applied sciences

Article

A Decision Support System with Artificial Intelligence and Natural Language Processing to Mitigate the Deduction Rate of Health Insurance Claims

Shey-Chiang Su [1], Chun-Che Huang [2,*], Roger R. Gung [2], Li-Kai Hsiung [1], Zhi-Wei Gao [1] and Cheng-En Tsai [2]

[1] Puli Christian Hospital, Puli 54546, Taiwan; 4267@mail.pch.org.tw (S.-C.S.); fc2m2s3adha@mail.pch.org.tw (L.-K.H.); 1718@mail.pch.org.tw (Z.-W.G.)

[2] Department of Information Management, National Chi Nan University, Puli 54561, Taiwan; Roger.Gung@phoenix.edu (R.R.G.); s109213511@mail1.ncnu.edu.tw (C.-E.T.)

* Correspondence: cchuang@ncnu.edu.tw

Abstract: Globally, 20% to 40% of medical resources are wasted, which could be avoided through professional audit of health insurance claims. The professional audit can pinpoint excessive use of unnecessary medicines and medical examinations. Taiwan's National Health Insurance Bureau (TNHIB) deducts the weight that medical resources carry if regarded as unnecessary or abused when examining health insurance claims. The ratio of the deducted weight to the total weight claimed by a hospital is defined as the health insurance claim deduction rate (HICDR). A high HICDR increases the operating expenses of the hospital. In addition, it takes the hospital many resources to prepare and file appeals for the deduction. This study aims to: (1) minimize the weight deducted by the TNHIB for a hospital; and (2) facilitate efficient appeals to claim denials. It is expected that HICDR will be reduced through big data analytics. In this study, evidence-based medicine (EBM) is involved to clarify the debate, dilemmas, conflicts of interests in examining health insurance claims. A natural language method—latent Dirichlet allocation (LDA), was used to analyze patients' medical records. The topics derived from the LDA are used as factors in the logistic regression model to estimate the probability of each claim to be deducted. The experimental results on various medical departments show that the proposed predictive model can produce accurate results, and lead to more than 41.7% reduction to the deduction of the health insurance claims. It is equivalent to more than a 750 thousand NT dollars saving per year. The efficiency of application is validated compared to the manual process that is time-consuming and labor intensive. Moreover, it is expected that this study will supplement the insufficiency of traditional methods and propose a new and effective solution to reduce the deduction rate.

Keywords: deduction rate of health insurance claims; evidence-based medicine; big data analytics; logistic regression; latent Dirichlet allocation

check for updates

Citation: Su, S.-C.; Huang, C.-C.; Gung, R.R.; Hsiung, L.-K.; Gao, Z.-W.; Tsai, C.-E. A Decision Support System with Artificial Intelligence and Natural Language Processing to Mitigate the Deduction Rate of Health Insurance Claims. *Appl. Sci.* **2021**, *11*, 11623. https://doi.org/10.3390/app112411623

Academic Editor: Valentino Santucci

Received: 27 October 2021
Accepted: 1 December 2021
Published: 7 December 2021

1. Introduction

The World Health Organization (WHO) pointed out in its annual report that high percentage of medical resources are wasted globally [1]. For example, in 2016, the United States spends more on health care than any other country, with costs approaching 18% of the gross domestic product (GDP) as well as approximately 30% of health care spending may be considered waste during 2012–2019 [2]. The amount of medical waste in health care remains an important issue in medical healthcare finance. Healthcare professionals perceive substance abuse patients as individuals suffering from self-inflicted illness that are unworthy of medical treatment, a burden to the medical system, and a waste of medical expense [3].

In Taiwan, a payment system for associated groups in medical diagnosis has been implemented in Taiwan since 2010 to use medical resources effectively. In most cases, after hospitals have applied for weight claims, every quarter, the Taiwan's National Health Insurance Bureau (TNHIB) conducts professional audit for random selected sample by its audit committee that consists of medical professionals. By checking whether the prescribed drugs and patients' medical records are consistent with the requirements of the TNHIB, the audit committee ensures the appropriateness of medical services provided by hospitals to patients. If the service is considered unnecessary, the hospital expense will not be reimbursed. Under Taiwan's payment system, each medical service carries a certain weight; the corresponding weight is deducted if the medical service is regarded as "wasted and unnecessary" by the audit committee. The deduction rate is the deducted weight divided by the total weight of a hospital claim. The medical records are usually described in a free text format. The audit committee is unaware of the relationship between the medical records and the prescribed drugs immediately after they read the records. However, much information in the medical records is closely related to the medical examination and weight deduction. Therefore, it is very important to determine how to find out the characteristics of medical records for later analysis so as to help hospitals discern health insurance claims that would be possibly denied by the committee before application so as to reduce the deduction rate, and to develop a model that could be used by the TNHIB's for preliminary audit for a two-way management.

To lower the deduction rate, the following impacts should be taken into consideration:

- Recognizing key words, phrases, and sentences which were written without specific formats, and understanding descriptive topics in patient records;
- The development of an effective mechanism and strategy from an economic perspective that will yield high success rates with minimum human works;
- The key factors that influence the weight deduction, including examination items, types of prescribed drugs, and patients' symptoms.

In addition, the challenges are with soundness of diverse group-specific focuses that may lead to debate, dilemmas and conflicts of interests. It is necessary to clarify the debate, dilemmas, conflicts of interests in examining the claims. Evidence-based medicine (EBM) is the focus in this study, to resolve the deficiency of the diversity, difference, and political bias. EBM could lead to various research methods, that could not only integrate collaborative and participatory approaches, but also define problems and consensus-based solutions.

To analyze patient records, text mining is used in this study to analyze the cases and determine the rules about how the weight deduction is obtained and to establish a database of the words used in medical records. In this way, the doctors can quickly and easily follow the rules to solve weight deduction problems. The objectives of this study to help hospitals minimize the deducted weight by establishing a database of successful and failed cases through big data analytics technologies, which could help avoid wrong or unsuitable weight claims and provide supporting information for an appeal to a medical claim denial.

This study focuses on three morbid entities with highest HICDR: E11 (type 2 diabetes mellitus), N18 (chronic kidney disease), K21 (gastro-esophageal reflux disease) and excludes the hospitalized records. Section 3 surveys the literature related to the claim examination and approaches to reduce the rates of health insurance claims. The proposed solution approach is presented in Section 3. The case of Pu-Chi Hospital is studied in Section 4. Section 5 concludes this study. The contribution of this study to the hospitals is to provide a solution to reduce the deduction rate of the health insurance claims and avoid medical resource waste.

