Research using TMUCRD Data

Jason C. Hsu

Chair, Local Host Committee of the 2022 OHDSI APAC Symposium
Associate Professor, Taipei Medical University, Taiwan
Director, Clinical Data Center, Office of Data Science, Taipei Medical University
Outline

1. TMUCRD: Introduction

2. TMUCRD: Development

3. TMUCRD: Applications
1. TMUCRD: Introduction
TMU’s Three Affiliated Hospitals

- Taipei Medical University Hospital (T)
- Wan Fang Hospital (W)
- Shuang-Ho Hospital (S)
Taipei Medical University Clinical Research Database (TMUCRD)
Composition of TMUCRD
Time Period and Case Number

1996-2021 (25 years)
N=4,125,097 (18.7% of Taiwan's population)
Features of TMUCRD (compared to NHIRD)

- Laboratory Data
- Insurers
- PHR Vital Signs
- Radiology Reports
- EMR
- Hospitals
- Physicians, Clinicians

- Health Examination
- New Drugs
TMUCRD was linked to cancer registry and death registry databases

1. De-ID
2. Matching

Cancer Patients’ ID → Cancer Registry → Death Registry

All Patients’ ID → Center for Management and Development

TMUCRD
2. TMUCRD: Development
High Quality Database

- Large Case Number
- Long Time Period
- Personal Privacy and Data Governance
- Data Correctness
- Data Accessibility
- Data Diversity
- Data Standardization (format, encoding)
- Data Integrity
Structure unstructured text reports and move towards automated reporting

<table>
<thead>
<tr>
<th>ORGH SOAP</th>
<th>organ</th>
<th>locat</th>
<th>operation</th>
<th>diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>left</td>
<td>partial mastectomy</td>
<td>Invasive ductal carcinoma</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORGH Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histologic type</td>
</tr>
<tr>
<td>invasive ductal carcinoma</td>
</tr>
<tr>
<td>pT</td>
</tr>
<tr>
<td>pT1c: Tumor &gt;10 mm but &lt; or =20 mm in greatest dimension.</td>
</tr>
<tr>
<td>pN</td>
</tr>
<tr>
<td>pN(sn)0 (i-): No sentinel lymph node metastasis identified histologically, negative IHC</td>
</tr>
<tr>
<td>pM</td>
</tr>
<tr>
<td>Not applicable.</td>
</tr>
<tr>
<td>pTNM</td>
</tr>
<tr>
<td>T1 N0 M0 IA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ER</th>
<th>PR</th>
<th>Her2</th>
<th>Ki-67</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative</td>
<td>positive</td>
<td>positive (Score 3+)</td>
<td>30%</td>
</tr>
</tbody>
</table>
Create New Type of Data

- Cardiac Ultrasound Report
- Cardiac Calcification Index
- Cancer Brain Metastasis Report
- Nutrition Report
- Mental Examination Scale
- Bacterial Test Report

New Type of Data
Build Specialized Databases

- Emergency
- COVID-19
- Dementia
- Specialized Databases
- Diabetes
- Ischemic Stroke
- Lung Cancer
Cross-Institutional Data Research Cooperation

Diabetes

CKD

Emergency

Dementia
Academic-Industry Collaboration

GCSF and Febrile Neutropenia

Inflammatory Bowel Disease (IBD)

Renal Dysfunction
3. TMUCRD: Applications
Rich research performance from the use of TMUCRD

Large Research Projects (9)

Theses and Dissertations (23)

Academic Publications (25)
To assess postoperative complications and mortality in patients with diabetes.

Diabetic patients undergoing surgery have higher risk of infectious complications and hospitalization and mortality compared with non-diabetic patients undergoing similar major surgery.
The results indicated that the addition of bevacizumab on gefitinib treatment could suppress MAPF-induced angiogenesis in lung adenocarcinoma patients.
To compare the efficacy and safety of long-term and short-term dual antiplatelet therapy (DAPT) after coronary stenting in patients with CKD.

