2 December 2019

Japanese-German-French Forum on AI and Healthcare –Quality Standards for AI Applications in Healthcare and Joint Database for Medical Data

Session II: Joint Database for Medical Data

# Benefits and Limitations of Large-Scale Health Databases in Japan

Hideo Yasunaga

Department of Clinical Epidemiology and Health Economics

Graduate School of Medicine, The University of Tokyo

THE UNIVERSITY OF TOKYO

# Topics

1. Overview of large healthcare databases in Japan
2. Studies using large healthcare databases
3. Future perspectives: linkage between multiple databases

# 1. Overview of large healthcare databases in Japan

# What are *large* healthcare data (or health-related *big* data)?

Large healthcare data are electronic data on any records of individual people's events related to health and healthcare, which can be searched, browsed, synthesized, and statistically analyzed.

# For example,

When one is born→Birth Certificate
When one is dead→Death Certificate
⇨ Vital Statistics
人口動態統計

Health checkups for workers
⇨ Specific health checkup database
特定健診データベース

Clinic visit or hospital admission
⇨ National Health Insurance Databases
レセプト・データベース
Electronic Medical Records
電子カルテ

When one is diagnosed with cancer
⇨ Cancer registry
がん登録

When one received long-term care
⇨ Long-term Care Benefit Expenditures
介護給付費実態統計

# Randomized controlled study

-Randomized Controlled Trial (RCT) is the gold standard for clinical and epidemiological studies.

-However, RCT is not always feasible due to ethical and financial problems.

-Observational studies using large healthcare databases can be a feasible alternative to RCT.

# Limitations of large healthcare databases

The observational design and resulting statistical control of **confounding factors** provides a weaker framework for **internal validity** and especially **causal inference** of exposure-disease or treatment-effect associations than experimental designs.

Some **quasi-experimental methods**, such as **propensity scores analysis** and **instrumental variable method**, can partially address this issue.

# National health insurance databases in Japan

(1)National DataBase of administrative claims (NDB):

Administrative claims data for all the inpatients and outpatients across Japan

(2) Diagnosis Procedure Combination (DPC) database:

Nationwide inpatient database including administrative claims data, discharge abstract, and some clinical data of approx. 8 million inpatients/year from approx. 1000 acute-care hospitals

# 2. Studies using large healthcare databases

# *Evaluating effectiveness of drugs using large healthcare databases*

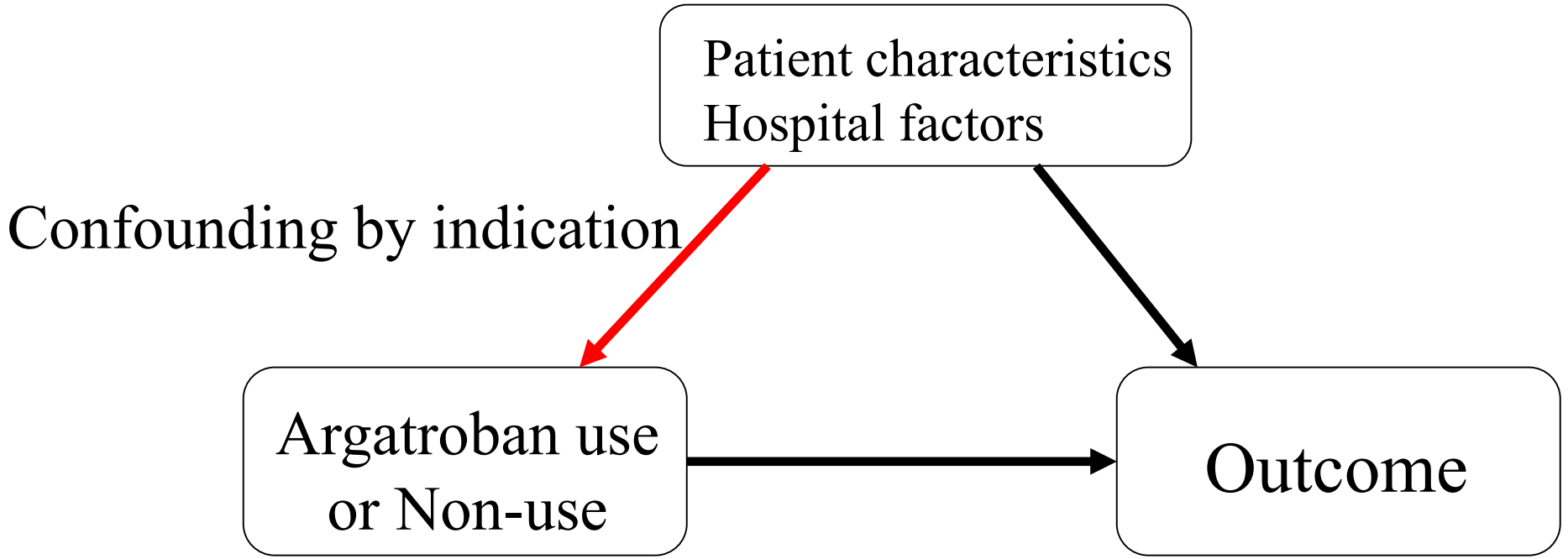
Argatroban Treatment in Patients with Atherothrombotic Stroke
(***Stroke*** 2016 ;47:471-6)

