1 Dataset Statistic Description

Table 1: Demographics and clinical characteristics for train vs. test cohorts of the dataset.

Variable	**Training cohort** (no. patients=49.870) (no. visits=78,142)	**Testing cohort** (no. patients=5,542) (no. visits=13,100)
Inpatient data		
LOS, days median [IQR]	3.0 [1–8]	4.0 [2–10]
In-hospital mortality, n (%)	1354 (2.3%)	109(1.6%)
Isolated room, n (%)	13272 (26.6%)	1843 (33.3%)
Readmission next 30 days, n (%)	3912 (7.8%)	569 (10.3%)
No. screens per CPE+ patient, median [IQR]	1.0 [0.0–2.0]	2.0 [1.0–3.0]
No. diagnosis per admission, median [IQR]	6.0 [4.0–11.0]	8.0 [5.0–19.0]
No. procedures per admission, median [IQR]	1.0 [1.0–2.0]	1.0 [1.0–3.0]
No. admissions per patient, median [IQR]	1.0 [1.0–2.0]	1.0 [1.0–3.0]
Sex, n (%)		
Male	22812 (45.7%)	2533 (45.7%)
Female	27057 (54.3%)	3009 (54.3%)
Area of residence, n (%)		
America	103 (0.2%)	3 (0.1%)
European	284 (0.6%)	13 (0.2%)
Irish	49223 (98.7%)	5500 (99.2%)
Others	260 (0.5%)	26 (0.5%)
Billing codes, n (%)		
Number of unique ICD-10-CM (full)	6,707	3,768
Number of unique ICD-10-PCS (full)	2,309	1,105
Number of unique DRG codes	710	82
Billing codes Grouped, n (%)		
Number of unique ICD-10-CM (grouped)	1,604	1,198
Number of unique ICD-10-PCS (grouped)	1,436	768
Transfer data, n (%)		
Emergency ward admission, n (%)	1843 (3.69%)	247 (4.456%)
ICU admission, n (%)	766 (1.5%)	108 (1.9%)

Table 2: Comparison of Network Features Between Non-Infected and Infected Individuals

Feature	**Non-Infected Mean (SD)**	**Infected Mean (SD)**
Degree	25.997 (7.806)	26.334 (7.964)
Degree Centrality	0.0394 (0.0120)	0.0393 (0.0121)
Infected Degree	0.1090 (0.3632)	0.2521 (0.5421)
Infected Contact Days	0.1090 (0.3632)	0.2521 (0.5421)
Infected Contact Days Centrality	0.0002 (0.0005)	0.0004 (0.0008)
Clustering Coefficient	0.9972 (0.0514)	0.9939 (0.0778)
Closeness Centrality	0.0395 (0.0120)	0.0393 (0.0121)
Infected Closeness Centrality	0.0000 (0.0000)	0.0647 (0.1718)
Betweenness Centrality	0.0000 (0.0001)	0.0000 (0.0000)
PageRank	0.0015 (0.0003)	0.0015 (0.0002)
K-core Number	25.988 (7.791)	26.334 (7.964)

2 Description of the used data features

Patient demographics
Gender `[C]`
Patient's sex. Male, Female or Other

Area of Residence `[C]`
The patient's place of living, includes Dublin Districts, other Ireland counties, and a small amount of other European countries.

Patient episode information
No. readmissions so far `[N]`
Total number of readmissions prior to the current episode

Is Isolated `[C]`
Boolean value whether the patient was in an isolation room in the current episode

Is screened `[C]`
Boolean value whether patient is screened for CPE during the episode

LOS `[N]`
Length of stay of the current episode

Discharge Description `[C]`
Episode's outcome

Discharge Specialty `[C]`
Episode's discharge ward specialty

Clinical code information
No. Diagnosis `[N]`
Number of diagnoses that patient received during the episode

No. Procedures `[N]`
Number of procedures that patient received during the episode

Diagnosis Codes `[T]`
All diagnosis codes received by the patient, following the Australian Modification (AM) of ICD-10

Procedure Codes `[T]`
All procedure codes received by the patient, following Irish Coding Standards (ICS)

Ward transition information
Admission Ward `[C]`
The codename of the ward to which the patient was admitted

Discharge Ward `[C]`
The codename of the ward on which the patient was discharged

No. Acute Department visits `[N]`
Number of acute departments that the patient visited during the current episode. In our study, we refer to wards that admit patients directly from the Emergency Department as Acute Departments.

