The method of automatic assignment ICD codes based on semantic information

M. ROMALDOWSKI

marcin.romaldowski@wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Kaliskiego 2, 00-908 Warsaw, Poland

The paper presents the method of automatic assignment of ICD codes based on semantic information contained in clinical reports of the MIMIC-III database. It is showing the possibility of using multi-criteria optimization methods for simple classifiers fusion in a more precise classifiers complex. ICD code assignment is important in the modern hospital, more accurate automation of assigning codes will make the clinical process more efficient and can help clinicians carry out better diagnostics and effectively improve medical care systems.

Keywords: MIMIC III, ICD, TFIDF, word2vec, classification, machine learning, natural language processing, classifiers synthesis, Pareto filter.

DOI: 10.5604/01.3001.0014.4435

1. Introduction

For last year's we are watching very fast growth of medical data in hospitals. EHR (Electronic Health Record) data contain various clinical information about the patient, such as medical history, vital signs, laboratory test results, clinical notes, etc., they create a continuous flow of information between the doctor's and the patients. A large part of the health card data is recorded with unstructured text, e.g. clinical observations include the patient's medical history, comments on the doctor's interaction with patients.

International Classification of Diseases (ICD) is a healthcare classification system developed by the World Health Organization. It provides a hierarchy of diagnostic codes of diseases, disorders, injuries, signs, symptoms, etc. Assigning the code is important at many levels in a modern hospital, from providing the exact invoice process to creating a patient care history. However, the coding process is tedious, subjective and requires specialist knowledge. Clinical coders need to extract key information from EHR and assign correct codes based on category, anatomic site, laterality, severity and etiology [8–12].

The article proposes the method that can automatically performs ICD coding given the clinical notes of patients. The method is based on use simple classifiers built on two document representation techniques, the computed similarity score of these classifiers will be synthesized to get a more accurate assignment of ICD codes.

2. MIMIC-III and ICD-9 codes

MIMIC-III is a large, publicly available database containing health-related data, which contains approximately 58,000 hospital admissions of 47,000 patients who stayed in the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001-2012. The database contains information such as: demographic data, measurements of vital signs at the bedside, results of laboratory tests, procedures, medications, doctor and nurse notes, procedure and diagnostic codes (ICD), imaging reports and mortality outside the hospital [13].

ICD-9 – the Ninth Revision of the ICD is a system of about 15,000 numeric codes representing diagnoses and procedures. These codes are used by health care institutions to facilitate and organize the performance of procedures, diagnostic and treatment procedures. The codes consist of 3–4 characters, of which the first two are headlines code groups, and the third and fourth are a clarification. In Poland, coding procedures are used to determine the cost of treatments or operations. The codes are included as attachments to the invoices issued by the hospital and based on the National Health Fund decides to grant a refund for a given treatment [14].

3. Preliminary analysis and data cleaning

The purpose of this article is to present the method of automatic assignment of ICD-9 codes

based on semantic information contained in clinical reports of patients. The dataset used for this study is MIMIC-III. Six main tables from the MIMIC III dataset can be distinguished [13, 28]:

1. ADMISSIONS – contain all information regarding a patient admission, including a preliminary diagnose.
2. LABEVENTS – contains all laboratory measurements.
3. MICROBIOLOGYEVENTS – contains microbiology information such as whether an organism tested negative or positive in the culture.
4. CHARTEVENTS – contains all charted data including patients' routine vital signs and other information related to their health.
5. DIAGNOSES_ICD – contains information about the ICD codes assigned to the patient in the hospital.
6. NOTEEVENTS – contains all notes including nursing and physician notes, echocardiography reports, and discharge summaries.

Each EHR has a clinical note called discharge summary, which contains multiple sections of information, such as 'discharge diagnosis', 'past medical history', 'family history', 'allergies', 'admission exam', 'history of present illnesses'. For these semi-structured text data, we will construct machine learning models to analyze semantic similarities between diagnosis descriptions and ICD code descriptions.

Data analysis showed that the TOP-10 and TOP-50 ICD codes assigned to clinical notes cover over 76,9% and 93,6% of data available in the MIMIC-III database [9]. For the purposes of this paper we will consider codes from TOP-10 and those patients to whom those codes are assigned.

