Simulation and Detection of Healthcare Fraud in German Inpatient Claims Data

Bernhard Schrupp[1,2], Kai Klede[1][0000−0002−1284−541X], René Raab[1][0000−0003−2035−3332], and Björn Eskofier[1][0000−0002−0417−0336]

[1] Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
{kai.klede,rene.raab,bjoern.eskofier}@fau.de
[2] AOK Bayern - Die Gesundheitskasse, Munich, Germany
bernhard.schrupp@by.aok.de

Abstract. The German Federal Criminal Police Office (BKA) reported damages of 72.6 million euros due to billing fraud in the German healthcare system in 2022, an increase of 25 % from the previous year. However, existing literature on automated healthcare fraud detection focuses on US, Taiwanese, or private data, and detection approaches based on individual claims are virtually nonexistent. In this work, we develop machine learning methods that detect fraud in German hospital billing data.
The lack of publicly available and labeled datasets limits the development of such methods. Therefore, we simulated inpatient treatments based on publicly available statistics on main and secondary diagnoses, operations and demographic information. We injected different types of fraud that were identified from the literature. This is the first complete simulator for inpatient care data, enabling further research in inpatient care.
We trained and compared several Machine Learning models on the simulated dataset. Gradient Boosting and Random Forest achieved the best results with a weighted F1 measure of approximately 80 %. An in-depth analysis of the presented methods shows they excel at detecting compensation-related fraud, such as DRG upcoding. An impact analysis on private inpatient claims data of a big German health insurance company revealed that up to 12 % of all treatments were identified as potentially fraudulent.

Keywords: Healthcare · Inpatient Claims · Healthcare Fraud · Fraud Detection · Data Generation · Inpatient Claims Simulation.

1 Introduction

Public health insurance companies in Germany spend roughly one-third of their budget each year on hospital treatments [1]. However, only about 10 % of detected fraudulent billing claims are accounted for hospital treatments, with most cases uncovered due to tips from insiders or patients [2]. This suggests a significant number of potentially fraudulent cases within the inpatient care system.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_22

The international research on fraud detection in healthcare focuses on identifying fraudulently acting participants, while in Germany, proving fraud for each claim separately is necessary. Limited accessibility to German inpatient claims data due to legal constraints poses a challenge for research. Addressing these issues, this study aims to identify potentially fraudulent claims in German inpatient billing data and generate simulated inpatient claims data to facilitate Machine Learning.

The project code is available at https://github.com/mad-lab-fau/inpatient-claims-simulator.

2 Related Work

Fraud in inpatient claims can be performed in various ways. Jürges and Köberlein [3] noted a significant decrease in reported newborn weights after the introduction of Germany's inpatient treatment billing system. Hospitals may manipulate weight following thresholds defined in the Diagnosis-Related Group (DRG) catalog to receive higher compensation [4]. Similar manipulation occurs with ventilation hours, where higher numbers lead to increased financial compensation [5].

Changes in the order of diagnoses are a more elaborate way to manipulate the billing of services. Shifting a profitable secondary condition (secondary = not the cause of the current hospitalization) to a primary diagnosis (primary = the cause of the current hospitalization) can be a rewarding fraudulent practice [6], often requiring access to medical documents not readily available to paying organizations.

Discharging a patient a few days earlier than medically necessary is another form of fraud, as these additional days would not increase the financial compensation, apart from the additional reimbursement for care [7].

Performing healthcare without medical indication is also considered fraudulent. A vivid example of this pattern of unnecessarily incurred hospital costs is the increase in cesarean deliveries compared with vaginal deliveries. Performing a cesarean section can be highly advantageous for hospitals as it almost doubles revenue while requiring fewer resources than complicated births [7]. A second example, also performed on newborns, is identifying the necessity of parental care throughout a hospital stay, meaning the child cannot stay there without one parent being nearby [5].

These fraud patterns will be injected in the simulated dataset and used to train fraud detection models (see Table 1).

3 Data Generation

As public inpatient claims billing data on an individual claims level is unavailable, a data simulation approach based on publicly available information is presented to facilitate Machine Learning development. The model operates under

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_22

Table 1. After simulating regular inpatient claims, we inject the following fraud patterns derived from the literature for 3.07 % of all records.

