2018 IEEE International Conference on Information Reuse and Integration for Data Science

A Survey of Medicare Data Processing and Integration for Fraud Detection

Richard A. Bauder
College of Engineering & Computer Science
Florida Atlantic University
Boca Raton, Florida, USA
rbauder2014@fau.edu

Taghi M. Khoshgoftaar
College of Engineering & Computer Science
Florida Atlantic University
Boca Raton, Florida, USA
khoshgof@fau.edu

***Abstract*—**

Healthcare is an important aspect in everyday life, with quality and affordable care being essential for a population's well-being and life expectancy. Even so, associated costs for medical services continue to rise. One aspect contributing to increased costs in healthcare is waste and fraud. In particular, with the rapidly rising elderly population in the United States, programs like Medicare are subject to high losses due to fraud. Therefore, fraud detection approaches are critical in lessening these losses. Even so, many studies using Medicare data do not provide sufficient details regarding data processing and/or integration making it potentially more difficult to understand the experimental results and challenging to reproduce the experiments. In this paper, we present current research using Medicare data to detect fraud, focusing on data processing and/or integration, and assess any gaps in the provided data-related details. We then present discussions on important details to look for when processing and merging different Medicare datasets indicating opportunities for future work.

Keywords: Medicare, Data Processing, Data Integration, Fraud Detection

I. Introduction

One of the most important aspects of data mining and machine learning is access to quality data which include necessary details for processing and/or integrating datasets. Clean data is a critical component for the successful deployment of data mining and machine learning solutions in any industry [27]. One such industry affecting the majority of United States (U.S.) citizens is healthcare. The quantity of healthcare-related data continues to increase and comes in a variety of forms that include electronic health records and medical insurance claims [25]. With such large and disparate data sources, data mining and machine learning approaches become increasingly important and necessary to extract meaningful information from big data sources [17], [24]. With these approaches, an understanding of how the data is processed can be critical for data analysis and model configuration. For example, questions such as what data is excluded or left missing and how missing data should be handled can help in generating proper input datasets for machine learning algorithms. Additionally, integrating two or more data sources can increase the amount of useful information, beyond the individual datasets, to include items such as claims payments for services and drug prescriptions by the same physician. An area where having quality data, alone or integrated, can add value is for the detection of fraudulent or wasteful activities. Fraud and waste are an ongoing problem in healthcare, contributing to millions of dollars in losses, with current methods doing little to reduce fraud losses [1].

To provide context to the impact of healthcare on the U.S. economy, healthcare spending increased in 2015 by 5.8%, totaling over $3.2 trillion [2]. One aspect of waste in healthcare is fraud, which accounts for 3-10% ($19 billion to $65 billion) of all medical claims [21]. One of the largest U.S. healthcare programs is Medicare, a government program with over 54.3 million beneficiaries, which reimburses hospitals and physicians for medical care provided to people over the age of 65 and to younger individuals with specific medical conditions and disabilities [5]. In this work, we focus on Medicare data for two reasons. The first is the increasing population of elderly individuals, which rose 28% from 2004 to 2015 versus an increase of just 6.5% for those under 65 years of age [3]. The second reason is Medicare accounted for 20% ($646 billion) of U.S. healthcare spending in 2015 [2]. The interested reader can find additional information on Medicare and Medicare fraud in [7].

Given the importance of data in detecting Medicare fraud and abuse, we survey the use of Medicare data, outside of our research group, with regards to fraud detection. In particular, we focus on investigating what datasets were used and how they were processed and/or integrated. Any gaps in data handling and processing are discussed in order to detail not only the current state but, more importantly, to present opportunities for future research to help increase the effectiveness of Medicare fraud detection. To the best of our knowledge, there are no other studies that assess the current state of Medicare data processing and integration and, more importantly, provide discussions on improvement opportunities.

The rest of the paper is organized as follows. In Section II, we present background information on the Medicare-related data sources. Section III discusses Medicare-related fraud research, with a focus on the details regarding data processing and/or integration. Section IV summarizes the gaps in the current research and thoughts on Medicare data processing and/or integration. Finally, we conclude our survey and discuss future work in Section V.

