# ˅ Predicting Health Insurance Fraud Using Machine Learning

## ˅ Group 2: Nhan Nguyen, Tan Nguyen, Andre Serna

## I. Introduction

Healthcare fraud is a pervasive issue that has significant financial, ethical, and operational implications, affecting government programs, insurance companies, healthcare providers, and patients alike. Fraudulent activities in healthcare result in billions of dollars in financial losses annually, placing a burden on both public and private insurers while increasing overall healthcare costs. The **National Health Care Anti-Fraud Association (NHCAA)** estimates that healthcare fraud costs the United States **tens of billions of dollars each year**, affecting the efficiency and sustainability of healthcare programs.

Fraud in health insurance can manifest in various forms, including **phantom billing** (charging for services never provided), **upcoding** (billing for more expensive procedures than those performed), **unbundling** (separating services that should be billed together), **duplicate billing**, and **identity fraud**. These fraudulent activities exploit loopholes in the billing system, making them difficult to detect through traditional rule-based fraud detection techniques. Most existing fraud detection methods rely on manual audits, predefined rules, and expert-based reviews, which are labor-intensive, time-consuming, and ineffective against evolving fraudulent schemes.

Machine learning (ML) and data-driven analytics have emerged as effective solutions for fraud detection in healthcare by leveraging large datasets to identify patterns indicative of fraudulent behavior. Unlike static rule-based systems, ML models can dynamically analyze vast amounts of healthcare claims data, recognize hidden anomalies, and detect fraud more accurately over time. ML-based fraud detection systems also reduce false positives, allowing investigators to focus on high-risk claims while minimizing disruptions for legitimate providers.

This project aims to build a **robust fraud detection model** that can accurately distinguish fraudulent from non-fraudulent claims using a **combination of supervised and unsupervised machine learning techniques**. The analysis will be conducted using **three key datasets**:

- **CMS Medicare Data** – A large dataset of Medicare claims providing insight into billing practices.
- **Kaggle Healthcare Fraud Dataset** – A labeled dataset containing known cases of fraudulent providers.
- **Synthea Synthetic Data** – A simulated dataset modeling real-world healthcare scenarios to enhance training and testing.

By integrating these datasets, we aim to develop an **intelligent fraud detection system** that utilizes **statistical learning techniques from An Introduction to Statistical Learning (ISL)** to enhance fraud identification. The results of this study will contribute to more effective fraud prevention strategies, reducing financial losses and improving healthcare system integrity.

# II. Data

This study aims to **develop predictive models to identify fraudulent health insurance claims using machine learning techniques**. To accomplish this, we will integrate data from three key sources that provide a holistic view of provider behavior, billing trends, and synthetic healthcare claims:

## 1. Medicare Physician & Other Practitioners Data (CMS)

- **Source:** [CMS Medicare Data](#)

## Observations

- Estimated millions of records per year.

## Data Description

- This dataset provides aggregated Medicare billing data, categorized by provider geography (state or national level) and specific services rendered.
- It includes key Medicare payment metrics, allowing analysis of billing patterns across different regions and providers.
- **The dataset is organized at two levels:**
  - **State-level aggregation** – Billing summarized per state.
  - **National-level aggregation** – Billing summarized across all states.

## Key Variables from the CMS Data Dictionary

- **Rndrng_Prvdr_Geo_Lvl** – Geographic level (State/National).
- **Rndrng_Prvdr_Geo_Cd** – FIPS code for the provider's state.
- **Rndrng_Prvdr_Geo_Desc** – State name where the provider is located.
- **HCPCS_Cd** – HCPCS medical service code (CPT codes included).
- **HCPCS_Desc** – Description of the medical service provided.
- **HCPCS_Drug_Ind** – Indicator for Medicare Part B drugs.
- **Place_Of_Srvc** – Facility type (F = Facility, O = Office).
- **Tot_Rndrng_Prvdrs** – Number of providers rendering the service.
- **Tot_Srvcs** – Total services performed.
- **Tot_Benes** – Number of distinct Medicare beneficiaries served.
- **Tot_Bene_Day_Srvcs** – Unique services per beneficiary per day.
- **Avg_Sbmtd_Chrg** – Average submitted charge amount.
- **Avg_Mdcr_Alowd_Amt** – Average Medicare allowed amount.
- **Avg_Mdcr_Pymt_Amt** – Average Medicare payment amount.
- **Avg_Mdcr_Stdzd_Amt** – Standardized Medicare payment (adjusted for location differences).

