Imperial College London

Learning a meaningful latent space representation for patient risk stratification: model development and validation for dengue

On behalf of the Vietnam ICU Translational Applications Laboratory (VITAL) investigators (vital.oucru.org)

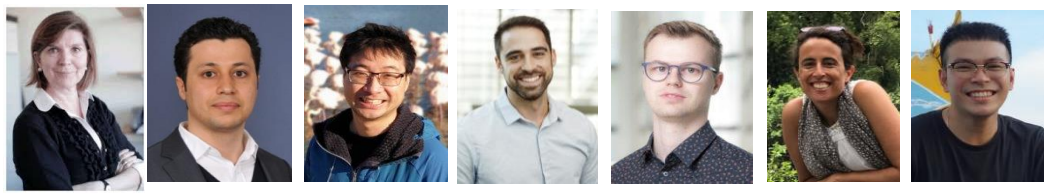
Bernard Hernandez
b.hernandez-perez@imperial.ac.uk
Center for Bio-Inspired Technology
Imperial College London

23rd of April 2022

32nd ECCMID EUROPEAN CONGRESS OF CLINICAL MICROBIOLOGY AND INFECTIOUS DISEASES
Lisbon, Portugal
23–26 April 2022

INTRODUCTION

"The aim of the work package is the **development and roll-out of a clinical decision support system** (CDSS) for the **early identification of severe dengue** in hospital settings, namely dengue shock and other relevant clinical outcomes. The CDSS module will be based on previous work done by the group for supporting antimicrobial prescribing in acute care and will **utilise machine-learning based algorithms** in order to support different aspects of decision-making in dengue management."



+ more!

Imperial College London

GRAPHICAL ABSTRACT

AIM: Development of a clinical decision support tool to support clinical management of patients with (or under suspicion of) dengue using unsupervised techniques to reduce data complexity and facilitate visualisation.

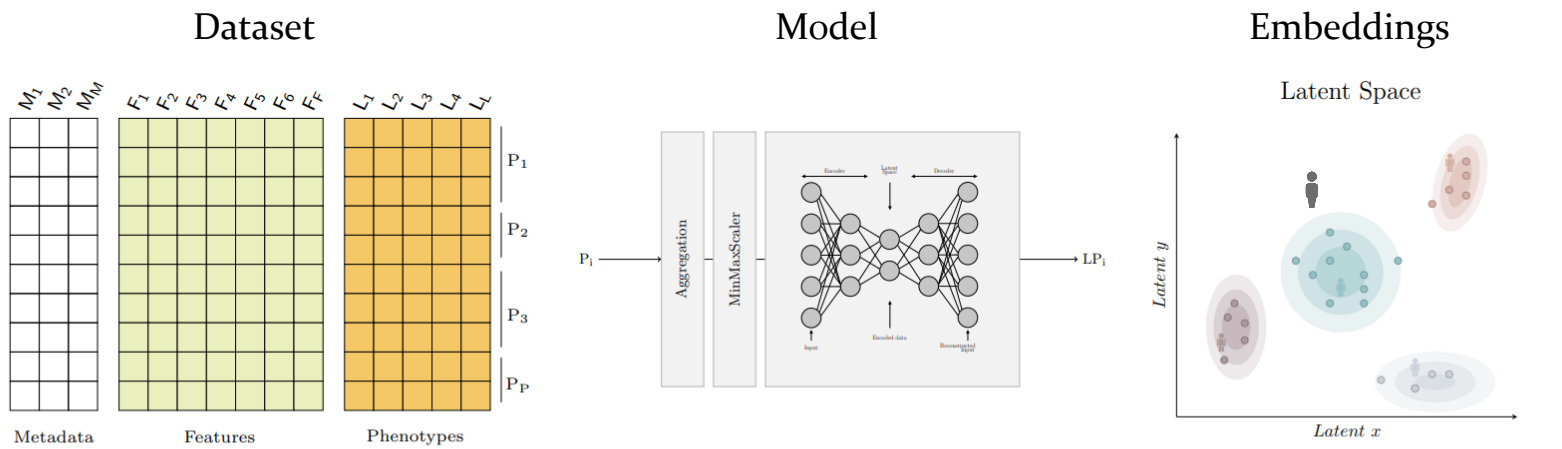


Figure 1: Graphical abstract. On the left, the dataset with metadata, features and phenotypes where each row represents a daily patient profile. In the middle, the model that transforms a patient stay with one or more daily profiles (P_i) into a two dimensional embedding (LP_i) for visualisation. The aggregation step is used to describe the worst patient status using the aggregation functions shown in Table 1. The embeddings are obtained using autoencoders. On the right, the latent space where similar patients are grouped together. Each point represents a patient and the shaded areas represent the density distribution; that is, the concentration of patients for which the phenotype of interest occurs. Note that the latent space can be used to visualise any feature or phenotype of interest.

THE DATASET

The dataset used in the study consists of an **aggregation** of prospective clinical data conducted at the Hospital of Tropical Diseases (HTD) and collaborator hospitals in Ho Chi Minh City, Vietnam by Oxford University Clinical Research Unit (OUCRU) between 2000 and 2021.

Code	Year	Population	Type of care	# patients
06DX	2009-2011	A&C	Inpatient	318
13DX	2010-2014	Children	Outpatient	8107
32DX	2013-2016	A&C	Inpatient and ICU	75
42DX	2016-2018	A&C	Inpatient and ICU	664
DF	1999-2009	Children	Inpatient and PICU	1719
DR	2005-2008	Children	Outpatient	1165
FL	2006-2009	A&C	Inpatient	740
MD	2001-2009	Children	Inpatient	3044
01NVA	2020-2021	Children	Inpatient	150*

[1] Only children (under 18 years old) have been considered since they were the most commonly represented in the datasets and there are separate paediatric and adult dengue guidelines.

[2] Dengue diagnosis was defined as one of i) a positive NS1 point of care assay or NS1 ELISA, ii) positive reverse transcriptase-polymerase chain reaction (RT-PCR), iii) positive dengue IgM through acute serology, iv) or seroconversion of paired IgM samples where available.

12,884
patients[1]

19,516
complete daily profiles

4,344 (33.7%)
diagnosed with dengue[2]

1. **Vital-oucru-clinical** documentation (https://bahp.github.io/vital-oucru-clinical/)

Imperial College London

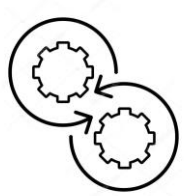
SELECTING THE APPROACH

Supervised vs Unsupervised: important **considerations**

Needs the **ground truth** or **label** for which the model is optimised.

It returns a probability which might be difficult to **interpretate**.

Needs to be **(re)trained for each** label or **phenotype**.



Outputs from **different algorithms** (e.g. shock and fluid accumulation) **must be consistent**.

Lacks **explainability**; that is, what are the underlying reasons motivating such probabilities.

