Patient-Level Data Integration of De-Identified Healthcare Databases to Support Improved Predictive Analytics

Yang Yang, Reza Sharifi Sedeh, Min Xue, Nandini Raghavan, Daniel Elgort

Contact: yang.yang_1@philips.com; reza.sharifi.sedeh@philips.com; daniel.elgort@philips.com

1, Introduction

Various types of de-identified healthcare databases, from clinical and administrative to utilization, have emerged recently, which enable researchers to perform analyses in each individual domain. However, in the absence of patient identifying data features, current methods do not allow for patient record level integration across these de-identified databases.

In this paper, we propose a novel approach to overcome this limitation and integrate multiple de-identified databases on the patient record level so that inter-domain research problems become addressable. In addition, we have developed a scalable healthcare data analytics pipeline, which incorporates multiple machine learning methods, including penalized and splined linear models, logistic regression, random forest, and survival models. Based on the nature of the integrated database and the analytics purpose, users are provided with options to use any combination of the available machine learning methods in a timely manner. Adopting this strategy, users could obtain more meaningful findings from the integrated dataset compared with using a single database or relying on a single analytical method.

2, Data Integration Approach

Many aggregated healthcare databases are strictly de-identified that all of the hospital and patient identifiers are removed before any secondary use by researchers [Meystre et al., 2010]. This fact makes the integration across databases highly challenging. Recently we developed a hierarchical approach to integrate de-identified databases on the patient level using non-uniquely identifying patient features. For example, age, sex, weight, primary diagnosis and length of hospital stay. The general approach follows:

- generate UID from features for each patient.
- calculate patient rarity score for each patient.
- use rarest patients to identify the same hospitals across databases.
- match patients belonging to the same hospital across databases and repeat it for all matched hospitals.
- categorize matching results into confident, impossible and possible matches.

Below is an example for a patient with UID 1.5.1.122.18, and his calculated rarity score.

gender	race	mortality	LOS	age	Rarity Score
45%	0.1%	10%	0.01%	1%	$4.5*10^{-11}$
Percentage of male patients	Percentage of Native Americans	Percentage of the dead	Percentage of patients with LOS >= 122 days	Percentage of patients with Age <= 18	=0.45*0.001*0.1*0.0001*0.01

Table 1.: Calculating rarity score for the Native American, 18-year-old, male patient who has a LOS of 122 days and has died in hospital.

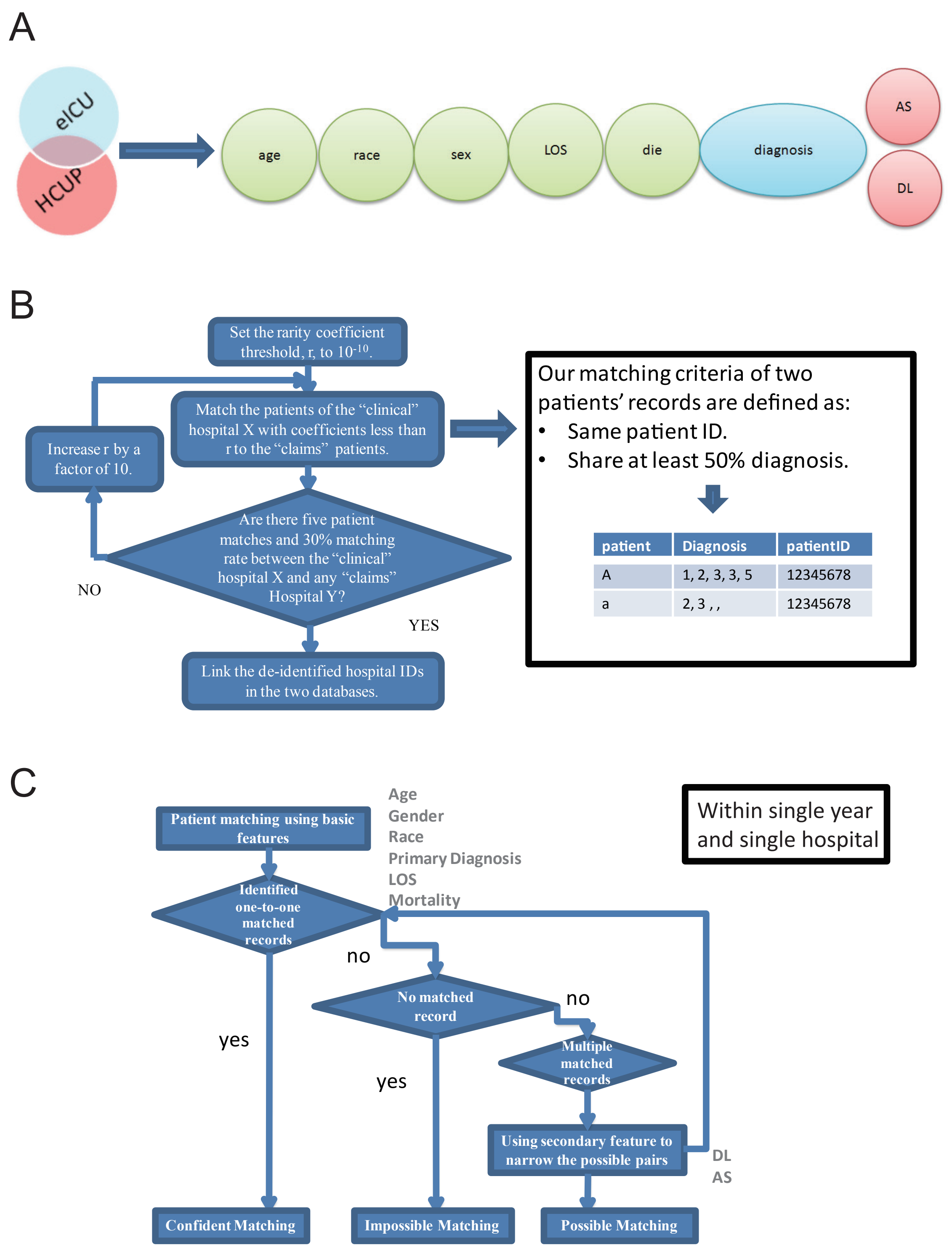
The rarity score $4.5 * 10^{-11}$ can be interpreted as, in every 22 billion patients from the hospital population, there is only one patient with the same UID as him.