The pre-processing step aims to provide a clean and standardized input for the machine learning model. The following techniques related to text preprocessing can be specified [7], [16]:

1. Remove all irrelevant characters such as any non-alphanumeric characters.
2. Tokenize text by separating it into individual words.
3. Remove words that are not relevant, such as "@", URLs, numbers.
4. Convert all characters to lowercase.
5. Stemming – is a process of reduction words to their word stem, base of root form.
6. Lemmatization – is to reduce inflectional forms to a common base form.

Tab. 1. Top-10 ICD codes in MIMICIII

ICD code	Description	Admissions
4019	Hypertension	20046
4280	Congestive heart failure	12842
42731	Atrial fibrillation	12589
41401	Coronary atherosclerosis	12178
5849	Acute kidney failure	8906
25000	Diabetes Type II	8783
2724	Hyperlipidemia	8503
51881	Acute respiratory failure	7249
5990	Urinary tract infection	6442
53081	Esophageal reflux	6154

4. Clinical reports representation

Here are some notations what will be used throughout the paper:

$V \in \mathbb{N}$ – number of words in vocabulary;
$M \in \mathbb{N}$ – number of patient's documentation;
$L \in \mathbb{N}$ – number of ICD codes;
$N_i \in \mathbb{N}$ – number of words in note;
$i \in \{1,...,M\}$ – index of patient notes;
$j \in \{1,...,V\}$ – index of words;
$l \in \{1,...,L\}$ – index of ICD code;
$X^{(t)} \in \mathbb{R}^{MxV}$ – vector space model for patients' documentations represented by TF-IDF;
$X^{(w)} \in \mathbb{R}^{MxV}$ – vector space model for patient documentations represented by word2vec;
$x \in \mathbb{R}^{Vx1}$ – one-hot encoded vector where $x_j = 1$ if word t_j does appear in the clinical note and then all other words take the form $x_{j'} = 0$, $j \neq j'$.

Creating a text vector representation means transforming a set of tokens $T^{(V)} = \{t_1,...,t_j,...,t_V\}$ into a vocabulary in which numbers represent words [8–12], [18]. There is a possibility to obtain different variations of representation vector-space for documents $d_i = \{w_{i1},...,w_{ij}\}^T$. The popular used in practice include: BoW and its term-frequency based variants [18], [22], language model-based methods [16], [23], topic models [24] and distributed vector representations [25–28]. For the purpose of the article, TF-IDF and word2vec methods will be used for feature extraction of patient's notes $D = \{d_1,...,d_i,...,d_M\}$.

$$T^{(V)} = \begin{cases} t_1 = Sore \\ t_2 = Throat \\ t_3 = Wheezing \\ t_4 = Pain \\ t_5 = Headaches \\ t_6 = Dizziness \\ t_7 = Cough \\ t_8 = Chest\ pain \end{cases}$$

Fig. 1. Sample text vector

TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [18]. The TF–IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general [18]. The TF-IDF values are calculated from the formula:

$$w_{ij}^{(t)} = (tf - idf)_{ij} = tf_{ij} \times idf_j \tag{1}$$

Where tf_{ij} is called "Term frequency", expressed by the formula [18]:

$$tf_{ij} = \frac{n_{ij}}{N_i} \tag{2}$$

Where n_{ij} is the number of occurrences of the word t_j in the document d_i, and the denominator is the sum of the occurrences of all words in the document d_i [18]. idf_j this "inverse document frequency", expressed by the formula:

$$idf_j = \frac{|D|}{|\{d : t_j \in d\}|} \tag{3}$$

$|D|$ – number of documents in the corpus,

$|\{d : t_j \in d\}|$ – the number of documents containing at least one occurrence of a given term.

$$X = \begin{pmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{pmatrix},\ X \in \mathbb{R}^{M \times V}$$

Fig. 2. Sample vector space model

Word2vec is a group of related models that are used to produce word embedding. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words [30]. Word2vec can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of--words (CBOW) or continuous Skip-Gram [30]. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. In the continuous Skip-Gram architecture, the model uses the current word to predict the surrounding window of context words [30]. For the purpose of the article, the Skip-Gram model will be used [21].

