# Fraud Detection with Medicare Provider Data

**Executive Summary**

*Team Bloom*: Austin Knies and Jonathan Leslie

Department of Economics, Indiana University, Bloomington, IN, USA

December 9, 2022

Health care fraud is a massive problem, contributing to billions of dollars in losses for insurers, health care systems, and taxpayers each year. Reducing and preventing health care fraud would lead to significant savings, but fraud can be difficult to identify through claims and other medical data. For this project, we perform anomaly detection to identify and classify fraud using publicly available Medicare data on health care providers and services. We contribute two valuable machine learning processes: first, we conduct an unsupervised learning routine on a large, unlabeled dataset of medical providers to identify fraud, and then we implement a supervised learning model that classifies whether a new provider's data appears fraudulent.

The dataset we use for this project comes from the 2013-2020 Medicare Physician & Other Practitioners Public Use Files (PUFs) through the Centers for Medicare and Medicaid Services (CMS). This data includes information on health care utilization, charges, and reimbursements by provider and service, and it is derived from administrative claims data for Medicare beneficiaries. We subset the data to a provider-year-service code level of just over 80 million observations and select, for each combination, service shares, the average submitted charge by the provider, and the average payment paid by Medicare to identify fraud through unnecessary utilization and overcharging.

With this data, our goal is to create a binary classifier that can predict whether or not a given provider is potentially fraudulent. This task presents two main issues. First, the original data does not include fraud labels. Second, fraudulent providers are expected to represent a tiny fraction of total providers thereby making it challenging for a classifier model to predict fraud since there will be very few instances of fraud to learn from. To overcome these issues, our machine learning process consists of three parts. We first develop an Autoencoder Neural Network to designate what providers in our dataset are potentially fraudulent. We then construct a Generative Adversarial Network (GAN) to create a set of fake fraudulent observations to help the binary classifier recognize fraud. Finally, we train a Convolutional Neural Network on the resulting data augmented with the fake observations in order to predict whether a new provider is potentially fraudulent or not. The resulting algorithm is capable of predicting fraud given as-yet-unseen provider data even though fraud is assumed to be extremely rare. The inclusion of the fake fraud data produced by the GAN improves fraud prediction allowing the algorithm to identify the majority of potentially fraudulent providers. With future improvements to prediction accuracy via hyperparameter optimization, our flexible methodology can be used to help catch and prevent medical fraud in order to reduce its costs on society as well as analyze trends in fraudulent behavior.