Relatively **easy evaluation** through standard metrics such as ROC, SENS or SPEC.

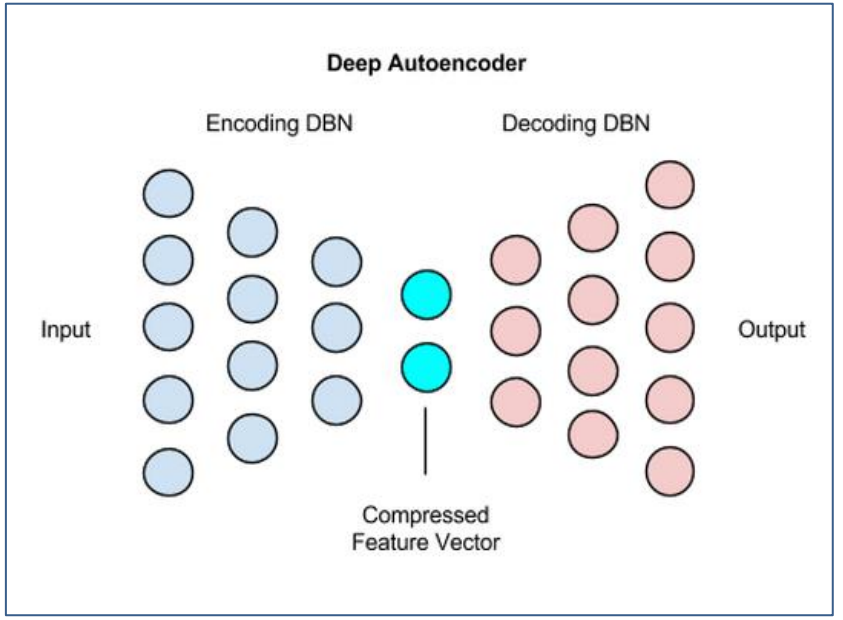
SELECTING THE MODEL

The **input features** selected are:
- Age (years)
- Weight (Kg)
- Body Temperature (°C)
- Platelets (k/μ L)
- Haematocrit (%)

- Features were consistently recorded in all clinical studies.
- Provide information on the course of dengue illness (WHO).
- Feasibility of prospective collection

Algorithm	Type	Metrics	Comments
PCA	Parametric	Good distance metrics. Density metrics are decent but inferior to SOM/AE.	Performance of PCA is likely to decrease as dimensionality increases due to its linear nature.
t-SNE	Non-parametric	Performance highly dependent on hyperparameters	High computationally expensive to create embeddings for **unseen data.** Can create completely separated clusters.
SOM	Non-parametric	Poor distance preservation. Good density metrics.	Limitations imposed by the **discrete space** limit its usability for similarity retrieval.
AE	Parametric	Good performance for both distance and density metrics.	Quite a flexible approach, possible to use with time-series signals or even images.

[1] UMAP might be also a promising algorithm.



PERFORMANCE METRICS / RESULTS

Table 3
Evaluation metrics.

Type	Metric	Aim	
Distance	Sheppard	distance preservation	na
	Pearson	distance preservation	↑
	Spearman	similarity retrieval	↑
	Procrustes	information loss	↓
Density	convex hull ratio	good visualisation	↓
	concave hull ratio	good visualisation	↓
	GMM ratio	good visualisation	↓

GMM: Gaussian Mixture Model
↑ Higher values are better
↓ Lower values are better

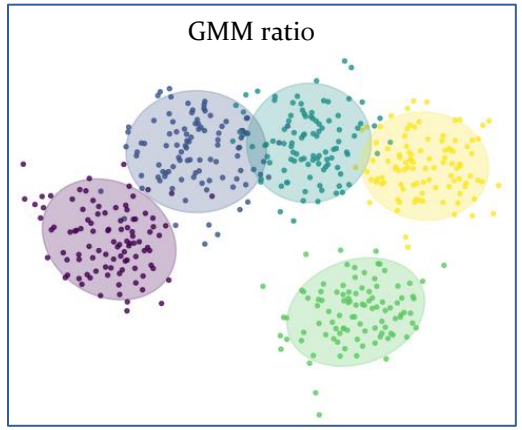
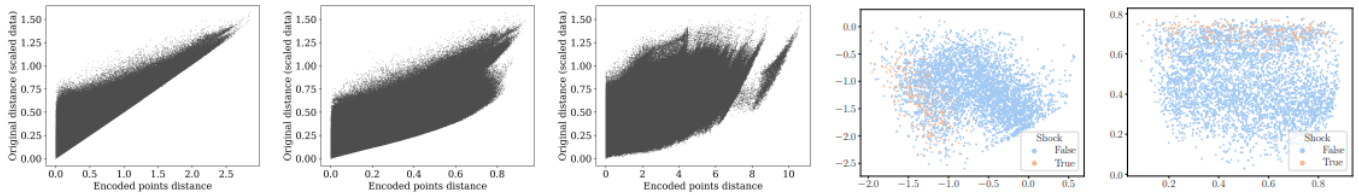


Table 6
Evaluation metrics for various representative hyperparameter configurations

Layers	Activation	Pearson	Spearman	Procrustes	GMM	Comments
-	-	0·916	0·896	0·272	0·814	PCA
[ ]	ReLU	0·940	0·920	0·226	0·695	The approximate linearity of the ReLU activation function of this model favours distance preservation.
[ ]	Sigmoid	0·917	0·906	0·240	0·543	The non-linearity of the Sigmoid activation affects distance metrics slightly and improves density metrics.
[3]	Sigmoid	0·840	0·830	0·301	0·321	It balances distance preservation and density metric results.
[5,4,3]	ReLU	0·635	0·622	0·505	0·104	It is a complex model with good density metric results but produces dense points in the latent dimension not apt for visualisation of patient trajectories over time. In addition, distance metric results show that distances are not preserved and therefore it is inadequate for similarity-based retrieval.

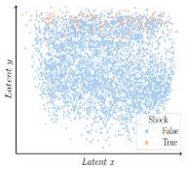


(a) () ReLU **(b)** (3) Sigmoid **(c)** (5,4,3) ReLU **(d)** (3) ReLU **(e)** (3) Sigmoid

Figure 3: Sheppard diagrams (left) and shock label projections (right). On the left, Sheppard diagrams obtained for autoencoders with three different configurations. On the right, distribution of patients in the latent space with (orange) and without (blue) shock.

EVALUATION (I) - EMBEDDINGS

Allows easy and complete **visualisation** of all the patients over the latent space.