3, Data Integration Approach con't

After generating UIDs for patients, we further added the diagnosis codes to reduce duplicated patient matches. The ICD9 [for Medicare et al., 2011] codes was collapsed to Clinical Classifications [Cost et al., 2010] for better accuracy and robustness of the matching. The general rules of matching two patients can be summarized:

- the patients have the exact, same patient UID.
- the patients share at least 50% of the diagnosis codes of the patient with less number of diagnosis codes. For example, if six and ten diagnosis codes have been assigned to Patient A in the "clinical" database and Patient B in the "claims" database respectively, then Patient A and Patient B must share at least three diagnosis codes to convince us there is a match.

Finally, we summarized the hierarchical matching algorithm into the following flowchart:



Set the rarity coefficient threshold, r, to 10^{-10}.

Match the patients of the "clinical" hospital X with coefficients less than r to the "claims" patients.

Are there five patient matches and 30% matching rate between the "clinical" hospital X and any "claims" Hospital Y?

Increase r by a factor of 10.

Link the de-identified hospital IDs in the two databases.

Our matching criteria of two patients' records are defined as:

- Same patient ID.
- Share at least 50% diagnosis.

patient	Diagnosis	patientID
A	1, 2, 3, 3, 5	12345678
a	2, 3 , ,	12345678

Patient matching using basic features

Identified one-to-one matched records

No matched record

Multiple matched records

Using secondary feature to narrow the possible pairs

Confident Matching

Impossible Matching

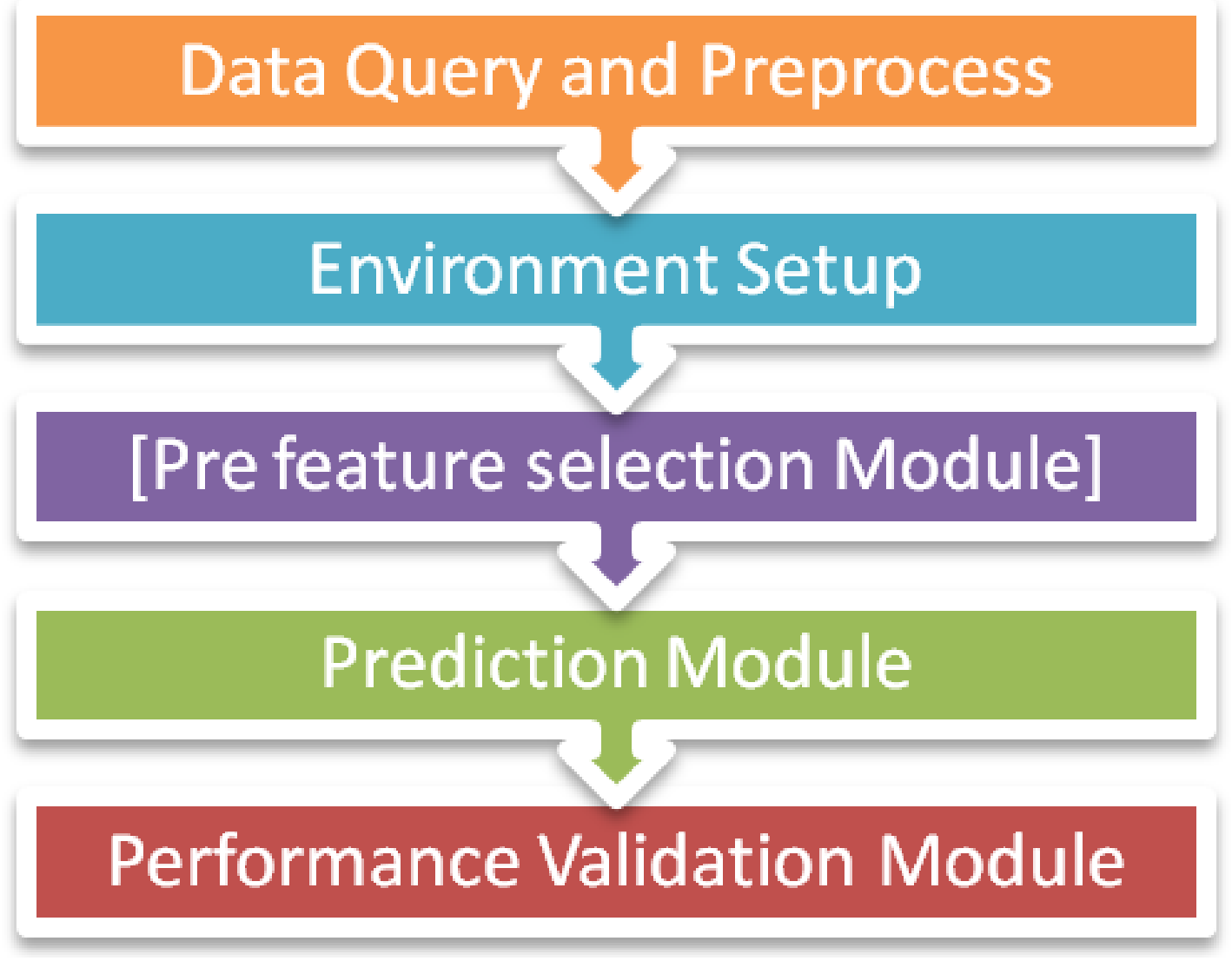
Possible Matching

Within single year and single hospital

Figure 1.: A. Integration of eICU and HCUP using provided common features (eICU and HCUP are two different healthcare databases. See Section 5. Data Application); B. Hospital matching algorithm flowchart; C. Individual patient matching algorithm flowchart.

4, Analytics Pipeline