2. Literature Review

There are some major studies of claim examination and analysis, for example, Sacks et al. [4] think that rising health insurance premiums represent a rapidly increasing burden on employer-sponsors of health insurance and their employees. A Web-Based Nutrition Program is proposed to reduce health care costs in employees with cardiac risk factors. The America's Institute of Medicine (IOM) reported that prescribing wrong medication is a big problem, and the effects can sometimes be fatal. To address this problem, Miller, Mansingh, and OptiPres designed and implemented a distributed intelligent mobile agent-based system called "OptiPres" [5]. It is to assist doctors in making more informed decisions by either choosing the optimal solution from processing a repository of past decisions, or by presenting a set of possible drugs and using a specific criterion to identify the optimal prescription. Huang et al. succeeded performed the analysis of a probabilistic model for reducing medication errors with various thresholds to reduce inappropriate prescriptions [6]. Sung et al. analyze bibliometric and text mining on PubMed using Taiwan's National Health Insurance claims data [7] and Chen et al. applied data mining technology to build a prediction model for this purpose and compared the performance of the model with logistic regression to the model with a neural network. The sensitivity analysis showed that an overall check-up on administrative audit can greatly reduce the load of professional audit. Henceforth, the load of auditing is reduced [8]. Maass et al. validate that the timely access to up-to-date diagnostic information reduced redundant clinical re-appointments, repeated tests, and mail orders for missing data [9].

Chien et al. [10] developed a model that informs doctors of the most suitable items with high unexpected costs and the medical expenses that may exceed the payment quota when doctors write on medical records. This model improves the writing of medical records, provides supplemental information for the main diagnosis and possible complications. It even includes detailed reasons for the high costs of a service or item, thus avoiding cost deduction in health insurance sample review. Chen discussed the influence of health insurance weight claim deduction on outpatient medical practices and interviewed those who worked in a regional teaching hospital in Taiwan through qualitative research [11]. Huang investigated the cases of approved insurance coverage for emergency treatments and analyzed the denied cases through a mixed method that combined a quantitative method of using the modified Delphi questionnaires and a qualitative method with using in-depth interviews to increase the success rate of the health insurance coverage claims [12]. Yi-Ting Cheng used medical records to produce retrospective professional judgements, and interviewed doctors for their opinions on medical records with high outpatient rates and low deduction rate. It provides reference for increasing quality index of outpatient medical records and reducing medical cost deduction [13].

Most of the above-mentioned research was conducted via expert interview and questionnaire survey. Only a few used big data systems and text mining to determine the topics and words related to the deducted items. Moreover, doctors do not maintain medical records in a uniform format. The audit committee will judge whether a medical treatment is reasonable according to the correlation between patients' medical records and prescribed drugs. In this perspective, research in recognizing and identifying important words and topics in medical records is in a great need. The technology of text mining can be used to pinpoint the important words and phrases, which can then be used as logistic regression model variables to investigate whether significant correlation exists in these variables and other medical variables [14–16]. The text content analysis model and logistic regression model are proposed in Section 3.

In addition, the emergence of evidence-based medicine in the early 1990s, which is defined as: 'the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients, based on skills which allow the doctor to evaluate both personal experience and external evidence in a systematic and objective manner' [17]. Therefore, this study develops an evidence-based tool that can crawl/collect topics based on the following five core inspirations of evidence-based medicine: question formulation, evidence search, critical appraisal, evidence application, and outcome evaluation [17,18] as references to decision makers.

3. Methodology

The research procedure of this study is shown in Figure 1. After data preprocessing, evidence of documents is crawled and used for supporting the topic recognition. Structural topic modeling was used to dig latent topic from medical text data. The results of the topic model can be used to train and test the professional audit prediction model. Lastly, the factors influencing the weight deduction were determined based on the test results.

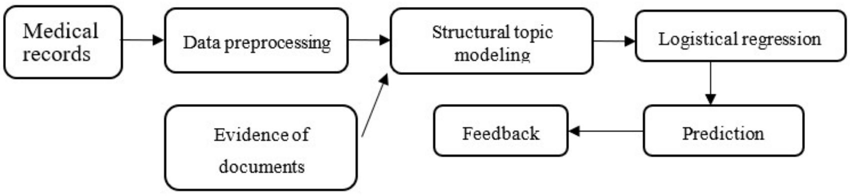


Figure 1. Solution Procedure.

Only accurate data lead to good analysis results; that is why original data needs to be preprocessed to ensure its accuracy and conformity with the research needs before research is conducted. In this study, data preprocessing consisted of four parts:

1. Data format conversion: Exported from different systems, raw data were need to be converted into a uniform format for subsequent data cleaning, integration and analysis. The common data storage formats include xml, csv, json, kml, xls, pdf, ods, txt, zip, and xls. In this study, the health care expenditure declaration data are stored in the format xml. Research team transfer all data into xlsx format for convenience of subsequent data integration and analysis.
2. Data integration: In order to ensure the integrity and accuracy of the aggregated data, a unique ID across all different data source should be selected to perform data integration. After data integration, the data should have following properties:
   - Uniqueness: a serial number for a primary key;
   - Integrity: there is no missing value in a column or row;
   - Validity: for example, the value of months in a spreadsheet should be 1 to 12.
3. Building a thesaurus database: The doctors' orders served as original data in this study. However, many abbreviations and wrongly written words due to doctors' different habits were found. This made it necessary to compile the key words into a thesaurus by way of manual review, which was later submitted to the doctors to confirm its accuracy.
4. Text cleaning: The text cleaning is an important step in this study, which can facilitate subsequent analysis of LDA. The Step 1 of text cleaning involves changing all the letters in the text to lowercase letters; Step 2 entails importing the reviewed correct key words to replace the synonyms; and Step 3 consists of deleting all special characters and numbers in the text and combine specific words.

3.1. Evidence Based Medicine

The research framework of evidence-based medicine emphasizes the knowledge, instead of position or privilege. This study will apply this research framework in order to explore the relationship between factors and have an effective evaluation, and to sublime the target problem into the cognition and interpretation of the nature of evidence by the perspective of evidence-based medicine. This is for the sense of responsibility in the professional cognition of the research object.

The strengths of evidence-based medicine are the opportunity that affords to directly compare the effectiveness of different interventions, and this development of the system could express the evidence to different research objects based on the background conditions, relationships, and interactions, also called the process of art in [19]. In original evidence-based medicine (EBM), the clinicians need to develop skills to evaluate research (critical appraisal skills) and keep up to date with research findings [20]. Five core actions, such as question formulation, evidence search, critical appraisal, evidence application, and outcome evaluation [17,18], are regard as the five explicit steps of evidence-based practice. Based on the five steps of EBM [18,20], this study proposes five major/practical steps to develop an evidence-based approach to support claim examination (Figure 2). Specifically, in the task of searching evidence, such as the definition, description, type, and measurement of indicators, this study is based on the concept of "the 5S evolution of information services" [21]. The 5S (studies, syntheses, synopses, summaries, systems) pyramid model is used to search the best empirical literature evidence.