Argatroban is a selective thrombin inhibitor, used for patients with atherothrombotic stroke.
However, effectiveness of this drug on stroke outcomes remains uncertain.

# Retrospective observational study

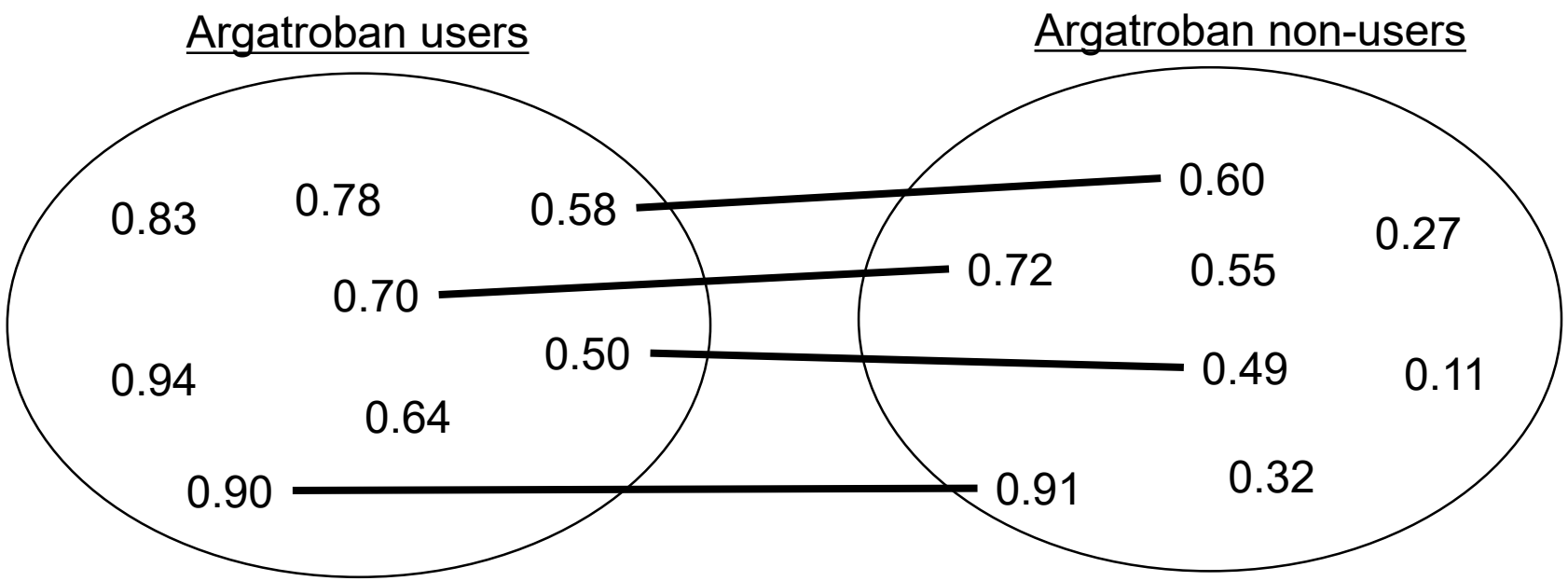
-Non-randomized

-Confounding factors



# Propensity score matching

-The log odds of the probability that a patient received argatroban was modeled for potential confounders.
-A one-to-one matched analysis using nearest-neighbor matching.



# RESULTS

2289 propensity-score-matched pairs

