eMethods

Union Hospital Dataset

The Union Hospital dataset consists of two main tables: (1) hospitalization records containing patient demographics, admission and discharge dates, and primary diagnoses; and (2) timestamped laboratory test results.

All timestamps were truncated to daily granularity. To facilitate analysis, we transformed the long-format laboratory test table into a wide-format representation. Specifically, for each patient and each day, all laboratory tests were aggregated into a single row, with each test item becoming a separate feature column (e.g., Patient ID, Test Date, Test Item 1, Test Item 2, ...). If a patient did not undergo a particular test on a given day, the corresponding value remained missing. During modeling, these missing values were explicitly encoded as -999 to avoid introducing bias while preserving the missingness pattern.

Clinical Timeline Construction and Record Alignment

We assumed that patients make hospitalization decisions based on their test results and socioeconomic conditions. Therefore, in the behavioral timeline, laboratory tests typically precede hospital admissions. Most patients had multiple tests and multiple admissions. However, a small number of records showed admissions without preceding tests, likely due to transfers or emergency cases. These outliers were removed to maintain consistency with the modeling assumption.

Each test record was aligned to the next available hospital admission, and the time gap (in days) between the test date and the admission date was computed as the target

variable. If no subsequent admission was found, the test record was excluded. Additionally, we extracted features from the most recent past admission (e.g., length of stay, department, and diagnosis) for each test record by applying a temporal shift. If no historical admissions existed prior to the test, these values were left missing. This alignment ensured that all predictors temporally preceded the admission decision.

Local economy (LE) and Local medical resources (LMR) Integration

We incorporated five socioeconomic indicators from the China City Statistical Yearbook (**Appendix 1**): per capita gross domestic product (GDP), number of hospitals, number of hospital beds, number of licensed physicians, and average annual wage.

Patient addresses were standardized to a "Province-City" format. For patients with multiple addresses, we selected the earliest recorded one. Socioeconomic indicators were matched to each patient using their standardized address and test year. When city-level data was missing, corresponding provincial-level averages were used as substitutes.

Laboratory Test Normalization

Each laboratory test value was normalized based on its clinical reference range. Specifically:

- Values within the reference range $[min_val, max_val]$were linearly mapped to the interval [0, 1].
- Values above the upper bound were mapped to values greater than 1, with higher values indicating greater deviation.

- Values below the lower bound were mapped to values less than 0, with lower values indicating stronger negative deviation.

The full normalization formula is:

$$\text{Normalized Value}(x) = \begin{cases} \dfrac{x - \text{min\_val}}{\text{max\_val} - \text{min\_val}} & \text{if min\_val} \le x \le \text{max\_val} \\ 1 + \dfrac{x - \text{max\_val}}{\text{max\_val} - \text{min\_val}} & \text{if } x > \text{max\_val} \\ \dfrac{x - \text{min\_val}}{\text{max\_val} - \text{min\_val}} & \text{if } x < \text{min\_val} \end{cases}$$

This transformation preserved the semantic direction of abnormality (e.g., hypo- vs. hyper-conditions) while ensuring all lab test features were on a comparable scale, making them suitable for downstream machine learning tasks.

All socioeconomic indicators were standardized using Z-score normalization. The Z-score for a value was computed as:

$$Z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the indicator in the matched population.

Categorical features with standardized vocabularies—such as primary diagnoses, admission departments, and discharge departments recorded in the encounter records—were encoded using categorical encoding. All other continuous features were standardized using Z-score normalization.

Note: A complete list of all feature names used in the Union Hospital dataset is provided in **Appendix 2**.

Same-day Admission Classification Task (Union Hospital)

Due to the large number of laboratory test indicators and the fact that most patients

undergo different sets of tests at each visit, directly including all test indicators as features would result in an extremely sparse feature matrix. Moreover, conventional imputation methods are not suitable in this context, as imputing missing test values could introduce medical bias and distort the modeling process. To address this, we first designed a binary classification task as a feature selection step, aimed at identifying the laboratory tests most relevant to immediate hospitalization decisions. Specifically, we modeled whether a patient would be admitted on the same day as their diagnostic tests, with positive labels assigned to test records that coincided with admission on the same calendar day (i.e., time gap = 0). The model included only demographic variables (e.g., age, sex) and normalized laboratory test features. To isolate the influence of laboratory tests on hospitalization behavior, we deliberately excluded all encounter-level variables (e.g., prior admission history, department information) and socioeconomic indicators (LE and LMR statistics) from this task.

