binary, categorical, or continuous). Where applicable, the median and missingness of each variable are reported after outlier clipping. Group Code IDs used for variable extraction are available on the QResearch website at `https://www.qresearch.org/data/qcode-group-library/`, which includes SNOMED-CT and ICD-10 mappings. Group codes used in multiple variables are listed as: Multiple^1: 561, 6282, 6285, 17064–17070; and Multiple^2: 2238, 2239, 2242–2244, 7527. ALT - alanine aminotransferase test. CRP - C-reactive protein. ESR - erythrocyte sedimentation rate. GGT - gamma-glutamyl transferase. MCV - mean corpuscular volume. WBC - white blood cell.

Baseline Characteristics

Variable	Group Code ID	Nature	Category	Median(Range)	Missingness [%]
Age	N/A	Static	Continuous	68.0 (20 – 98)	0.00
Sex (Male)	N/A	Static	Binary	58.83 %	0.00
Ethnicity	N/A	Static	Categorical	N/A	36.85
Study Practice	N/A	Static	Categorical	N/A	0.00
Strategic Health Authority (SHA)	N/A	Static	Categorical	N/A	0.00
Townsend deprivation quantile	N/A	Static	Categorical	2.0 (1.0 – 4.0)	0.22

Risk Factors

Variable	Group Code ID	Nature	Category	Median(Range)	Missingness [%]
Alcohol Intake	Multiple ^{1}	Time-varying	Continuous	2.0±13.0 (0.0 – 279.6)	59.97
Body Mass Index (BMI)	200	Time-varying	Continuous	27.7±5.6 (8.3 – 47.1)	25.07
Crohn's disease	45	Static	Binary	871 (0.41%)	-
Irritable Bowel Syndrome (IBS)	17179	Time-varying	Binary	3,010	-
Family History of non-GI Cancer	2527	Static	Binary	1,809	-
Family History of GI cancer	1345	Static	Binary	20,403	-
Smoking Category	Multiple^2	Time-varying	Categorical	-	15.49
Ulcerative Colitis (UC)	46	Time-varying	Binary	1,846	-