No significant differences in modified Rankin Scale at discharge between the argatroban and the control groups (adjusted odds ratio, 1.01; 95% confidence interval, 0.88-1.16).

No significant differences in the occurrence of hemorrhagic complications between the argatroban and the control groups (3.5% versus 3.8%; P=0.58).

# CONCLUSIONS

Argatroban was safe, but had no added benefit in early outcomes after acute atherothrombotic stroke.

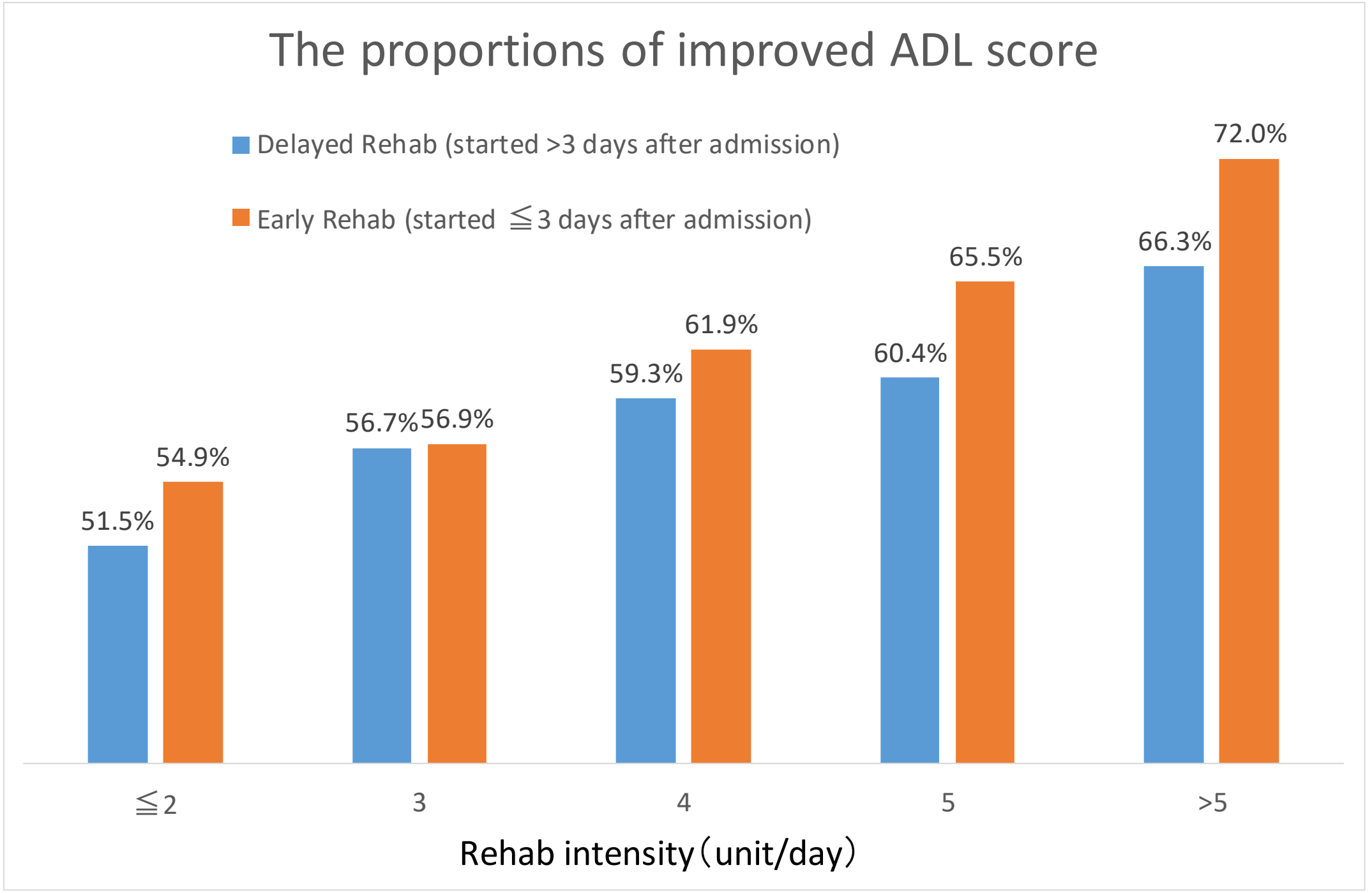
# *Evaluating effectiveness of treatments using large healthcare databases*