No. Emergency Department visits `[N]`
Number of emergency departments that patient visited during the current episode

No. Wards visits `[N]`
Number of ward transitions during the episode

3 Description of the used contact network data features

Contact-network variables

Infected degree `[N]`
Total number of direct contacts with CPE+ patients.

Infected degree centrality `[N]`
Days spent in direct contact with CPE+ patients, normalised by the total number of patients in hospital.

Infected closeness centrality `[N]`
Measure of network distance relative to all CPE+ patients, normalised by the total number of patients in hospital.

Degree `[N]`
Total number of direct contacts.

Degree centrality `[N]`
Days spent in direct contact with other patients, normalised by the total number of patients in hospital.

Closeness centrality `[N]`
Measure of network distance relative to all other patients, normalised by the total number of patients in hospital.

Clustering coefficient `[N]`
Measure of likelihood that a patient is also connected to the contacts of their immediate contacts.

Betweenness centrality `[N]`
Centrality of a patient with respect to the shortest paths through the contact network.

Page rank `[N]`
Centrality measure of a patient in the contact network given by the importance of their neighbours in the network.

K-core number `[N]`
Measure of how central a patient is to the most connected region of the graph

4 Temporal drift analysis

Temporal drift—shifts in data distribution or model performance over time—is a critical concern in clinical machine learning models, especially for deployment in real-world hospital environments. To assess the stability and generalizability of our models across time, we conduct a temporal drift analysis using chronologically split data segments. Specifically, admissions from January 2018 to December 2020 are used to train the model, and admissions from Q1 2021 are used as a validation set. We then divide the remaining data into four consecutive, non-overlapping quarterly test sets. We report the results of 3 models (TabTransformer, TabNet, and ResNet) on 3 prediction tasks in Figure 1. It can be seen that the performance is quite stable across timeframe, indicating that the timedrift distribution does not have meaningful impact.

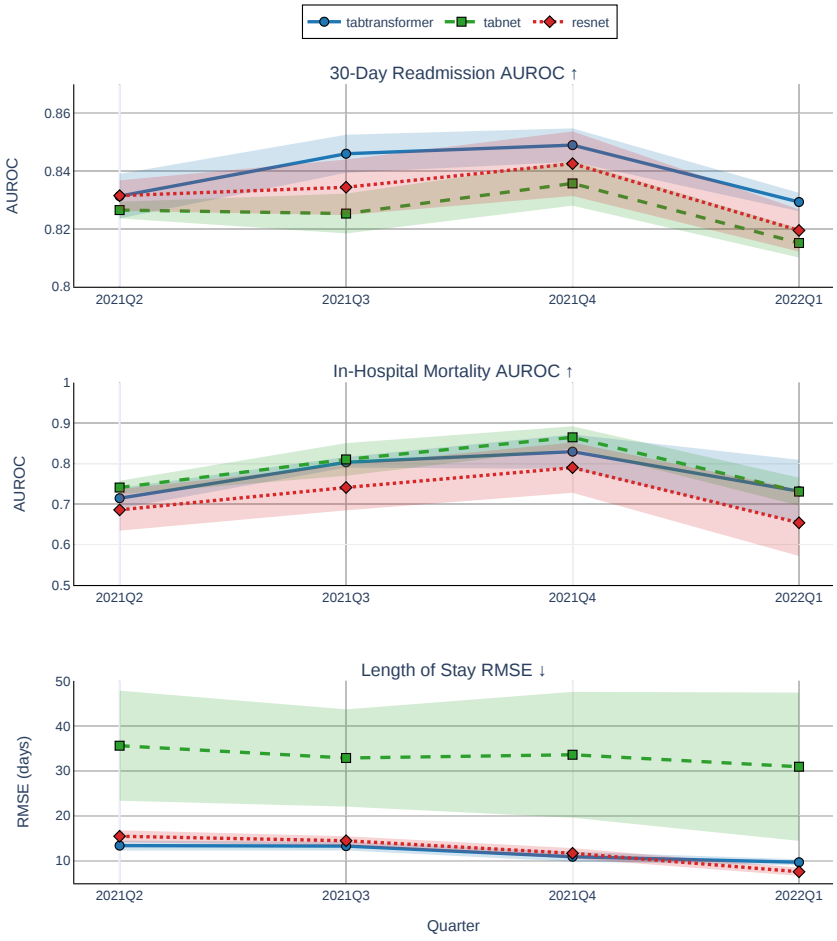

Figure 1: Temporal drift analysis showing model performance across four quarterly time intervals (Q2 2021 to Q1 2022). Each line represents one model, with performance measured by AUROC for classification and RMSE for regression.