Fraud Pattern	Source	Share in Dataset
Changing order of ICD codes	[6]	1.73 %
Reducing duration of stay	[6]	0.85 %
Adding need for personal care when treating newborns	[5]	0.34 %
Changing a vaginal birth to a cesarean section	[7]	0.05 %
Increasing number of ventilation hours to reach next threshold	[5]	0.06 %
Decreasing reported weight of newborns to reach next threshold	[3]	0.05 %

assumptions such as treatments solely occurring in the main department, uninterrupted treatments, and patients not being transferred between hospitals or departments within the same hospital. Admissions are based solely on referral, and discharges are always medically justified.

3.1 Inpatient Claims Modeling

The simulation process was initialized with patients and hospitals, defined by unique IDs and locations (zip code). The zip code was sampled by following the German population density and hospital addresses [11]. Patient attributes include age and gender, randomly sampled to reflect observed distributions in primary ICD statistics. Hospitals and patients are matched based on distance, and each treatment is defined by primary and secondary ICD codes (diagnosis codes; version ICD-10-GM-2021) and OPS codes.

The treatment defining parameters primary ICD, a list of up to 20 secondary ICD codes [8, 9], and up to 20 OPS codes (treatments carried out, including imaging, surgeries and medication) [10] were generated by sampling according to publicly available statistics and demographics. When simulating treatment codes, constraints in form of relationships of OPS codes with diagnoses were formed, excluding combinations causing errors in a DRG Grouper [12].

Next, length of stay was determined using thresholds from relevant DRGs and a Gaussian distribution. The duration was added to a randomly selected admission date (between 01.01.2021 - 31.12.2021) to calculate the discharge date.

To complete the simulation, for cases with OPS codes referencing to ventilation we sampled the value according to a power-law distribution. If the patient is a newborn (age $<$ 1 year), it is necessary to determine the weight, which was sampled according to WHO information [13, 14] using a Gaussian distribution.

Fraudulent behavior was injected by randomly selecting 20 % of all claims for adjustments. Where possible, the values for ventilation hours and weight at birth were altered accordingly. If a birth is coded in the billing information, it was changed to a cesarean section. In case a newborn is hospitalized, the need for assistance with personal care was added to the list of secondary diagnoses. To achieve the pattern of exchanged primary ICDs, the combination of diagnoses

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_22

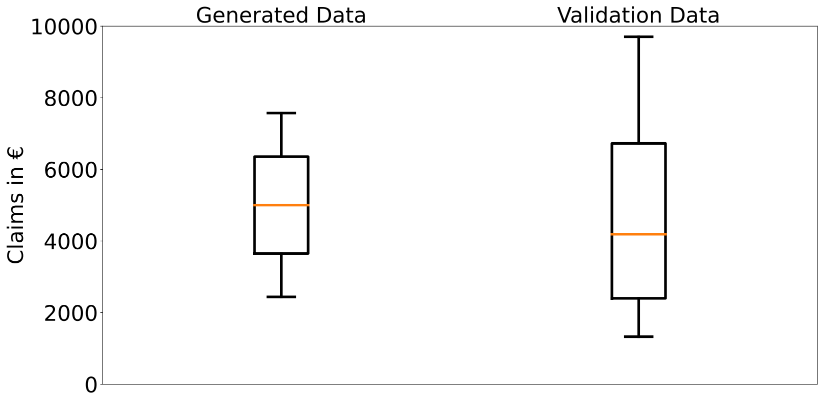

Fig. 1. Comparing the generated inpatient claims with the validation data, the median (highlighted in both plots) is 5,005 Euros in the generated data and 4,199 Euros in the validation data. Although the box plot clearly shows the lack of outliers in both directions (whiskers indicate the 10th and 90th percentiles), this dataset is the first and only simulation of inpatient treatments, and future research could likely improve the model.

with the highest relative weight was chosen (ignoring combinations of primary ICDs starting with 'Z'). If the duration exceeds the lower threshold defined [4], it is reduced by 1/4, but no less than 1 and no more than five days. After the fraud creation and injection process, the final claim was calculated.