978-1-5386-2659-7/18/$31.00 ©2018 IEEE
DOI 10.1109/IRI.2018.00010

IEEE computer society

II. Background

In this section, we provide the necessary background information on the Medicare datasets and the provider exclusion database as used in the current research. We focus on physician (or provider) claims data, excluding sources such as chronic condition reports which do not directly involve Medicare claims. The three datasets of interest, provided by the Centers for Medicare and Medicaid Services (CMS) [14], are the *Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B)*, *Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D)*, and *Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS)*. Additionally, we discuss the *List of Excluded Individuals/Entities (LEIE) database* [20] which contains information usable as fraud labels. Medicare Part B [11] is available for the 2012, 2013, 2014, and 2015 calendar years, with the latest year being released in mid-2017. This dataset summarizes services and procedures provided to Medicare beneficiaries by providers such as physicians and other healthcare professionals. Part D [10] contains information about prescription drugs administered by physicians under the Medicare Part D Prescription Drug Program. This dataset is available for 2013, 2014, and 2015. The DMEPOS dataset is also available for 2013, 2014, and 2015 and presents provider claims pertaining to durable medical equipment, prosthetics, orthotics and supplies to include corresponding services.

The information in all three Medicare datasets is recorded after claims payments were made [15] and is assumed to have been cleansed and reviewed by CMS, with any modification, such as imputation, noted where appropriate. With each dataset, the provider or physician is denoted by his or her unique National Provider Identifier (NPI) [13] for each claim. Each claim represents how many times per year a provider is billed for specific services, procedures, drug prescriptions, or equipment. Payments are made by a provider for a specific service rendered. This is an important detail especially when trying to integrate other data sources or map fraud labels. Other sources of information may not be at the procedure-level but rather at the provider- or NPI-level, thus having consistent levels are critical in making sure the linked data makes sense together. For example, the LEIE lists provider exclusions without any information regarding which procedures or prescriptions led to being placed on the exclusion list, thus would be considered an NPI-level data source. Lastly, to protect the privacy of Medicare beneficiaries, any records, which are derived from 10 or fewer claims, are excluded.

The Medicare datasets do not include labels indicating fraud; however, a list of physicians and other healthcare entities that are excluded from participation in federally funded healthcare programs, such as Medicare, for a certain period of time, can be obtained from the Office of Inspector General's LEIE database [20]. These exclusions are authorized in accordance with Sections 1128 and 1156 of the Social Security Act [22]. A physician could be on the exclusion list due to several different categories, such as 1128(a)(1) and 1128(b)(4), which could indicate writing medically unnecessary prescriptions and having his or her license suspended, respectively. It is important to note that even though providers are listed in the LEIE database, 38% with fraud convictions continue to practice medicine and 21% were not suspended from medical practice despite their convictions [23]. Table I lists the mandatory exclusions indicating the severity of the crime and punishment.

TABLE I: LEIE mandatory exclusions

Rule Number	Description
1128(a)(1)	Conviction of program-related crimes.
1128(a)(2)	Conviction for patient abuse or neglect.
1128(a)(3)	Felony conviction due to healthcare fraud.
1128(b)(4)	License revocation or suspension.
1128(c)(3)(g)(i)	Conviction of 2 mandatory offenses.
1128(c)(3)(g)(ii)	Conviction on 3+ mandatory offenses.

III. Medicare Data Processing and Integration

Because our primary interest is how Medicare data was processed and merged, we focus this section on works, excluding our current research, that use Medicare data sources (Part B, Part D, and DMEPOS) and/or the LEIE database to detect fraud or other aberrant provider activities. We detail how the data was handled, the merging of datasets, and discuss any gaps in the research regarding data processing. The gaps in these works make it difficult to recreate the input data and reproduce the experiments. Even so, these limitations present opportunities for future research.