Impact of Rehabilitation on Outcomes in Patients with Ischemic Stroke

(***Stroke*** 2017;48:740-746)

Using the DPC database, we analyzed patients with ischemic stroke who received rehabilitation (n=100,719) from 2012 to 2014.

We examined the association of early and intensive rehabilitation with the proportion of improved activities of daily living (ADL) among patients with ischemic stroke.


The proportions of improved ADL score
Delayed Rehab (started >3 days after admission)
Early Rehab (started ≦3 days after admission)
51.5%
54.9%
56.7%
56.9%
59.3%
61.9%
60.4%
65.5%
66.3%
72.0%
≦2
3
4
5
>5
Rehab intensity（unit/day）

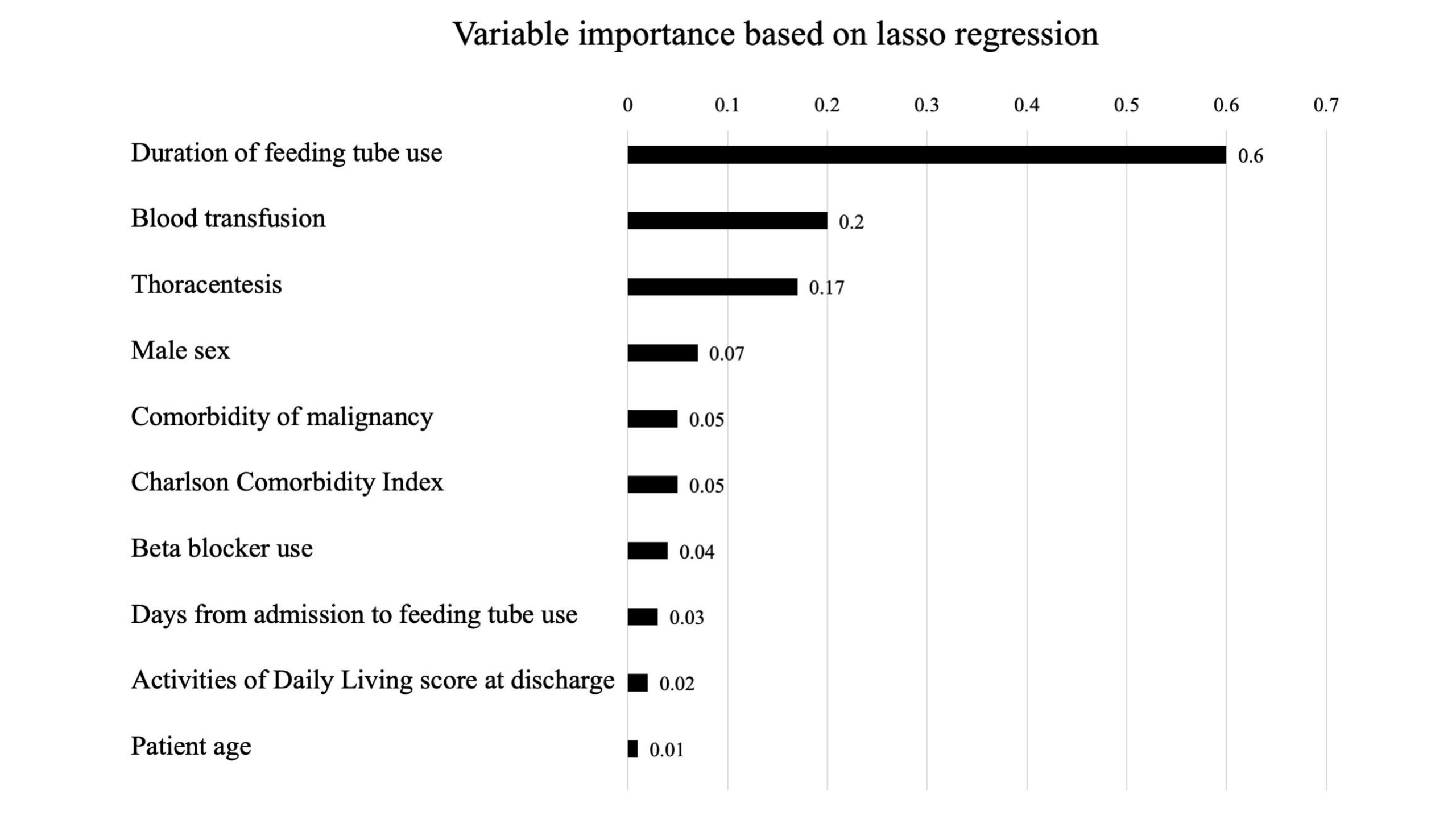
# Conclusion

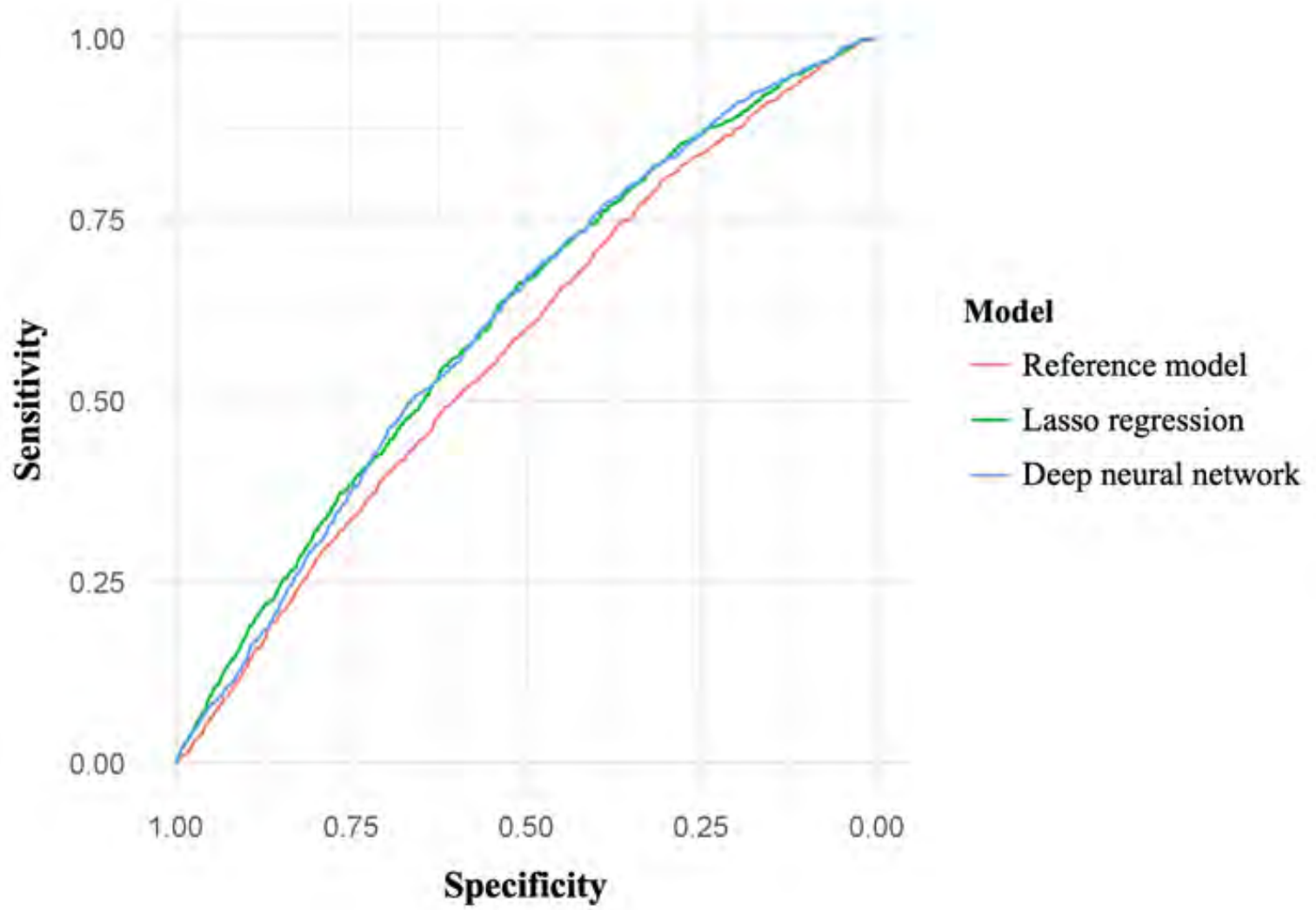
Early and intensive rehabilitation improved ADL during hospitalization in patients with ischemic stroke.