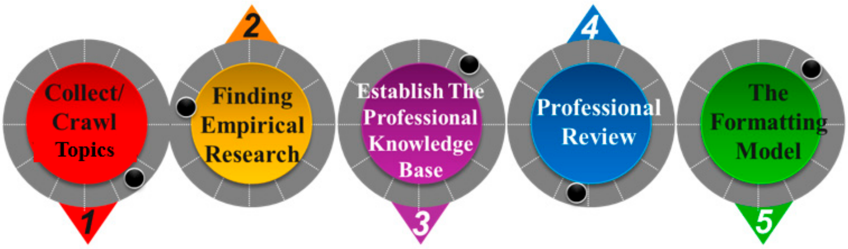

Collect/ Crawl Topics
1
2
Finding Empirical Research
Establish The Professional Knowledge Base
3
4
Professional Review
The Formatting Model
5

Figure 2. Five steps of EBM.

Step 1: Collect/crawl topics.

Step 2: Searching the best empirical literature evidence.

According to the hierarchy concept of the 5S (studies, syntheses, synopses, summaries, systems) pyramid model [21], the process of searching the best empirical literature evidence is applied. The 5S in this study refer to the following five stages (Figure 3):

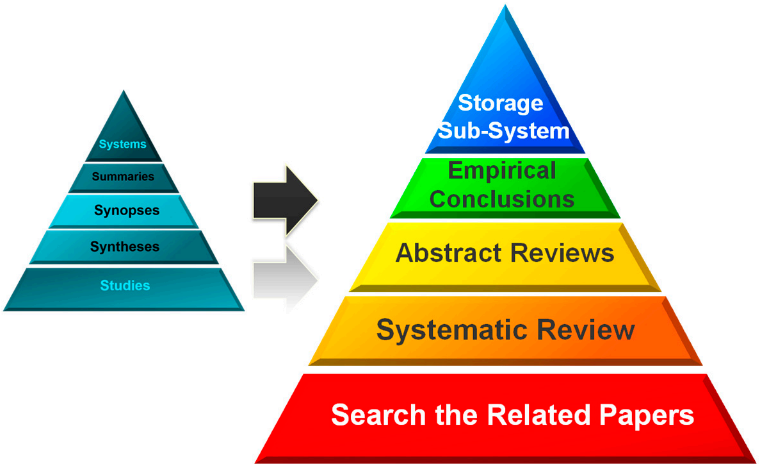


Figure 3. The 5S Pyramid Model.

- Studies: search the related papers;
- Syntheses: synthesize those studies in a systematic review;
- Synopses: abstract reviews of individual research and retrospective documents;
- Summaries: obtain empirical conclusions problem/topic;
- Storage sub-system: develop a data lake [22] to store the evidence.

Step 3: Analyze the literature to develop the topic base;

Step 4: Professional review;

Step 5: Modeling prediction model based on the evidence.

3.2. Topic Model

Topic model is an unsupervised generation model, which is widely used in word frequency analysis and text classification. The latent Dirichlet allocation (LDA) proposed by Blei et al. [23] is one of the most classic topic models [24]. This study considered the background information of each document in text as important for research analysis. Incorporating the background information into the model as covariates can effectively improve the performance of the model and better understand the model results. The LDA allows researchers to analyze the relationship between a topic and the background information of a text. Therefore, the LDA was employed as a probabilistic topic model for this research to better explore the latent topics in the text data.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document w in a corpus D:

1. Choose N ~ Poisson(ξ);
2. Choose θ ~ Dir(α);
3. For each of the N words w_n:
    - i. Choose a topic z_n ~ Multinomial(θ);
    - ii. Choose a word w_n from p(w_n | z_n,β), a multinomial probability conditioned on the topic z_n.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality *k* of the Dirichlet distribution (and thus the dimensionality of the topic variable *z*) is assumed known and fixed. Second, the word probabilities are parameterized by a k × V matrix β where $\beta_{ij} = p(w_j = 1 \big| \, z_i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows, and more realistic document length distributions can be used as needed. Furthermore, note that *N* is independent of all the other data generating variables (θ and *z*). It is thus an ancillary variable, and we will generally ignore its randomness in the subsequent development.

A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (a *k*-vector θ lies in the $(k-1)$-simplex if $\theta_i \geq 0$, $\sum k \; i = 1 \; \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \tag{1}$$

where the parameter α is a *k*-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the gamma function. The Dirichlet is a convenient distribution on the simplex—it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 5, these properties will facilitate the development of inference and parameter estimation algorithms for LDA. Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics *z*, and a set of *N* words *w* is given by

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

3.3. Regression Analysis

Logistic regression (LG) model is an efficient and commonly used classification model for binary outcomes, first proposed by Cox [25]. Logistic regression is widely used, e.g., in analyzing the effect of patient-physician relationship [26], the effect of web-based exercise as an effective complementary treatment for patients [27], etc. Fuurthermore, a regression model typically can serve very well in predicting a dependent variable, when a good number of independent variables are avaiable. In this research, logistic regression models are developed to predict the probability of a certain health insurance claim becoming denied. The independent variables used to perform the logistic regression analysis include the topics derived from the LDA model and the medical information, such as treatments, medicines, and prices. A list of binary variables is defined based on the results of the LDA model. For each binary variable, the value one represents that a case with a certain topic. To determine if a topic is covered in the document, the number of high-frequency words of the topic utilized in the document is measured. The common thresholds are three or four, which varied across different departments.

Like LDA, the logistic regression model is built for each department. Due to the large number of predictive variables, a variable selection algorithm is used to enhance the regression model. Zou and Hastie proposed a regularization and variable selection method for the general linear regression called elastic net [28]. This algorithm can be used to solve several common issues such as overfitting and collinearity.

To realize the regression model, this research builds a generalized linear model with the elastic-net penalty by using R package glmnet [29]. The function of glmnet is to solve the following problem:

$$\text{in}\frac{1}{N}\sum_{i=1}^{N} l\Big(y_i,\ \beta_0+\beta^T x_i\Big)+\lambda\Big[(1-\alpha)||\beta||_2^2+\alpha||\beta||_1\Big] \tag{3}$$

in this study, where:

y_i is the outcome of the case i,

x_i is the vector of predictor variables for case i, such as topic, disease and drug, and l is the negative log-likelihood function.

The hyper-parameter λ in Equation (1) controls the overall strength of the penalty. This research performed a cross validation to find the best λ which gives a minimum mean misclassification error. The graph in Figure 4 with the department of family medicine as example shows the binomial deviance which varies as the λ changes.