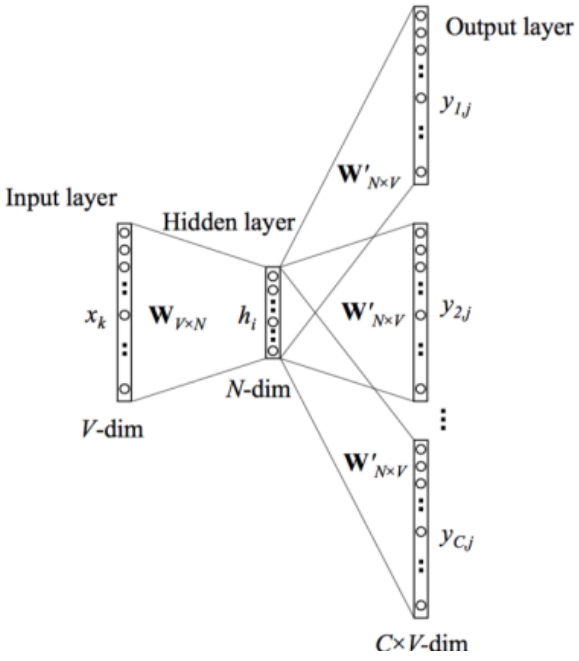

Fig. 1. Skip-Gram model [21]

The input layer is represented by a one-hot encoded vector x of dimension V. The hidden layer, h, is defined by a vector of dimension N. The output layer is a vector of dimension V. The weights between the input and the hidden layer are represented by a matrix W, of dimension $V \times N$.

The hidden layer h is calculated using the following formula:

$$h = W^T x_j := v_{t_j} \tag{4}$$

Where v_{t_j} is the vector representation of the input word t_j. Similarly, the weights between the hidden and the output layer are represented by a matrix *W′*, of dimension *N* × *V*. Using these weights, we can compute a score u_j for each word in the vocabulary:

$$u_j = v'^T_{t_j} \cdot h \tag{5}$$

Where $v'^T_{t_j}$ is output vector of the the *j-th* word t_j of the matrix W′. Then we can use a log-linear classification model, to obtain the posterior distribution of words:

$$y_{t_j} = \frac{\exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j'})} \tag{6}$$

After training we will have word embedding $U = \{y_{t_1}, ..., y_{t_j}, ..., y_{t_V}\}$ and these embedding will be used as feature input to machine learning models. Each patient documentation will be represented as an average of the embedding of the words in the document:

$$d_i^{(w)} := \frac{1}{N_i} \sum_{j \in d_i} y_{t_j} \tag{7}$$

Both (TF-IDF, word2vec) serve different purposes in natural language processing. Word2vec helps in going deeper into the document, measure syntactic and semantic similarities between sentences, helps to derive relations between a word and its contextual words. Whereas TF-IDF helps in visualizing important words in document and topic modelling by using the importance score of words.

5. Classification

In this type of task, the computer program is asked to specify which of l categories some input belongs to. To solve this task, the learning algorithm is usually asked to produce a function [29]:

$$C : D \to L \tag{8}$$

The following concepts can be distinguished: supervised classification and unsupervised classification. In unsupervised learning, data points have no labels associated with them. Instead, the goal of an unsupervised learning algorithm is to organize the data in some way or to describe its structure. This can mean grouping it into clusters or finding different ways of looking at complex data so that it appears simpler or more organized [29]. Supervised learning algorithms make predictions based on a set of examples. For instance, historical data about patients and ICD codes assigned to their cards can be used to predict ICD codes for new patients. Each example used for training is labeled with the value of interest – in this case the ICD codes. A supervised learning algorithm looks for patterns in those value labels. May use all information that may be relevant and included in clinical notes and each algorithm looks for different types of patterns. After the algorithm has found the best pattern it can, it uses that pattern to make predictions for unlabeled testing data [29].