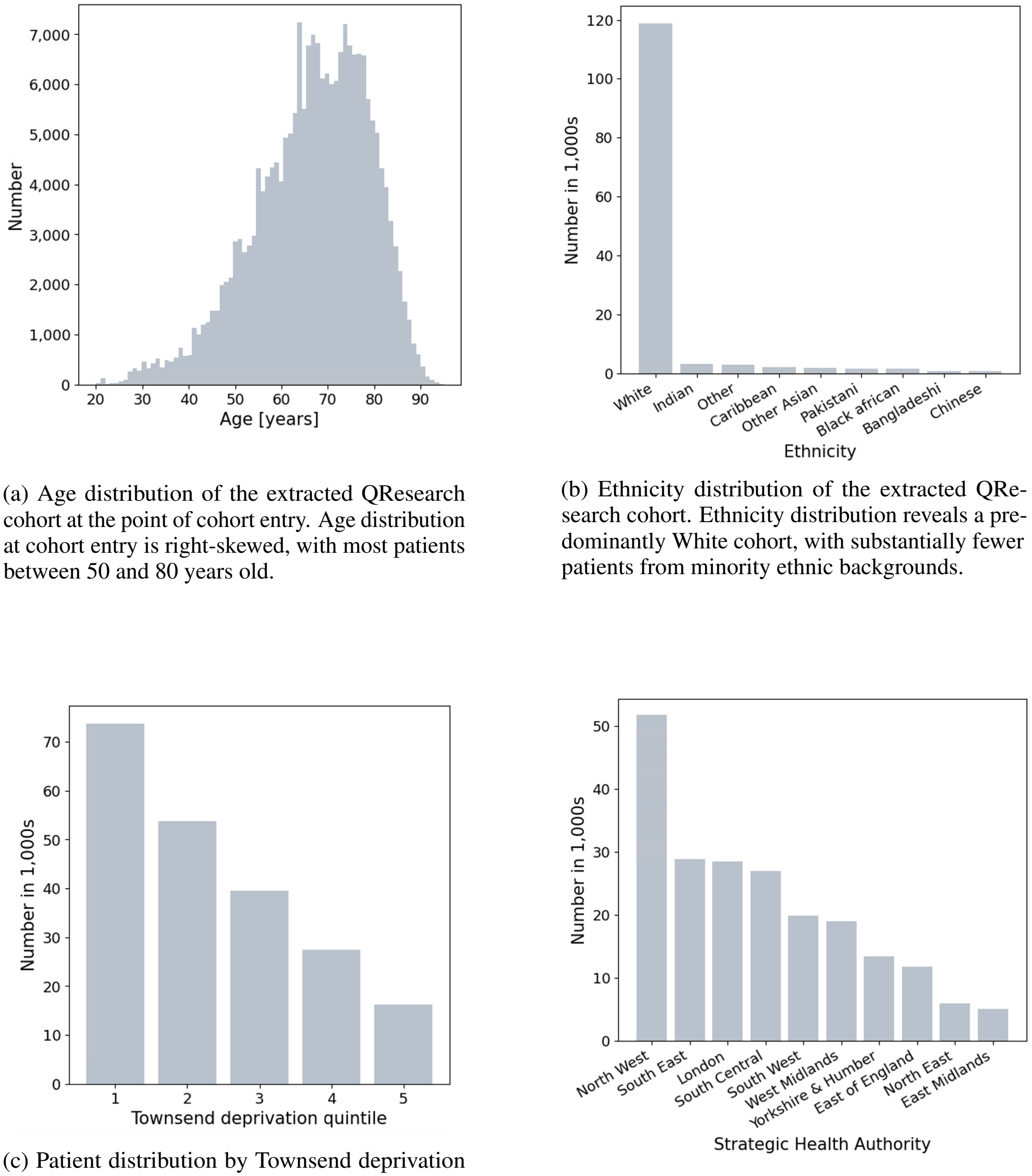
Symptoms

Variable	Group Code ID	[%] with $\geq$1 entry per observation window	Red Flag Symptom (YES/NO)
Abdominal Mass	4968	0.66	NO
Abdominal Pain	135	14.55	YES
Appetite Loss	1393	0.98	YES
Back Pain (Non-Sciatica)	2374	22.47	NO
Bowel Change	1845	2.49	NO
Constipation	141	7.62	NO
Diarrhea	107	8.56	NO
Dysphagia	1385	1.76	YES
Heartburn	2065	9.97	YES
Indigestion	2066	4.79	YES
Nausea	2375	2.07	NO
Pelvic Pain	2376	0.32	NO
Rectal Bleed	279	4.02	YES
Sciatica	2374	4.26	NO
Tiredness	605	10.82	NO
Weight Loss	1397	2.67	YES

Blood Test Results

Variable	Group Code ID	Median (Range)	Missingess [%]
Albumin	4990	41.0±3.9 (25.6 – 54.0)	20.90
ALT	1297	20.0±11.18 (4.0 – 84.0)	20.06
Bilirubin	1446	9.0±4.8 (2.0 – 32.8)	18.13
CRP	2367	5.0±16.3 (2.0 – 114.5)	68.41
ESR	2366	10.7±14.9 (2.0 – 85.8)	65.11
Ferritin	1400	76.0±115.0 (2.0 – 718.0)	65.33
GGT	1299	28.0±45.0 (3.0 – 380.0)	68.62
Haemoglobin (Hb)	1410	137.0±15.2 (75.0 – 197.7)	19.31
Iron Level	17236	13.5±6.4 (0.8 – 40.5)	93.19
MCV	1411	91.1±5.5 (68.2 – 114.0)	19.39
Platelet	1447	237.5±67.15 (21.0 – 781.0)	19.35
Cholesterol	405	3.3±1.0 (1.0 – 7.9)	25.40
WBC	2069	6.9±1.9 (1.5 – 16.0)	19.33

C Summary Statistics



(a) Age distribution of the extracted QResearch cohort at the point of cohort entry. Age distribution at cohort entry is right-skewed, with most patients between 50 and 80 years old.

(b) Ethnicity distribution of the extracted QResearch cohort. Ethnicity distribution reveals a predominantly White cohort, with substantially fewer patients from minority ethnic backgrounds.

(c) Patient distribution by Townsend deprivation quintile. 1 indicates the least deprived and 5 the most deprived. Townsend deprivation quintiles indicate that the cohort skews toward lower deprivation, with the majority of patients falling into the least deprived categories (1 and 2).