We split the cleaned dataset into 80% training and 20% testing subsets using stratified sampling based on the binary outcome. To handle class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique, imbalanced-learn v0.10.1) todoto the training set. Features with variance below 0.1 were removed using VarianceThreshold.

A random forest classifier (scikit-learn v1.2.2) todowas trained with n_estimators=100 and random_state=42. Model performance was evaluated using precision, recall, F1-score, and AUC.

To estimate 95% confidence intervals for the AUC, we used bootstrap resampling

with iterations. For each bootstrap sample, the AUC was computed and sorted, and the confidence interval was given by:

$$\mathrm{CI}_{\mathrm{lower}} = \mathrm{Percentile}_{0.025}\left(\mathrm{AUC}^{(1)}, \mathrm{AUC}^{(2)}, \ldots, \mathrm{AUC}^{(1000)}\right)$$

$$\mathrm{CI}_{\mathrm{upper}} = \mathrm{Percentile}_{0.975}\left(\mathrm{AUC}^{(1)}, \mathrm{AUC}^{(2)}, \ldots, \mathrm{AUC}^{(1000)}\right)$$

Temporal Stratification and Feature Selection Based on SHAP

We hypothesized that factors influencing patients' hospitalization decisions may vary across different time periods. To explore temporal dynamics, we divided the dataset into 16 non-overlapping half-year intervals based on test dates, covering the time span from March 12, 2012 to March 12, 2020. A separate random forest model was trained on each of these temporal subsets, and performance results are summarized in **eFigure 1**.

SHAP values were computed using the shap package (v0.41.0)[1], and TreeExplainer was applied to each trained model to compute feature-level attributions for individual samples. The global importance of each feature was defined as the mean absolute SHAP value across all samples:

In the full dataset analysis, we retained only features who's global SHAP importance exceeded 0.01 (see **Appendix 3**). These selected features were subsequently used in downstream regression modeling tasks.

Hospital Readmission Regression Task (Union Hospital)

For the Union Hospital dataset, we also framed a regression task to predict the number of days from a given test to the patient's next hospital admission. We retained only samples with positive labels (i.e., future admissions existed), and excluded those

with large outlier gaps (>1500 days). All continuous features were standardized, and missing values were filled with -999.

We used group-based data splitting, ensuring that all visits from the same patient ID were assigned to either the training or test set, but not both. We then trained a wide range of models:

- **Ridge Regression** (scikit-learn v1.2.2)[2]:

  standard linear models with L2 regularization applied in the former.

- **Support Vector Regression (SVR)**:

  used RBF kernel, with parameters C=1.0, epsilon=0.1, and gamma='scale'.

- **Random Forest Regression**:

  ensemble of 100 decision trees with max_depth=None and random_state=42.

- **Gradient Boosting Models**:

  - **XGBoost** (xgboost v1.7.6)[3] todo: n_estimators=100, max_depth=6, learning_rate=0.1, objective='reg:squarederror'.
  - **LightGBM** (lightgbm v3.3.2)[4] todo: same settings as XGBoost with objective='regression'.
  - **HistGradientBoostingRegressor** (scikit-learn v1.2.2) todo: max_iter=100, learning_rate=0.1, l2_regularization=1.0, max_bins=255.

- **Deep Learning Models**:

  - All (LSTM, GRU, Transformer-based) were trained with hidden_size=64, batch_size=32, learning_rate=0.001, num_layers=2,

and epochs=100.

- **Transformer Regressor**:

  used two encoder layers with 8 attention heads and feedforward dimension 128.

All models received standardized feature inputs, and hyperparameters were selected either from defaults or based on preliminary validation performance.

