Document Title:	Using Physician Behavioral Big Data for High Precision Fraud Prediction and Detection
Author(s):	Sally S. Simpson, Ph.D., Ritu Agarwal, Ph.D., Guodong (Gordon) Gao, Ph.D.
Document Number:	306161
Date Received:	March 2023
Award Number:	2019-R2-CX-0016

Summary Report

Using Physician Behavioral Big Data for High Precision Fraud Prediction and Detection (# 2019 R2 CX 0016)

Dr. Sally S. Simpson (PI)
Department of Criminology and Criminal Justice
University of Maryland

Dr. Ritu Agarwal (Co-PI)[1]
Distinguished University Professor
Department of Decisions, Operations, and Information Technology
Robert H. Smith School of Business
University of Maryland
and

Dr. Guodong (Gordon) Gao (Co-PI)[2]
Department of Decisions, Operations, and Information Technology
Robert H. Smith School of Business
University of Maryland

[1] Current affiliation, Wm. Polk Carey Distinguished Professor at Johns Hopkins Carey Business School.

[2] Current affiliation, Professor at the Carey Business School of Johns Hopkins University.

Using Physician Behavioral Big Data for High Precision Fraud Prediction and Detection (# 2019 R2 CX 0016)

1. Introduction

In the United States, experts estimate that between 3 and 10% of all healthcare spending is fraudulent (Berwick & Hackbarth, 2012; Federal Bureau of Investigation, 2019; Institute of Medicine, 2013). Considering that 4.1 trillion dollars are spent on healthcare yearly (Rama, 2020), the costs of fraud to society are immense. Among those acutely impacted are state and federal governments, insurers and legitimate health care providers, US taxpayers, and patients--who pay more for health care than they should while often experiencing additional adverse consequences including unnecessary and potentially harmful care.

Although the costs of fraud are well-understood, the way in which fraud typically is detected, via a "pay and chase" model, is far from ideal. The "pay and chase" approach operates *ex post facto*. The government and other major payers pay nearly every expense that is billed to them by medical service providers and suppliers. If a physician or a particular encounter is later flagged for potential fraud, that money has to be recovered after funds have already been disbursed. This process is much more labor-intensive and time-consuming than preventing improper payments in the first place (Government Accountability Office (GAO), 2017) and is associated with high false positives and false negatives (2015 MA-6-Massachusetts_MassHealth, Preventing Healthcare Fraud through Predictive Modeling).

In this project, we aimed to provide new policy-relevant applications by using state-of-the-art data science to improve risk assessment of physician engagement in fraud. Specifically, we sought answers to the following research questions:

- Can models using big data on non-clinical physician behavior (e.g., illegal behavior, consumer complaints and malpractice, other disciplinary action, conspicuous spending, and life stressors) successfully predict engagement in fraud in the near-term future (1-5 years)?
- Of these behavioral factors, which ones represent the greatest risk for fraud engagement?
- Which machine learning algorithm is most accurate in predicting a physician's risk of engaging in fraud?

To answer these questions, we rely on techniques that use behavioral big data and deploy state-of-the-art data analytic tools to detect the risk of Medicare fraud early on, before payment takes place, and potentially early enough to prevent it from occurring in the first place. Our approach utilizes a matched set of physicians—one set (472) excluded from Medicare participation between 2015 through 2019 due to fraudulent activity and a control group of matched non-fraudulent physicians. We collected extensive publicly available information on both sets of physicians and then, leveraging the non-clinical physician behavioral data into models of fraud risk assessment, we developed a machine learning-based algorithm to predict a physician's risk of engaging in fraud.

In our analysis, we investigated two kinds of models: (1) Predictive and (2) Explanatory. Our main predictive modeling findings reveal a high degree of performance accuracy up to five years prior to the exclusion year, as defined in the LEIE list (List of Excluded Individuals/Entities) from the Office of Inspector General of HHS. Prediction accuracy rates are 89.25% (one-year ahead of the exclusion year), 86.02% (two-year ahead), 82.80% (three-year ahead), 79.57% (four-year ahead), and 77.42% (five-year ahead). In the explanatory models (where we utilize survival modeling and use the exclusion event as the failure point), we found that factors associated with high risks of fraud included a prior criminal case; tax liens, property purchases, gifts from companies, 1-star online reviews. Factors associated with a low fraud risk include 5-star physician reviews and holding a DEA license. Both sets of results point to the utility of using nonclinical data in fraud prevention and control efforts.