## How It Was Collected

- This dataset is compiled from Medicare claims, submitted by healthcare providers.
- It is updated annually and includes detailed payment structures and billing codes.

## Why This Dataset Was Chosen

- It allows us to analyze provider billing trends and detect anomalies.
- Helps in identifying potential fraud indicators, such as excessive billing, duplicate charges, or unusual regional patterns.

## 2. Healthcare Provider Fraud Detection Analysis (Kaggle)

- **Source:** [Kaggle Dataset](#)

## Observations

- The dataset contains **8 CSV files**, separated into **training and test datasets** for model development and evaluation.
- It consists of **167 columns**, representing different variables related to healthcare provider claims, patient conditions, and reimbursement details.
- The dataset includes **104 string variables**, **46 integer variables**, **14 datetime variables**, and **3 other types of variables**.

## Data Description

- This dataset is designed to support the analysis of **fraudulent behaviors among healthcare providers** by examining various factors such as billing patterns, service types, patient conditions, and claim amounts.
- **It provides labeled data**, making it suitable for supervised machine learning approaches to fraud detection.
- The dataset contains **a mix of categorical, numerical, and time-based variables**, allowing for a comprehensive analysis of fraudulent trends over time.

## Variable Summary

- **String Variables (104)** – Includes categorical data such as provider names, patient demographics, and diagnosis codes.
- **Integer Variables (46)** – Contains numerical counts such as claim amounts, reimbursement values, and service counts.
- **Datetime Variables (14)** – Represents timestamps related to admission, discharge, and claim submission dates.
- **Other Variables (3)** – Includes specialized data types used for internal processing.

## How It Was Collected

- This dataset aggregates information from **real-world healthcare provider claims** and **fraud investigations**, making it a valuable resource for developing machine learning models.
- The dataset includes labeled fraud cases, allowing researchers to study **patterns of fraudulent behavior based on past investigations**.

- Data was collected from **insurance claims and medical records**, covering a broad range of patient-provider interactions.

## Why This Dataset Was Chosen

- **Supervised Learning Capability**: The dataset includes a `PotentialFraud` label, making it useful for training **classification models** to detect fraud.
- **Rich Feature Set**: With **167 variables**, the dataset provides diverse information, allowing for deep fraud analysis across **billing behaviors, medical conditions, and provider characteristics**.
- **Real-World Relevance**: The dataset reflects actual insurance claims and fraud detection practices, making it applicable to **healthcare fraud prevention strategies** in the real world.

---

This dataset plays a **crucial role** in our project, serving as the **primary labeled dataset** for training machine learning models to **identify fraudulent healthcare providers**.

## ⌄  3. Synthea Synthetic Patient Data

- **Source:** [Synthea Downloads](Synthea Downloads)

## Observations

- The dataset consists of multiple CSV files, each representing different aspects of **synthetic patient health records**.
- It includes **a comprehensive set of simulated healthcare data**, covering **millions of patient records**.
- The dataset contains various data types, including **categorical, numerical, and time-based information**.

## Data Description

- Synthea generates **realistic but fully synthetic** healthcare records, ensuring privacy compliance while enabling large-scale health data analysis.
- The dataset provides a **complete longitudinal history** for each patient, simulating events from birth to death.
- It is structured to mirror real-world **Electronic Health Records (EHRs)** and includes multiple categories of healthcare interactions.

## Variable Summary

- **Demographics** – Patient age, gender, race, ethnicity, and social determinants of health.
- **Encounters** – Records of patient visits to healthcare providers, including visit dates and types.
- **Conditions** – Diagnosed diseases and medical conditions with onset and resolution dates.
- **Medications** – Prescribed medications, including dosages and administration timelines.
- **Procedures** – Medical procedures performed, along with associated billing codes.
- **Observations** – Clinical measures such as lab test results and vital signs.
- **Allergies** – Documented allergic reactions and substances.
- **Immunizations** – Vaccination records for each patient.
- **Care Plans** – Treatment strategies and follow-up care details.

- **Time-Based Information** – Admission and discharge dates, procedure timestamps.