PhilipsHealthcareBDS is an automated pipeline which gives the user opportunities to execute a range of statistical/machine learning models on a specific dataset in a neat and fast manner. The whole pipeline is written in R language and it is a Linux command-line based program. The pipeline contains five modules in a flowchart (Figure 2). The pipeline features flexible parallel/serial scheme, flexible model parameter tuning, robustness to different datasets with mixed types of explanatory and response variables, complete logging and error collecting system, and the ease to add more models in the future. Currently the pipeline contains ten models/algorithms, including Generalized Linear Model with stepwise variable selection; Lasso, Ridge and Elastic Net algorithm; Group Elastic Net algorithm; SCAD/MCP algorithm; Random Forest; Random Survival Forest; Quantile Regression and Normal-Probit Bivariate Model.



Data Query and Preprocess

Environment Setup

[Pre feature selection Module]

Prediction Module

Performance Validation Module

Figure 2.: Flowchart of PhilipsHealthcareBDS. Module within square parentheses is optional.

5, Data Application

We integrated patients from Philips eICU database and Healthcare Cost and Utilization Project (HCUP[a]) State Inpatient Database (SID) for Massachusetts between 2008 and 2011. From this full dataset, by "DX1" (primary diagnosis ICD-9 code) values we further extracted those with Heart Disease (i.e., Heart Failure and Cardiovascular Myocardial Infarction). The variables available are clinical variables, utilization variables, billing variables, demographic variables and hospital characteristics.

We selected and applied five analytical methods on the real data including: 1, Linear regression with stepwise variable selection by AIC criteria; 2, Penalized linear model such as elastic net, SCAD and MCP; 3, Group based penalized linear model; 4, Random Forest; 5, Quantile Regression.

[a] Disclaimer: Study design, Data sources, analysis and findings described in this paper were executed in compliance with the Data Use Agreement of HCUP.

References

Healthcare Cost, Utilization Project, et al. Clinical classifications software (ccs) for icd-9-cm. *Rockville, MD: Agency for Healthcare Research and Quality*, 2010.