A. Does Medical School Training Relate to Practice? Evidence from Big Data

In a paper by Feldman et al. [16], the authors analyze healthcare information to identify differences in the number of procedures performed, average charges, and average payments for physicians based on their respective medical schools. Their study does not include the application of machine learning methods, but rather employs primarily descriptive statistics. The authors provide some information regarding data processing using the following datasets from 2012: CMS Physician Compare and CMS Medicare Part B. Additionally, only U.S. medical schools, from the Physician Compare dataset, are included and information concerning school locations are filled in, as needed, for geographic analysis. They add or update zip codes for each medical school using sources such as newspaper articles and school announcements. If a school has no zip code or is no longer active, the authors use the zip code for the city of the defunct school. Information from the 2012 Association of American Medical Colleges Tuition and Student Fees Reports [6] is used for medical school tuition costs. Besides the inclusion of the additional medical school information, there is no other discussion on cleansing or processing of the Physician Compare or Part B data. The authors integrate and aggregate the data as follows:

- Link the Physician Compare and Part B datasets by matching unique NPI values.
- Group the data by medical school name and procedure code and aggregated over physicians.

Each instance in the final merged dataset shows the code, for the procedure performed, and medical school name with corresponding procedure cost and count information (*line_srvc_cnt*, *average_submitted_chrg_amt*, and *average_Medicare_payment_amt*), as well as school tuition and location. With this dataset, the authors could potentially detect fraudulent physicians or flag physicians early in their careers who are at-risk for future fraud. The study is limited by only using one year of data and is missing fraud validation, which could be done by including known fraudulent physicians from the LEIE database.

B. Physician Medicare fraud: characteristics and consequences

Pande et al. [23] employ descriptive statistics and data analysis to find patterns and make recommendations on Medicare-related fraud, like using predictive models for claims fraud detection. The authors aim to provide answers to who commits Medicare fraud and what happens after they get caught. Their primary data source is the LEIE dated October 6, 2011. From this data, the authors use excluded providers based on subsection 1128(a)(1) only, indicating a conviction for Medicare or Medicaid fraud. After selecting those individuals with a Medical Degree (MD), they ended up with 795 physicians for their study. They did not include any doctor of osteopathic medicine (DO) degrees which may have limited the potential number of physicians in their study. The authors do not provide details on exclusion time periods or whether waiver or reinstatement dates were taken into consideration. Any additional data sources that may have been used by the authors in their study are not discussed.

C. Graph Analytics for Healthcare Fraud Risk Estimation

In a study by Branting et al. [8], the authors present a method for pinpointing fraudulent behavior by determining the fraud risk through the application of graph algorithms and detecting fraudulent providers using these graph-based features and a decision tree learner. For their research, they use CMS Medicare Part B (2012 to 2014) and Part D (2013) data, as well as the LEIE database for fraud labels. The authors perform the following data processing steps:

- Link the three datasets by matching NPI values.
- Additional linking done using fuzzy-string matching on provider names and other identity-related criteria, such as requiring that a matching set of providers have the addresses in the same state.
- Generate a graph integrating the provider, prescription, and procedure data sources used to represent provider activities and behaviors, with nodes to include NPI and HCPCS and edges indicating behaviors, locations, etc.

Given these steps for processing and integrating the data, the authors state that only 10-15% of the providers in the LEIE without an NPI could be confidently matched. With regards to the mapping of LEIE labels, it is unclear as to which specific exclusion rules were used when matching LEIE excluded providers, if the exclusion period itself was included or excluded, and if any waiver or reinstatement dates were taken into account. Additionally, it is unclear how the authors integrated the Part B and Part D datasets with three years for Part B and only one year for Part D. This lack of data processing clarity makes it difficult to reproduce the authors' research.

D. Detection of Fraudulent Claims Using Hierarchical Cluster Analysis

Khurjekar et al. [18] propose a two-step unsupervised approach to detecting fraud using residuals from a multivariate regression model. They identify suspicious claims based on a residual threshold of $500 and apply clustering to these residuals to find fraud based on average cluster distances. They use 2012 Medicare Part B data only with the following features: *hcpcs_code*, *line_srvc_cnt*, *bene_day_srvc_cnt* and *avg_medicare_payment_amt* (response variable). The authors do not discuss data processing, so the assumption is that they simply used the 2012 dataset as is and subset the aforementioned features for analysis. Additionally, another limitation of their work is that only 285 Medicare claims were used in their analysis, though no explanation of this limitation is presented.