## How It Was Collected

- The dataset is generated using **Synthea**, an open-source patient data simulation tool.
- It is based on **clinical models and publicly available health data sources**, such as CDC and NIH guidelines.
- The simulation process replicates **disease progression, medical interventions, and patient-provider interactions**.
- Since the data is **entirely synthetic**, it does not contain any **real patient information**, ensuring privacy and unrestricted research use.

## Why This Dataset Was Chosen

- **Privacy-Safe Data**: Unlike real-world medical records, this dataset poses **no privacy concerns**, making it ideal for unrestricted analysis.
- **Comprehensive Medical Records**: The dataset includes a **full spectrum of patient health information**, covering conditions, treatments, and billing events.
- **Standardized Formats**: Available in CSV, FHIR, and C-CDA formats, ensuring compatibility with various healthcare data analysis tools.
- **Scalability**: The dataset can simulate **millions of patients**, allowing researchers to test models on large-scale medical datasets.

---

This dataset plays a **critical role** in our project, providing **a synthetic yet realistic representation of patient-provider interactions**, which can be used to **train and evaluate fraud detection models** without violating privacy laws.

Each dataset will be preprocessed and merged based on key attributes such as NPI (National Provider Identifier) and geographic information.

# III. Methods

## ⌄ **Overview of Approach**

To build an effective fraud detection model, we adopt a structured methodology that follows the techniques covered in **An Introduction to Statistical Learning (ISL)**. This approach ensures our methodology is statistically robust while remaining interpretable and effective in detecting fraudulent claims.

The following steps outline the methodology for fraud detection:

## 3.1 Data Preprocessing & Preparation

Before applying machine learning models, we need to ensure that the datasets are **cleaned, merged, and transformed** into a structured format for analysis.

- **Data Cleaning & Merging**:

  - Remove duplicates, missing values, and inconsistencies.
  - Merge datasets from **CMS, Kaggle, and Synthea** to create a unified dataset.
  - Standardize numerical values (e.g., claim amounts, reimbursement fees) and encode categorical variables (e.g., provider type, patient demographics).

- **Train-Test Split & Resampling**:

  - The cleaned dataset will be split into **training (80%) and test (20%)** sets.
  - Since fraud cases are often **imbalanced**, we will apply **resampling techniques** such as **oversampling (SMOTE) or undersampling** to ensure a balanced distribution of fraudulent and non-fraudulent claims.

## 3.2 Initial Statistical Analysis & Regression Models

To establish a baseline, we will first apply **linear models and generalized linear models (GLMs)** before progressing to advanced methods.

- **Logistic Regression** (Chapter 4, ISL):

  - Models fraud as a **binary classification problem**.
  - Provides interpretable coefficients to identify **risk factors** for fraud.

- **Multiple Logistic Regression**:

  - Extends logistic regression by incorporating multiple predictors, allowing a deeper understanding of fraud patterns.
  - Identifies **significant variables contributing to fraudulent claims**.

- **Generalized Additive Models (GAMs)** (Chapter 7, ISL):

  - Handles **non-linear relationships** between fraud likelihood and billing trends.
  - Allows smooth, interpretable variable interactions for improved prediction accuracy.

## 3.3 Advanced Machine Learning Models

Once the initial models provide insight, we move toward **more sophisticated models** for better fraud classification:

- **Decision Trees & Bagging (Random Forests)** (Chapter 8, ISL):

  - Decision Trees identify key fraud indicators **through feature importance analysis**.
  - Random Forests aggregate multiple decision trees to improve **robustness and accuracy**.

- **Support Vector Machines (SVM)** (Chapter 9, ISL):

  - Effective for handling **high-dimensional fraud classification problems**.
  - Uses **kernel tricks** to detect non-linear patterns in fraud.

- **Boosting Methods (Gradient Boosting, AdaBoost)**:

  - Enhances model performance by reducing **bias and variance**.
  - XGBoost optimizes decision tree ensembles, while AdaBoost improves weak learners.

## 3.4 Clustering for Fraud Detection

To further refine fraud detection, we incorporate **unsupervised learning techniques** from ISL:

- **K-Means Clustering** (Chapter 10, ISL):
  - Groups similar providers and claims to detect **unusual billing patterns**.
  - Useful for **unsupervised fraud detection** when labeled fraud data is unavailable.
- **Hierarchical Clustering**:
  - Identifies provider groups with **similar fraud risk characteristics**.