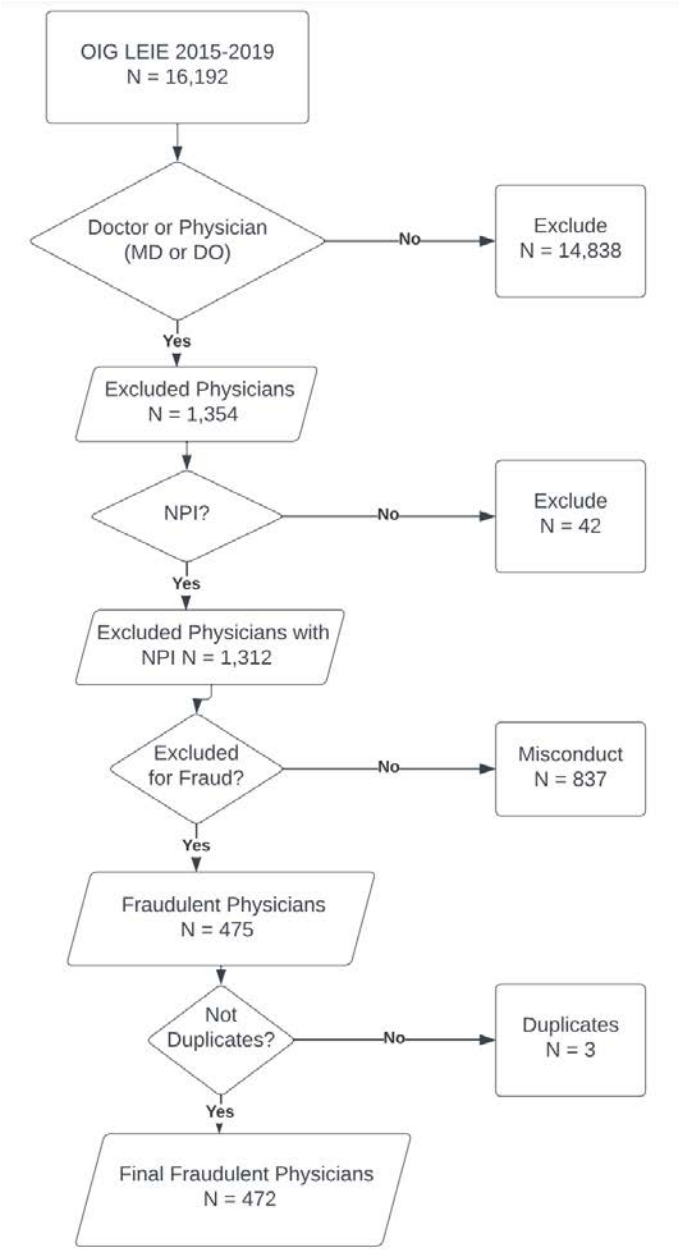
In the next section, we provide more detail about the project methods including the physician selection process, the general kinds/types of variables collected and their sources, along with the models utilized.

2. Methods: Physician Sample, Variables, Models

2.1 Physician Sample

We created a sample of **fraud physicians** using Medicare *yearly* exclusion criteria (2015 through 2019) *by practice* (internal medicine, family medicine, or general practice). The Office of the Inspector General (OIG) has the authority to exclude individuals and entities from Federally funded health care programs for a variety of reasons, including a conviction for Medicare or Medicaid fraud. A list of physicians who, in any given year, are excluded (List of Excluded Individuals/Entities or LEIE) is kept by the OIG. We relied on this exclusion list to identify our sample of 472 fraudulent physicians (see Figure 1).

Figure 1: Physician Fraud Flowchart



To identify a control group of **matched physicians**, we initially utilized five characteristics based on information available from the National Provider Identifier (NPI) registry: gender, primary practice zip code (the same or similar determined by the Area Deprivation Index or ADI), primary taxonomy, whether the physician was the singular owner of the practice (sole proprietorship), and degree credentials (e.g., MD, DO), later adding age (+ or – 5 years). To get a complete control sample, we loosened the matching criteria somewhat, reclassifying some taxonomies (the NPI Registry term for medical specialties) to be slightly broader (e.g., 'Internal Medicine' instead of 'Internal Medicine Cardiovascular Disease') and making additional adjustments. For those remaining unmatched fraud physicians, we loosened the matching criteria with the following sequence:

- Relax the age difference to +-7 years
- Relax the zip code to similar zip codes
- Relax credential/general taxonomy

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

- Relax the age difference by choosing the closest age match
- Relax sole proprietorship
- Match on gender and similar zip codes, and randomly select physician

These matching procedures resulted in a matched control sample of 472 physicians that was mostly balanced. In Table 1, we show the distribution of physicians and selected profiles in our sample by each exclusion year.