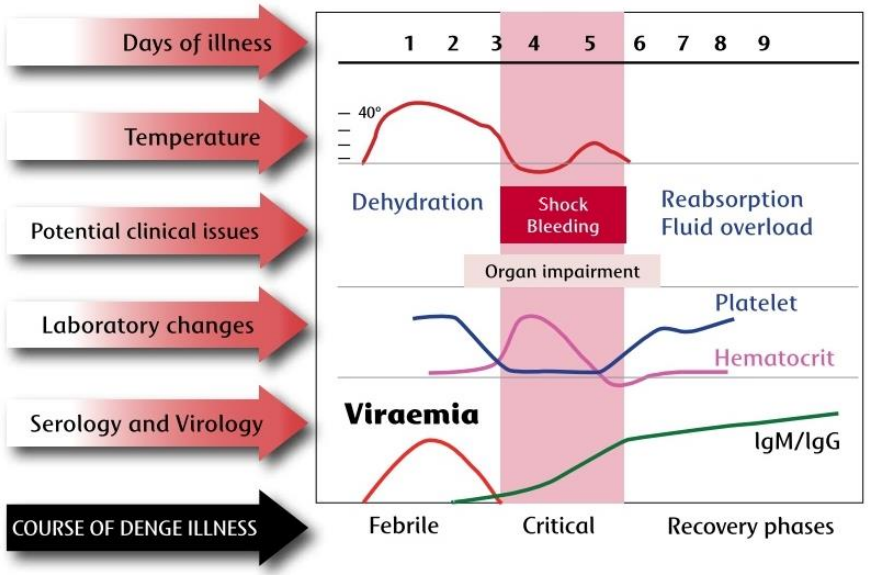


Figure 1. The course of dengue illness diagram. The figure, which has been adapted from WCL Yip, et all 1980 [28], presents phases, lab results, and associated problems.

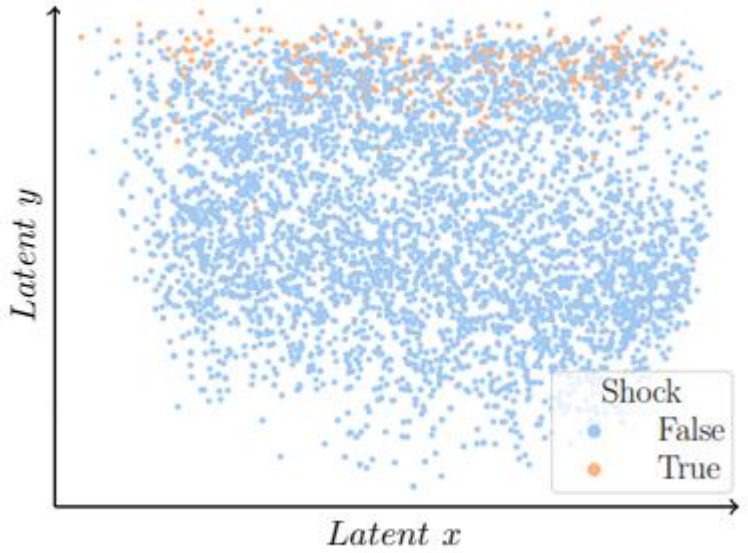


Figure 2. Latent space: embeddings. The worst patient status for all the patients has been projected into the latent space with the shock phenotype.

EVALUATION (II) - FEATURES

Aligns with established **characteristics of disease progression** such as increase in haematocrit levels, decrease in platelet levels and decrease in body temperature from febrile to critical phase.

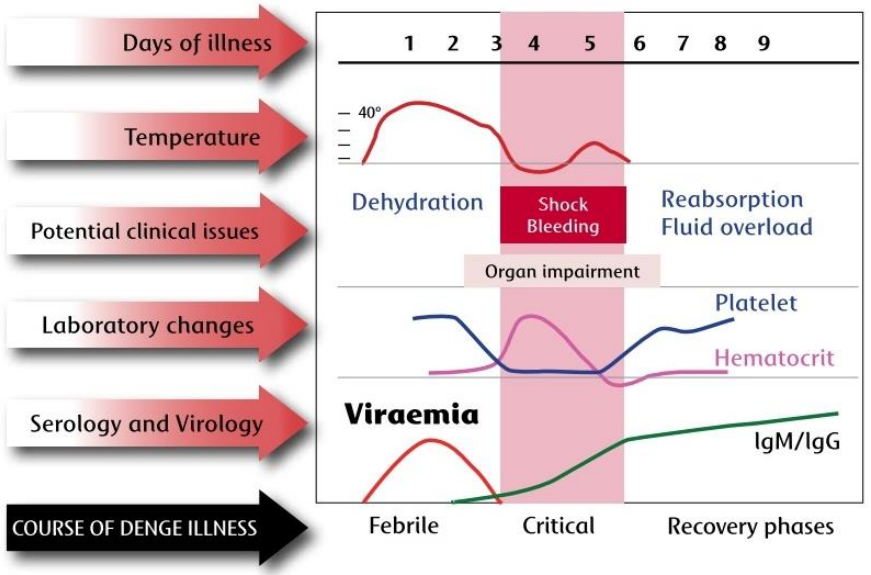


Figure 1. The course of dengue illness diagram. The figure, which has been adapted from WCL Yip, et all 1980 [28], presents phases, lab results, and associated problems.

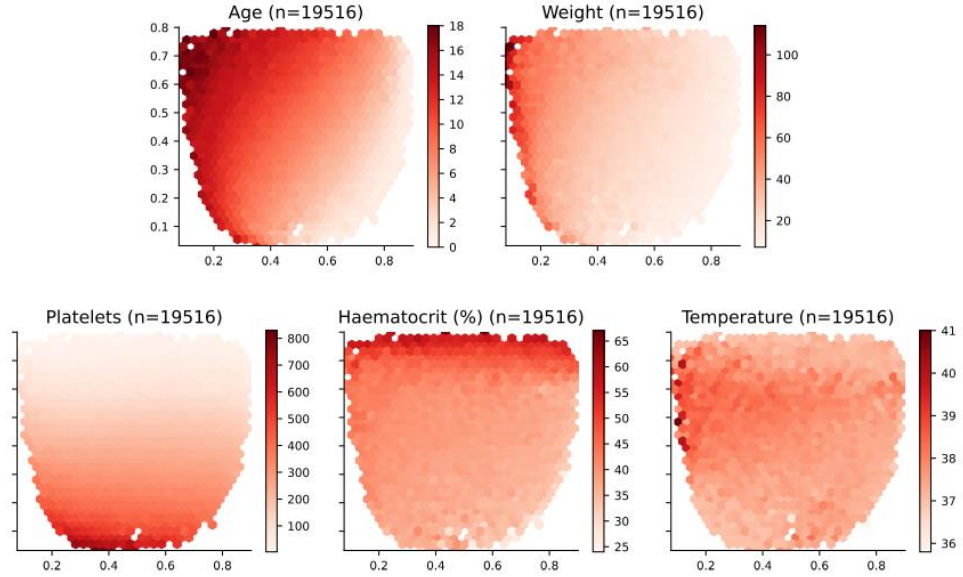
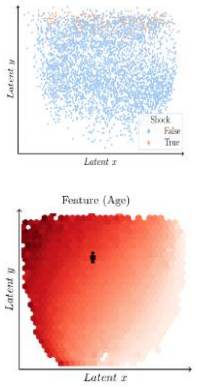


Figure 3. Latent space description: Features. The graphs represent the density distribution using hexagonal binning over the latent space.