# *Machine learning using large healthcare databases*

Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease
(*COPD: Journal Of Chronic Obstructive Pulmonary Disease* 2019 in press)

We identified 44,929 patients aged ≥40 years with unplanned hospitalization for COPD in the DPC database from 2011 through 2016.
Of them, 3,413 (7%) were readmitted within 30 days after discharge.
In the training set (70% of sample), patient characteristics and inpatient care data were used as predictors to derive a conventional logistic regression and two machine learning models (lasso logistic regression and deep neural network). In the test set (remaining 30% of sample), the prediction performances of the machine learning models were examined.


Variable importance based on lasso regression
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Duration of feeding tube use
0.6
Blood transfusion
0.2
Thoracentesis
0.17
Male sex
0.07
Comorbidity of malignancy
0.05
Charlson Comorbidity Index
0.05
Beta blocker use
0.04
Days from admission to feeding tube use
0.03
Activities of Daily Living score at discharge
0.02
Patient age
0.01




| | C-statistic | p- value* |
|---|---|---|
| Reference model | 0.57 (0.56–0.59) | Reference |
| Lasso logistic regression | 0.61 (0.59–0.62) | 0.004 |
| Deep neural network | 0.61 (0.59–0.63) | 0.007 |

# 3. Future perspectives: linkage between multiple databases

# Privacy protection regarding health and healthcare data

Health and healthcare data are handled under the legal framework for personal data, including **Act on the Protection of Personal Information** and related guidelines.

These legislation have two important aspects:

(i) patient consent for collecting and using routinely collected data

(ii) de-identification of the routinely collected data

# Data linkage of multiple databases

**Next-generation Healthcare Infrastructure Act** (or Healthcare Big Data Act) was put into force in 2018. The purposes of this new law include:

(i) "certified operators for de-identifying medical data" would be entrusted with managing patients' personal information.

(ii) medical institutions are required to post up a notice announcing that anonymized patient data will be secondarily used for research purposes.

Under this law, various health and healthcare databases can be linked together by "certified operators for de-identifying medical data", and researchers can be provided with the linked data.

*Thank you for your attention.*