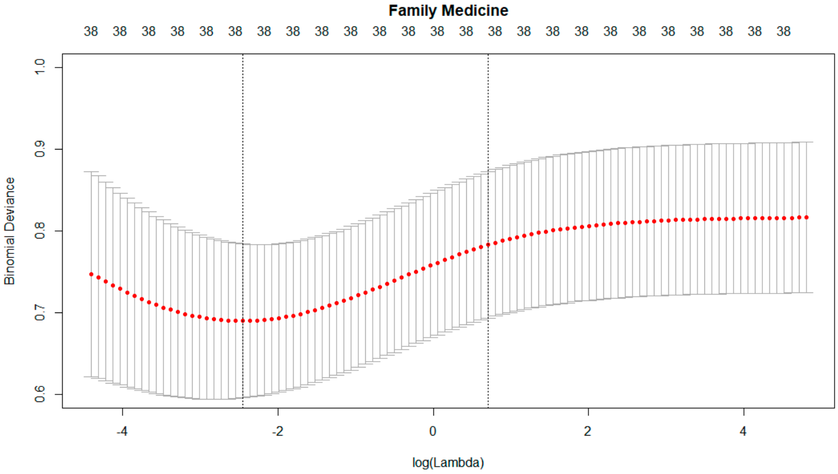


Figure 4. The hyper-parameter λ of the Department of Family Medicine.

The elastic-net penalty controlled by α in the equation bridges the gap between lasso ($\alpha = 1$) and ridge ($\alpha = 0$). Based on the knowledge that ridge penalty shrinks the coefficients of correlated predictors towards each other while the lasso tends to pick one of them and discard the others, this research tried many different values to find the best α that can minimize the overfitting problem.

The predictive model was built on training dataset and validated by testing dataset. The performance of the model on the testing dataset including precision, recall and area under curve (AUC). "Precision" represents the proportion of actual denied GC under the predicted denials, which was calculated as precision = TP/(TP + FP), where TP is the true positive count, and FP is the false positive count. Furthermore, "recall" represents the proportion of predicted denied GC under the actual denials, which was calculated as recall = TP/(TP + FN), where FN is the false negative count. A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied and the sensitivity and specificity are used as the coordinates to construct the curve. The area under curve (AUC) is a statistical index that indicates the classification power of the predictive model, larger AUC, the better classification. Besides ROC and its AUC, the F_β score is the harmonic mean of "precision" and "recall", which was calculated as $F_\beta = \frac{(\beta^2+1)PR.}{\beta^2 P+R}$, where β is the weight of Recall that indicates recall is β times as important as precision [30]. F_β-score is not only used to measure the model performance with the given classification threshold, but also can be used to determine the optimal threshold for decision making. Once the weight β is determined, the optimal threshold can be set to the level where yield the largest F_β-score.

4. Results

4.1. Data

In this study, the data of 2016–2019 provided by the Pu-Ch Hospital are analyzed. The three morbid entities with the most Health Insurance Claim Deduction Rates: N18 (chronic kidney disease), K21 (gastro-esophageal reflux disease), E11 (type 2 diabetes mellitus) are focused.

Firstly, the medical records are preprocessed to eliminate duplicated data, ignore symbols and numbers, and transfer uppercases to lowercases for the topic model, for example of morbid entity E11 (Table 1).

Table 1. Results of the data preprocess.

	Content	
	Subject	Object
1	Euthyroidism Stable l	Free T TSH GLUAC AC CHO. Free T TSH T P
2	Common cold recently	One touch glucose—AC HBA C keep.
3	A case of DM recently	DMP body weight PCKD body weight.
4	Hyperlipidemia symptoms	Body height (cm) body weight (BMI)
5	A case of DM recently	One touch glucose—AC HBA LDL cholesterol taper lantus to UHS still left
6	Toothache no discomfort right edema. Referred from Doctor Young for poor renal function	Heart rate edema bun creatinine EGFR CKD-stages–Stage B albumin potassium HBA C GLUAC AC
7	Type DM years HTN hyperlipidemia. Symptom free SMBG FBS around PC. Each meal—two bowls of steamed rice.	Body weight—GLUAC no injection. Last night, poor memory—daughter complaint
8	Common cold recently	One touch glucose AC HBA C keep

Secondly, we crawled large of documents related to each morbid entity to capture keywords. Based on the frequency of these keywords, the weight of for each document is computed by the LDA model presented next. The highest weighted documents are used in this study.

4.2. Topic Model

The topic model, LDA is applied to the documents crawled in Section 4.2 to produce exclusive topics. After numerous try and error, topic number is determined as six based on topic number and perplexity, for example, E11 in Figure 5.

The topic model is developed by LDA and the ggplot2 package of R is used. The most 20 frequent words of six topics are show, for example of E11 in Figures 6–8.

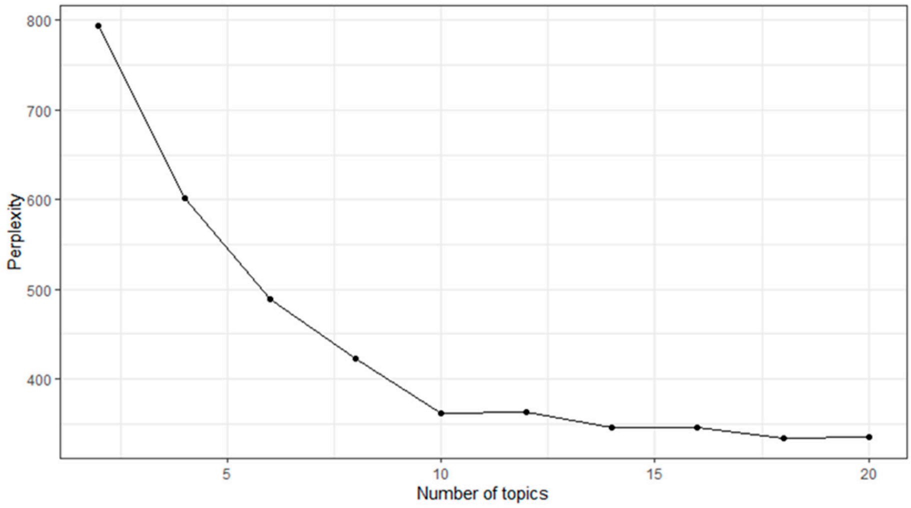


Figure 5. Topic number and perplexity (E11).

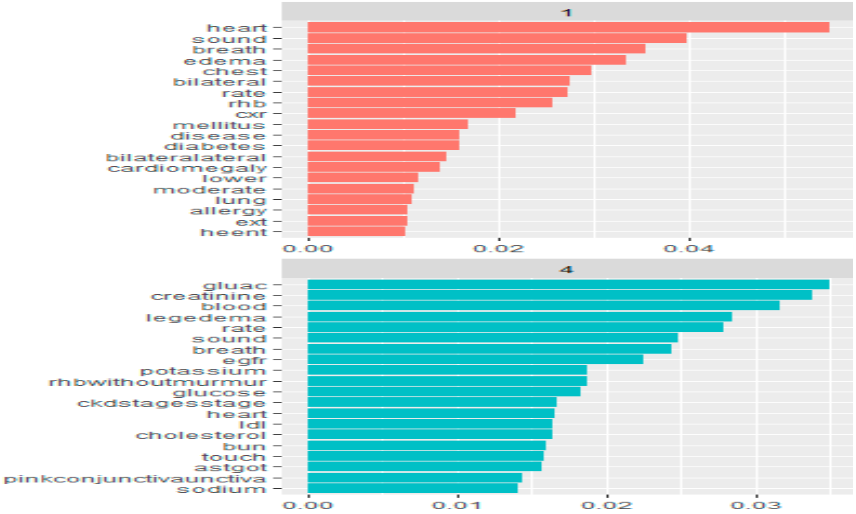


Figure 6. Word frequency distribution of 20 words in six topics (E11 Topics 1 and 4).