5 Feature Ablation Study

We conducted an ablation study to assess the contribution of key modeling components, including pretraining, contact network-derived features, class balancing via sampling, and Focal Loss. Using a fixed model and a representative fold, we systematically removed each component to quantify its individual impact on predictive performance.

Table 3 presents results from an ablation study conducted to isolate the effect of four key modeling components: contact network features, data sampling for class balance, and Focal Loss. We trained three representative deep models (TabTransformer, TabNet, and ResNet) under different ablation settings, using a fixed validation fold for consistency. Each row shows the model's AUROC performance when one component is removed, compared to the full setup ("All features"). The Δ values indicate the change in performance relative to the full-feature baseline.

Table 3: Feature ablation results: change in performance Δ relative to full model setup for each component. Results are from a single validation fold. '↓' indicates performance drop, '↑' indicates improvement. Only for task 30-day readmission.

Ablation Setting	**30-day Readmission (ΔAUROC)**		
	TabTransformer	TabNet	ResNet
All features	**0.834**	**0.827**	**0.830**
wo Contact Network	0.827 (Δ = -0.7%)	0.801 (Δ = -2.6%)	0.822 (Δ = -0.8%)
wo Data sampling	0.803 (Δ = -3.0%)	0.797 (Δ = -3.0%)	0.816 (Δ = -1.4%)
wo Focal Loss	0.830 (Δ = -0.4%)	0.822 (Δ = -0.5%)	0.825 (Δ = -0.5%)

Overall, we observe that removing data sampling has the most substantial negative impact on performance across all models (up to -3.0% AUROC), highlighting its critical role in addressing class imbalance for 30-day readmission prediction. The removal of contact network features also leads to performance degradation, especially for TabNet (-2.6%), suggesting these features provide valuable signals related to transmission dynamics. Interestingly, Focal Loss contributes modest but consistent gains, reflecting its ability to manage skewed label distributions. These findings underscore the importance of model training strategies and structural inputs, particularly in settings with imbalanced clinical data and sparse infection events.

We conducted an ablation study to evaluate the relative contribution of different feature groups to model performance on the 30-day readmission prediction task. Table 4 presents the AUROC scores for the TabTransformer

model under seven distinct feature configurations, with features incrementally added to a base setup comprising only patient demographics.

Table 4: Comparison of effectiveness between features used in training the TabTransformer. The results are models trained using cross-validation and benchmarked on our data testset.

ID	Patient Information	Clinical Codes	Ward Transition	History Episode	Contact Network	AUROC
1	✓					0.620
2	✓	✓				0.782
3	✓		✓			0.622
4	✓	✓	✓			0.788
5	✓	✓		✓		0.823
6	✓	✓	✓	✓		0.827
7	✓	✓	✓	✓	✓	**0.834**

The results indicate that incorporating clinical codes leads to the most substantial performance improvement. Adding this feature group to the patient-only baseline (Model 1 vs. Model 2) increases AUROC from 0.620 to 0.782, highlighting the strong predictive value of structured diagnostic information. This finding aligns with prior research demonstrating the clinical relevance of ICD-coded data for capturing patient comorbidities and risk profiles.

The addition of historical hospitalization data (Model 5) yields a further increase in AUROC, rising to 0.823. This suggests that longitudinal patient records—such as prior diagnoses, procedures, and hospital utilization—play a significant role in predicting readmission risk, likely by capturing temporal disease progression or healthcare behavior patterns. The inclusion of ward transition features alone (Model 3) offers minimal gain, and only modest improvement when combined with clinical codes (Model 4), indicating that ward movement dynamics contribute limited predictive signal in isolation.

Finally, contact network-derived features provide an incremental boost when added to the full feature set (Model 6 vs. Model 7), resulting in the best overall AUROC score of 0.834. While this gain is relatively modest (+0.7%), it underscores the complementary value of modeling in-hospital patient interactions and potential exposure patterns.

Overall, these findings demonstrate that clinical codes and patient history are the most impactful feature sets, while contact network and ward transition features provide additional signal in our modeling framework.

6 CPE-related ICD10-CM codes

Table 5: List of ICD10CM codes for antimicrobial agents that are potential indicator of CPE acquisition.

Diagnosis Code	**Code Description**
Z22	Carrier of infectious disease
B968	Other specified bacterial agents as the cause of diseases classified elsewhere