Fall 9-11-2020

Data Mining Pipeline for Performing Decision Tree Analysis On Mortality Dataset With ICD-10 Codes

Prapti Burkul
University of North Florida, N01440630@unf.edu

Karthikeyan Umapathy
University of North Florida, k.umapathy@unf.edu

Asai Asaithambi
University of North Florida, asai.asaithambi@unf.edu

Haiyan Huang
Flagler College, hhuang@flagler.edu

DATA MINING PIPELINE FOR PERFORMING DECISION TREE ANALYSIS ON MORTALITY DATASET WITH ICD-10 CODES

Prapti Burkul
University of North Florida
n01440630@unf.edu

Karthikeyan Umapathy
University of North Florida
k.umapathy@unf.edu

Asai Asaithambi
University of North Florida
asai.asaithambi@unf.edu

Haiyan Huang
Flagler College
HHuang@flagler.edu

ABSTRACT

Modernization of the healthcare sector has led to the introduction of wider and newer varieties of medical devices in hospitals. Consequently, there are increasing numbers of infectious complications related to medical devices. However, managing and monitoring the risk of medical devices are difficult and costly. The hospitals and the healthcare device service providers require effective means to manage the healthcare device maintenance to provide better patient care. To address this issue, we propose a data mining pipeline to classify medical devices based on mortality rates and ICD-10 codes. We utilize the decision tree grouping method to build a connection between the mortality dataset and ICD-10 codes. We anticipate that the results of this study will assist with healthcare providers identify risks associated with medical devices based on how many deaths are caused due to the improper use or use of faulty medical instruments during the treatment.

KEYWORDS

Decision trees, ICD-10 codes, medical devices, mortality rates, data mining pipeline

INTRODUCTION

Modernization of healthcare systems involves the employment of diagnostic classification standards for reporting diseases and health conditions for clinical and research purposes. The medical classification standards assist with data management tasks such as patient care, record storage and retrieval, statistical analysis, insurance, and billing. One of the widely used standards is the International Classification of Diseases (ICD) (WHO-ICD 2019), which is the official system of assigning codes to diagnoses and procedures associated with hospital utilization (Escorpizo et al. 2013). The World Health Organization (WHO) maintains and updates the ICD standard with the advancement of medical technologies and practices, and version 10 is currently in use. After a patient's clinical treatment, ICD-10 codes are assigned to that patient's health record by general health care personnel or by specially trained medical coders (Subotin and Davis 2016). In addition to the information recorded such as diagnosis, external cause of injury, and procedure codes; also coded in ICD-10 data are medical devices used for patient treatment and diagnosis.

With a consistently growing number of newer and a broader variety of inserted devices during patient care and treatments, there is a more substantial likelihood of escalation for the occurrence of infectious complications related to medical devices (Kojic and Darouiche 2004). The ramifications of medical device-related infections can be traumatic and often create complicated health issues for patients. Occurrence of adverse incidents during patient care in hospital setting can lead to morbidity and increased health care costs (Hougland et al. 2008). Hospitals need to monitor diseases, and other device-associated risks and remove malfunctioning devices before they cause potentially life-threatening systemic infections. Given the inherent susceptibility of medical devices, hospitals are looking for effective ways of identifying and managing devices that are associated with infections and potentially cause mortality. To reduce the occurrence of adverse incidents such as infections and death, hospitals and device service providers would need to identify specific risk factors. Based on the associated factors, hospitals and service providers can develop appropriate interventions to prevent adverse incident occurrences.

Given the growth of data management tasks within the healthcare sector, it is interesting to explore how data mining can help to achieve the goal of providing better services to patients in need. Analyzing mortality data is vital to understand the complex circumstances of death across the country. Since healthcare systems follow the ICD-10

standard, classifying patient deaths based on the causes, devices used for treatment, and early diagnosis of major illness has become a possibility. The healthcare instrument providers need practical approaches to monitor and provide proper diagnostics for the instruments. Such monitoring of devices will ultimately reduce the deaths associated with using defective healthcare instruments in the treatment of the patients. However, we do not have a classification of medical devices based on the risks of mortality or factors that are associated with adverse incidents caused due to usage of medical devices during treatment. Previous studies that examined ICD-9 coded datasets for detecting inpatient complications produced mixed results and focused on specific population and patient characteristics (Hougland et al. 2008). Thus, we need to conduct a design science research to build a data science solution to address the problem. As a first step, in this research-in-progress paper, we present a data mining pipeline to address the research problem of classifying medical devices based on the risk of adverse incidents. The mortality dataset released by the Center for Disease Control and Prevention (CDC), along with ICD-10 codes, will be used as the training data. We will analyze the number of deaths that occurred due to a defective healthcare instrument/device used in the treatment of the diseases. As a research outcome, we intend to develop a decision tree that can be used by hospitals and device service providers. The decision tree would provide an assessment of a given device's risk and help with identifying associated factors.