Evaluation metrics included:

- **RMSE** (Root Mean Squared Error) $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- **MAE** (Mean Absolute Error) $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **R²** (Coefficient of Determination) $1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- **MAE/STD** (MAE-to-Standard-Deviation ratio) $= \frac{\text{MAE}}{\text{STD}(y)}$
- **MASE** (Mean Absolute Scaled Error), comparing each month's error to a naive baseline from the training set: $= \frac{\text{MAE}}{\text{MAE}_{\text{naive}}}$
- **SMAPE (Symmetric Mean Absolute Percentage Error):**

$$\text{SMAPE} = \frac{100\backslash\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

The best model was selected based on the lowest MAE/STD.

Monthly Evaluation Metrics

To examine the robustness of our hospital readmission regression model over time, we performed a comprehensive **temporal evaluation** using the best-performing regressor (XGBoost). This evaluation was carried out on the Union Hospital dataset, using time-aware metrics and monthly segmentation of the test data.

For each calendar month present in the test set, we computed: RMSE, MAE, R^2

MAE/STD.

Temporal Feature Importance with SHAP

We computed monthly and sliding-window SHAP values using TreeExplainer from the shap library (v0.41.0). SHAP values were computed on up to 1000 samples per window to ensure tractable computation. For each time window (**eFigure 6**):

- Mean absolute SHAP values were computed across all features.
- Fixed and sliding windows were both evaluated:
  - **Fixed windows**: each calendar month was treated as an independent segment.
  - **Sliding windows**: overlapping windows of length 1 month were generated, stepping forward monthly.

Special Temporal Analysis

During monthly evaluations, we observed clear phase-dependent patterns in both feature importance and model performance metrics. We hypothesize that these trends may reflect the impact of major healthcare reforms in China and the COVID-19 pandemic. Therefore, we selected two key timepoints—December 31, 2015, and August 31, 2018—as segmentation points, dividing the timeline into three distinct periods (see **eFigure 7**).

For each time segment sub dataset, we independently trained an XGBoost model. Data preprocessing, model configuration, and evaluation metrics were kept consistent with prior analyses.

MIMIC-III Dataset

Data preprocessing

We accessed the publicly available MIMIC-III database via PhysioNet, following established data use protocols. MIMIC-III includes de-identified health-related data from over 40,000 intensive care unit (ICU) patients. For our study, we extracted structured tables including ADMISSIONS, PATIENTS, LABEVENTS, DIAGNOSES_ICD, D_ICD_DIAGNOSES, D_LABITEMS, D_ITEMS, INPUTEVENTS_CV, INPUTEVENTS_MV, and OUTPUTEVENTS.

Data preprocessing began with filtering patients who had more than one hospital admission. Admissions were chronologically ordered for each patient, and we computed the number of days between discharge and the next admission as a continuous target variable: TIME_TO_NEXT_ADMISSION. If no subsequent admission was found, we assigned a default value of 9999.

365-day readmission classification task

To conduct the 365-day readmission classification task, we binarized the target variable: patients with TIME_TO_NEXT_ADMISSION $\leq$ 365 were labeled as 1 (positive), and those with TIME_TO_NEXT_ADMISSION > 365 or without any subsequent admission were labeled as 0 (negative). We excluded records with missing age and capped age values at 120. The feature matrix included demographics, diagnostic categories, laboratory results, input/output events, and derived statistics such as diagnosis count and prior admission frequency.

We split the dataset using an 80/20 train-test stratified split. To mitigate class imbalance, SMOTE was applied to the training data. Low-variance features (variance

< 0.1), highly missing features (>50% missing), weakly correlated features (|r| < 0.01), and highly collinear features (Pearson r > 0.95) were removed. We evaluated multiple classifiers for the final prediction task, including:

We evaluated a comprehensive set of classification models using **scikit-learn v1.2.2**, **xgboost v1.7.6**, **lightgbm v3.3.2**, and custom PyTorch-based neural network modules. The hyperparameters were either explicitly set or used their respective default values from the libraries. Details are as follows:

- **Logistic Regression** (scikit-learn v1.2.2):
  solver='lbfgs', penalty='l2', C=1.0, max_iter=1000, random_state=42.
- **Decision Tree Classifier** (scikit-learn v1.2.2):
  criterion='gini', max_depth=None, min_samples_split=2, random_state=42.
- **Random Forest Classifier** (scikit-learn v1.2.2):
  n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, random_state=42, n_jobs=-1.
- **Gradient Boosting Classifier** (scikit-learn v1.2.2):
  n_estimators=100, learning_rate=0.1, loss='log_loss', random_state=42.
- **XGBoost Classifier** (xgboost v1.7.6):
  n_estimators=100, max_depth=6, learning_rate=0.1, objective='binary:logistic', use_label_encoder=False, eval_metric='logloss', random_state=42.
- **LightGBM Classifier** (lightgbm v3.3.2):
  n_estimators=100, learning_rate=0.1, objective='binary', random_state=42.

- **Support Vector Machine (SVC)** (scikit-learn v1.2.2):

  kernel='rbf', C=1.0, gamma='scale', random_state=42.

- **Recurrent Neural Network (RNN)**:

  input_size set to the number of input features, hidden_size=64, num_layers=2, output_size = number of classes, trained using batch_size=32, learning_rate=0.001, num_epochs=200.

- **Long Short-Term Memory (LSTM)**[5]:

  same settings as RNN.

- **Gated Recurrent Unit (GRU)**:

  same settings as RNN.

- **Transformer-based Classifier**[6]:

  input_size, hidden_size=64, num_layers=2, output_size, batch_size=32, learning_rate=0.001, num_epochs=200. The Transformer encoder used 2 layers, 8 attention heads, and a feedforward dimension of 128.

All deep learning models were trained using standardized features, and categorical variables were encoded prior to model fitting.

All neural models (RNN, LSTM, GRU, Transformer) were implemented with hidden size = 64, number of layers = 2, batch size = 32, learning rate = 0.001, and trained for 200 epochs.

Performance metrics included accuracy, precision, recall, F1-score, and AUC.

The summary of model performance across all classifiers is provided in **eTable 1**

Shandong Cancer Hospital dataset

For the Shandong Cancer Hospital dataset, we only had access to hospitalization records, which included demographic details, admission and discharge timestamps, and diagnostic codes. The full list of available features is provided in Appendix 4. Due to the absence of reliable residential address information, we could not incorporate city-level socioeconomic indicators from the China City Statistical Yearbook. However, the hospitalization records contained patients' health insurance categories, which we used as a proxy indicator for socioeconomic status.

We chronologically sorted each patient's hospitalization records to build a temporal sequence of admissions. Initially, we framed a regression task to predict the time gap (in days) between a patient's previous and next hospital admissions using only features from the prior admission. However, because the available features lacked sufficient clinical depth to explain or model the decision to readmit, the regression models showed poor performance and failed to generalize.

As a result, we reframed the task as a **binary classification problem**, aiming to predict whether a patient would be readmitted **within 15 days** after discharge. We adopted a similar modeling pipeline as described for the Union Hospital dataset, including feature preprocessing and temporal alignment.

- **Logistic Regression** (scikit-learn v1.2.2):

  solver='lbfgs', penalty='l2', C=1.0, max_iter=1000, random_state=42
- **Ridge Classifier** (scikit-learn v1.2.2):

  alpha=1.0, solver='auto', class_weight=None, random_state=42
- **Support Vector Machine (SVC)** (scikit-learn v1.2.2):

kernel='rbf', C=1.0, probability=True, random_state=42

- **Random Forest Classifier** (scikit-learn v1.2.2):

  n_estimators=100, max_depth=None, min_samples_leaf=1,

  class_weight='balanced', n_jobs=-1, random_state=42

- **Histogram-based Gradient Boosting Classifier**

  (HistGradientBoostingClassifier, scikit-learn v1.2.2):

  max_iter=100, learning_rate=0.1, l2_regularization=0.0, max_bins=255,

  class_weight='balanced', random_state=42

- **XGBoost Classifier** (xgboost v1.7.6):