Centers for Medicare, Medicaid Services, et al. Icd-9-cm official guidelines for coding and reporting. *US GPO, Washington, DC*, 2011.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70, 2010.

6, Results

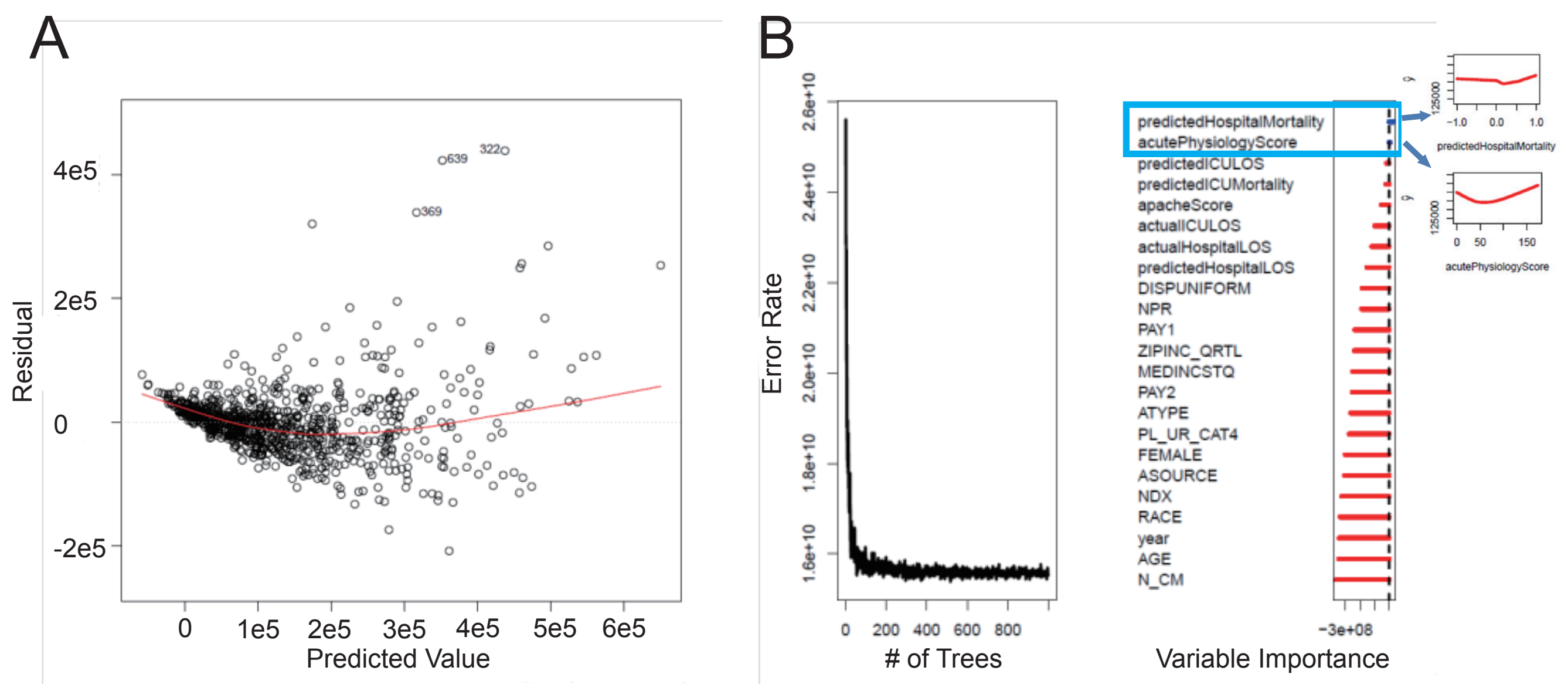


Figure 3.: (A) Linear regression residual plot of in-hospital expenditure. (B) Random forest tree error rate (left panel) and variable importance rank (right panel). The two variables with the highest effects (blue box) on in-hospital expenditure were plotted versus the in-hospital expenditure (the two rightmost plots).

Summary of conclusions

- We found a significant correlation between the actually observed values of mortality or length of stay (from eICU) and the in-hospital expenditure (from HCUP).
- We learned that the in-hospital expenditures (HCUP) of the patients who died in hospital (eICU) are higher than those alive.
- We found the patients in either extremely bad condition or excellent condition, inferred from "Predicted Hospital/ICU mortality" or "Acute Physiology Score" (eICU), have higher in-hospital expenditures than patients with moderate condition. These two variables were ranked as the top two predictors of expenditure by a random forest method (Figure 3B).

In addition, there are several other findings: Asian or Pacific Islander patients paid *more*; patients with *more* interventional procedures paid *more*; patients with *longer* actual hospital/ICU lengths of stay paid *more*; patients admitted from other health facilities paid *less*.