BACKGROUND

An ICD-10 code indicates a classification of a disease, disorder, injury, and other health conditions (WHO-ICD 2019). The initial purpose of the ICD classification system was to monitor mortality causes. The historical backdrop of ICD goes back to 1763, when French doctor Francois Bossier de Lacroix, looking to help his kindred doctors distributed a grouping framework consisting of 10 significant classes of maladies and 2400 individual diseases (Sundararajan et al. 2004). The WHO turned into the caretaker of ICD in 1948 and extended it to incorporate additional ICD coding through several revisions (WHO-ICD 2019). The current ICD standard in effect is version 10. ICD-10 data is used by physicians, nurses, hospital administrators, insurance companies, researchers, and policymakers. The Department of Health and Human Services distributed a guideline on January 16, 2009, that required the supplanting of ICD-9 with ICD-10 starting on 1 October 2013. The ICD-10 codes are divided based on the various diagnosis and procedures. While ICD-9 and ICD-10 have a similar chain of command in their structures, ICD-10 is progressively intricate and fuses various changes. The migration of the coding system to ICD-10 allows healthcare providers to categorize diseases and track healthcare and medical complications effectively. The migration has increased the number of codes from 17,000 to 141,000, which eventually has helped the healthcare industry to capture more granular information (CDC-IDC 2015). The characteristics for the ICD-10 codes are the listing of the titles of the cause of death and related codes; the cause of death titles and the corresponding inclusion and exclusion terms; alphabetic index to diseases and nature of injury; injury and their relevant external causes; table containing information about the drugs and chemicals; and description, guidelines, and coding rules. The National Vital Statistics System overseen by CDC maintains nationwide mortality data containing ICD codes of medical information on death certificates issued in the United States (NVSS - Mortality 2019). This mortality data source is the primary dataset for this research.

RELATED WORK

ICD-10 codes and their fundamental value of provider-based and point-of-care coded assessment of diseases are essential as they provide valuable insights to patient care (Weiner 2018). There is significant amount of research reported in the field of healthcare addressing mortality issues using the ICD-10 coded datasets and data mining approaches. Koopman et al. (2015a) developed a two-level hierarchy Supply Vector Machine (SVM) model for identifying cancer-related causes of mortality from an Australian death certificate dataset. The first level of the SVM model was a binary classifier for identifying the presence of cancer, and the second level was classifier for identifying the type of cancer using the ICD-10 classification system. Researchers measured performances of the models using the F1 score, which is a weighted average of precision (fraction of relevant items retrieved) and recall (ratio of the number of relevant items and total relevant items). The macro-averaged F1 score is calculated by computing precision and recall values independently and then taking average. The SVM models had a macro-averaged F1-score of 0.94; thus, they were considered highly effective at identifying cancer as the underlying cause of death. However, the model had a low F1-score of 0.12 when classifying rare cancers and ambiguous cases like cancer in the stomach region due to limited training data for such cases. In another study, Koopman et al. (2015b) conducted a comparative evaluation of a rule-based classification method against to SVM model to classify death certificates of high impact diseases such as diabetes, influenza, pneumonia, and HIV. Both classification models were trained against 340,000 death certificates issued in Australia. Rule-based classified had a macro-averaged F1-score of 0.95 and 0.94 for the SVM classification method. Mujtaba et al. (2017) applied random forest and J48 decision tree classification models against 2,200 autopsy reports obtained from a hospital in Kuala Lumpur, Malaysia. The study aimed to assist pathologists in determining

ICD-codes accurately based on the causes of death from autopsy findings. Researchers adopted expert-driven feature selection to build their decision tree classifiers which yielded 85% to 90% accuracy for feature subset size of 30. Lavergne et al. (2016) described a large-scale dataset prepared for training machine learning models for ICD-10 coding. The dataset was developed using 93,000 French death certificates referring to 3450 ICD-10 codes. Zweigenbaum and Lavergne (2016) developed a hybrid method combining dictionary linking with SVM classifier for ICD-10 coding of death certificates using (Lavergne et al. 2016) dataset. The results indicate that the hybrid model received a macro-averaged F1-score of 0.85. Jay et al. (2013) produced morbidity profiles of breast cancer patients reported in a French national casemix system. The dataset contains ICD-10 codes for main diagnoses for hospitalization and mortality. Researchers performed Formal Concept Analysis to identify the cluster of trajectories of care for breast cancer patients and develop the morbidity profiles. Sundararajan et al. (2004) mapped the Deyo algorithm developed for ICD-9 to be used with ICD-10 codes for predicting in-hospital mortality. In comparison to the performance of the Deyo algorithm with ICD-9 codes (0.865), the algorithm modifications for ICD-10 (0.855) had little difference in terms of accuracy.

Therefore, it is evident that several researchers have worked with ICD-10 coded datasets and mortality issues from a variety of diseases and diagnoses. However, we did not come across any literature on addressing mortality caused due to medical devices. The goal of this research proposal is to address this gap, and towards that, we describe a data mining pipeline in this paper.