Table 1: Physician EXCLUDED and Physician MATCH each year.

			Differences between Fraud and Non-Fraudulent Physicians					
Exclusion Year	Fraudulent	Matched Non-Fraudulent	Age at Exclusion	Age at Graduation	Female Rate	MD Credential Rate	Sole Proprietorship	ADI National Rank Mean of Physician's Zipcode
2015	92	92	57.43 vs 57.48	27.16 vs 26.74	9.78% vs 8.70%	91.30% vs 90.22%	48.91% vs 47.82%	45.73 vs 46.07
2016	98	98	61.42 vs 60.37	27.47 vs 27.26	9.18% vs 9.18%	89.80% vs 87.76%	46.94% vs 48.98%	46.50 vs 46.15
2017	105	105	59.13 vs 57.92	27.16 vs 27.50	15.24 vs 15.24%	89.52% vs 86.67%	45.71% vs 45.71%	45.94 vs 46.50
2018	84	84	60.08 vs 60.62	27.17 vs 27.04	8.33% vs 8.33%	89.29% vs 91.67%	42.86% vs 42.86%	45.29 vs 45.33
2019	93	93	61.90 vs 61.16	27.14 vs 27.57	8.6% vs 8.6%	91.40% vs 92.47%	49.46% vs 49.46%	48.51 vs 48.57
Total	472	472	59.99 vs 59.47	27.22 vs 27.23	10.38% vs 10.17%	90.25% vs 89.62	46.82% vs 47.03%	46.33 vs 46.52

2.2 Variables

The project team paired fraudulent physicians with non-fraudulent physicians based on matching criteria described in the last section. To build a predictive model, we acquired data on static and dynamic variables about these two groups of physicians to contribute to modeling accuracy. A variable that does not change over time or only has a single record with no timestamp is defined as a static variable and collected in 2020. In comparison, a variable with multiple records with a timestamp is defined as a dynamic variable collected from 2000 to 2019.

The static variables include information about the physician's demographics, practices, geolocations, and background checks (single record). Additionally, dynamic variables include information about the physician's online reviews, political donations, transaction records, criminal records, services to Medicare beneficiaries, and background checks (multiple records). Table. 2 shows the value of static and dynamic variables.

These variables were collected from a variety of different sources. Two of the most important data sources include (1) the Physician Public Use File from the Centers for Medicare and Medicaid Services (CMS), which contains information on services and procedures physicians provide to Medicare beneficiaries and (2) Public Access to Court Records (PACER), which includes criminal, civil, and bankruptcy information of physicians collected from their federal court records. In addition, we also use the data from official organizations such as the Federal Election Commission, Office of Inspector General. External public data sources included Area Deprivation Index, Healthgrades, Vitals, and RateMDs. We also purchased background data from Find Out the Truth (FOTT) and physician-company transaction data from Dollars for Docs.

Table 2: Variables Sources, Types, Samples, and Statistical Description

Variable Type	Variable Group	Variable Items	Selected Variables	Whole Sample (Average)	Fraudulent Physicians (Average)	Matched Non-Fraudulent Physicians (Average)
Static: Demographics or Records without timestamp collected in 2020	Demographics	Date of Birth, Graduation Year, Gender,	Age at Exclusion	59.72 (944)	59.99 (472)	59.46 (472)
			Age at Graduation	27.23(944)	27.22 (472)	27.23 (472)
			Gender (female =1, male = 0)	10.28% (944)	10.38% (472)	10.17% (472)
	Practice	Credential, Specialty, Sole	MD Credential Rate	89.93% (944)	90.25% (472)	89.62% (472)
			Sole Proprietorship Rate	46.93% (944)	46.82% (472)	47.03% (472)
	Geolocation Information	Practice Address, Zipcode, State, ADI	ADI National Rank Mean	46.43 (944)	46.33 (472)	46.52 (472)
			ADI Population	1,464,012 (944)	1,492,225 (472)	1,435,800 (472)
	FOTT Background (Static)	DEA License	DEA License Average Number	0.51 (944)	0.45 (472)	0.57 (472)
		Property Records	Number of Property Records	3.41 (944)	3.68 (472)	3.14 (472)
		Corporate History	Number of Corporates	3.93 (944)	6.15 (472)	1.71 (472)

