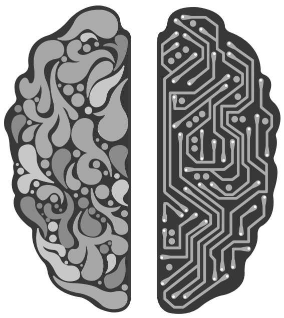
# How Artificial Intelligence and Machine Learning Can Help Answer Many Questions in New Ways

*AIM-AHEAD is a program created by the National Institutes of Health to increase the participation and representation of researchers and communities currently underrepresented in use of AI/ML to understand and address their community's health needs.*

## What is Artificial Intelligence and Machine Learning?

Artificial intelligence (AI) is a broad array of analyses in which a machine (e.g., computer, phone) completes tasks based on a set of specified rules, or algorithms. Machine learning (ML), a type of AI, is a method that uses computer software to automatically identify patterns in existing data and develop algorithms in order to be able to predict specific outcomes. For example, ML may identify patterns in data on associations between a person's characteristics and their risk of heart disease, which may then be employed to predict an individual's risk of heart disease onset based on those characteristics.



Examples of ML applications in health research are provided below.

## How can AI/ML increase our ability to address health risks, enhance health service delivery, and increase the effectiveness of limited resources?

Data on a population's health needs, service utilization, and expenditures are available through:

- National, regional, and local surveys;
- Hospital and clinic health records like patient charts or electronic health records (EHRs);
- Administrative records (e.g., data warehouses, registries for specific conditions) that include data from EHR and other data systems; and
- Health service payment data (or "claims data") from private health insurers (e.g., Kaiser, Anthem), Medicare, Medicaid, and other payers.

These data typically include information on a person's age, sex, location, health insurance, health status, service use, and costs. Researchers conduct analyses of these data to improve our understanding of a population's health risks, health service utilization (e.g., treatment, follow-up), health expenditures, and mortality. The analyses may be conducted using traditional parametric statistical techniques and/or through ML approaches.

Logistic and Cox proportional hazard regression models (parametric models) have been used to develop models to predict health outcomes (for example, predicting onset of diabetes) for a long time. Parametric statistical techniques rely on parametric assumptions about the data. For example, a logistic regression requires that the log odds of the outcome (e.g., having diabetes) be a linear combination of one or more independent predictor variables (e.g., patient characteristics). However, these assumptions may not be met in real-world data analyses. A systematic review of research findings found the linearity assumption was only verified in 5 of 56 Logistic and Cox regression models examined.[1] If key assumptions of the model are not met, the model results can be biased.

In addition, some persons may have missing data for important characteristics, and some characteristics may be related to each other (referred to as multicollinearity) in a way that potentially biases the results. ML may be used to predict the outcome without parametric model assumptions and address missing data and multicollinearity effectively. Additionally, AI/ML requires less pre-processing of predictor variables (e.g., making new measures from existing measures in order to meet the linearity assumption) and has higher computation efficiency, compared to parametric regression models, when analyzing large amounts of data. Thus, AI/ML could more efficiently result in better prediction models, which may be used by health providers, planners, and policy experts to provide improved and coordinated care for those at risk for a specific condition, being admitted to the hospital, or requiring a wide array of health resources.

# ML Applications in Health Research Studies

## Study 1: Patient Risk Factors and Onset of Diabetes[2]

**Study population and data:** 138,146 adults aged 31 and older who participated in the 2014 Behavioral Risk Factor Surveillance System (BRFSS).

**Outcome measure:** Self-reported onset of type 2 diabetes among the adults.

**Analysis:** The authors compared the performance of 8 ML predictive models using R software.

**Results:** Among the adults, 14.8% had diabetes. The neural network model had the best performance in predicting diabetes onset when the ML results for the 8 models were compared to the actual data on diabetes onset. The analyses confirmed patient risk factors for diabetes reported in previous research (age, body mass index, sex, race/ethnicity) and identified 2 new potential risk factors using the BRFSS data. Less frequent visits for medical check-ups and having slept 9 or more hours per day were associated with higher risk of diabetes onset.

**Implications:** Such models could help facilitate early diagnosis and intervention for diabetes and reduce medical costs in the long-term.

## Study 2: Understanding Racial Bias in Electronic Health Records (EHR)[3]

**Study population and data:** 2020 EHR data for 18,459 patients obtaining services at an urban academic medical center.

**Outcome measure:** One or more uses of stigmatizing language recorded in the EHR clinical history and physical notes (defined as 1 of 15 types of negative descriptors).