Patients notes $D = \{d_1, ..., d_i, ..., d_M\}$ may have multiple ICD-9 codes assigned $l \in L$, so this is a classification task with multiple labels. The classification of many labels in this case has an additional degree of difficulty, because the number of correct labels for each patient is unknown [2], [9–12]. In the described method, logistic regression model will be used as classifiers. The corpus will be defined $X' \subset X^{(t)}$ or $X' \subset X^{(w)}$ as a set of observed patients notes that should be used to train and test the classifier. A separate classifier will be trained for each ICD code and for each feature vectors, they predict independent values from the range 0-1 (if the higher value then the higher probability of assigning the ICD code). Logistic regression works on the basis of a function called a logistic function or more often called a sigmoid. This function is responsible for predicting or classifying input data. The function is defined as [5]:

$$z = \beta^T X' \tag{9}$$

$$sigmoid(z) = \frac{1}{1 + e^{-(z)}} \tag{10}$$

Weights (represented by β in our record) are an important part of logistic regression algorithms and other machine learning algorithms, we should find the best values for them. At the beginning, we will choose random values and we need measure how well the algorithm uses these random weights. This measure is calculated using the loss function.

$$h = sigmoid(X'\beta) \tag{11}$$

$$J(\beta) = \frac{1}{m} \cdot [-l^T \log(h) - (1-l)^T \log(1-h)] \quad (12)$$

were:

m – number of samples in X',

l – target ICD code.

The goal is to minimize losses by increasing or decreasing the weight, which is commonly called fitting. Which scales should be larger, and which should be smaller, this can be determined using gradient methods. Gradient is a derivative of the loss function in relation to its weight.

$$\frac{\delta J(\beta)}{\delta \beta} = \frac{1}{m} \cdot X'^T (h-l) \quad (13)$$

The weights are updated as below:

$$\beta = \beta - \alpha \cdot \frac{\delta J(\beta)}{\delta \beta} \quad (14)$$

Where: α – it is usually 0.1.

6. Classifier synthesis

The output values from classifiers can be interpreted as the similarity of patient documentation $d_i \in D$ to the ICD code $l \in L$ and saved in the following form [3]:

$C_1^l(d_i) = f_{d_i}^1(l)$ – similarity rate for classifiers based on TFIDF.

$C_2^l(d_i) = f_{d_i}^2(l)$ – similarity rate for classifiers based on word2vec.

Tab. 2. Sample similarity ranking

L	1	2	3	4	5	6	7	8	9	10
f_d^1	0.31	0.42	0.51	0.62	0.62	0.54	0.54	0.31	0.22	0.51
f_d^2	0.62	0.73	0.62	0.42	0.33	0.23	0.15	0.11	0.11	0.51

Patients clinical notes may have multiple ICD codes assigned. As we can see from the example in table 2, different observations are identified differently by classification functions, it means that each classifier has a different information potential and there is variation between classifiers. This feature is useful in combining classifiers [6] [14]. An interesting proposal for solving this type of problems is offered by the multi-criteria optimization theory [1–4]. By creating an appropriate R synthesis model, we can define a task in the form (Y_{d_i}, R). Set Y_{d_i} will be the ranking image of the set L for observation $d_i \in D$, given by function f_{d_i}.

$$Y_{d_i} = f_{d_i}(L) = \{y = f_{d_i}(l) \in R^2\} \quad (15)$$

Element $y \in f_{d_i}(L)$ is the image of the label l in the sense of its evaluation by all functions $f_{d_i}^n(l)$ understood as the level of similarity of observations $d_i \in D$ to the ICD code $l \in L$.

The synthesis relation shall be the following relation:

$$R \subset f_{d_i}(L) \times f_{d_i}(L) = Y_{d_i} \times Y_{d_i} \quad (16)$$

Defined as follows:

$$R = \begin{Bmatrix} (y,z) \in Y_{d_i} \times Y_{d_i} \mid \\ \text{"committe prefers } y \text{ then } z\text{"} \end{Bmatrix} \quad (17)$$

The solution of this task will be the Pareto set, i.e. the set of these ICDs from the pre--estimate set from which there are no more probable, this set will be marked with the symbol [2–4]:

$$Y_{d_i}^{RN} = \begin{Bmatrix} y \in Y_{d_i} \mid \text{does not exist } z \in Y_{d_i} \\ z \in Y_{d_i} - \{y\}, \text{such } y \le z \end{Bmatrix} \quad (18)$$