(d) Patient distribution per Strategic Health Authority (SHA). Distribution by Strategic Health Authority (SHA) shows most patients are registered in practices located in the North West, South East, and London regions.

Figure 5: Summary of cohort demographics and socioeconomic characteristics. Descriptive statistics for the extracted QResearch cohort. Patients missing ethnicity or Townsend deprivation quintile are omitted from the plots. No patients were missing age or Strategic Health Authority (SHA).

D Hyper-Parameters

Model Training. Standard Benchmarks. The complete list of hyperparameters used for the grid search optimisation of the benchmarking models is shown in Table 4. The best parameters are shown in **bold**.

Table 4: Grid search hyperparameter ranges for the benchmark models. The table presents the parameters used for TSKM and XGBoost during model tuning. Best-performing values are shown in bold.

Parameter	TSKM	XGB
γ	/	{**0.1**, 0.2, 0.5}
kernel	{**"soft-DTW"**, "DTW" "eucl"}	/
init	{"random", **"km++"**}	/
C	/	/
method	/	{"per feat", **"all"**}
n-estimators	/	{50, 100, **200**}
depth	/	{1, 3, 5, **10**}
min-child-weight	/	{1, 2, 5, **7**}
K	{**5**, 6, 7, 8}	/

Model Training. Deep learning models. The full list of hyperparameters used for grid search optimisation of LSTMClassifier, CAMELOT, and GI-Clust is shown in Table 5. The best parameters are shown in **bold**. We run the clustering models for 50 epochs and use Early Stopping with patience of 3 epochs and tolerance Δ=0.0001. The models were optimised using the Adam optimiser and the ReduceLROnPlateau learning rate scheduler, monitoring the validation loss. In addition, we conducted an ablation study to determine the optimal training parameters. We tested batch sizes bs ∈ {64, 128, 256, 512}, learning rates lr_init ∈ {1e-6, 1e-5, 1e-4, 1e-3, 1e-2}, and initialisation learning rates lr ∈ {1e-5, 1e-4, 1e-3, 1e-2}. We trained the models for the number of classes K ∈ $\{5, 6, 7, 8\}$ and latent dimension latent_dim ∈ $\{32, 64\}$, hidden layers hidden_layers ∈ $\{1, 2, 3\}$. The loss coefficients α_1, α_2, α_3, and β were chosen from {0.5, 0.1, 0.05, 0.01, 0.005, 0.001}. The best values are indicated in the table below.

Table 5: Final selected hyperparameters for the deep learning models. Hyperparameter configurations for LSTMClassifier, CAMELOT, and GI-Clust models after grid search optimisation. These include learning rates, architecture depths, and loss weighting coefficients.

Parameter	LSTMClassifier	CAMELOT	GI-Clust
bs	512	512	64
lr_init	/	0.001	0.0001
lr	0.001	1e-5	1e-5
latent_dim	8	32	32
hidden_layers (LSTM-Encoder)	1	2	1
hidden_nodes (LSTM-Encoder)	8	32	16
hidden_layers (Baseline Encoder)	/	/	2
hidden_nodes (Baseline Encoder)	/	/	8
hidden_layers (Fusion)	/	/	1
hidden_nodes (Fusion)	/	/	16
α_1	/	0.5	0.01
α_2	/	0.01	0.05
α_3	/	0.5	0.05
α_4	/	0.01	0.001
K	/	5	6

Table 6: Top five features with the highest attention scores in the final 6-month window per cluster. Table shows the five feature–time pairs with the highest attention scores in the 6-month window prior to cohort exit, listed for each cluster. Attention scores are derived from the cluster-specific attention maps in Figure 6. Cluster indices (K) correspond to those shown in the attention plot. Dysphagia was excluded from the table due to its consistent high ranking across all clusters. Abbreviations for clinical variables are provided in Table 3. Higher attention indicates that changes in those feature–time pairs played a greater role in the model's decision to assign patients to that cluster.