  n_estimators=100, max_depth=6, learning_rate=0.1,

  objective='binary:logistic', use_label_encoder=False, eval_metric='logloss',

  random_state=42

- **LightGBM Classifier** (lightgbm v3.3.2):

  n_estimators=100, max_depth=-1, learning_rate=0.1,

  objective='binary', class_weight='balanced', boosting_type='gbdt', verbose=-1,

  random_state=42, n_jobs=-1

- **LSTM / GRU / Classifiers** (PyTorch):

  input_dim = feature dimension, hidden_size = 64, num_layers = 2,

  output_size = 1, learning_rate=0.001, batch_size=64, num_epochs=10,

  task='binary'

We also applied both full-dataset and semi-annual temporal splits to evaluate model performance on the readmission prediction task (see **eFigure 2, eFigure 3**). In

addition, we computed feature importance scores for each period. This heatmap illustrates the contribution of each feature to the model's output, with importance scores standardized using the standard deviation (STD) score (see **eFigure 4**).

The summary of model performance across all classifiers is provided in **eTable 1 and Table 2.**

References:

1. Lundberg SMaL, Su-In. A Unified Approach to Interpreting Model Predictions. In: Garnett IGaUVLaSBaHWaRFaSVaR, editor. Advances in Neural Information Processing Systems: Curran Associates, Inc.; 2017.

2. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011; **12**: 2825-30.

3. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016; 2016. p. 785-94.

4. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 2017; **30**.

5. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation* 1997; **9**(8): 1735-80.

6. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems* 2017; **30**.

Appendix 1. China Urban Statistical Yearbook by the National Bureau of Statistics used in this study. LE and LMR indicators were obtained from the China City Statistical Yearbook, published by the Urban Socioeconomic Survey Department of the National Bureau of Statistics of China.

References website:
https://olap.epsnet.com.cn/

Appendix 2. Label Data (Post-Mapping Format). 485 Labels Used in Wuhan Union Hospital Data for This Study.

medical record number	IFN-γ	leucine aminopeptidase	nucleated red blood cells	lipoprotein(a)
inspection time	INR	human cytomegalovirus DNA	nucleated red blood cells/100WBC	lipoprotein-associated phospholipase A2
Procollagen type 1 N-terminal propeptide	IgE	human growth hormone 120min	total nucleated cell count	cerebrospinal fluid ADA
17-KS	Type I collagen N-terminal propeptide	human growth hormone 30min	immature granulocytes	cerebrospinal fluid AST
17-OH	Type I collagen carboxy-terminal peptide	human growth hormone 60min	weed pollen mix	cerebrospinal fluid CK
2019nCov antibody IgG	K	human growth hormone 90min	Syphilis TRUST	cerebrospinal fluid LDH
2019nCov antibody IgM	K/L ratio	human chorionic gonadotropin β-subunit	Syphilis specific antibody	cerebrospinal fluid chloride
24h urine cortisol	K light chain	human epididymis protein 4	Chlorine	cerebrospinal fluid glucose
24h urine protein	L/P	low molecular weight heparin concentration	Prolactin	cerebrospinal fluid protein
24h urine protein (for calculation)	LDL	low-density lipoprotein cholesterol	Serum Amyloid A	adenosine deaminase
24h urine volume	LH Luteinizing Hormone	low fluorescence RET	Amylase	glucose