The L_d^{RN} set is a counter image of the Pareto Y_d^{RN} set, the most probable ICD codes, based on the examples from Table 2, are:

$$L_{d_i}^{RN} = f_{d_i}^{-1}(Y_{d_i}^{RN}) = \left\{ l \in L \mid f_{d_i}(l) \in Y_{d_i}^{RN} \right\} = \\ = \{l_2, l_3, l_4\} \quad (19)$$

By calculating the distance of images of these ICD codes from $y^* = (y_1^*, y_2^*)$ so-called ideal point, we can create a ranking of codes for further classification [3]. The coordinates of point y^* should be determined as follows:

$$y_1^* = \max f_{d_i}^1(l), y_2^* = \max f_{d_i}^2(l) \quad (20)$$

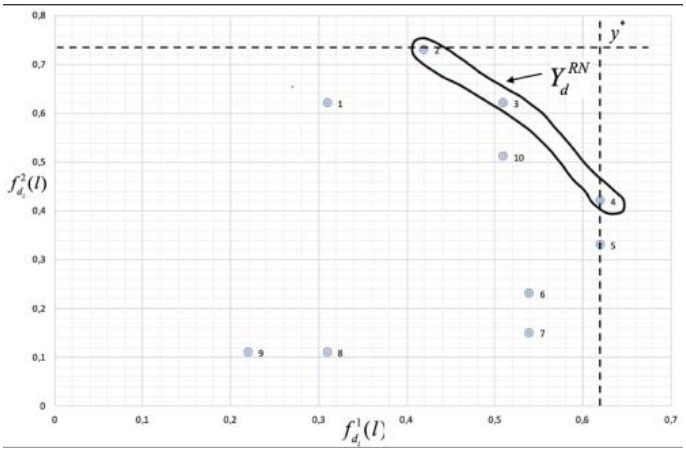

Fig. 2. Front Pareto

7. Conclusion

The paper presents the method of automatic assignment of ICD codes based on semantic

information contained in clinical reports of the MIMIC-III database. The method uses classifiers based on the logistic regression model and two different extraction feature method. In clause 6 has been show the possibility of using multi--criteria optimization methods for sample classifiers fusion in a more precise classifiers complex. My future research will focus on studying a wider class of classifiers and then trying to merge them, thanks so that I will try to get even more accurate attribution of ICD codes for clinical notes. Even more accurate automation of assigning ICD codes will make the clinical process more efficient and can help clinicians carry out better diagnostics and effectively improve medical care systems.

8. Bibliography

[1] Ameljańczyk A., *Optymalizacja wielokryterialna*, WAT, Warszawa 1986.

[2] Ameljańczyk A., "Properties of the Algorithm for Determining an Initial Medical Diagnosis Based on a Two-Criteria Similarity Model", *Computer Science and Mathematical Modeling*, No. 8, 9–16 (2011).

[3] Ameljańczyk A., "Property analysis of multi-label classifiers in the example of determining the initial medical diagnosis", *Computer Science and Mathematical Modeling*, No. 1, 11–16 (2015).

[4] Ameljańczyk A., "Pareto filter in the process of multi-label classifier synthesis in medical diagnostics support algorithms", *Computer Science and Mathematical Modeling*, No. 1, 5–10 (2015).

[5] Krzyśko M., Wołyński W., Górecki T., Skorzybut M., *Systemy uczące się*, WNT, 2008.

[6] Kuncheva L. I., *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Inc. 2004.

[7] Mirończuk M., "Przegląd metod i technik eksploracji danych tekstowych", *Studia i Materiały Informatyki Stosowanej*, Tom 4, Nr 6, 25–42 (2012).

[8] Farkas R., Szarvas G., "Automatic construction of rule-based ICD-9-CM coding systems", *BMC Bioinformatics*, Luty 2008.

[9] Huang J., Osorio C., Wicent Sy L., "An Empirical Evaluation of Deep Learning for ICD-9 Code Assignment using MIMIC-III Clinical Notes", 7 lutego 2018, https://arxiv.org/abs/1802.02311.