Long-term DAPT was associated with similar risk of MACE (HR: 1.05, 95% CI: 0.65–1.70, P = 0.83) compared with short-term DAPT. Different CKD risk did not modify the risk of MACE. There was also no significant difference in all-cause mortality (HR: 1.10, 95% CI: 0.75–1.61, P = 0.63) and TIMI bleeding (HR 1.19, 95% CI: 0.86–1.63, P = 0.30) between groups.
Our study indicated that the use of ICSI does not associated with higher risk of neurodevelopmental disorders in the offspring. But male sex, and ICU admission do have increased risk of neurodevelopmental disorders.
The goal of the study is to analyze the incidence and associated factors of lactic acidosis between metformin user and non-user with advanced CKD.

Conclusions: Metformin was associated with a significant increased risk of laboratory-defined lactic acidosis (p=0.0204) even after adjusting confounder such as age, sex and underlying comorbidities. This “REMIND” study reminds us that metformin-associated lactic acidosis is mainly caused by decreased drug renal elimination other than underlying comorbidities in advanced CKD patients.
The time interval between ED arrival and brain CT was significantly shortened during P1 and P2 compared with the pre-pandemic interval, and no significant delay between ED arrival and surgical management was found, indicating increased treatment efficiency for TBI in the ED during the COVID-19 pandemic.
Compared with the control group, the case group demonstrated more comorbidities and higher proportions of hypertension, diabetes mellitus, gout, dyslipidaemia, heart disease and cerebrovascular disease, but had lower risk of progression to CKD stage IIIb before and (HR 0.72; 95% CI 0.61 to 0.85) and after (adjusted HR (aHR) 0.67; 95% CI 0.55 to 0.81) adjustments.
Machine-Learning Monitoring System for Predicting Mortality Among Patients With Noncancer End-Stage Liver Disease: Retrospective Study

Yu-Jinn Lan\(^1\), MD; Ray-Jade Chen\(^2\), MD, MSc; Jui-Hsiang Tang\(^3\), MD; Cheng-Sheng Yu\(^{1,4}\), PhD; Jenny L Wu\(^1\), BSc; Li-Chuan Chen\(^{5,7}\), MSN; Shy-Shin Chang\(^{1,2,9}\), MD, PhD

\(^1\)Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
\(^2\)Department of Family Medicine, Taipei Medical University Hospital, Taipei, Taiwan
\(^3\)Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
\(^4\)Division of General Surgery, Department of Surgery, Taipei Medical University Hospital, Taipei, Taiwan
\(^5\)Division of Gastroenterology and Hepatology, Department of Internal Medicine, Taipei Medical University Hospital, Taipei, Taiwan
\(^6\)Department of Community and Preventive Medicine, Taipei Medical University Hospital, Taipei, Taiwan
\(^7\)School of Gerontology Health Management, College of Nursing, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Shy-Shin Chang, MD, PhD
Department of Family Medicine
School of Medicine, College of Medicine
Taipei Medical University
256 Wuxing Street
Taipei, 11031

AUCROC = 0.852 (Random Forest)
To construct a medical database system from electronic medical records (EMRs) of subjects who have undergone health examination.

This system aims to provide online self-health evaluation to clinicians and patients, enabling personalized health and preventive health.

AUCROC = 0.904 (predict metabolic syndrome)
AUCROC = 0.982 (predict CKD)
From TMUCRD study to OHDSI study
Observational Study

Abstract

IMPORATANCE More than 1 billion adults have hypertension globally, of whom 70% cannot achieve their hypertension control goal with monotherapy alone. Data are lacking on clinical use patterns of dual combination therapies prescribed to patients who escalate from monotherapy.

OBJECTIVE To investigate the most common dual combinations prescribed for treatment escalation in different countries and how treatment use varies by age, sex, and history of cardiovascular disease.

DESIGN, SETTING, AND PARTICIPANTS This cohort study used data from 11 electronic health records from 8 countries and regions between January 2006 and December 2018. Included participants were adult patients (ages ≥18 years) who newly started dual combination treatment and had a prior year of monotherapy.

Key Points

Question What are the most common antihypertensive dual combinations prescribed to patients who escalate from monotherapy in clinical practice, and how do the combinations differ by country and patient demographic subgroup?