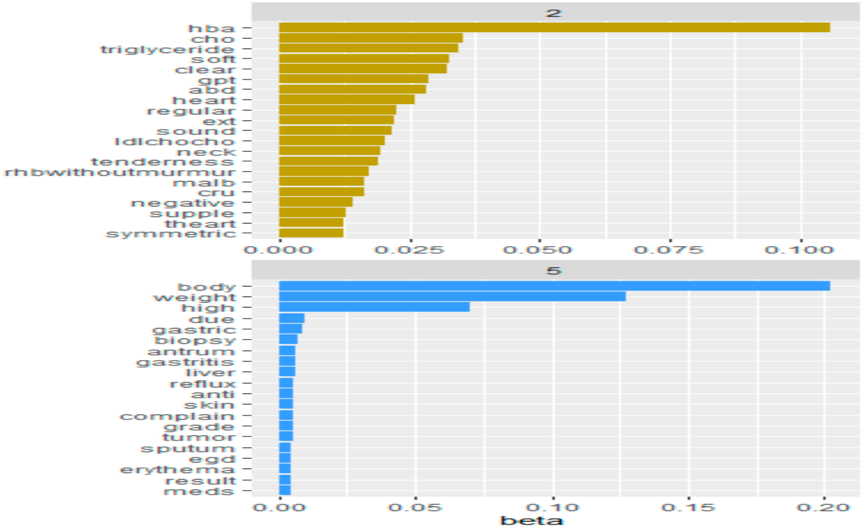


Figure 7. Word frequency distribution of 20 words in six topics (E11 Topics 2 and 5).

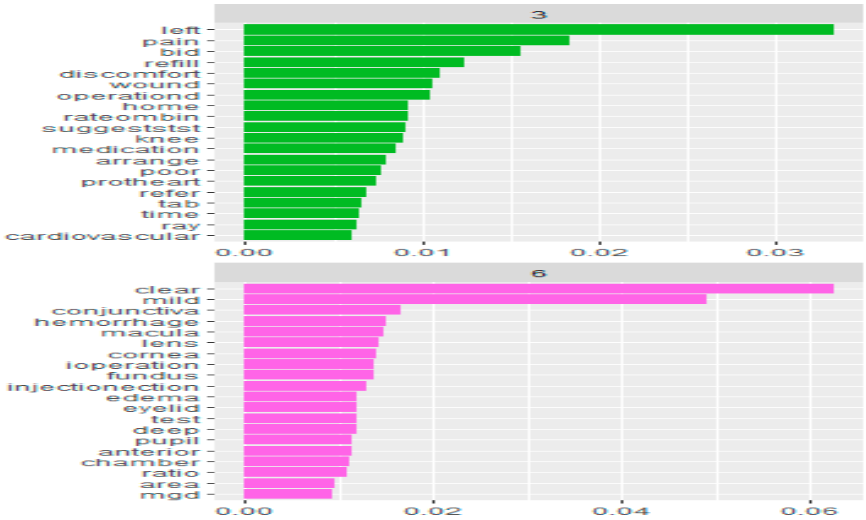
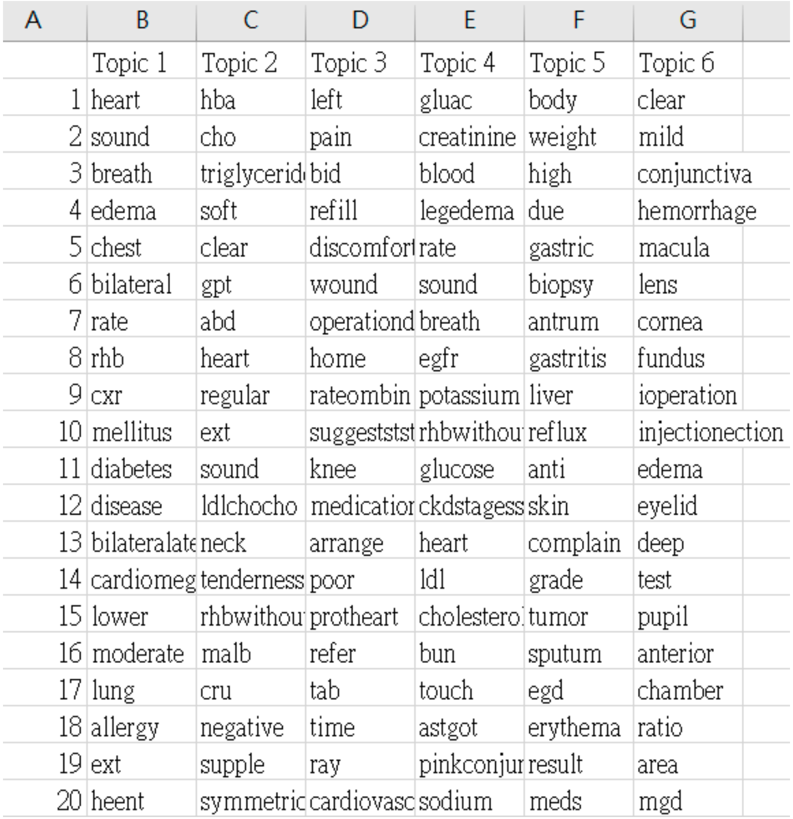


Figure 8. Word frequency distribution of 20 words in six topics (E11 Topic 3 and 6).