EVALUATION (III) - PHENOTYPES

Similar **phenotypes** are presented close to each other.

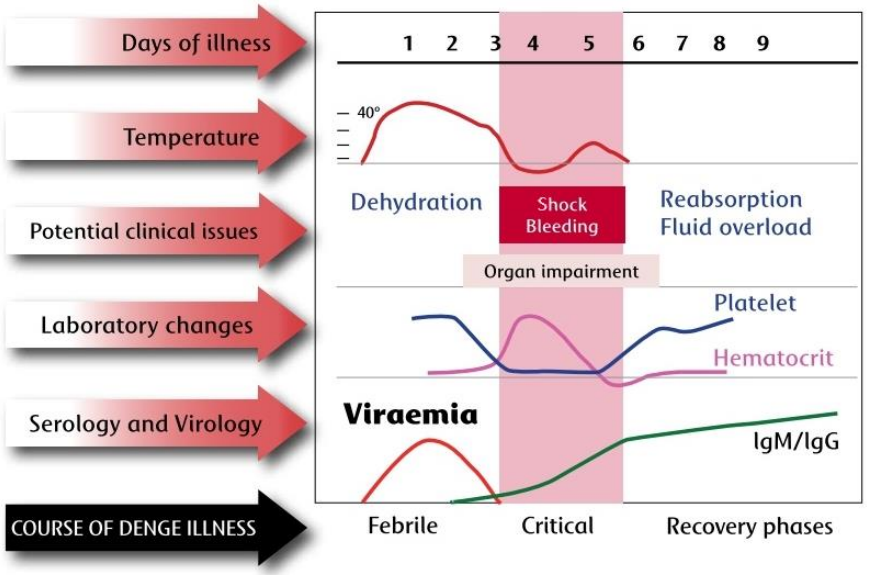


Figure 1. The course of dengue illness diagram. The figure, which has been adapted from WCL Yip, et all 1980 [28], presents phases, lab results, and associated problems.

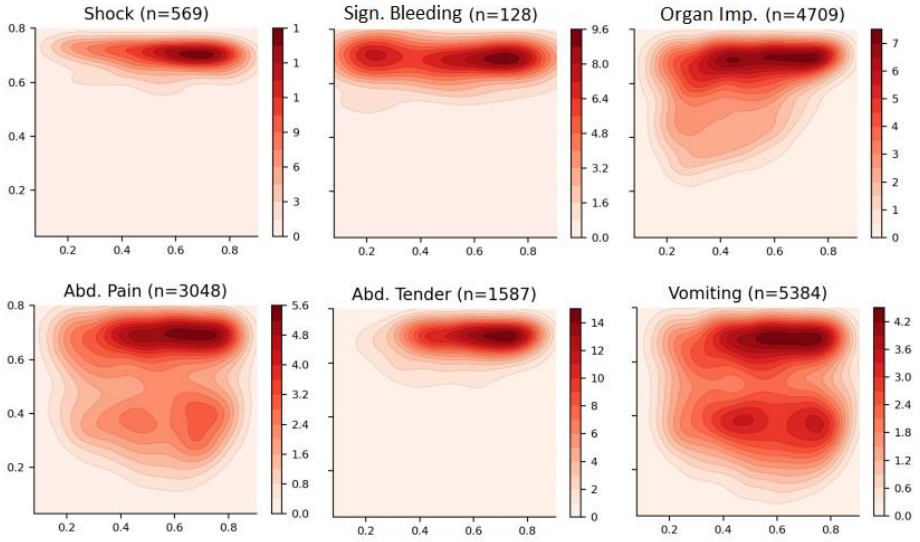
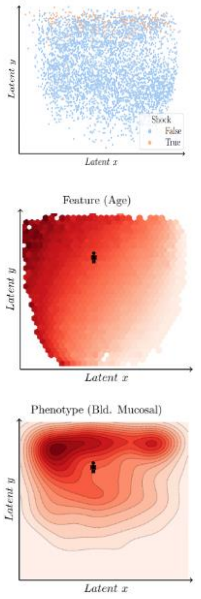


Figure 4. Latent space description: Phenotypes. The graphs represent the density distribution using contour lines using a Gaussian kernel.



EVALUATION (IV) - CATEGORIES

The approach allows to **create categories** as a compendium of phenotypes and facilitates its visualisation. In addition, for the current scenario, the defined categories A, B, and C are consistent with the WHO 2009 guidelines.

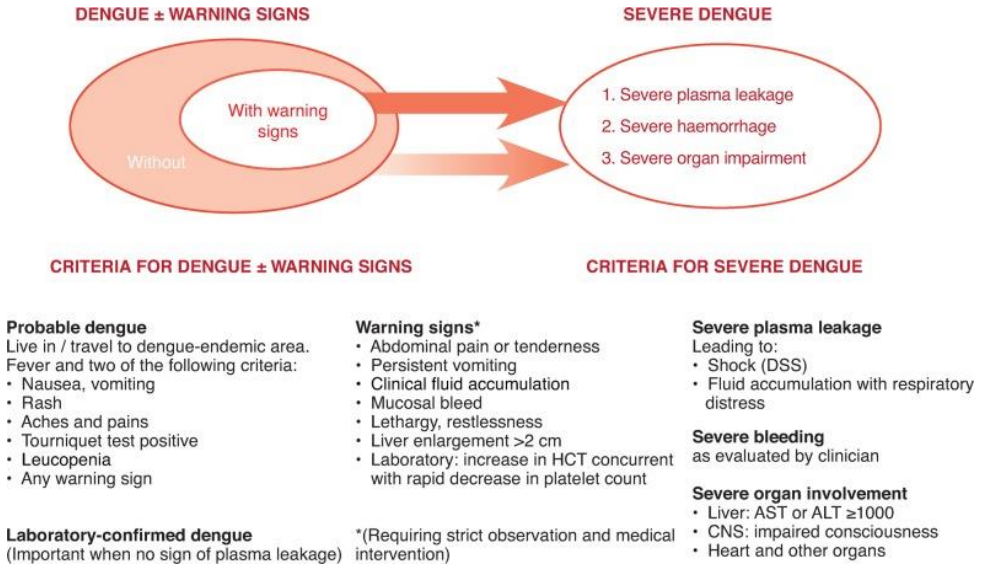


Figure 5. The 2009 revised dengue case classification. Diagram with the 2009 dengue classification system proposed by WHO for the diagnosis and management of dengue [6].