		Tax Lien	Tax Lien Amount	441,970.4 (944)	855,014.7 (472)	28,926.2 (472)
Dynamic: Records with timestamps between 2000 to 2020	Online Reviews	From Online Physician Platforms	Number of Total Reviews from 2000 to 2019	6.53 (944)	5.03 (472)	8.02 (472)
			Number of 5-Star Total Reviews from 2000 to 2019	4.13 (944)	2.86 (472)	5.39 (472)
	Political Donation	From the Political Donation Website	Number of Political Donations from 2000 to 2019	1211.38 (944)	1703.22 (472)	719.53 (472)
	Dollar for Docs	Payment to Physician from Industry	Total Amount from 2000 to 2019 (in dollars)	1877.20 (944)	789.41 (472)	2964.98 (472)
			Number of Distinct Companies from 2000 to 2019	6.45 (944)	2.49 (472)	10.42 (472)
			Number of Transactions from 2000 to 2019	38.95 (944)	12.42 (472)	65.49 (472)
	PACER	Criminal, Civil, Bankruptcy	Number of Criminal cases in Federal Courts from 2013 to 2019*	0.39 (944)	0.78 (472)	0.01 (472)
			Number of Civil cases in Federal Courts from 2013 to 2019*	0.19 (944)	0.34 (472)	0.03 (472)
			Number of Bankruptcy cases in Federal Courts 2013 to 2019*	0.06 (944)	0.10 (472)	0.03 (472)
	FOTT Background (Dynamic)	Property, Marriage, Criminal and Traffic Records	Number of Total Properties from 2000 to 2019	0.72 (944)	1.28 (472)	0.16 (472)
			Number of Marriages from 2000 to 2019	0.014 (944)	0.008 (472)	0.019 (472)
	PUF	PUF Item List from CMS	Average Age of Patients	51.10 (762)	37.07 (370)	64.34 (392)
			Proportion of Patients with Diabetes	25.69 % (764)	18.46% (370)	32.51% (392)
			Proportion of Patients with Depression	23.81% (764)	19.81 % (370)	27.60% (392)

Note: Values in parentheses reflect the sample size. * PACER search collected information about cases prior to 2013 only when an ongoing case was referenced in the physician record between 2013 and 2019, but it had not concluded before 2013.

2.3 Models

2.3.1 Machine Learning for Prediction

Aim: we use the features collected between the graduation year and the prediction year to predict whether the physician will be excluded in the result year.

Q-Year Look Ahead Model

We first group physicians by the year of exclusion. For example, in the table, C2018 refers to the set of fraudulent physicians who were excluded by OIG in 2018 (C2018-Pos) and the corresponding matched-non-fraudulent physicians (C2018-Neg).

Our data include cohorts from 2011 to 2019; we use cohorts from 2011 to 2018 as the training dataset and cohort 2019 as the test dataset.

Our goal was to train a Q=1 Year Look Ahead model based on cohorts from 2015 to 2018.

- For example, for cohort 2018, we set the year we make the prediction (PREDYear) as 2017, and the resulting event year (RSLTYear) as 2018. In other words, in 2017, we predict whether the physician will be excluded due to fraud in 2018.
- Our machine learning (ML) model uses features before 2017 to learn whether the physician will be excluded in 2018.
- Similarly, for every other cohort Y, the SAME model utilizes features before PREDYear = Y - Q (where Q = 1) to learn whether the physician will be excluded in RSLTYear = Y.
- In other words, for each RSLTYear, we only consider the cohort of the corresponding year (Cohort RSLTYear) as the physician list.
- The selected physician lists for each PREYear-RESLYear pair is shown in Table. 3.