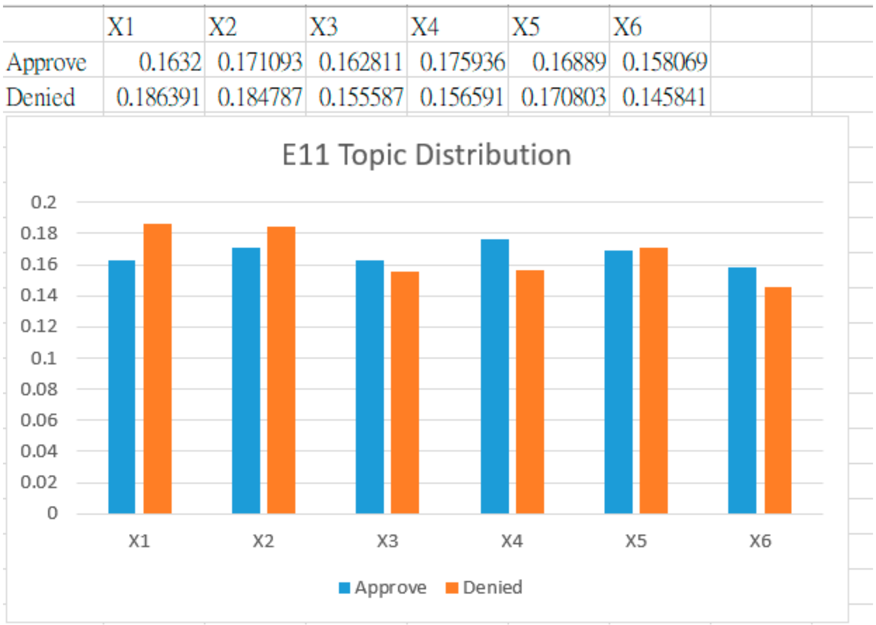
The topic and words by LDA are output to csv files. The files could be used in the prediction model next, for example of E 111 in Figure 9.



A	B	C	D	E	F	G
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	heart	hba	left	gluac	body	clear
2	sound	cho	pain	creatinine	weight	mild
3	breath	triglycerid	bid	blood	high	conjunctiva
4	edema	soft	refill	legedema	due	hemorrhage
5	chest	clear	discomfort	rate	gastric	macula
6	bilateral	gpt	wound	sound	biopsy	lens
7	rate	abd	operationd	breath	antrum	cornea
8	rhb	heart	home	egfr	gastritis	fundus
9	cxr	regular	rateombin	potassium	liver	ioperation
10	mellitus	ext	suggeststst	rhbwithou	reflux	injectionection
11	diabetes	sound	knee	glucose	anti	edema
12	disease	ldlchocho	medicatior	ckdstagess	skin	eyelid
13	bilateralate	neck	arrange	heart	complain	deep
14	cardiomeg	tenderness	poor	ldl	grade	test
15	lower	rhbwithou	protheart	cholestero	tumor	pupil
16	moderate	malb	refer	bun	sputum	anterior
17	lung	cru	tab	touch	egd	chamber
18	allergy	negative	time	astgot	erythema	ratio
19	ext	supple	ray	pinkconjur	result	area
20	heent	symmetric	cardiovasc	sodium	meds	mgd

Figure 9. List of words for each topic (E11).

The distribution of topics in the deducted cases, as well as in the approved cases, are provided by the LDA model, for example, E11 in Figure 10. This aims to understand the relationship among topics in deducted cases and approved cases.



	X1	X2	X3	X4	X5	X6
Approve	0.1632	0.171093	0.162811	0.175936	0.16889	0.158069
Denied	0.186391	0.184787	0.155587	0.156591	0.170803	0.145841

Figure 10. Topic deducted (approved/denied) distribution (E11).

4.3. Prediction Model

The medical records are classified as 70% for training date set and 30% for testing data set. For example, of E11—type 2 diabetes mellitus, 3154 records are classified as training data and 1351 records as testing data. This study aims to predict the medical records which are judged as deducted since the medical service is regarded as “wasted and unnecessary” by the audit committee. However, the deducted medical records are with a low rate of 5.22% of all records in Pu-Chi Hospital. A small number of deducted records vs.. large number of approved records result in low accuracy in data training since deducted records are with a low probability. The synthetic minority oversampling technique (SMOTE) is therefore used to increase the percentage of deducted cases [31]. In the case of E11—type 2 diabetes mellitus, the AUC is enhanced to 0.782 from 0.677 after applying the SMOTE procedure.

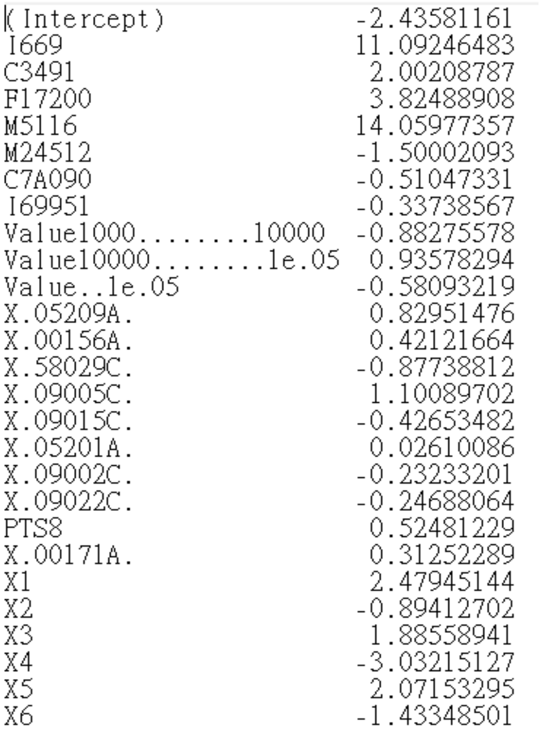
Table 2 shows the five factors that may led to weight deduction obtained from the results of regression analysis. (1) Object (data of doctors’ judgement and examinations), (2) D19 (primary diagnostic code), (3) D20~d23 (secondary diagnostic code), (4) p4, (codes of drugs and examinations) and (5) price (total weight). For text-based information such as “object”, research team transfer it into binary variable and put into the regression model for prediction. Moreover, D19 (primary diagnostic code) and D20–d23 (secondary diagnostic code) were used to observe whether a disease that has a great possibility of being deducted (far higher than the deduction rate of outpatient cases at Puli Christian Hospital) is relevant to weight deduction. The price (total weight) was utilized to observe whether the denial of a medical claim is related to the amount involved in the medical claim. Lastly, p4 (codes of drugs and examinations) was employed to see whether a drug has a high denial rate of medical claims.

After the SMOTE was used and the prediction model was trained, the 30% data set was tested. The regression coefficients were obtained, where positive/negative refers to deducted or not, respectively, for example E11 in Figure 11.

The factors that led to medical claim denial were selected for logistic regression modeling. The model training results are shown: the area under the curve (AUC) of the model is large enough for both training set and testing set. This indicates that the discrimination power of the ROC curve was enough to be adopted.

Table 2. Definition of Variables.