5'nucleotidase	LY30	low fluorescence RET%	Lymphocyte	shrimp
AA inhibition rate	L light chain	thyroid stimulating hormone receptor antibody	Lymphocyte %	blood NT-ProBNP
ACTH adrenocorticotropic hormone	MA	immunoglobulin A	Osmotic pressure	blood β-hydroxybutyrate
ADP inhibition rate	MA A	immunoglobulin G	Free prostate specific antigen	blood pyruvate
ANTI-TPO	MA ADP	immunoglobulin G4	Cyclic citrullinated peptide antibody	blood lactic acid
APTT	MCH	immunoglobulin M	Globulin	platelet
ATIII	MCHC	intact N-terminal procollagen type I propeptide	Triglycerides	platelet (optical method)
ATPO thyroid microsomal antibody	MCV	thrombin-antithrombin complex	Parathyroid hormone	plateletcrit
ATPO peroxidase antibody	NK lymphocytes	prostate specific antigen	Thyroglobulin	mean platelet volume
ATg thyroglobulin antibody	NK cells (CD3-CD16+56+)	prealbumin	Alpha-fetoprotein	platelet distribution width
Alpha 1	PCO2(T)	animal dander mix	Pathological cast	thrombomodulin
Alpha 2	PGI/PGII ratio	mononuclear leukocytes%	Pathological cast (sediment)	blood ammonia
Andro Androstenedione	PO2(T)	monocytes	Carcinoembryonic antigen	blood amylase
Angle	PRL Pituitary Prolactin	monocytes%	Albumin/Globulin Ratio	serum retinol-

				binding protein
Angle (Quick Method)	PT	follicle stimulating hormone	White blood cell	serum transferrin
Beta	PTA	soluble transferrin receptor	Interleukin-10	blood glucose
BuBc	PTH Parathyroid Hormone	folic acid	Interleukin-2	hemoglobin
B-type natriuretic peptide	Prog Progesterone	homocysteine	Interleukin-4	blood lipase
B lymphocyte	R	basophils	Interleukin-6	complement C3
B lymphocyte (CD3-CD19+)	R(Heparin)	basophils%	Total white blood cell count	complement C4
C8000HLDL	RET_He	eosinophils	Cortisol	gamma-glutamyl transpeptidase
CD3+ total T lymphocyte	SCT Confirmation	eosinophils%	Cortisol 0 o'clock	high-sensitivity C-reactive protein
CD4+ T lymphocyte	SCT Confirmation Ratio	Remarks	Cortisol AM 12 o'clock	high-sensitivity troponin I
CD4/CD8 ratio	SCT Screening	Remarks 2	Cortisol AM 4 o'clock	transferrin
CD8+ T lymphocyte	SCT Screening/Confirmation	polymorphonuclear leukocytes%	Cortisol AM 8 o'clock	apolipoprotein AI
CK-MB activity measurement	SCT Screening Ratio	stool microscopy	Cortisol PM 4 o'clock	apolipoprotein B
CK-MB mass measurement	TEG-ACT	large platelet ratio	Direct bilirubin	helper/inducer T lymphocytes (CD3+CD4+)

CMV-DNA (whole blood/intracellular)	TG Thyroglobulin	aspartate aminotransferase	Testosterone	hyaline cast (sediment)
CMV-DNA-intracellular	TNF-α	progesterone	Alkaline phosphatase	calcium
CSF-IgA	TNI	small dense low-density lipoprotein	Phosphorus	sodium
CSF-IgG	TRAb Anti-TSHR	urine PH	Neuron-specific enolase	potassium
CSF-IgM	TSH	urine α1-microglobulin	Fasting blood glucose	ferritin
C-peptide	TSH Thyroid Stimulating Hormone	urine β2 microglobulin	Rheumatoid factor	ceruloplasmin
C-peptide 0min	TT	urine epithelial cells	Rheumatoid factor IgA	ceruloplasmin assay
C-peptide 10min	Testo Testosterone	urine epithelial cells quantitative	Rheumatoid factor IgG	magnesium
C-peptide 120min	T lymphocytes (CD3+)	urine immunoglobulin G	Rheumatoid factor IgM	anion gap
C-peptide 150min	VCA-IgA Ratio:	urine specific gravity	Dust mite	procalcitonin
C-peptide 180min	VMA	urine chloride	Glycated serum protein	random urine creatinine
C-peptide 240min	ZnT8A Zinc Transporter 8 Antibody	urine amylase	Glycated hemoglobin	random blood glucose
C-peptide 2min	cCa++	urine leukocytes	Carbohydrate antigen 72-4	occult blood (chemical method)
C-peptide 300min	cCl-	urine leukocytes quantitative	Carbohydrate antigen CA125	estradiol
C-peptide 30min	cH	urine cortisol	Carbohydrate antigen CA15-3	fungal mix
C-peptide 360min	cH(T)	urine iodine concentration	Carbohydrate antigen CA153	unconjugated bilirubin