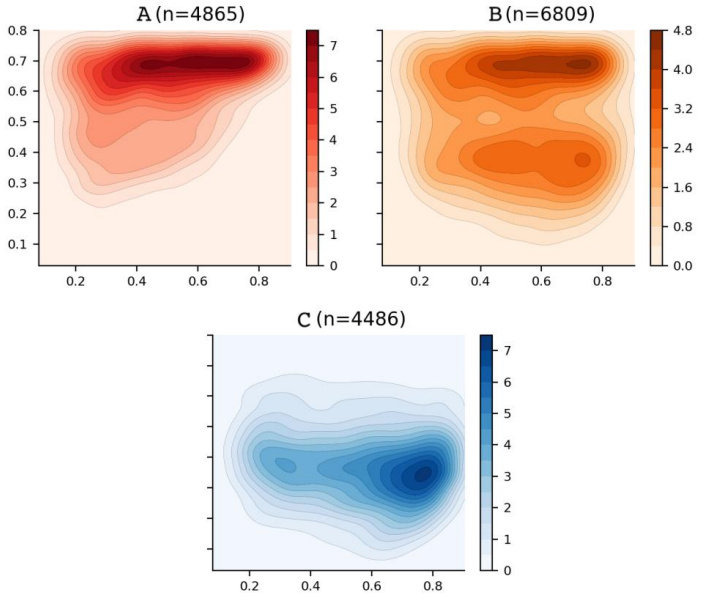
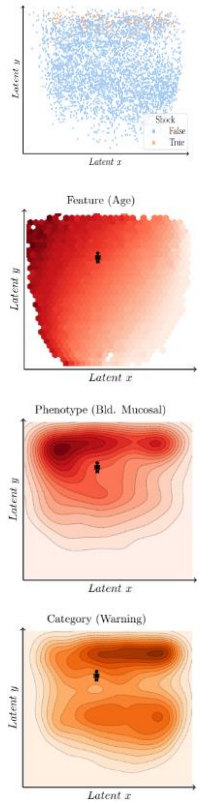


Figure 6. Latent space description: Categories. The graphs represent the density distribution over the latent space for three categories defined as a compendium of various phenotypes.



EVALUATION (V) – TRAJECTORIES AND RETRIEVAL

Allows to visualise **patient trajectories** over time.
Allows to retrieve similar patients more efficiently (and easier to understand).

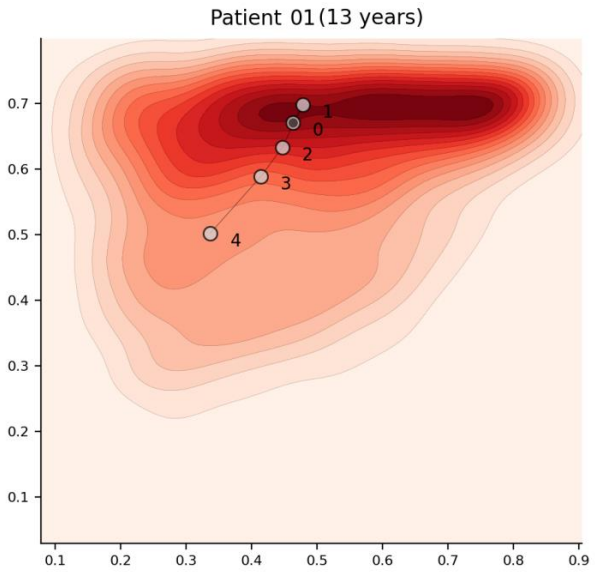


Figure 7. Latent space description: Trajectories. The graphs represent the trajectory of a patient over Category A. The number represents the day from admission.

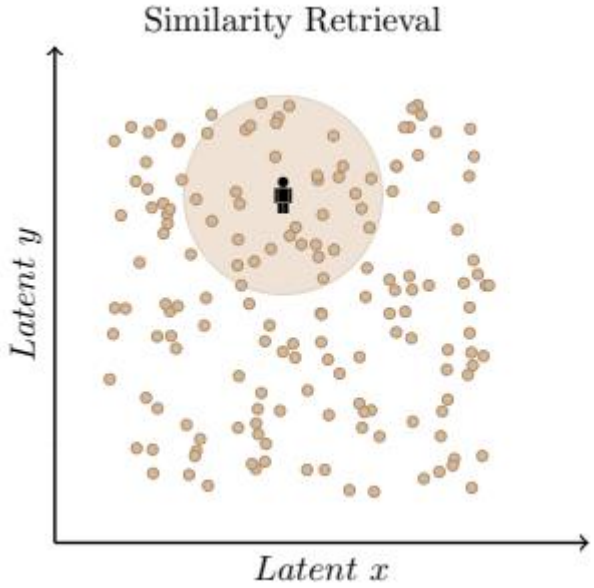
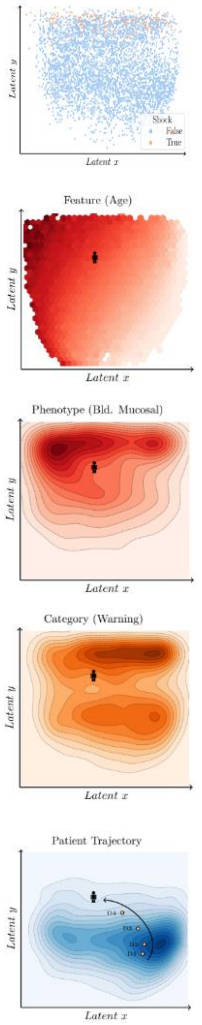
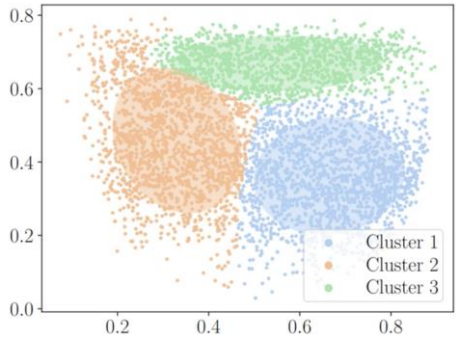


Figure 8. Latent space description: Similarity retrieval. The diagram represents a query patient and the area of interest for which patients, which are deemed to be similar, should be retrieved.