General Variable	Definition	Type	Value
Price	Total weight	Integer	[0~1000, 1000~10,000, 10,000~100,000, 100,000+]
LDA Topic variables	Definition		
Topic 1~Topic *n*	Structural topic model	Decimal	[0, 1]
Professional variable of medical fields	Definition		
Object	Data of doctors' judgement and examinations	Binary integer	[0, 1]
D19	Primary diagnostic code	Binary integer	[0, 1]
D20~d23	Secondary diagnostic code	Binary integer	[0, 1]
P4	Codes of drugs and examinations	Binary integer	[0, 1]



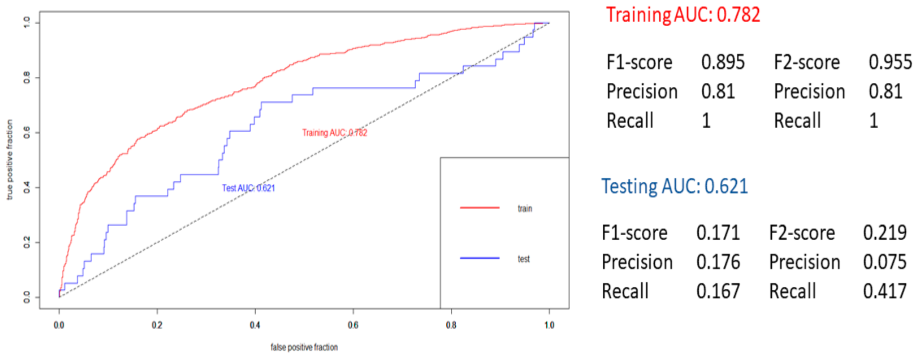
```
(Intercept)                  -2.43581161
I669                         11.09246483
C3491                         2.00208787
F17200                        3.82488908
M5116                        14.05977357
M24512                       -1.50002093
C7A090                       -0.51047331
I69951                       -0.33738567
Value1000........10000       -0.88275578
Value10000........1e.05       0.93578294
Value..1e.05                 -0.58093219
X.05209A.                     0.82951476
X.00156A.                     0.42121664
X.58029C.                    -0.87738812
X.09005C.                     1.10089702
X.09015C.                    -0.42653482
X.05201A.                     0.02610086
X.09002C.                    -0.23233201
X.09022C.                    -0.24688064
PTS8                          0.52481229
X.00171A.                     0.31252289
X1                            2.47945144
X2                           -0.89412702
X3                            1.88558941
X4                           -3.03215127
X5                            2.07153295
X6                           -1.43348501
```

Figure 11. E11 regression coefficients.

In Figure 12, the E11 recall increased from 0.167 to 0.417, which shows that the proposed prediction model improved 25% of deducted cases.

4.4. Model Performance

After building the logistic regression model, the research team evaluated model performance for all morbid entities by using the test dataset. Table 3 refers to the model performance. The recalls were 0.955, 0.955, and 0.849, and AUCs are 0.782, 0.776, and 0.852 respectively for E11, N18, and K21. Recall indicates the percentage of deducted cases were identified by the model, while AUC indicates the power of classification on a deducted case and an approved case. The results show the model performs very well in predicting the binary outcomes.



Training AUC: 0.782

F1-score	0.895	F2-score	0.955
Precision	0.81	Precision	0.81
Recall	1	Recall	1

Testing AUC: 0.621

F1-score	0.171	F2-score	0.219
Precision	0.176	Precision	0.075
Recall	0.167	Recall	0.417

Figure 12. E11 logistic regression.

Table 3. Model Performance.

Morbid Entity	AUC	Recall
E11	0.782	0.955
N18	0.776	0.955
K21	0.852	0.849

The benefits of the proposed solution approach are compared to the traditional manual approach in Table 4.

Table 4. The comparison of the proposed solution approach to traditional manual approach.

	Morbid Entity	The Proposed Solution Approach	Traditional Manual Approach
% Records to be fixed	E11	41.7%	Less than 10%
	N18	89.3%	Less than 10%
	K21	90.3%	Less than 10%
Amount money to be saved *	>NT 300 thousands dollars		Less than NT thousand dollars

* Amount of money saved = $\sum_{all\ diseases}$ records × recall × NT $ 0.9.

5. Discussion

The model established in this study can be used to predict the probability that a medical claim with item weights will be deducted during professional audit. Through the preliminary analysis of the system, if a claim is identified as having a low probability of being denied or weight deduction, the claim could be sent to the committee of TNHIB without further review. Otherwise, the claim must be reviewed and modified based on the system's suggestions before sending to the committee, to prevent claim denial. This study finds that not only the proposed system could provide modification suggestions pertinent to all weight deduction, but also a rule base for record writing guidance could give critical suggestions to increase the chance of passing the professional audit.

In this study, the topics explored by the LDA and medical treatments are used as the variables for the logistic regression model. The coefficients of the variables are used to analyze the topics, diagnostic results, and the influence of drugs on the denied medical claims. A variable with a positive coefficient indicated the existence of the variable can increase the probability of weight deduction, while a variable with a negative coefficient indicates the reverse effect. For example, the variables in E11, except Topic 2, all show significantly positive/negative to deducted or not. The topic model plays an important role in the perdition modeling.

In addition, the results regarding Disease I669 diagnosis shows that medical records M5116 are at greater risk of weight deduction. In this respect, it is recommended that doctors provide more supporting information when writing medical records that contain such topics, drugs and diagnoses while little information is related to main diagnosis code or requires high cost of drugs and examinations.

Moreover, it was found that the disease classification code (Codes for International Classification of Diseases II) is also significant to the probability. It can be speculated that the disease classification code exerts an impact on the audit committee's approval of a claim. Apparently, only when a known disease classification exists the symptoms of a disease can be described or recorded properly based on the doctor's diagnosis.

6. Conclusions

The professional audit and weight deduction by TNHIB are performed to achieve sustainable development of health insurance. However, this has been a significant burden to the management of the medical institutions. Effectively recording the medical process and evaluating the impact of examination data on the weight deduction in details help lessen this burden. This study identifies the key factors that can lead to weight deduction through the investigation of the denied medical claims using patients' reports, doctors' objective statements, examination items, prescribed medicines, and patient symptoms. After preprocessing of unformatted texts, an LDA is developed for each department to classify the text-based medical records into different topics. These topics are combined with other variables, such as weight, medicine bottles, and diseases, as binary variables to establish a logistic regression model. In the regression model, the variables with significant impact on weight deduction are determined.

To achieve the optimization of the deducted rates, further studies are required:

1. Extend beyond the three morbid entities to formalize the medical records to prevent doctors from mistakes in documentation.
2. Extend beyond the Pu-Chi Hospital to all local hospitals to develop a platform to reduce the waste of medicine resource through medicine.
3. The data from Pu-Chi Hospital are not of a sufficient quantity, which also have an impact on accuracy and machine learning. Extending all hospital medical records is a possible way to perform deep learning to improve the prediction modeling.