E. Variability in Medicare Utilization and Payment Among Urologists

In a work by Ko et al. [19], the authors focus on analyzing the variability among urologists using service utilization, such as the number of office visits, and payment to determine the estimated savings from a standardized service utilization. More specifically, they employ linear regression to model these relationships and look at predicted payment values versus actual Medicare payment amounts to compare urologists with their peers. The authors use the 2012 Medicare Part B data and filter for urologist provider types only. This led to a dataset which consisted of 8,792 urologists. Additionally, the number of patient visits as indicated by HCPCS codes for new patient visits (99201, 99202, 99203, 99204, and 99205) and for return visits (99211, 99212, 99213, 99214, and 99215) was totaled for each urologist. As with [18], the work by Ko et al. has little discussion on data processing and it is assumed that Medicare data was used as is.

F. Knowledge Discovery from Massive Healthcare Claims Data

Chandola et al. [9] present a general coverage paper employing different machine learning methods for fraud detection to include social network analysis, text mining, and temporal analysis. Moreover, the authors discuss typical treatment profiles based on procedures performed. These profiles indicate the normal activity of physicians which are used to compare against other providers to determine possible concerns or abuses in procedures. The authors use claims data for 48 million U.S. beneficiaries, but it is unclear as to whether this is Medicare, specifically Part D, or Medicaid data. There is little discussion on the initial data sources. Another dataset is composed of provider enrollment information which was obtained from several private organizations. In order to have fraud labels, they use a list of excluded providers from the Texas Office of Inspector General's exclusion database. In this exploratory study, the authors provide minimal discussion on the data and no information regarding data processing or integration. This lack of detail with regard to the data makes it difficult to generate reproducible results.

G. *Mining anomalies in Medicare big data using patient rule induction method*

In a study by Sadiq et al. [26], the authors detect anomalies in Medicare data using an unsupervised method known as Patient Rule Induction Method based bump hunting technique, which attempts to determine peak anomalies by spotting spaces of higher modes and masses within the dataset. The authors use 2014 CMS data which included Part B (Physician and Other Supplier Data), Part D (Prescriber Data), and DMEPOS. It is unclear which features are used from any of the datasets or whether these features are related. Additionally, there is no discussion on data processing or integration making it difficult to assess whether these datasets were used separately or merged. With these gaps in how the data was processed, it is difficult to reproduce the experiments in their study.

IV. Discussion

The gaps and weaknesses exhibited in the current research with regards to processing and/or integrating Medicare datasets are apparent but can be lessened by thorough and detailed data processing and integration discussions. One key to reproducible research is to present enough detail on what data was used and how it was processed and integrated. The addition of these details can help to increase the understanding of the research methodology and results. This, though, requires analyzing and understanding the data, as well as any provided documentation. There are many possible issues which include, but are not limited to, missing values, assumptions on the meaning of features, data-related flags and filters, and features to merge data from different sources. Most of these data processing topics have not been adequately discussed in the current research. In this section, we provide discussions relating to data processing and the integration of multiple datasets, along with the incorporation of LEIE fraud labels.

For privacy reasons, information in the Medicare data is left out if there are less than 10 beneficiaries. From our perspective, this does not impact any data analysis or usage but should still be noted as it is part of the data generation process by CMS. All three datasets (Part B, D, and DMEPOS) have this type of removed information. For the Part D and DMEPOS datasets, there are suppression flags that exclude values based on beneficiary and claims counts being less than 11. Instances with a suppression flag have values for the count and payment features that are purposefully made blank or missing. For example, a suppression flag of "*" is assigned if the *generic_claim_count* is between 1 and 10, forcing the claim count feature to be blank for that particular instance. These blanks then are missing values, but they are distinct from values being missing because of clerical errors or an individual not filling in a particular field, such as forgetting to enter a gender or middle name. The latter are truly missing and cannot be reasonably imputed. For instance, if a provider's gender is missing, would one impute a male or female? The first and last names could be used but these are often wrought with ambiguity. In the case of gender, the best approach is just to leave it as a missing value and let the machine learning algorithm handle it or encode gender to more clearly represent missing values. For the Medicare data with the suppression flags, values are made blank but still have underlying context, i.e. claims counts are known to be between 1 and 10. The Medicare methodology documentation [10], [12] discusses these suppression flags and gives suggestions for handling these missing values. The documentation notes that leaving these as blanks can often lead to underestimating the true values. Furthermore, a suggestion is made to impute a value of 5 for the suppressed claim and beneficiary entries. It is thus reasonable to use imputation for these suppressed values to produce more accurate modeling results.