C-peptide 4min	cHCO3(P)	urine phosphorus	Carbohydrate antigen CA19-9	non-high-density lipoprotein cholesterol
C-peptide 60min	cHCO3(Pst)	urine cast quantitative	Carbohydrate antigen CA199	1-hour postprandial blood glucose
C-peptide 6min	cK+	urea	Red blood cell	2-hour postprandial blood glucose
C-peptide 8min	cLac	blood urea nitrogen	Red blood cell distribution width coefficient of variation	3-hour postprandial blood glucose
C-peptide 90min	cNa+	urine red blood cells	Red blood cell distribution width	0.5-hour postprandial blood glucose
D-dimer	ctCO2(B)	urine red blood cells quantitative	Red blood cell distribution width standard deviation	osteocalcin N-terminal mid-molecule fragment
DHEA Dehydroepiandrosterone	ctHb	urine vitamin C	Hematocrit	osteocalcin N-terminal fragment
Delta bilirubin	ctO2	urine creatinine	Erythrocyte sedimentation rate	osteocalcin Oc
E2 Estradiol	dRVVT Confirmation	urine urobilinogen	Tissue plasminogen activator-inhibitor 1 complex	high-density lipoprotein cholesterol
EA-IgA ratio:	dRVVT Confirmation Ratio	urine retinol-binding protein	Total cell count	high fluorescence RET
EB-DNA-intracellular	dRVVT Screening	urine transferrin	Cytokeratin 19 fragment	high fluorescence RET%

EB-DNA-extracellular	dRVVT Screening/Confirmation	urine calcium	Conjugated bilirubin	squamous cell carcinoma related antigen
EBV-DNA (whole blood/intracellular)	dRVVT Screening Ratio	urine sodium	Vitamin B12	bmi
EBV-DNA	eGFR	urine potassium	Reticulocyte	systolic pressure
EBV early antigen antibody IgA	hGH Human Growth Hormone	urine magnesium	Reticulocyte %	diastolic pressure
EBV early antigen antibody IgG	p50(st)	bunya virus RNA	Myoglobin	heart rate
EBV nuclear antigen-IgG	pCO2	immature RET ratio	Creatinine	body temperature
EBV viral capsid antigen IgA	pH	immature RET ratio%	Creatinine (picric acid method)	respiration
EBV viral capsid antigen IgG	pH(T)	microalbumin	Creatinine (enzymatic method)	GDP per capita (yuan)
EBV viral capsid antigen IgM	pH(st)	German cockroach	Creatine kinase	number of hospitals and health centers (units)
EPL	pO2	cardiolipin antibody IgA	Gastrin-17	number of beds in hospitals and health centers (beds)
FDP	pO2(A-a)	cardiolipin antibody IgG	Cholinesterase	number of doctors (people)
FIB	sO2	cardiolipin antibody IgM	Insulin	average wage of employees (yuan)
FS51HLDL	B2 Glycoprotein 1 IgA	total 25-hydroxyvitamin D	Insulin 0min	gender

FSH Follicle-stimulating hormone	B2 Glycoprotein 1 IgG	total IgE	Insulin 10min	date of birth
FT3	B2 Glycoprotein 1 IgM	total PINP	Insulin 120min	age
FT3 Free triiodothyronine	α-Hydroxybutyrate Dehydrogenase	total carbon dioxide	Insulin 150min	admission department
FT4	α1-Antitrypsin	total cholesterol	Insulin 180min	discharge department
FT4 Free thyroxine	β-HCG	total bile acid	Insulin 240min	main diagnosis
G	β2-Microglobulin	total bilirubin	Insulin 2min	admission time
GADA Glutamic acid decarboxylase antibody	βCrossLap	total protein	Insulin 300min	discharge time
GADA Gamma	γ-Glutamyl Transferase	house dust mite	Insulin 30min	last diagnosis
HCV Ab	Alanine Aminotransferase	suppressor/cytotoxic T lymphocytes (CD3+CD8+)	Insulin 360min	last admission department
HCV-Ab	Neutrophils	anti-MPO antibody	Insulin 4min	last admission time
HCV-RNA-intracellular	Neutrophils %	anti-RA33 antibody	Insulin 60min	last discharge time
HCV-RNA-extracellular	Medium Fluorescence RET	anti-hepatitis C virus antibody	Insulin 6min	time difference
HCV-RNA quantitative	Medium Fluorescence RET%	spot count of antigen A	Insulin 8min	last hospitalization days
HCV-RNA quantitative-intracellular	Hepatitis B e Antibody	spot count of antigen B	Insulin 90min	fee type