K	Variables
1	Tiredness, Platelet, CRP, Bilirubin, Hb
2	Hb, Platelet, Bowel Change, IBS, Back Pain
3	Platelet, Hb, ESR, Back Pain, CRP
4	Bowel Change, IBS, WBC, BMI, Sciatica
5	WBC, Indigestion, UC, Cholesterol, IBS

E Visualisations

E.1 GI-Clust

To further examine GI-Clust, we visualise the associated attention heatmaps, shown in Figure 6. Averaged attention heatmaps reveal that the most informative signals often occur in the last six months before cohort exit. We extract the five most important feature-time pairs for each cluster from these heatmaps and summarise them in Table 6. For example, haemoglobin receives high attention across clusters: persistently low Hb values mark cancer subgroups, while normal values help distinguish non-cancer cases with similar symptoms. Thus, attention maps provide interpretable insight into which clinical signals guide stratification, even when the direction of values differs.

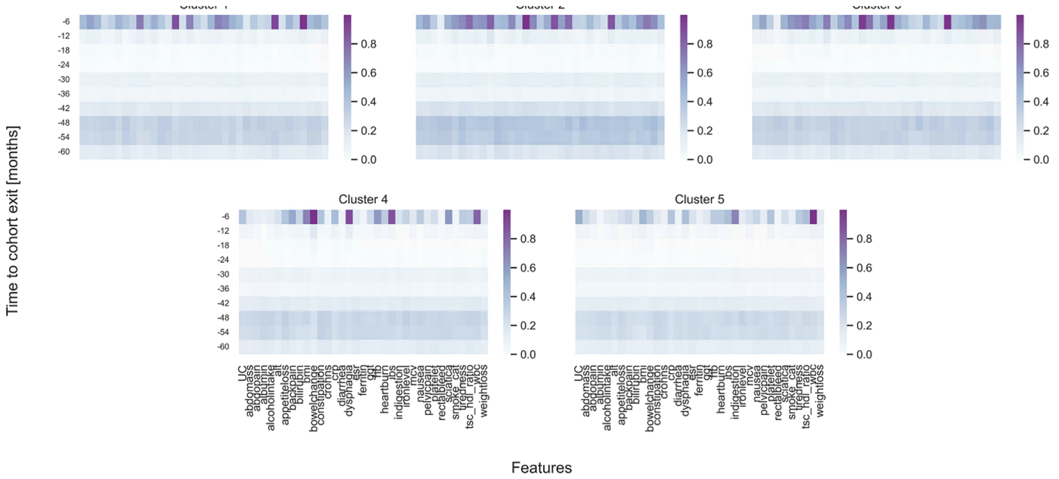


Figure 6: Attention heatmaps for each cluster learnt by GI-Clust. For each cluster, the attention maps show the class-weighted averaged attention scores across time-varying features and time windows. The weights are normalised to a 0.0 to 1.0 range. Higher attention indicates that changes in those feature–time pairs played a greater role in the model's decision to assign patients to that cluster. The x-axis shows clinical features, the y-axis shows time to cohort exit in 6-month intervals (top = most recent). The most highly attended features for each cluster are summarised in Table 6

E.2 CAMELOT

For completeness, we present the visualisations for our baseline comparison, CAMELOT. Figure 7 displays the distribution of clinical outcomes across the identified clusters (phenotypes), revealing two clusters (0 and 4) that contain few cancer patients. Clusters 1, 2, and 3 show a heterogeneous mix of all five outcomes, reflecting uncertainty in their feature profiles. This is further supported by the attention heatmaps in Figure 8, which show consistent temporal feature importance patterns across clusters, with highest attention values concentrated in the months immediately prior to cohort exit.