Table 3: Q-Year Look Ahead Prediction Model and Train and Test Settings

		Training Cohort for Each Year and Cohort Size				**Testing**
RSLTYear	**PREDYear**	**C2015 (184)**	**C2016 (196)**	**C2017 (210)**	**C2018 (168)**	**C2019 (186)**
2015	2015 - Q	Yes	-	-	-	
2016	2016 - Q	-	Yes	-	-	
2017	2017 - Q	-	-	Yes	-	
2018	2018 - Q	-	-	-	Yes	
2019	2019 - Q					**Yes**

After the ML model is trained, we apply it to the new and hitherto unused physician list of cohort 2019. Here the PREDYear is 2018 and the RSLTYear is 2019. For these physicians, our model uses the features before 2018 to predict whether they will be excluded from OIG in 2019. Then we verify and compute the accuracy by comparing the model's predictions with the ground truth (i.e., physicians who actually committed fraud).

Following the same logic, we also develop models for Q Years Look Ahead with Q = 2, 3, 4, 5. The training and testing approach is similar to Q = 1. For the Q-year look ahead model, given the RSLTYear, the corresponding PREDYear will be RSLTYear - Q.

Machine Learning Algorithms

We use three different machine learning algorithms: Logistic Regression, Random Forest, and XGBoost to determine their comparative performance.

- Logistic Regression (Tolles and Meurer 2016): We can learn the probability of a fraud occurring (a categorical event) by determining the probabilistic value which lies between 0 and 1 and then calculating the log-odds (the probability of success divided by the probability of failure).
- Random Forest (Breiman 2001): Random forest algorithm is an ensemble learning method for classification as a combination of tree predictors. In the random forest, each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.
- XGBoost: XGBoost is an optimized distributed gradient boosting library designed under the Gradient Boosting framework. It is widely used by data scientists to achieve state-of-the-art results on many machine learning challenges.

Which Machine Learning Algorithm Performs the Best?

Table 4: Algorithms in Different Q-Year Look Ahead Models

Q-Year Look Ahead Model	1-Year	2-Year	3-Year	4-Year	5-Year
Logistic Regression	63.44%	66.67%	68.82%	68.28%	59.68%
Random Forest	88.17%	84.41%	83.87%	74.73%	75.81%
XGBoost	**89.25%**	**86.02%**	**82.80%**	**79.57%**	**77.42%**

When we use all behavioral big data and apply different ML algorithms, XGBoost performs the best. Moreover, XGBoost Models predict fraud with high accuracy even if the look-ahead gap is large (see Table. 4).

Are Behavioral Big Data important for Prediction?

In this section, we examine whether behavioral big data have important predictive power. In the experiment, we use the XGBoost algorithm to train Q-Year Look Ahead models under different variable settings.

Table 5: Impacts of Behavioral Big Data on XGBoost Model Prediction Performances

Q-Year Look Ahead Model	1-Year	2-Year	3-Year	4-Year	5-Year
Total Features	**89.25%**	**86.02%**	**82.80%**	**79.57%**	**77.42%**

Drop PUF	86.56%	84.41%	78.49%	76.34%	73.66%
Drop FOTT Background	86.56%	77.42%	69.35%	68.28%	63.44%
Drop FOTT & PUF & Review	84.41%	82.80%	76.34%	78.49%	76.88%
Drop FOTT & PACER	69.89%	67.74%	64.52%	62.37%	61.29%

As shown in Table 5, after gradually dropping PUF, FOTT, PACER, and Online Review Data, we find that the predictive performances dropped significantly. When compared between “Total Features” and “Drop FOTT & PACER”, the performance of the 1-Year ahead model has a 19.36% decrease (from 89.25% to 69.89%). These big behavioral data are even more important for long-term prediction. The Q-Year Look Ahead models’ performances are closer to the “random guess” when Q is larger. This experiment highlights the importance of introducing behavioral big data into the fraud-prediction model.

Overall, our application of advanced data analytic methods revealed that XGBoost outperformed other machine learning methods, and that behavioral big data increased the algorithm’s prediction power.

2.3.2 Survival Analysis for Explanatory Model

In the time elapsed between the GRDYear (physician medical school graduation year) to the EXCLYear (physician exclusion from Medicare participation year), physicians can experience many events. How do these events increase the hazard of engaging in future fraud? We set the exclusion event as the failure points and used survival analysis to model time to failure (exclusion). As reported in Table 5, we use the time-dependent Cox proportional hazards model (Fisher and Lin 1999) to identify factors that are positively related to fraud risk and those that lower it. We select this survival model because many of our factors change over time, such as online reviews, gifts received from companies, criminal records, and so forth.