Author Contributions: Conceptualization, S.-C.S.; methodology development, C.-C.H.; coding, C.-E.T.; data preparation and analysis, R.R.G.; writing—review and editing, C.-E.T. and C.-C.H.; data collection and annotation, Z.-W.G. and L.-K.H.; supervision, L.-K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the joint funding of Puli Christian Hospital and National Chi Nan University under the grant number (109-PuChi-AIR-005) and the Ministry of Science and Technology (MOST 110-2410-H-260-016-MY2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study did not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. *World Health Report*; World Health Organization: Geneva, Switzerland, 2010.
2. Shrank, W.H.; Rogstad, T.L.; Parekh, N. Waste in the US health care system: Estimated costs and potential for savings. *JAMA* **2019**, *322*, 1501–1509. [CrossRef]
3. Pagano, A.; Robinson, K.; Ricketts, C.; Cundy, J.J.; Henderson, L.; Cartwright, W.; Batt, A. MEmpathy levels in Canadian paramedic students: A longitudinal study. *Fac. Staff Publ. Public Saf.* **2019**, *11*, 1492–1498.

4. Sacks, N.; Cabral, H.; Kazis, L.; Jarrett, K.; Vetter, D.; Richmond, R.; Moore, T. A web-based nutrition program reduces health care costs in employees with cardiac risk factors: Before and after cost analysis. *J. Med. Internet Res.* **2009**, *11*, e43. [CrossRef]
5. Miller, K.; Mansingh, G. OptiPres: A distributed mobile agent decision support system for optimal patient drug prescription. *Inf. Syst. Front.* **2017**, *19*, 129–148. [CrossRef]
6. Huang, C.-Y.; Nguyen, P.-A.; Yang, H.-C.; Islam, M.-M.; Liang, C.-W.; Lee, F.-P.; Li, Y.-C.J. A probabilistic model for reducing medication errors: A sensitivity analysis using electronic health records data. *Comput. Methods Programs Biomed.* **2019**, *170*, 31–38. [CrossRef] [PubMed]
7. Sung, S.F.; Hsieh, C.Y.; Hu, Y.H. Two decades of research using Taiwan's National Health Insurance claims data: Bibliometric and text mining analysis on pubmed. *J. Med. Internet Res.* **2020**, *22*, e18457. [CrossRef]
8. Jian-Shen Chen, C.-H.L.; Chen, M.-C.; Wang, A.-P. The study on auditing the expenditure of National Health Insurance. *Chaoyang Bus. Manag. Rev.* **2006**, *5*, 111–130.
9. Maass, M.C.; Asikainen, P.; Maenpa, T.; Wanne, O.; Suominen, T. Usefulness of a Regional Health Care Information System in primary care: A case study. *Comput. Methods Programs Biomed.* **2008**, *91*, 175–181. [CrossRef] [PubMed]
10. Tsair-Wei Chien, W.-C.W.; Chen, N.-S.; Lin, H.-J. Prediction and management of medical fees when patients receiving cares in hosiptals under DRGs. *J. Taiwan Assoc. Med. Inform.* **2007**, *16*, 13–24.
11. Chen, W.-Y. Influence of Health Insurance Subtract on Surgery Outpatient Medical Practices—An Example of a Regional Teaching Hospital in Taiwan. Master's Thesis, National Yunlin University of Science and Technology, Douliu, Taiwan, 2015.
12. Huang, Y.-W. Research on the Correlation Causes of Health Insurance Review and Verification and the Success of Claiming. Master's Thesis, National Yunlin University of Science and Technology, Douliu, Taiwan, 2018.
13. Cheng, Y.-T. *The Study of between Outpatient Health Insurance Claim and Medical Record—A Regional Hospital in Taiwan*; I-SHOU University: Kaohsiung, Taiwan, 2010.
14. Zhu, C.; Idemudia, C.U.; Feng, W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Inform. Med. Unlocked.* **2019**, *17*, 100179. [CrossRef]
15. Ng, S.-K.; Tawiah, R.; McLachlan, G.J. Unsupervised pattern recognition of mixed data structures with numerical and categorical features using a mixture regression modelling framework. *Pattern Recognit.* **2019**, *88*, 261–271. [CrossRef]
16. Chan, K.S.; Jiao, F.; Mikulski, M.A.; Gerke, A.; Guo, J.; Newell, J.D., Jr.; Hoffman, E.A.; Thompson, B.; Lee, C.H.; Fuortes, L.J. Novel logistic regression model of chest CT attenuation coefficient distributions for the automated detection of abnormal (Emphysema or ILD) versus normal lung. *Acad. Radiol.* **2016**, *23*, 304–314. [CrossRef] [PubMed]
17. Sackett, D.L.; Rosenberg, W.; Gray, J.; Haynes, R.B.; Richardson, W.S. Evidence based medicine: What it is and what it isn't. *Clin. Orthop. Relat. Res.* **1996**, *455*, 3. [CrossRef]
18. Haynes, R.B.; Sackett, D.L.; Richardson, W.S.; Rosenberg, W.; Langley, G.R. Evidence-based medicine: How to practice & teach EBM. *Can. Med. Assoc. J.* **1997**, *157*, 788.
19. Pollio, D.E. The art of evidence-based practice. *Res. Soc. Work Pract.* **2006**, *16*, 224–232. [CrossRef]
20. Reynolds, S. The anatomy of evidence-based practice: Principles and methods. In *Evidence-Based Practice: A Critical Appraisal*; Blackwell Science Ltd.: Hoboken, NJ, USA, 2000; pp. 17–34.
21. Haynes, R.B. Of studies, syntheses, synopses, summaries, and systems: The "5S" evolution of information services for evidence-based healthcare decisions. *BMJ Evid.-Based Med.* **2006**, *11*, 162–164. [CrossRef]
22. O'Leary, D.E. Embedding AI and crowdsourcing in the big data lake. *IEEE Intell. Syst.* **2014**, *29*, 70–73. [CrossRef]
23. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
24. Zafari, B.; Ekin, T. Topic modelling for medical prescription fraud and abuse detection. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2019**, *68*, 751–769. [CrossRef]
25. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B* **1958**, *20*, 215–242. [CrossRef]
26. Yang, X.; Parton, J.; Lewis, D.; Yang, N.; Hudnall, M. Effect of patient-physician relationship on withholding information behavior: Analysis of health information national trends survey (2011–2018) data. *J. Med. Internet Res.* **2020**, *22*, e16713. [CrossRef] [PubMed]
27. Pfirrmann, D.; Huber, Y.; Schattenberg, J.M.; Simon, P. Web-based exercise as an effective complementary treatment for patients with nonalcoholic fatty liver disease: Intervention study. *J. Med. Internet Res.* **2019**, *21*, e11250. [CrossRef]
28. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
29. Hastie, T.; Qian, J. Glmnet Vignette. Available online: http://www.web.stanford.edu/hastie/Papers/GlmnetVignette.pdf (accessed on 13 September 2014).
30. Sasaki, Y.; Fellow, R. *The Truth of the F-measure, Manchester: MIB-School of Computer Science*; University of Manchester: Manchester, UK, 2007.
31. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]