Findings In this cohort study of 970,335 individuals from 11 large databases, 12 dual combinations of antihypertensive medications were identified.
Development of Lung Cancer Survival Prediction Models Based on Real-world Data and Machine Learning

Jason C. Hsu; Phung-Anh Nguyen; Phan Thanh Phuc; Tsai-Chih Lo; Min-Huei Hsu; Chi-Tsun Cheng; Tzu-Hao Chang; Cheng-Yu Chen
Taipei Medical University, Taiwan

Abstract

Background

The development of disease risk and prognosis prediction models using machine learning or deep learning algorithms with big data is a major area of academic research based on the results of multiple clinical studies and the findings of big data research. This study involved using machine learning or deep learning algorithms to develop lung cancer risk and prognosis prediction models.

Methods

This study used Taipei Medical University Clinical Research Database (TMUCRD) to link data from 3 hospitals as the data source. The data were mapped to OHDSI OMOP CRM. We selected a very small lung cancer patients with a retrospective data collection at TMUCRD and Taiwan Cancer Registry between January 2008 and December 2018. All patients were maintained from the index date of cancer diagnosis until the event of death or the last visit to hospitals. Variables included demographics, comorbidities, medications, laboratory, and pathology tests of patients were retrieved and used to develop the machine learning models. Five machine learning algorithms with various models (i.e., integrating different variables) were used to develop the prediction model. The performance of the algorithms was evaluated using both internal validation and external validation. The performance of the models was assessed using area under the ROC, accuracy, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) (Prediction), and F-score.

Results

In total, 1,751 patients were included (1,129 cases for the training dataset and 622 for the testing dataset). The artificial neural network (ANN) model with all available data achieved the highest score of 0.89. The best performance of the ANN model was achieved when integrating all variables with the AUC, accuracy, precision, recall, and F-score of 0.86, 0.85, 0.80, 0.85, and 0.80, respectively. The most important features were the cancer stage, cancer site, diagnosis age, smoking status, ECOG status, and body mass index.

Conclusion

In this evaluation of lung cancer survival, the ANN model exhibited a better positive performance with high AUC, precision, and recall when integrating different data types. Further research is necessary to determine the feasibility of applying the algorithm in the clinical setting and explore whether using this tool could improve care and outcomes. This study is expected to be developed into a multinational comparative research using electronic health records in the future.

Methods

Study Design and Data Source

We conducted a retrospective study to which we obtained the data from the Taipei Medical University Clinical Research Database (TMUCRD), which were mapped to OHDSI OMOP CRM. The TMUCRD retrieved data from various electronic medical records (EMR) of three hospitals, Taipei Medical University Hospital (TMUH), Tzu Chi Hospital (TCH), and Shin Kong Hospital (SHK). The database covers the electronic medical record data of 3.6 million people accumulated from 2008 to 2018. This study has been approved by the Joint Institute Review Board of Taipei Medical University (TMU JIRB).

Cohort Selection

This study selected patients with lung cancer (ICD-9-O under CR0, CR1) from 2008 to 2018 in the TMU database. Patients included in the analysis were those who did not have any medical history in the three hospitals (TMUH, TCH, and SHK). The 3,714 patients were included in this study, including 960 patients from TMUH, 1,156 from TCH, and 1,698 from SHK.
Development of Machine Learning models for Cancer Survival among Lung cancer patients with Tyrosine Kinase Inhibitors (TKIs) treatment

Alex P.A. Nguyen1, Phu T. Pham1, Min-Huei Hsu1, Jason C. Hsu1,2,3*
1 Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan;
2 International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan;
3 Department of Biomedical Science and Biotechnology, College of Management, Taipei Medical University, Taipei, Taiwan;

Background
Lung cancer is the most common cause of cancer death worldwide, including in Taiwan. Mutation in the EGFR gene is a driver in lung adenocarcinoma, as this gene is overexpressed in more than 50% of non-small cell lung cancer (NSCLC) in Asia. Most patients benefited from TKI therapies, but 5%-10% of patients did not achieve disease control when administered EGFR-TKIs and therefore acquired drug resistance within 10-12 months.

In this study, we aimed to develop prediction models for lung cancer survival among patients with TKI treatment using a larger number of samples, different data types, and various machine learning algorithms.

Methods

Study Design and Data Source
We conducted a retrospective study in which we obtained the data from the Taiwan Cancer Registry (TCR) database and the Taipei Medical University Clinical Research Database (TMAURCD).