Imperial College London

EVALUATION (VI) – DEMOGRAPHICS TABLE

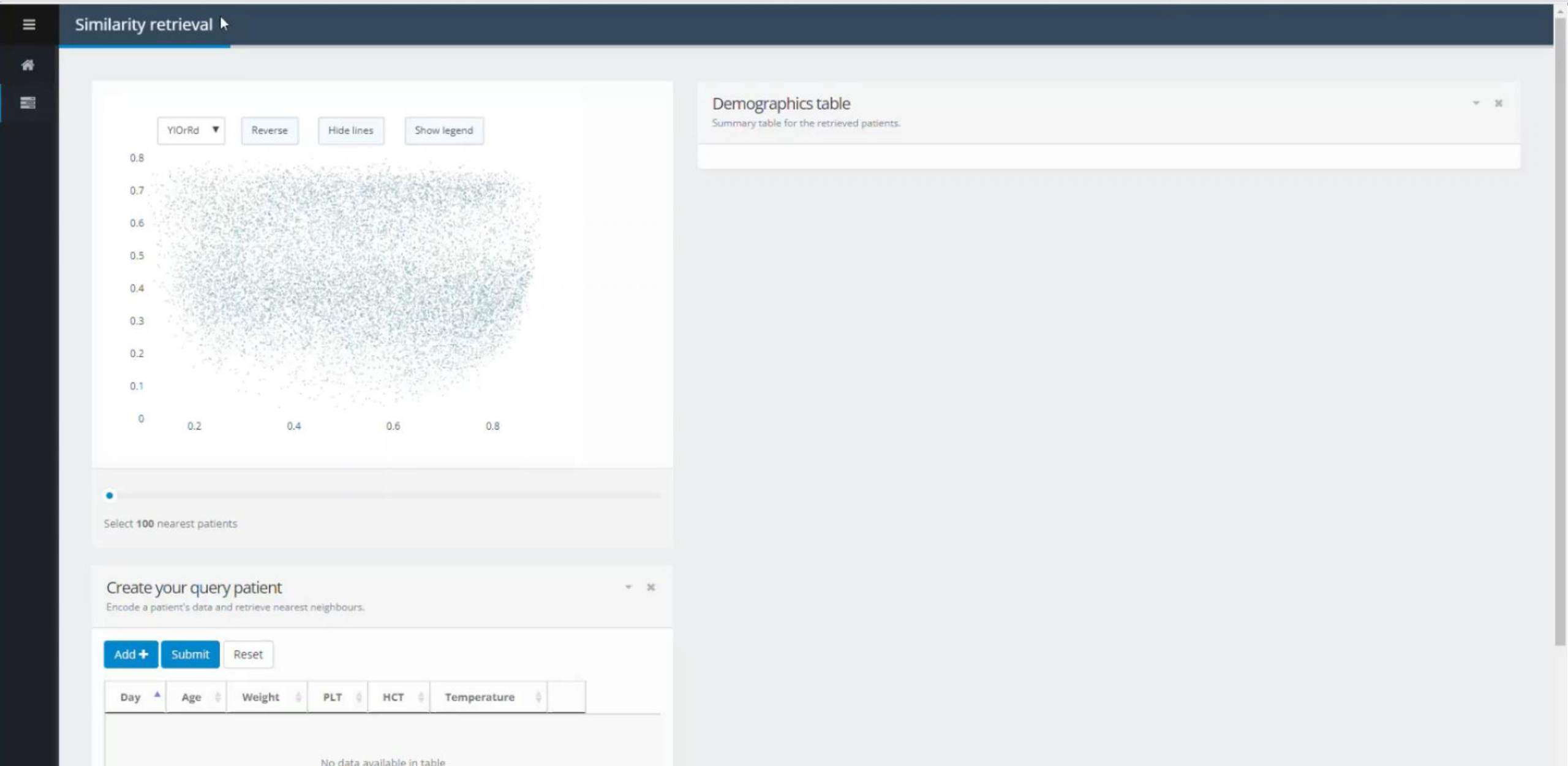


		Overall	Cluster 1	Cluster 2	Cluster 3
n		14484	5588	5017	3879
abdominal_pain, n (%)	False	9878 (68.2)	4533 (81.1)	3780 (75.3)	1565 (40.3)
	True	4606 (31.8)	1055 (18.9)	1237 (24.7)	2314 (59.7)
ascites, n (%)	False	12153 (83.9)	5147 (92.1)	4001 (79.7)	3005 (77.5)
	True	2331 (16.1)	441 (7.9)	1016 (20.3)	874 (22.5)
bleeding, n (%)	False	10760 (74.3)	5133 (91.9)	3831 (76.4)	1796 (46.3)
	True	3724 (25.7)	455 (8.1)	1186 (23.6)	2083 (53.7)
bleeding_gum, n (%)	False	12895 (89.0)	4844 (86.7)	4372 (87.1)	3679 (94.8)
	True	1589 (11.0)	744 (13.3)	645 (12.9)	200 (5.2)
bleeding_mucosal, n (%)	False	11818 (81.6)	5387 (96.4)	4432 (88.3)	1999 (51.5)
	True	2666 (18.4)	201 (3.6)	585 (11.7)	1880 (48.5)
bleeding_skin, n (%)	False	7864 (54.3)	4820 (86.3)	2490 (49.6)	554 (14.3)
	True	6620 (45.7)	768 (13.7)	2527 (50.4)	3325 (85.7)
gender, n (%)	Female	6327 (43.7)	2563 (45.9)	1922 (38.3)	1842 (47.5)
	Male	8157 (56.3)	3025 (54.1)	3095 (61.7)	2037 (52.5)
shock, n (%)	False	13783 (95.2)	5576 (99.8)	4905 (97.8)	3302 (85.1)
	True	701 (4.8)	12 (0.2)	112 (2.2)	577 (14.9)
age, median [Q1,Q3]*		8.0 [5.0,11.0]	4.0 [3.0,6.0]	11.0 [10.0,13.0]	10.0 [8.0,12.0]
temperature, median [Q1,Q3]*		37.6 [37.2,38.3]	37.4 [37.2,37.8]	37.9 [37.4,38.5]	38.0 [37.0,38.8]
hct, median [Q1,Q3]*		40.3 [37.2,45.0]	37.2 [35.1,39.2]	41.0 [38.6,44.0]	46.4 [43.0,50.0]
plt, median [Q1,Q3]*		169.0 [71.0,243.0]	229.0 [182.0,279.0]	182.0 [109.0,243.0]	46.0 [30.0,69.0]
weight, median [Q1,Q3]*		26.0 [19.0,37.0]	18.0 [14.0,22.0]	38.0 [30.0,46.0]	29.0 [22.0,36.0]

1. Work carried out by **Oliver Stiff (MEng)**, clusters identified through **GMMs.**
2. * indicates attributes used in training.