[10] Nigam P., "Applying Deep Learning to ICD-9 Multi-label Classification from Medical Records", Stanford University, https://cs224d.stanford.edu/reports/priyanka.pdf.

[11] Xie P., Shi H., Zhang M., Xing E. P., "A Neural Architecture for Automated ICD Coding", http://aclweb.org/anthology/P18-1098.

[12] Li M., Fei Z., Zeng M., Wu F., Li Y., Pan Y., Wang J., "Automated ICD-9 Coding via A Deep Learning Approach", 20 marca 2018, IEEE, https://ieeexplore.ieee.org/document/8320340.

[13] Johnson A.E.W., Pollard T.J., Shen L., Lehman L., Feng M., Ghassemi M., Moody B., Szolovits P., Celi L.A., Mark R.G., "MIMIC-III, a freely accessible critical care database", *Scientific Data*, 3:160035 (2016).

[14] https://en.wikipedia.org/wiki/International_Statistical_Classification_of_Diseases_and_Related_Health_Problems.

[15] Ćwiklińska Jurkowska M., "Klasyfikatory pojedyncze i zintegrowane jako narzędzie wspomagania medycyny", *StatSoft*, 31–46 (2013).

[16] Mikolov T., Chen K., Corrado G., Dean J., *Efficient estimation of word representations in vector space*, CoRR, abs/1301.3781, 2013.

[17] https://www.nlm.nih.gov/research/umls/.

[18] https://pl.wikipedia.org/wiki/TFIDF.

[19] https://en.wikipedia.org/wiki/Natural_language_processing.

[20] https://en.wikipedia.org/wiki/Vector_space_model.

[21] Xin Rong, "word2vec Parameter Learning Explained" https://arxiv.org/pdf/1411.2738.pdf

[22] Salton G. and Buckley C., "Term--weighting approaches in automatic text retrieval", *Information Processing & Management*, 24(5), 513–523 (1988).

[23] Mikolov T. and Dean J., "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems* (NIPS 2013).

[24] Huang E.H., Socher R., Manning C.D., Ng A.Y., "Improving word representations via global context and multiple word prototypes", in: Proceedings of ACL, 2012.

[25] Le Q.V, Mikolov T., "Distributed representations of sentences and documents", in: Proceedings of ICML, 2014.

[26] Kim Y., Jernite Y., Sontag D., Rush A.M., "Character-aware neural language models", arXiv preprint arXiv:1508.06615, December 2015.
[27] Blei D.M., Ng A.Y., Jordan M.I., "Latent dirichlet allocation", *Journal of Machine Learning Research*, 3, 993–1022 (2003).
[28] Xu K., Lam M., Pang J., Gao X., Band Ch., Mathur P., Papay F., Khanna A.K., Cywinski J.B., Maheshwari K., Xie P., Xing E., "Multimodal Machine Learning for Automated ICD Coding", CoRR abs/1810.13348 (2018).
[29] Goodfellow I., Bengio Y., Courville A., *Deep Learning*, MIT Press 2016.
[30] https://en.wikipedia.org/wiki/Word2vec.

Metoda automatycznego przypisywania kodów ICD na podstawie informacji semantycznych

M. ROMALDOWSKI

W artykule przedstawiono metodę automatycznego przypisywania kodów ICD-9 na podstawie informacji semantycznych zawartych w raportach klinicznych pacjentów bazy MIMIC-III. Została pokazana możliwość wykorzystania metod optymalizacji wielokryterialnej do budowy fuzji klasyfikatorów w celu utworzenia bardziej precyzyjnych klasyfikatorów. Przypisanie kodu ICD jest ważne na wielu poziomach w nowoczesnym szpitalu, dokładniejsza automatyzacja przypisywania kodów sprawi, że proces kliniczny stanie się bardziej wydajny i może pomóc klinicystom w przeprowadzeniu lepszej diagnostyki i skutecznej poprawie systemów opieki medycznej.

Słowa kluczowe: klasyfikatory, TFIDF, word2vec, kody ICD, MIMIC III, fuzja, synteza, filtr Pareto, przetwarzanie języka naturalnego.