Cohort Selection
Patients with lung cancer (ICD-O-3 code: C33, C34.1) from 2008 to 2018 in the TCR database. Exclusion criteria included individuals under 20, small cell lung cancer (SCLC) patients, and patients who did not receive lung cancer treatment in the three hospitals. Following that, only cancer patients who were undergoing TKIs (i.e., patients using EGFR-TKIs, ATO code L01EB) were included in our study cohorts.

Outcome Measurement
The outcome of this study was death within two years following diagnosis. Data were censored at the date of death or loss to follow-up, insurance termination, or the study’s end on December 31, 2020.

Feature Selection
The selected features were as follows: (1) Demographic information; (2) Cancer condition; (3) Comorbidities; (4) Current medications use; and (5) Laboratory test results. All the features were defined before the time patients were prescribed TKI drugs.

Developing the Machine Learning models
Six machine learning algorithms were used including Logistic Regression (LR), bootstrap aggregation (bagging), gradient boosting machine (GBM), AdaBoost, random forest (RF), and extreme gradient boosting (XGBoost), to develop the prediction models.

The training set, containing the data of Taipei Medical University Hospital and Wang Fang Hospital. The testing set, including the data of Shuang Ho Hospital, was used to validate the models. The 5-fold cross-validation was applied.

Figure 1. Receiver Operating Characteristic (ROC) Curve of various models

Conclusions
Random forest was observed as the best model when using all features. Moreover, while choosing the top 30 features, Gradient Boosting Classifier was found with the highest AUC of 0.94.

In summary, the model developed using the Gradient Boosting Classifier algorithm had the highest AUC regardless of the model and was the most suitable tool for NSCLC survival prediction among patients who underwent TKI treatment. In addition, using more types of data (especially laboratory and genomic test results) led to better predictive performance. Cancer stage, cancer size, gender, diagnosis age, and body mass index were the essential features for NSCLC survival prediction.
Machine Learning Study (Prediction)

Development of Breast Cancer Survival Prediction Models Based on Real-world Data and Machine Learning

Chun-Jung Wang, Min-Huei Hsu, Ruo-Kai Lin, Chin-Sheng Hung, Nei-Hui Kuo, Yu-Wen Cheng, Phung-Anh Nguyen, Phan Thanh Phuc, Chi-Tsun Cheng, Jason C. Hsu

Background

Breast cancer is the with the highest incidence and mortality among women in most countries. There are approximately 2.3 million newly diagnosed cases worldwide each year, and approximately 690,000 deaths annually. In the past, many researchers solved various clinical big data and machine learning algorithms to predict survival models for breast cancer diagnosis and prognosis, respectively, to assist medical decision-making and improve treatment outcomes. However, the parameters and accuracy of such prediction models may vary due to differences in race, geographic location, or other ethnic or individual factors, so it is necessary to use various data sources to develop various prediction models.

Objectives

This study aims to use clinical real-world data with multiple attributes and multiple machine learning algorithms to determine the key factors that affect overall survival when the patient is diagnosed with breast cancer and establish a prediction model which can act as a supporting decision aid for physician by modifying the magnitude of treatment.

Methods

Taipei Medical University Clinical Research Database (TMU-CRD) was the data source of this study, which contains the electronic medical records of three hospitals in Taiwan, including Taipei Medical University Hospital (TMUH), Wan-Fang Hospital (WFH), and Shuang-Ho Hospital (SHH). All the data was mapped to OHDSI GAOPD.COM. We selected breast cancer female patients whose ICD-0-3 code was C50.0-C50.9 from 2000 to 2019 as the study cohort, and non-primary breast cancer cases or cases with insufficient information on personal medical background and treatment were excluded. Neither do the patient nor the patient’s follow-up period less than one year. Patients from TMUH and WFH were the training dataset, and patients from SHH were used for external testing. The percentage of alive and death is around 87% and 13% in training data, and 84% and 16% in external testing data, respectively. The date of diagnosis of breast cancer for each patient was used as the index date, and death within five years after diagnosis was used as the outcome. All the information could be gathered on the index data. Totally, there were 45 features involved, including the patient’s basic demographic information, cancer condition, comorbidity, current medication, laboratory test result, were selected and used in the model generation. The comorbidities which occurred prior to breast cancer diagnosis were collected. And the lab values which recording within 6 months before and after the index date. Missing value of categorical data were classified as a new value with mean value. If the percentage of missing value more than 10%, the selected features are BUN, CA153, CEA, creatine kinase, LDL, HIV or viral infection. Finally, there are 57 features by logistic regression (LR), support vector machine (SVM), random forest (RF), and artificial neural network (ANN) were applied to build prediction models. Based on the external test results, the model with the largest area under the receiver operating characteristic curve (AUC) is the best model.