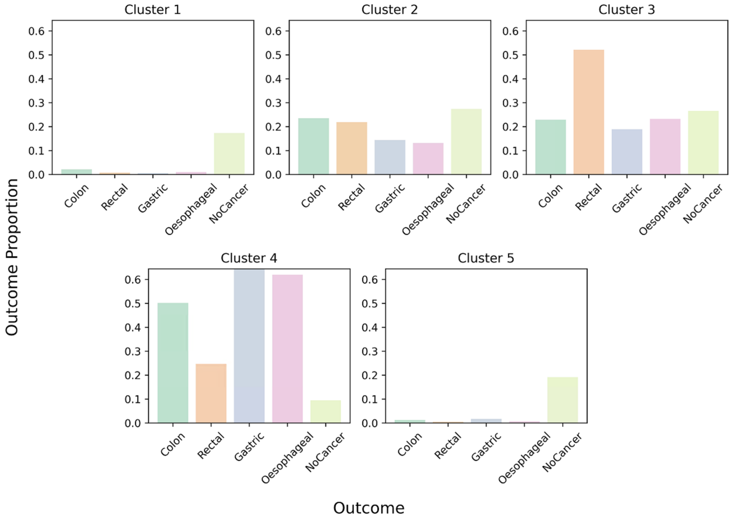


Figure 7: Distribution of clinical outcomes across identified clusters in CAMELOT. For each cluster, the bar plot represents the percentage of patients per each of the five outcome classes: Colon, Rectal, Gastric, Oesophageal, and No Cancer. Each subplot is labelled Cluster 1 through Cluster 5, corresponding to the cluster numbers assigned by the model. Clusters 0 and 4 show few cancer cases, while Clusters 1, 2, and 3 display a mixed distribution. Cluster 3 has the highest proportion of cancer to non-cancer cases, potentially indicating a patient group at elevated risk of gastrointestinal cancer. Similarly, Cluster 2 shows a higher proportion of rectal cancer cases relative to non-cancer cases, suggesting a subgroup at increased risk of rectal cancer.

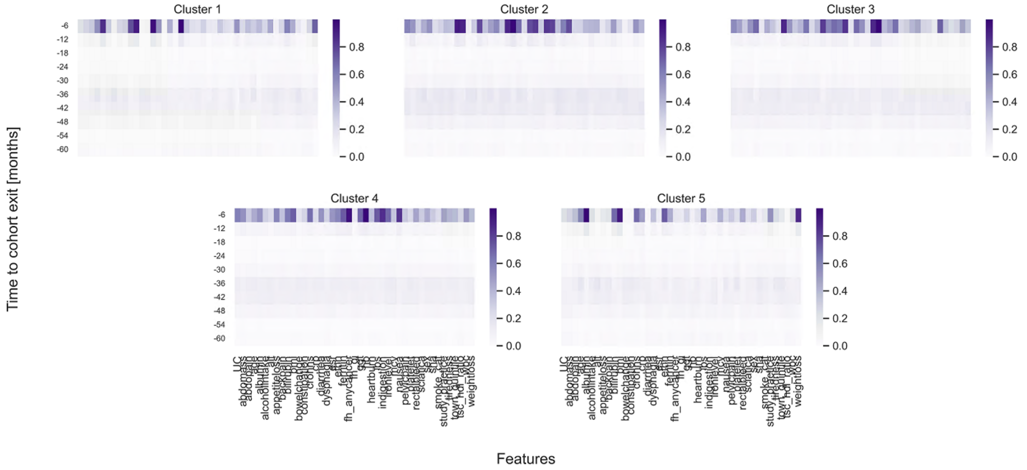


Figure 8: Attention heatmaps for each cluster learnt by CAMELOT. For each cluster, the attention maps show the class-weighted averaged attention scores across time-varying features and time windows. The weights are normalised to a 0.0 to 1.0 range. Higher attention indicates that changes in those feature–time pairs played a greater role in the model's decision to assign patients to that cluster. The x-axis shows clinical features, the y-axis shows time to cohort exit in 6-month intervals (top = most recent). Higher attention is concentrated in the final 6 months prior to exit. The most highly attended features are consistent across the five clusters and include Alcohol Intake, Age, Tiredness, Haemoglobin (Hb), and Crohn's disease.

F Cluster Contingency Matrices

Contingency Matrices. Table 8 and 7 show the empirical number of outcome distributions observed for each cluster learnt by the proposed model and CAMELOT, respectively. We suppress (*) counts where the number of patients is smaller than or equal to 5 for privacy and data protection reasons.

Table 7: Contingency matrix for each learnt cluster for GI-Clust.

Cluster	1	2	3	4	5
Colon	22	438	34	315	168
Rectal	7	180	*	45	183
Gastric	40	69	*	55	12
Oseophageal	134	121	6	29	7
No Cancer	582	26,805	17,713	1,135	593

Table 8: Contingency matrix for each learnt cluster for CAMELOT.

Cluster	1	2	3	4	5
Colon	21	230	224	490	12
Rectal	*	92	219	104	*
Gastric	*	26	34	116	*
Oesophageal	*	39	69	184	*
No Cancer	8,118	12,836	12,454	4,439	8,981