**Analysis:** Using Python's Natural Language ToolKit, the authors standardized the text data and split notes into sentences. ML methods were used to develop the model to analyze the resulting data.

**Results:** Compared with White patients, Black patients had 2.54 times the odds of having at least one negative descriptor in the EHR history and physical notes.

**Implications:** This study raised concerns about stigmatizing language in the EHR and its potential to exacerbate racial and ethnic health care disparities.

## Study 3: Predicting the Risk of Inpatient Hypoglycemia among Patients with Diabetes[4]

**Study population and data:** 2014-2018 EHR data for 17,658 adults with diabetes who had 32,758 admissions to a large university hospital.

**Outcome measure:** Having one or more hypoglycemic events (i.e., inappropriately low blood glucose) during a hospital stay. A 2017 study found that almost 1 in 5 people with diabetes experienced hypoglycemia during a hospital stay. Such hypoglycemia events are associated with higher mortality.

**Analysis:** The highest performing ML model was the XGBoost model. Analyses were performed using R version 3.3 and Python 3.6.

**Results:** Among inpatients with diabetes, predictive factors for having hypoglycemia during the hospital stay were previous hypoglycemia during an inpatient stay, albumin level, type 2 diabetes compared to type 1, medication use, and having certain types of procedures.

**Implications:** In this study, advanced ML models were superior to logistic regression models in predicting the risk of hypoglycemia among inpatients with diabetes. Future studies of such models should be conducted conducted to evaluate their utility in real time to reduce inpatient hypoglycemia among hospitalized adults with diabetes.

# ML Applications in Health Research Studies

## Study 4: Identifying Patient Risk Factors Associated with Preventable Hospital Readmissions[5]

**Study population and data:** Administrative claims for a network of hospitals in Florida for 2005 to 2012. Adults aged 18 and older with 1 of 5 conditions (type 2 diabetes, congestive heart failure, acute myocardial infarction, pneumonia, and chronic obstructive pulmonary disease).

**Outcome measure:** Preventable hospital readmissions within 30 days of discharge from a previous admission.

**Analysis:** ML models were developed for each of the 5 disease cohorts using functions available in the R package caret.

**Results:** Patient characteristics associated with high risk of being readmitted to the hospital for a possible preventable reason included a longer length of stay in the first admission, disease severity index, being discharged to another hospital, and primary language other than English. The predictive capabilities of the models varied by condition.

**Implications:** Improved predictive models for preventable hospital readmissions are needed since understanding patient risks for preventable hospital readmissions can point to the enhancement of interventions needed to improve health care quality and efficiency.

## Study 5: Identifying Patient Risk Factors Associated with High Expenditures after Surgery[6]

**Study population and data:** 2012-2016 Medicare outpatient and inpatient data for over 1 million adults aged 65 years and older who underwent one of six types of surgeries.

**Outcome measure:** Medicare inpatient and outpatient expenditures during the year following the surgical procedure were classified into four expenditure levels. Patients in the highest expenditure level were referred to as "super-utilizers."

**Analysis:** Logic Forest analysis, an ML algorithm, was performed using R, version 3.2.5, Logic Forest package.

**Results:** Nearly 5% of patients were classified as super-utilizers. Their Medicare expenditures averaged $72,800 during the 12-month period. The expenditures for super-utilizers represented 30% of all of the study populations' Medicare expenditures during that 12-month period. Preoperative risk factors associated with being a super-utilizer included hemiplegia/paraplegia (types of paralysis), chronic heart failure, chronic renal failure, and weight loss.

**Implications:** If patients at high risk for very high use of health services during the year following their surgery can be identified proactively and provided targeted services, it may be possible to improve their health outcomes and reduce expenditures.

AIM AHEAD
NORTH & MIDWEST HUB
University of Colorado Anschutz Medical Campus

Centers for American Indian & Alaska Native Health
colorado school of public health

This work is part of AIM-AHEAD (Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity). The Centers for American Indian and Alaska Native Health serve as the lead for the North and Midwest Hub of the AIM-AHEAD Leadership Core.

## Want to learn more?

**Contact Amanda Slezsak** (amanda.slezsak@cuanschutz.edu), AIM-AHEAD Consortium Partner Manager at the Centers for American Indian & Alaska Native Health

## Want to get involved?

**Join the AIM-AHEAD Consortium at aim-ahead.net**