Results

A total of 5,503 patients were included (4,071 for the training dataset, and 1,432 for the testing dataset). Based on the external test results, neural network model had the highest AUC (0.880), following by GB (AUC=0.864), RF (AUC=0.863), SVM (AUC=0.804), LR (AUC=0.798) and DT (AUC=0.675). The accuracy of all models is above 85%. In addition, this study also found that, according to the results of the best model (NN), tumor clinical stage, clinical lymph node stage, primary site, Charlson-Deyo Comorbidity Index, and dementia played the most important role in predicting the five-year survival of breast cancer.

Conclusions

This study successfully established an accurate 5-year survival predictive model for breast cancer patients. Furthermore, this study also found many key factors that may affect the survival of breast cancer patients in Taiwanese patients. The results of this study can be used as a reference for clinical practice of breast treatment.

References

Machine Learning to Predict the Ischemic Stroke among Type 2 Diabetes Mellitus Patients using Taipei Medical University Clinical Research Database

PHAN THANH PHUC 1, PHUNG ANH NGUYEN 2, JASON C. HSU 1,2,*
1 International Ph.D Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan;
2 Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan;

Background
Ischemic stroke has been recognized as a clinically important complication of type 2 diabetes (T2DM) patients. Risk prediction models for DM complications/comorbidities have substantial capacity to support the decision-making process regarding the patient’s clinical management. This study aims to develop machine learning algorithms to predict the risk of ischemic stroke among T2DM patients using various predictors such as patients’ characteristics, disease history, laboratory tests, and medication.

Methods
1. Data source and study population
The dataset was collected from the Taipei Medical University Clinical Research Database (TMU-CRD) in this study. Index 2008 data as wash-out-period—nearly diagnosed T2DM patients from 2009 to 2019 as our cohort study with ICD-9-CM codes 290.xx, and ICD-10-CM codes E11.xx.

2. Outcome
All patients were monitored from the date of taking antidiabetic drugs to the date the patients were admitted to hospitals with ischemic stroke (ICD-9-CM codes 433, 434, 438, and ICD-10-CM codes I60, I61, I63) during a one-year follow-up.

3. Features
The features were collected, including (i) patient characteristics (i.e., age, sex), (ii) comorbidities (i.e., any diagnoses before the date of taking antidiabetic drugs), (iii) other medication uses, and (iv) laboratory exams (i.e., Glucose, HbA1C, etc.).

4. Statistical analysis and Model development
The training set, containing the data of Taipei Medical University Hospital and Wang Fang Hospital. The testing set, including the data of Shuang Ho Hospital, was used to validate the models. The stratified 10-fold cross-validation was applied in the training set to assess different machine learning models’ performance and general errors. Machine learning techniques, such as Logistic Regression (LR), Linear Discriminant Analysis (DT), Gradient Boosting Machine (GBM), and Random Forest (RF), to develop the prediction models. The performance of the algorithms was measured by Area Under the Curve (AUC), sensitivity, specificity, and F1-score.

Results
Table 2. Model performance evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (CV)</th>
<th>AUC (Testing)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.88</td>
<td>0.85</td>
<td>0.819</td>
<td>0.748</td>
<td>0.16</td>
</tr>
<tr>
<td>LDA</td>
<td>0.88</td>
<td>0.85</td>
<td>0.721</td>
<td>0.802</td>
<td>0.161</td>
</tr>
<tr>
<td>GBM</td>
<td>0.91</td>
<td>0.85</td>
<td>0.744</td>
<td>0.813</td>
<td>0.164</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.93</td>
<td>0.84</td>
<td>0.811</td>
<td>0.692</td>
<td>0.133</td>
</tr>
</tbody>
</table>

Figure 1. Receiver Operating Characteristic (ROC) Curve to evaluate the model performance

Conclusions
We successfully developed machine learning models to predict the risk of ischemic stroke among T2DM. Our model performance improved from Random Forest to Gradient Boosting Machine. The top three important features executed from our best model are antiplatelet agent, age, and prior stroke.