Table 5: Risk Factors for the Exclusion Event

Variables	**Descriptions**	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
Accumulated Purchased **Properties**	(each year; from FOTT)	**0.0147**	1.0150	0.0037	3.9680	7.24e-05 ***
Accumulated **Gifts** from Companies	(k dollar; each year; from DollarDoc)	**0.0225**	1.0230	0.0113	1.9880	0.04677 *
Accumulated **Criminal** Cases	(each year; collected from PACER)	**0.3516**	1.4210	0.0191	18.4080	< 2e-16 ***
Accumulated **1-Star Reviews**	(each year; from online physician review platforms)	**0.0106**	1.0110	0.0043	2.5000	0.01242 *
Accumulated **2-Star Reviews** (each year)	(each year; from online physician review platforms)	-0.0119	0.9882	0.0268	-0.4440	0.65731

Accumulated **3-Star Reviews** (each year)	(each year; from online physician review platforms)	-0.0198	0.9804	0.0187	-1.0600	0.28911
Accumulated **4-Star Reviews** (each year)	(each year; from online physician review platforms)	-0.0096	0.9904	0.0116	-0.8320	0.40557
Accumulated **5-Star Reviews** (each year)	(each year; from online physician review platforms)	**-0.0037**	0.9963	0.0018	-2.0680	0.03865 *
Accumulated **Political Donation**	(each year; from Federal Election Commission)	0.0000	1.0000	0.0000	1.0990	0.27165
Have **Tax Lien**	(from FOTT)	**0.0129**	1.0130	0.0054	2.4160	0.01568 *
Have **DEA** License	(from FOTT)	**-0.1687**	0.8448	0.0647	-2.6090	0.00909 **
Material Consumption Style	(buying at least one vessel, aircraft, or 70k+ property; from FOTT)	-0.2139	0.8074	0.1120	-1.9100	0.05612 .
Other control variables: Age at Exclusion Year, Gender, Specialty, Sole Proprietorship, ADI, and other FOTT information. Please see our code for more information.						

Note: *** p<0.001, ** p<0.01, * p<0.05, ‘<0,10
Table. 5 shows the following findings:

Law-Breaker Experiences
Accumulated Criminal Case: An additional criminal case is associated with higher fraud risk hazards. A physician with an additional criminal case could be more likely to engage in fraud activities.

Gifts from Companies
Accumulated Gifts Case: Increased gifts from companies are associated with higher fraud risk hazards.

Online Review Platforms
Review Star 1: Receiving an additional 1-star review is associated with high fraud risk hazard. Review Star 5: Receiving an additional 5-star review associated with low fraud risk hazard. More 1-star reviews indicate the physician’s low-quality service, negative personality, or limited devoted efforts to patients. If a physician engages in fraudulent activities, s/he might provide low-quality care to patients where patients are extremely unhappy and hence give 1-star reviews. In contrast, if a physician doesn’t engage in fraudulent activities, s/he may provide high-quality healthcare to patients resulting in patients who are more likely to give 5-star reviews.

Practice and Monitor
Have **DEA** License: Having a Drug Enforcement Administration license is associated with lower fraud risk hazards. A DEA license might indicate additional supervision from the government which would lower the perceived opportunities for fraud and enhance deterrence.

Financial Burden
Tax Lien: Large Tax Lien cases are associated with high fraud risk hazards.

Accumulated Purchased Property: An additional purchased property is associated with the fraud risk hazard.
Both Tax Lien and additional purchased property indicate extra financial burden. This can act as the motivation for physicians to engage in fraudulent activities.

3. Discussion and Conclusion

Results from both our predictive and explanatory models yield promising insights for future Medicare fraud detection models. Regarding the first of our research questions, "Can models using big data on non-clinical physician behavior successfully predict engagement in fraud in the near-term future (1-5 years)", the answer is clearly yes. Our findings show that behavioral big data contains valuable information that, when included in prediction models, improves prediction accuracy. Regarding our second research question, "Of these behavioral factors, which represent the greatest risk for fraud engagement," several key variables are associated with an increased fraud risk: Property Purchases, Criminal case records, Tax Liens, Gifts from companies, and 1-star reviews in online platforms. In contrast, behavioral factors associated with a reduced fraud risk include having a DEA license and physician Five-Star Reviews. We also conclude, after comparing the performance of several machine learning algorithms, that the most accurate algorithm in predicting a physician's risk of engaging in fraud is XGBoost.