3.2 Evaluation of the Simulation Results

Simulation results are evaluated against publicly available information and proprietary data from AOK Bayern. That dataset comprises all 2021 inpatient treatments, including diagnoses, treatments, admission and discharge data, claim amounts, DRG codes, the hospital's ID, patient information, location, and over 1 million individual claims.

To evaluate the claims data generated, some preliminary assumptions were made to achieve comparability between the validation data and the generated data. Adding additional surcharge for extra charges, inflation and additional charges for particularly complex diagnoses (such as severe burns or hemophiliacs), the average claim rises to 4,620.65 Euro, still neglecting the inflation from 2018 to 2019. In 2021, the average claim for inpatient treatments observed in the validation data was 5,537.26 Euros, with a median absolute deviation (MAD) of 1,328.05 Euro. The MAD in the generated dataset is 1,187.86 Euro. While this indicates that simulation and validation data resemble each other, Figure 1 implies, that large outliers are skewing standard deviation and interquartile range.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_22

While the average length of stay in the simulated data is 0.39 days or 5.4 % shorter than in the statistical data (6.81 days in comparison to 7.2 days) [15], the ratio of short stays (within three days) is two percentage points smaller than the statistically observed numbers in 2021 (38 % compared to 40 %). Comparing the generated data with information from the validation data, these effects increase further. The median, which is less prone to outliers, is 5.0 days per inpatient stay in the generated data, while it is 6.86 days in the validation set, with a MAD of 3.0 days in both cases. These effects indicate either a slight underestimation of stays within three days or the missing representation of a few very long hospital stays. The longest hospital stay in the validation set is 730 days, four times longer than the longest generated case.

Finally, 15 cases were randomly selected for an expert review by a public health insurance's hospital controller to validate the hospital claims in their respective context. Consistently, the problem of too many codes for imaging (OPS chapter 3), and too few actual surgeries (OPS Chapter 5) was observed. This pattern was also visible in statistical comparisons (Chapter 3 occurred 16 % too often, while Chapter 5 occurred 32 % too rarely).

4 Fraud Detection

The primary objective of this paper is to identify instances of fraud in individual cases by employing Machine Learning (ML) models. To train these, the previously simulated inpatient claims were used. Initially, we adjusted the features to avoid hospital-based over-fitting by replacing the hospital ID with six individual attributes calculated for each hospital. These attributes include average claim per treatment, average number of ventilation hours, average rate of cesarean section deliveries, average weight of newborns, number of inpatient treatment cases, and average distance between the hospital and its patients.

For fraud detection using ML models, standard algorithms such as Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Multi-Layer Perceptron (MLP), and Logistic Regression (LR) from scikit-learn [16] were trained. Li et al. [17] noted that most research on healthcare fraud detection focuses on MLPs and DTs, suggesting their potential performance superiority after hyperparameter tuning.

4.1 Results

All five models perform well at classifying claims with no prevalent fraud, achieving a precision of over 98 %. However, slight differences exist. While RF, GB, and MLP have a recall of over 99 %, DT and LR labeled less than 99 % of all non-fraudulent cases correctly (see also Table 2).

Regarding the precision of models in predicting fraudulent claims, RF performs the best, correctly identifying 98.4 %, closely followed by GB with a precision of 91.2 %. Conversely, LR and DT (precision of 15.0 % and 4.6 %) perform the poorest. However, in terms of recall, LR outperforms others, detecting 67.4 %

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_22

Table 2. Performance comparison of Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Multi-Layer Perceptron (MLP), and Logistic Regression (LR) on the synthetic fraud detection dataset. As expected, all models detect the majority class (No Fraud) with high precision; however, the performance differs significantly for the minority class (Fraud). Random Forest and Gradient boosting achieve the highest weighted F1 score and are therefore selected for further studies.