## 3.5 Model Evaluation & Optimization

To ensure model effectiveness, we will use various **evaluation metrics** from ISL:

- **Confusion Matrix & Precision-Recall Curve** – Measures fraud detection accuracy and trade-offs.
- **Cross-validation (Chapter 5, ISL)** – Ensures model generalization.

## Conclusion

This structured methodology ensures a **comprehensive approach** to fraud detection, integrating **statistical learning techniques from ISL, machine learning methods, and clustering for anomaly detection**. By progressively refining our models, we aim to build a robust system capable of identifying fraudulent claims with **high precision and real-world applicability**.

# IV. Review of Earlier Work

Existing research in healthcare fraud detection has demonstrated that integrating multiple datasets and leveraging machine learning techniques significantly improves predictive performance. Various studies have explored the effectiveness of **supervised and unsupervised learning methods** in fraud classification, providing valuable insights into model selection and data handling challenges.

## ⌄ 4.1 Fraud Detection Using CMS Medicare Data

Several studies have focused on analyzing Medicare data to identify fraudulent billing patterns. A study titled **"Big Data Fraud Detection Using Multiple Medicare Data Sources"** explored the use of machine learning models on publicly available Medicare data to detect fraudulent providers. Researchers applied **decision trees, random forests, and ensemble learning methods** to analyze billing patterns, identifying anomalies indicative of fraudulent activities. The study concluded that **random forests provided the best trade-off between accuracy and interpretability** in fraud detection ([Springer Journal of Big Data](#)).

Another study, **"Medicare Fraud Detection Through Big Data and Machine Learning Models,"** focused on analyzing open data to predict and detect fraudulent Medicare providers. By examining provider claim patterns and integrating geo-demographic metrics, the study identified various fraud schemes, including excessive billing, unnecessary medical procedures, and prescription abuse ([GitHub CMS Fraud Detection](#)).

## 4.2 Fraud Detection Using Kaggle Healthcare Provider Data

The **"Healthcare Provider Fraud Detection Analysis"** dataset on Kaggle has been widely used to train fraud detection models. Researchers have applied **logistic regression, gradient boosting, and support vector machines (SVMs)** to classify providers as fraudulent or non-fraudulent. One such study, titled **"Medical Provider Fraud Detection Using Machine Learning,"** implemented multiple classification algorithms and found that **XGBoost outperformed other methods, achieving high precision and recall scores** ([Kaggle Medical Fraud Project](#)).

Another study, **"Healthcare Provider Fraud Detection and Analysis,"** utilized this dataset to compare feature selection methods and evaluated the impact of different preprocessing techniques on fraud classification performance. The research highlighted the importance of **feature engineering and resampling techniques such as SMOTE** to improve fraud detection accuracy ([Medium Research Post](#)).

## 4.3 Fraud Detection Using Synthetic Healthcare Data

Synthetic datasets have been employed in fraud detection to mitigate data privacy concerns while maintaining model generalizability. A study titled **"Optimizing Fraud Detection Models with Synthetic Data: Advancements and Challenges"** investigated how synthetic data enhances fraud detection models by improving training data diversity. Researchers found that **using synthetic claims data in combination with real-world datasets significantly improved fraud classification performance** ([ResearchGate](#)).

In another study, **"Enhancing AI Fraud Predictive Models with Synthetic Data: A Case Study,"** researchers utilized synthetic datasets to evaluate deep learning approaches for fraud detection. The study demonstrated that **combining synthetic and real-world data helped mitigate class imbalance issues, leading to higher accuracy in fraud classification** ([Dedomena AI Blog](#)).

## 4.4 Summary and Contribution of This Project

Building upon these studies, our project integrates **CMS Medicare data, the Kaggle Healthcare Fraud Dataset, and Synthea Synthetic Data** to develop a comprehensive fraud detection model. By combining these datasets, we aim to:

- Capture **a diverse range of fraudulent behaviors** across real-world and simulated claims.
- Leverage **machine learning techniques from An Introduction to Statistical Learning (ISL)** to enhance fraud classification.
- Address challenges such as **imbalanced datasets, feature selection, and interpretability** to build a scalable fraud detection system.

The integration of multiple datasets and the application of **robust statistical learning techniques** have been pivotal in advancing fraud detection research. Our project contributes to this growing body of work by developing an interpretable, high-accuracy fraud detection model that can be used by **healthcare providers, insurance companies, and regulatory bodies** to prevent fraudulent claims and reduce financial losses.