The project findings are suggestive that behavioral data can help identify physicians who are at a higher risk of engaging in fraud prior to the fraud's occurrence, thus triggering potential fraud prevention strategies that lower the risk-- hence improving on the *ex post facto* pay and chase fraud detection models. Predictive models are more efficient than traditional detection techniques. For example, if the current practice of fraud detection has a high false positive rate, then our model can help improve the precision rate, which can reduce the overall surveillance costs. More importantly, our model can predict fraud risk up to five years ahead of the exclusion case, this allows possible behavioral interventions to prevent the physician from committing fraud behavior down the road. This will reduce losses from paid illegitimate claims as well as the loss of a well-trained clinician. Wide scale adoption and utilization of predictive models may also deter physicians from engaging in fraud as they learn more about the predictive accuracy of algorithms, which reduces surveillance costs.

Limitations

Like most studies, ours has some weaknesses. Our methodology of matched cases allowed us to drill down on the specific mechanisms that increase or reduce the risk of fraud occurring. A larger sample using the same methodology would provide more robust findings. However, the matched design does not accurately represent the population rate of fraud and thus our study is apt to give inflated predictions.

Further, our study is hampered by inconsistent information available in administrative datasets. For instance, some criminal and case information included in PACER offers the alleged start date of the behavior under legal consideration. However, oftentimes the inception date is missing. It would be useful for this kind of information to be collected consistently—not just for our purposes but in order for court actors accurately to identify victims and to calculate the costs and consequences of the behavior.

In a similar vein, behavioral big data often have missing values and because the data are collected from a variety of different sources, calculating inferential statistics and recognizing biases can be challenging. "Studies can be of relatively little value if the large sample size is not representative of the population to which the results will be generalized or is missing a key information, especially on a nonrandom basis" (Kaplan, Chambers, and Glasgow, 2014: 342).

Finally, we would be negligent not to mention the potential ethical implications of physician surveillance BEFORE offending occurs. Surveillance has been described as an expression of control that raises concerns about privacy, data accuracy, potential discrimination, and nefarious data usage by information collectors including governments and companies (Auxier, Raine, Anderson, Perrin, Kumar, and Turner, 2019). Such issues are challenging where practical decisions about harm reduction are manifest. Are the possible harms associated with surveillance outweighed by the benefits of crime prevention, better detection and control? These are important issues that should be carefully weighed and considered moving forward.

References

Auxier, B., Raine, L. Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019, November 15). Americans and privacy: Concerned, confused and feeling lack of control over their personal information. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacyconcerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Breiman L. "Random Forests." Machine Learning 45, no. 1 (October 1, 2001): 5–32.

Berwick, D. M., & Hackbarth, A. D. (2012). Eliminating Waste in US Health Care. *JAMA*, *307*(14), 1513–1516. https://doi.org/10.1001/jama.2012.362

Chen T and Guestrin C. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. San Francisco California USA: ACM, 2016.

Federal Bureau of Investigation. (2019). Health Care Fraud. Retrieved March 22, 2019, from https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud

Fisher L and Lin DY. "Time-dependent covariates in the cox proportional-hazards regression model." *Annual Review of Public Health* 20, no. 1 (May 1999): 145–57.

Institute of Medicine. (2013). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. https://doi.org/10.17226/13444

Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. Clin Transl Sci. 2014 Aug;7(4):342-6. doi: 10.1111/cts.12178. Epub 2014 Jul 15. PMID: 25043853; PMCID: PMC5439816.

MassIT, Preventing Health Care Fraud Through Predictive Modeling. https://www.nascio.org/wp-content/uploads/2020/09/2015MA6-Massachusetts_MassHealth_Improving-State-Operations_6.1.2015.pdf

Rama, Apoorva. "Policy Research Perspectives: National Health Expenditures, 2020: Spending accelerates due to spike in federal government expenditures related to the COVID-19 pandemic." American Medical Association. 2020 https://www.ama-assn.org/system/files/prp-annual-spending-2020.pdf

Tolles J and Meurer WJ. "Logistic Regression: Relating Patient Characteristics to Outcomes." JAMA 316, no. 5 (August 2, 2016): 533–34. https://doi.org/10.1001/jama.2016.7653.

UKEssays. Ethical and Moral Issues in Surveillance Technology. November 2018. Retrieved from https://www.ukessays.com/essays/security/ethical-moral-issues-surveillance-7571.php?vref=1

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.