			Class 0 (No Fraud)		Class 1 (Fraud)	
Model	AUC	Weighted F1 score	Precision	Recall	Precision	Recall
DT	0.615	0.416	0.981	0.600	0.046	0.629
RF	0.697	0.777	0.982	**0.999**	**0.984**	0.394
GB	0.753	**0.824**	0.985	0.998	0.912	0.513
MLP	0.709	0.759	0.982	0.995	0.706	0.424
LR	**0.778**	0.589	**0.989**	0.882	0.150	**0.674**

of all fraudulent cases detected, tailed closely by DT at 62.9 %. GB detects every second fraudulent case, while MLP and RF (recall of 42.4 % and 39.4 %) perform slightly worse. Overall, RF and GB achieve the best results.

When analyzing recall and precision on the six included fraudulent patterns, differences between models and patterns are evident across all models and settings. RF and MLP fail to predict increases in ventilation hours and changes to cesarean sections. The fraud pattern best detected by all classifiers is changes in the order of ICD codes.

RF and GB, the top-performing models on the generated dataset, were applied to AOK's data. Among 899,610 cases of interest, RF classified 1 % of all claims as fraudulent, while GB did so at 12 %. Although these values vary widely and slightly exceed the internationally observed healthcare fraud rate of up to 10 % [18]. However, further validation is required.

4.2 Discussion

Implementing these models in practical applications can improve processes if the limitations are taken into account. RF and GB exhibit high precision scores, resulting in a lower false positive rate. However, this comes at the cost of a relatively low number of detected fraud cases in the dataset, with only 39 % (RF) and 51 % (GB). Despite this, they remain favorable compared to DT, MLP, and LR for practical applications.

As all models struggle with fraud patterns where only small changes occur, additional approaches are necessary. The structure of the data implies that outlier detection algorithms could be suitable. This is especially promising for parameters such as the number of ventilation hours. In our data an unexpectedly high number of occurrences can be observed at threshold values, indicating potentially fraudulent behavior.

The employment of RF and GB on the real-world dataset shows the potential of applying models only trained with simulated data. As the results lie within the expected range, the applicability of the proposed approach is supported. Nonetheless, the ML models' dependency on the quality of the inpatient claims simulation is conclusive. With improvements in this regard, an increased performance of the prediction algorithms seems to be assured. These results demonstrate the possibility of inpatient fraud detection based on previously generated data. Even though the method is not yet refined, sufficient results have been achieved, both on simulated as well as on real data.

5 Conclusion

We presented the first known approach to simulate inpatient billing data and developed machine learning methods to detect inpatient claims fraud based on it. We leverage publicly available statistics and inject known fraudulent behaviors into the dataset to enable the supervised training of fraud detection methods. While most inpatient care cases are simulated accurately, outliers are underrepresented in the simulation.

Nonetheless, our method achieved sufficient results for fraud detection in the next step. Based on the simulated data, Gradient Boosting and Random Forest were the most convincing models, with weighted F1 scores of around 80 %. To detect small changes in only one parameter, alternative approaches should be considered to improve performance. Outlier Detection methods are suspected to be promising. When applying this approach to uncover inpatient claims fraud on German health insurance data, the observed rate is up to 12 %. While this claim is backed by literature, further validations are necessary.

The proposed approach for data simulation may inspire further research in the inpatient care domain, specifically in fraud detection, hospital planning, and healthcare resource allocation.

References

1. Bundesministerium für Gesundheit: Vorläufige Finanzergebnisse der GKV für das Jahr 2021. https://www.bundesgesundheitsministerium.de/presse/pressemitteilungen/vorlaeufige-finanzergebnisse-gkv-2021.html. Last accessed 23 Dec 2023
2. AOK Bundesverband GbR: Fehlverhalten im Gesundheitswesen. Bericht über die Arbeit und die Ergebnisse der Stellen zur Bekämpfung von Fehlverhalten im Gesundheitswesen. 2021. https://aok-bv.de/imperia/md/aokbv/presse/pressemitteilungen/archiv/taetigkeitsbericht_fv_im_gesundheitswesen_2018-2019.pdf. Last accessed 23 Dec 2023
3. Jürges, H., Köberlein, J.: First do no harm. Then do not cheat: DRG upcoding in German neonatology. DIW Discussion Papers (2013)
4. Institut für das Entgeltsystem im Krankenhaus: Fallpauschalen-Katalog gem. § 17b Abs. 1 S. 4 KHG Katalog ergänzender Zusatzentgelte gem. § 17b Abs. 1 S. 7 KHG Pflegeerlöskatalog gem. § 17b Abs. 4 S. 5 KHG. https://www.g-drg.de/ag-drg-system-2021/fallpauschalen-katalog/fallpauschalen-katalog-2021. Last accessed 26 Dec 2023