Some of the Medicare features could be used for more than one purpose. For instance, the Medicare Part B dataset has a feature called *line_srvc_cnt* that can represent several different calculated values [4]. For the majority of claims, this feature indicates the number of services rendered or procedures performed. However, for Part B drug-related claims this feature represents the weight or volume of the drug. Furthermore, the counts can be bundled indicating a drug was used during a procedure. There is a flag that can filter out the Part B prescription claims information which is important for distinguishing between procedures and drugs in prediction of provider claims fraud. It can also be used for integration with other Medicare data, such as Part D which is focused on prescription claims, to avoid misleading or confounding information. There are other flags to consider that can impact analytics and machine learning. The DMEPOS dataset has an indicator flag for equipment rentals, and the Part B dataset has a flag for procedures performed in hospitals or physician's offices. The names for each of the provider types, or medical specialties, may not be the same or come from the same sources, but there is a flag indicating the use of Medicare-provided specialty names that are consistent across the Medicare datasets.

In addition to providing requisite details regarding data processing, discussions around data integration should also be included for completeness and reproducible experiments. The merging of different datasets can be based on several key features, or even fuzzy matching techniques between datasets, but regardless, the integration methods should be discussed in detail. The Part B, Part D, and DMEPOS datasets have several standardized features with common values, such as the provider's NPI, provider types, and HCPCS codes. The values used to join the data can greatly alter the final resulting dataset. Using other, inexact, methods to merge datasets, such as fuzzy string matching of names, can increase the number of usable instances but also introduces errors with incorrect matches. Data integration, if done smartly with a clear methodology, can provide additional "connective" information about a provider and the claims made across procedures, drugs, and medical equipment leading to more accurate fraud detection results.

Finally, the LEIE database can be leveraged as fraud labels for Medicare provider-related claims data. The mapping of fraud labels to other Medicare datasets needs to be clearly stated in order to understand any fraud detection results. The LEIE is updated monthly and includes provider exclusions going back to 1977, whereas the Medicare data goes back to 2012 or 2013 and ends in 2015. The years must align correctly between the information in the LEIE and the available years of the Medicare data. Moreover, the LEIE has daily exclusion dates, while the Medicare datasets are annual. This needs to be taken into account in order to successfully map fraud labels and avoid too many or too few fraud labels. For example, if a provider is on the LEIE starting January 1, 2008 with an exclusion period of 5 years, then this provider would no longer be considered excluded starting January 1, 2013. The fraud labels assigned to the Medicare data should only be for 2012, not for 2013 or any year thereafter. In addition to the exclusion start and end dates, any waivers or reinstatements should be taken into account which could reduce the overall exclusion period and number of fraud labels. Regardless of the mapping, there will be some discrepancies due to the differing time representations in each of the datasets.

V. Conclusion

Healthcare fraud continues to be a threat to our economy and general well-being. In particular, the elderly who use Medicare resources are increasingly susceptible to the effects of fraud with the increasing elderly population and Medicare costs that continue to rise. In order to provide effective fraud detection, the Medicare data used in research studies must be discussed in a way that provides enough detail on data processing and/or integration in order to better understand the results and to enable reproducible research. In this paper, we present current research on Medicare fraud analysis and detection with an emphasis on data processing and/or integration. Our contribution includes summarizing gaps and lessons learned in related studies, which can guide researchers in conducting future research in this area. The focus is on Medicare Part B, Part D, and DMEPOS claims data, as well as the LEIE database, since these describe medical procedures, drugs, and equipment used by providers. We found that the current research is lacking in discussion on both data processing and integration (where it applies), thus making it difficult to understand what datasets are used and how they are used to produce fraud detection results. Moreover, this lack of detail makes it very difficult to reproduce these experiments. We identify the gaps in these studies and provide some discussion on some important data processing issues, such as missing values and imputation and data filtering using provided flags. The LEIE mapping process is also discussed to include caveats like the matching of dates. Our study indicates that there are deficiencies in the current research. These gaps, however, provide ample opportunities for future work.