The strong association of diabetes with stroke has long been appreciated. To the best of our knowledge, there are limited studies in classifying and predicting ischemic stroke in the T2DM cohort by developing machine learning-based models. Therefore, our findings were essential to improve the accuracy of early detection, diagnosis, and prognosis of ischemic stroke to manage the risk of diabetes complications.

Contact: *Jason C. Hsu, International Ph.D. program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan; E-mail: jasonchsu@tmu.edu.tw.*
Post-Stroke Prediction on Cognitive Impairment Development: A Machine Learning Approach
Muhammad Sollhuddin Muhtar1, Faizul Hasan2, Alex P.A. Nguyen3, Jason C. Hsu1, Hsiao-Yen Chiu2, Min-Huei Hsu3
1Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
2School of Nursing, College of Nursing, Taipei Medical University, Taipei, Taiwan

Background
Cognitive impairment following stroke has wide prevalence ranging from 25% to 81%. Further, stroke and the subtypes, including ischemic stroke, transient ischemic attack and intracerebral hemorrhage, significantly increase the long-term risk of dementia after 5 and 10 years. The incidence rate of post-stroke dementia increases yearly, though the relative risk gradually decreases. The study aims to predict the dementia development one year after stroke diagnosis (index date).

Methods
The study conducted on TMUCRD from January 2004 to September 2017. The inclusion, exclusion and outcome criteria are selected based on ICD9 and ICD10 codes. We include all patient with history of stroke, insomnia, cognitive impairment and other codes related with the diseases (362.3, 433.x, 434.x, 436, 431.x, 430.x, 435.x, H34.1, H63.x, I64.x, I61.x, I60.x, G45.x).
We exclude psychiatric disorder, sleep apnea, traumatic brain injury, cancer, Parkinson’s disease, and cognitive impairment from the outcome (300.4, 296.2-296.3, 300, 293.84, 296.4-296.7, 295, 327.23, 800-804, 850.0, 850.1, 850.5, 850.9, 854.0, 959.01, 139.1, 332.0, F34.1, F32.9, F41.9, F31.x, F32, F32.x, F33, F33.x, F20, 9, G47, 33, 502.0, 502.1, 502.8, 502.9, 504.02, 504.03, 504.04, 506, 507.1, 774.4, 509.90, C80.1, G26).
The outcome are mild cognitive impairment, Senile dementia, uncomplicated, Senile dementia with delusional or depressive features, Senile dementia with delirium, Dementia in conditions classified elsewhere, Alzheimer’s disease, Frontotemporal dementia, and Senile degeneration of brain (331.83, 290.1, 290.2, 290.3, 294.1, 331.0, 331.1, 331.2, G31.84, R41.89, R41.84, R41.83, R41.82, R41.81, F00.0, F00.3, F03).
The patient with outcome at least one year after stroke index date are labelled by 1, and the rest without outcome labelled by 0. We use pycaret library to compare the performance of many different machine learning algorithms.

Results
We can see the performance in Figure 1. The true positive is still less than the false negative. This is to be expected since we use common features between two labels. The Figure 2 shows the optimal threshold is pretty low, 0.27, much lower than default 0.5 for binary classification. Figure 3 shows the distribution of the data based on labels. We can see that some positive labels are overlapping with the negative ones (have the same features).

Conclusions
The current model is able to determine whether a patient will develop cognitive impairment in the next year or not, though the probability is still very low, by using only gender, age and ICD code. Further features engineering will be conducted to improve the performance, such as adding medication or demographic features, especially to increase the true positive and to reduce the false negative numbers. Some hyperparameters may need to be adjusted to obtain better metrics, since the current model still uses the default pycaret parameters. We plan to run it on the CDM once the final result is reasonable.

References

Machine Learning Study (Prediction)

Contact: jasonhshu@tmu.edu.tw
Observational Study (Estimation)

Contact: jasonhsu@tmu.edu.tw
Future trends and prospects

EMR + Claims

Estimation + Prediction
(Data Science)

Multicenter Data Collaboration
(Domestic/International)
THANKS

Jason C. Hsu

Chair, Local Host Committee of the 2022 OHDSI APAC Symposium
Associate Professor, Taipei Medical University, Taiwan
Director, Clinical Data Center, Office of Data Science, Taipei Medical University