# V. Tentative Schedule & Task Distribution

The project will follow the schedule below:

To effectively manage this project, we have assigned tasks among the three team members: **Nhan, Tan, and Andre**. Each member is responsible for different phases of the report and future tasks, ensuring a structured approach to fraud detection research and implementation.

## 5.1 Task Assignments

### Nhan (Data & Preprocessing Lead)

- Collect, clean, and merge the datasets (**CMS, Kaggle, Synthea**).
- Conduct **exploratory data analysis (EDA)** to identify key fraud patterns.
- Handle **data balancing (SMOTE, undersampling)** to ensure a fair model.
- Generate **feature selection insights** based on statistical learning techniques.

### Tan (Model Development & Evaluation Lead)

- Implement **baseline models (Logistic Regression, Decision Trees, Random Forests).**
- Apply **advanced models (SVM, Boosting, Clustering) for fraud detection.**
- Tune hyperparameters using **cross-validation and performance metrics.**
- Evaluate model performance using **AUC-ROC, Precision-Recall Curves, and SHAP Analysis.**

### Andre (Documentation & Research Lead)

- Write and structure the report, ensuring clarity and academic integrity.
- Research and document prior works on healthcare fraud detection.
- Handle citations and references using proper academic formats.
- Assist in interpreting model results and preparing visuals.

## ⌄ 5.2 Timeline for the Project

(This is the current itme line, but then can still be adjust)

| Week | Tasks | Responsible Member(s) |
|------|-------|----------------------|
| **Week 1-2** | Data collection, preprocessing, and exploratory analysis | **Nhan** |
| **Week 3-4** | Feature selection, data balancing (SMOTE, undersampling) | **Nhan** |
| **Week 5-6** | Initial model training: Logistic Regression, Decision Trees | **Tan** |
| **Week 7** | Advanced modeling: SVM, Boosting, Clustering | **Tan** |
| **Week 8** | Model tuning and hyperparameter optimization | **Tan** |
| **Week 9** | Evaluation: Precision-Recall, ROC Curves, SHAP Analysis | **Tan, Nhan** |
| **Week 10** | Documentation & Literature Review | **Andre** |
| **Week 11-12** | Finalizing report, citations, and presentation preparation | **Andre, Nhan, Tan** |

# VI. Conclusion

This study aims to **develop a robust fraud detection model** by leveraging **statistical learning and machine learning techniques** on **CMS Medicare data, Kaggle Healthcare Fraud datasets, and Synthea synthetic data**. Through rigorous **data preprocessing, model development, and performance evaluation**, we strive to create a scalable and interpretable fraud detection framework that can be utilized by **healthcare providers, insurance companies, and regulatory bodies**.

By integrating multiple datasets, using **supervised and unsupervised learning methods**, and incorporating **real-world fraud cases**, this project will contribute to ongoing efforts in **preventing fraudulent claims and reducing financial losses** in the healthcare industry. Future work may explore **deep learning-based fraud detection techniques** and **real-time fraud monitoring systems**.

## References

1. **Big Data Fraud Detection Using Multiple Medicare Data Sources** – Springer Journal of Big Data. Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0138-3
2. **Medicare Fraud Detection Through Big Data and Machine Learning Models** – GitHub CMS Fraud Detection. Available at: https://github.com/Pyligent/CMS-Medicare-Data-FRAUD-Detection
3. **Medical Provider Fraud Detection Using Machine Learning** – Kaggle Project. Available at: https://www.kaggle.com/code/rohitrox/medical-provider-fraud-detection
4. **Healthcare Provider Fraud Detection and Analysis Using Machine Learning** – Medium Research Post. Available at: https://rohansoni-jssaten2019.medium.com/healthcare-provider-fraud-detection-and-analysis-machine-learning-6af6366caff2
5. **Optimizing Fraud Detection Models with Synthetic Data: Advancements and Challenges** – ResearchGate. Available at: https://www.researchgate.net/publication/383232341_Optimizing_Fraud_Detection_Models_with_Synthetic_Data_Advancements_and_Challenges
6. **Enhancing AI Fraud Predictive Models with Synthetic Data: A Case Study** – Dedomena AI Blog. Available at: https://dedomena.ai/blog/enhancing_ai_fraud_predictive_models_with_synthetic_data_a_case_study