Imperial College London

THE SUPPORT TOOL



QUESTIONS

Bernard Hernandez
b.hernandez-perez@imperial.ac.uk
Center for Bio-Inspired Technology
Imperial College London

23rd of April 2022

Imperial College London

APPENDIX - FEATURES

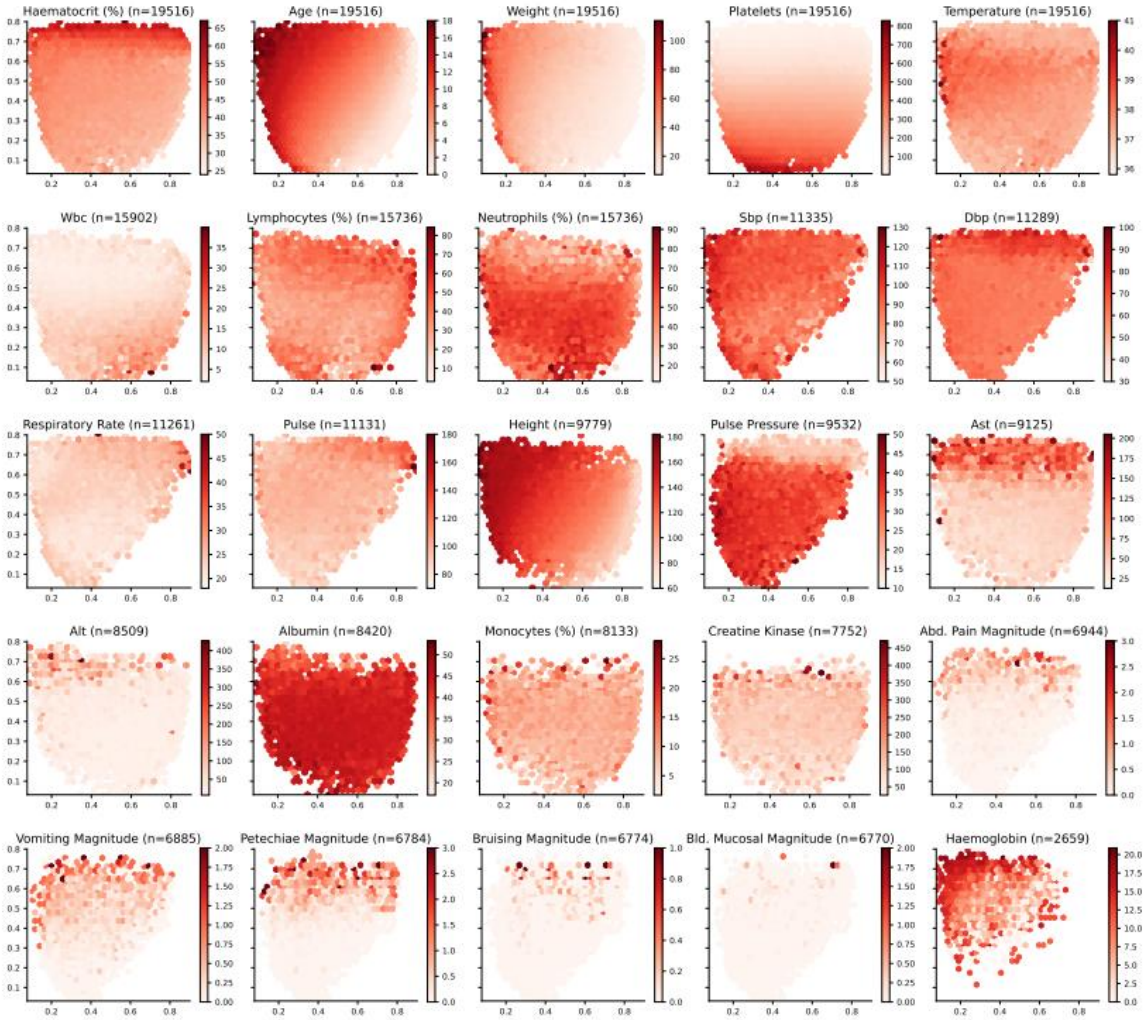


Figure 8: Latent space description: Biomarkers. The graphs represent the density distribution using hexagonal binning over the latent space for the most frequent features in the dataset. The title includes the name and the number of daily profiles. The value on each hexagonal bin represents the mean value of all the daily profiles that have been projected on that bin. The magnitude values are subject to the clinicians interpretation and have been included for reference.

*

APPENDIX - PHENOTYPES

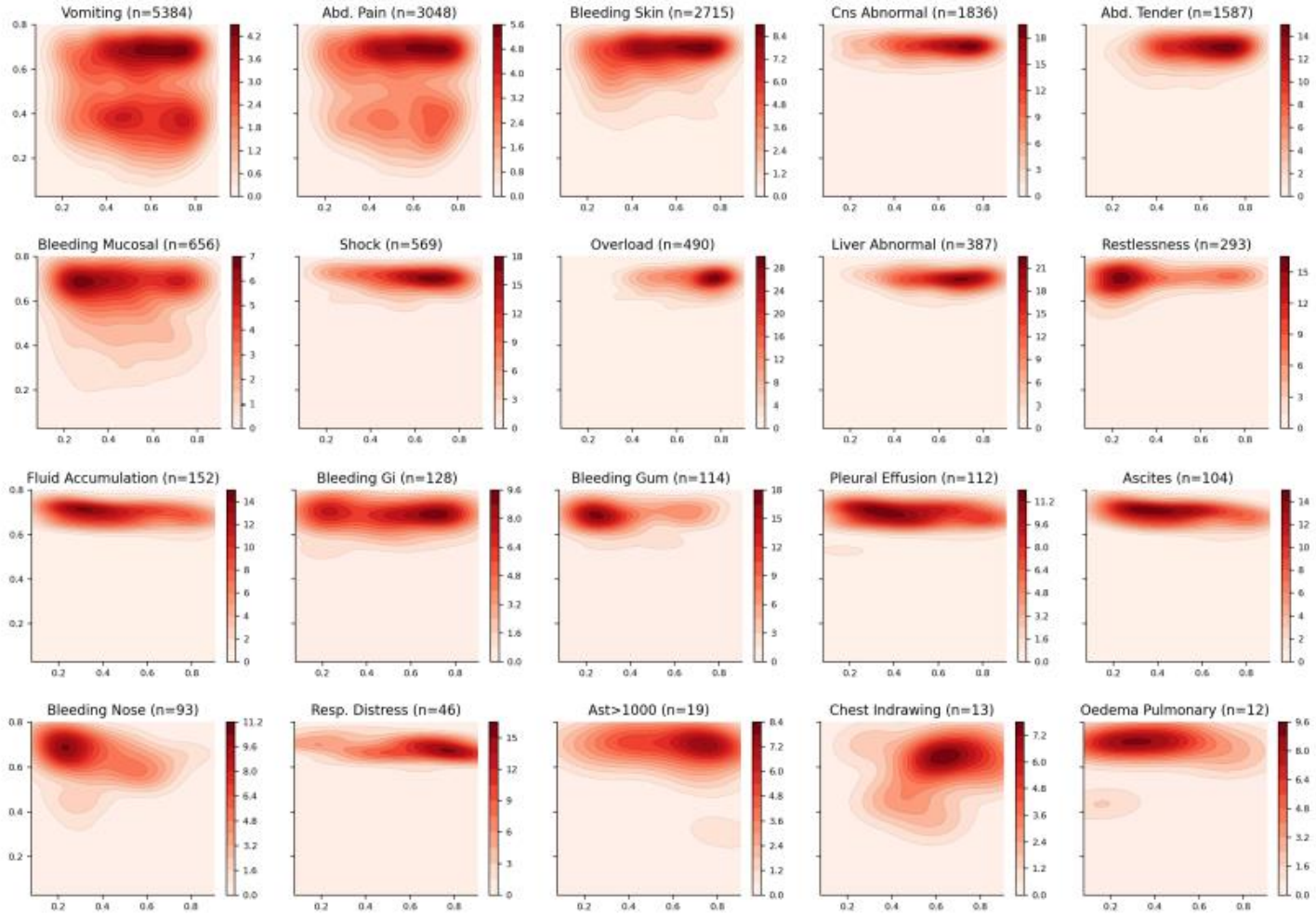


Figure 9: Latent space description: Phenotypes. The graphs represent the density distribution using contour lines estimated using a Gaussian kernel over the latent space. The title includes the phenotype and the number of patients in which it occurs.