5. Busse, R., Geissler, A., Aaviksoo, A.: Diagnosis related groups in Europe: moving towards transparency, efficiency, and quality in hospitals?. BMJ (Clinical research ed.) (2013). https://doi.org/10.1136/bmj.f3197
6. van Herwaarden, S., Wallenburg, I., Messelink, J.: Opening the black box of diagnosis-related groups (DRGs): unpacking the technical remuneration structure of the Dutch DRG system. Health Economics, Policy and Law (2020). https://doi.org/10.1017/S1744133118000324
7. Sievert, J.: Möglichkeiten der Abrechnungsmanipulation im Krankenhaus. Logos, Berlin (2011)
8. Statistisches Bundesamt: "23131-0003: Krankenhauspatienten: Deutschland, Jahre, Geschlecht, Altersgruppen, Wohnort des Patienten, Hauptdiagnose ICD-10 (1-3-Steller Hierarchie)" (2022). https://www-genesis.destatis.de/genesis/downloads/00/tables/23131-0003_00.csv. Last accessed 26 Dec 2023
9. Statistisches Bundesamt: "23141-0003: Nebendiagnosen der vollstationären Patienten: Deutschland, Jahre, Geschlecht, Altersgruppen, Wohnort des Patienten, Nebendiagnosen ICD-10 (1-3-Steller Hierarchie)" (2022). https://www-genesis.destatis.de/genesis//online?operation=table&code=23141-0003. Last accessed 26 Dec 2023
10. Statistisches Bundesamt: "23141-0111: Operationen und Prozeduren an vollstationären Patienten: Bundesländer, Jahre, Geschlecht, Altersgruppen, Operationen und Prozeduren (1-4-Steller Hierarchie)" (2022). https://www-genesis.destatis.de/genesis//online?operation=table&code=23141-0111. Last accessed 26 Dec 2023
11. Statistisches Bundesamt: Neues Krankenhausverzeichnis. (2021) https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/krankenhausverzeichnis.html. Last accessed 26 Dec 2023
12. IMC clinicon: IMC Navigator https://www.imc-clinicon.de/tools/imc-navigator/index_ger.html. Last accessed 26 Dec 2023
13. World Health Organization: Weight-for-age BOYS, https://cdn.who.int/media/docs/default-source/child-growth/child-growth-standards/indicators/weight-for-age/wfa-boys-0-13-zscores.pdf. Last accessed 26 Dec 2023
14. World Health Organization: Weight-for-age GIRLS, https://cdn.who.int/media/docs/default-source/child-growth/child-growth-standards/indicators/weight-for-age/wfa-girls-0-13-zscores.pdf. Last accessed 26 Dec 2023
15. Statistisches Bundesamt: Grunddaten der Krankenhäuser, https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Kranken haeuser/Publikationen/Downloads-Krankenhaeuser/grunddaten-krankenhaeuser-2120611217004.pdf. Last accessed 26 Dec 2023
16. Pedregosa, F., Varoquaux, G., Gramfort, A.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (2011)
17. Li, J., Huang, K.-Y., Jin, J.: A survey on statistical methods for health care fraud detection. Health Care Management Science (2008). https://doi.org/10.1007/s10729-007-9045-4
18. Gee, J., Button, M., Brooks, G.: The financial cost of Healthcare fraud. University of Portsmouth and Maclntyre Hudson LLP (2010) https://pure.port.ac.uk/ws/portalfiles/portal/1925942/The-Financial-Cost-of-Healthcare-Fraud—Final-%282%29.pdf. Last accessed 26 Dec 2023

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_22