HCV-RNA quantitative-extracellular	Hepatitis B e Antigen	anticardiolipin antibody	Insulin-like growth factor 1	last main surgical operation name
HDL+LDL	Hepatitis B Core Antibody	anti-cyclic citrullinated peptide antibody	Insulin-like growth factor 1-CT	last discharge main diagnosis name 1
HIV-Ab	Hepatitis B Surface Antibody	anti-thyroid peroxidase antibody	Cystatin C	last pathological diagnosis
Hct	Hepatitis B Surface Antigen	anti-proteinase 3 antibody	collagen degradation product	hospitalization times
IA-2A tyrosine phosphatase antibody	lactic dehydrogenase	Anti-Streptolysin O	fat globule	last hospitalization duration

Appendix 3. Selected Feature on Regression of Time to Next Admission Task

medical record number	Direct bilirubin	bmi
body temperature	high-density lipoprotein cholesterol	potassium
respiration	PT	Glycated hemoglobin
heart rate	White blood cell	Fasting blood glucose
systolic pressure	mean platelet volume	Osmotic pressure
Red blood cell distribution width standard deviation	Neutrophils	Alanine Aminotransferase
large platelet ratio	urine specific gravity	basophils%
monocytes	INR	sodium
leucine aminopeptidase	5'nucleotidase	magnesium
diastolic pressure	MCH	low-density lipoprotein cholesterol
Lymphocyte%	MCV	apolipoprotein AI
Neutrophils %	Phosphorus	α-Hydroxybutyrate Dehydrogenase
Creatinine (enzymatic method)	eosinophils	Alpha-fetoprotein
gamma-glutamyl transpeptidase	FT4	TSH
platelet distribution width	apolipoprotein B	FT3
eosinophils%	urine uric acid	Albumin/Globulin Ratio
monocytes%	total bilirubin	high-sensitivity C-reactive protein
small dense low-density lipoprotein	plateletcrit	lipoprotein(a)
hemoglobin	Chlorine	unconjugated bilirubin
Creatinine	total protein	urine PH
APTT	Creatine kinase	ferritin
prealbumin	calcium	cLac
age	FIB	urea
lactic dehydrogenase	Alkaline phosphatase	ATIII
Red blood cell	blood urea nitrogen	Amylase
Lymphocyte	Cystatin C	ctO2
platelet	γ-Glutamyl Transferase	pCO2
total bile acid	aspartate aminotransferase	Hct
Hematocrit	Albumin	cK+
MCHC	anion gap	complement C3
TT	total carbon dioxide	pO2
glucose	Globulin	cCl-

Appendix 4. Label Data (Post-Mapping Format). 27 Labels Used in Shandong Cancer Hospital Data for This Study.

Medical Record Number	Primary Discharge Diagnosis	Other Social Insurance
Number of Hospitalizations	Pathological Diagnosis	Fully Public Funded
Payment Type	Primary Procedure Name	Poverty Assistance
Gender	Urban Employee Medical Insurance	Basic Medical Insurance for Urban Employees
Age	New Rural Cooperative Medical Insurance	Basic Medical Insurance for Urban Residents
Admission Date	Commercial Health Insurance	New Rural Cooperative Medical Scheme
Discharge Date	Urban Resident Medical Insurance	Medical Restriction
Discharge Department (Front Page)	Fully Self-Paid	Commercial Medical Insurance
Length of Stay (Days)	Other	Referred from Other Institutions