Acknowledgment

We would like to thank the reviewers in the Data Mining and Machine Learning Laboratory at Florida Atlantic University. Additionally, we acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this paper are the authors' and do not reflect the views of the NSF.

References

[1] "The facts about rising health care costs," 2015. [Online]. Available: http://www.aetna.com/health-reform-connection/aetnas-vision/facts-about-costs.html

[2] "National Health Expenditures 2015 Highlights," 2015. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/highlights.pdf

[3] "Profile of older Americans: 2015," 2015. [Online]. Available: http://www.aoa.acl.gov/Aging_Statistics/Profile/2015/

[4] "Centers for Medicare and Medicaid Services: Research, Statistics, Data, and Systems," 2018. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html

[5] "US Medicare Program," 2018. [Online]. Available: https://www.medicare.gov

[6] Association of American Medical Colleges. Tuition and student fees reports (2012). [Online]. Available: https://www.aamc.org/data/tuitionandstudentfees/

[7] R. A. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, no. 1, pp. 31–55, 2017.

[8] L. K. Branting, F. Reeder, J. Gold, and T. Champney, "Graph analytics for healthcare fraud risk estimation," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 845–851.

[9] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1312–1320.

[10] CMS. Medicare provider utilization and payment data: Part d prescriber. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html

[11] CMS. Medicare provider utilization and payment data: Physician and other supplier. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html

[12] CMS. Medicare provider utilization and payment data: Referring durable medical equipment, prosthetics, orthotics and supplies. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/DME.html

[13] CMS. National provider identifier standard (npi). [Online]. Available: https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/

[14] CMS. (2018) Center for medicare and medicaid services. [Online]. Available: https://www.cms.gov/

[15] CMS Office of Enterprise Data and Analytics. (2017) Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf

[16] K. Feldman and N. V. Chawla, "Does medical school training relate to practice? evidence from big data," *Big data*, vol. 3, no. 2, pp. 103–113, 2015.

[17] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *Journal of Big Data*, vol. 1, no. 1, p. 2, 2014.

[18] N. Khurjekar, C.-A. Chou, and M. T. Khasawneh, "Detection of fraudulent claims using hierarchical cluster analysis," in *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2015, p. 2388.

[19] J. S. Ko, H. Chalfin, B. J. Trock, Z. Feng, E. Humphreys, S.-W. Park, H. B. Carter, K. D. Frick, and M. Han, "Variability in medicare utilization and payment among urologists," *Urology*, vol. 85, no. 5, pp. 1045–1051, 2015.

[20] LEIE. (2017) Office of inspector general leie downloadable databases. [Online]. Available: https://oig.hhs.gov/exclusions/index.asp

[21] L. Morris, "Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy," 2009. [Online]. Available: https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.28.5.1351

[22] OIG. (2018) Office of inspector general leie exclusion authorities. [Online]. Available: https://oig.hhs.gov/exclusions/authorities.asp

[23] V. Pande and W. Maas, "Physician medicare fraud: characteristics and consequences," *International Journal of Pharmaceutical and Healthcare Marketing*, vol. 7, no. 1, pp. 8–33, 2013.

[24] A. N. Richter, T. M. Khoshgoftaar, S. Landset, and T. Hasanin, "A multi-dimensional comparison of toolkits for machine learning with big data,"

in *Information Reuse and Integration (IRI), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–8.

[25] G. Roesems-Kerremans, “Big data in healthcare,” *Journal of Healthcare Communications*, 2016.

[26] S. Sadiq, Y. Tao, Y. Yan, and M.-L. Shyu, “Mining anomalies in medicare big data using patient rule induction method,” in *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*. IEEE, 2017, pp. 185–192.

[27] D. Tunkelang. (2017) Machine Learning: 10 Facts Everyone Needs to Understand. [Online]. Available: https://www.huffingtonpost.com/entry/machine-learning-10-facts-everyone-needs-to-understand_us_59af169ee4b0d0c16bb5285a