Imperial College London

APPENDIX - TRAJECTORIES

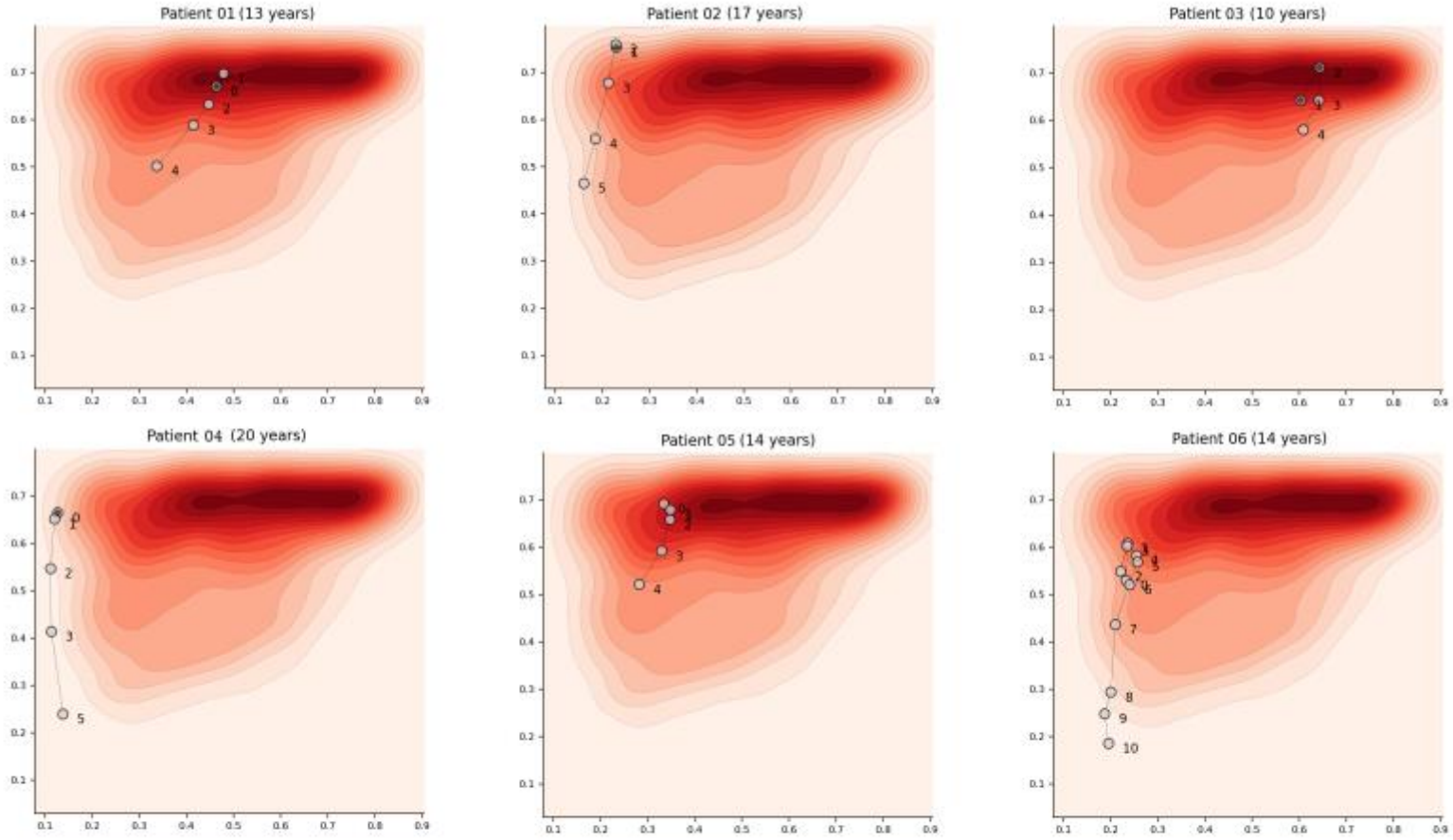


Figure 10: Latent space description: Trajectories. The graphs represent trajectory of patients over the latent space using the density distribution for the *Severe* category as a background reference. Each marker represents a daily profile where the number indicates the day from admission. Filled markers indicate days in which the patient suffered an episode of shock.

CBR MODULE

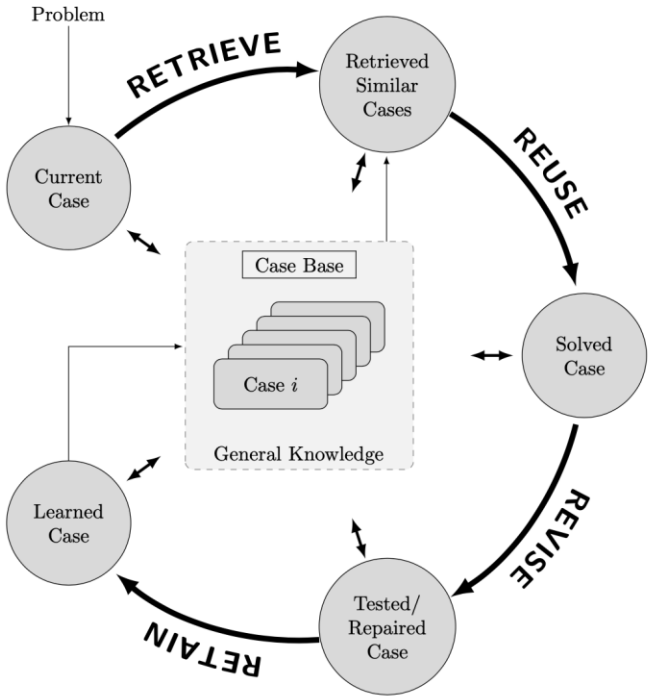


Figure 5.1: The CBR cycle. Diagram showing the different phases for a cycle within the case-based reasoning methodology as outlined by Aamodt and Plaza [7].

The E. Coli case study

145 patients (all received antibiotics)

Antimicrobial Spectrum index (ASI)

Physicians	- 83% appropriate
CBR	- 90% appropriate

*

B Hernandez et al *Data-driven web-based intelligent CDSS for infection management…* – Health Informatics (2017)
TM Rawson et al *A real-world evaluation of CBR to support antimicrobial prescribing…* - CID (2020) *