RESEARCH APPROACH

ICD-10 codes offer specific and increased ability to accommodate new findings and technologies (Cartwright 2013). These codes can help evaluate and improve the quality of patient care if appropriately used (Bowman 2008). In order for hospitals and device service providers to monitor and maintain the medical devices, they need devices classified based on the occurrence of adverse incidents such as infections and mortality. Thus, we intend to apply classification methods to categorize medical devices based on the causes of adverse incidents. Classification is a data-mining technique that organizes related entities into a given number of classes. As the data inside ICD-10 is already categorized, it is appropriate to apply classification techniques like a decision tree to find solutions to the research problem. We address the problem by following the design science research methodology (Hevner et al. 2004). Towards building a data science application solution for the problem, in this section, we describe the data mining pipeline that will be used to apply decision tree classification techniques to address the research problem.

The data mining pipeline consists of the following stages: data scrubbing, data preparation, and classification, as shown in figure 1. These data mining stages will help us achieve the goal of bringing the datasets and classification results in a format that can be fed into a reporting tool to build dashboards and visualization for better analysis. The pipeline will be built using tools and technologies that are open source or do not require any licensing costs. The research activities will be carried using a personal laptop and open source tools like Python and RapidMiner. Python and RapidMiner offer various functions to properly cleanse data and build analysis on top of the cleansed data.

RapidMiner is a user-friendly software tool for building data mining solutions as it provides access to data mining strategies and machine learning algorithms through template-based frameworks (Mierswa and Klinkenberg 2020). These template-based frameworks help to reduce errors (Naik and Samant 2016). One more advantage of using RapidMiner is that it eliminates the need to write code. It represents a standard approach to design even a very complicated issue (Naik and Samant 2016). RapidMiner has versatile operators for information input and output in several file formats. It contains various learning schemes for classification and clustering tasks.

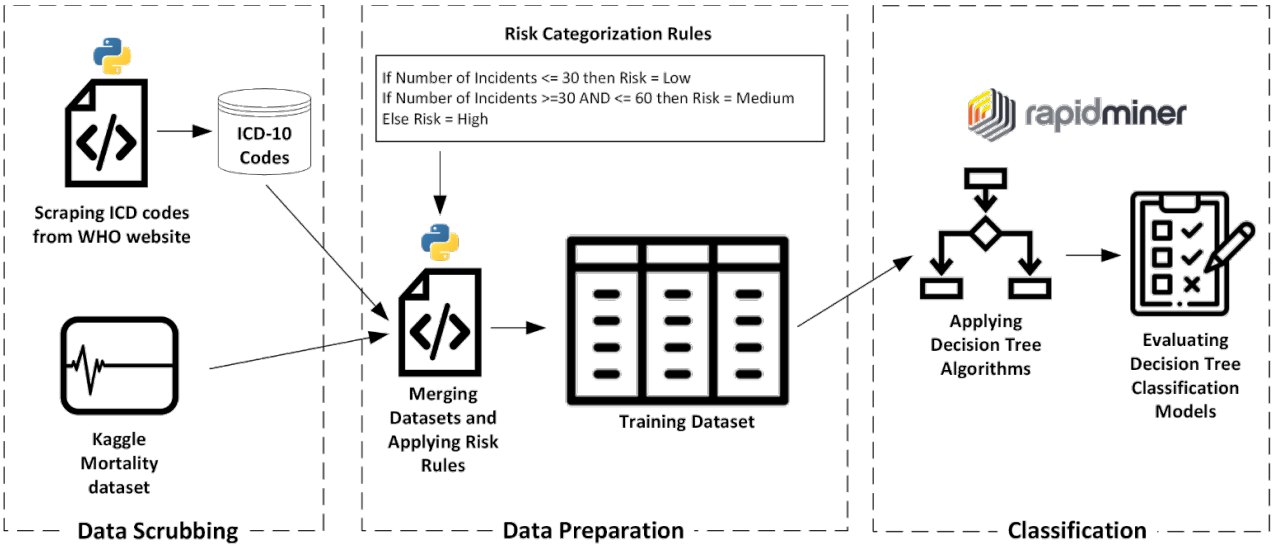

Figure 1. Data Mining Pipeline

Dataset Scrubbing

The data will be collected for various ICD codes related to the device diagnostic and monitoring for various procedures. The ICD-10 code descriptions dataset will be collected from the WHO website by data-scrubbing. Python (2020) programs will be written to scrub the data in the required format. Once the data is extracted, the data can be cleansed using Python or any other open-source tool. The mortality data set, along with ICD-10 codes, are available on the Kaggle website. The data set is available in pkl, CSV, and JSON formats. Kaggle has obtained the mortality data from CDC's National Vital Statistics Systems (Kaggle-Mortality 2017). The mortality dataset is a record of every death in the United States for 2005 through 2015, including detailed information about the causes of death and the demographic background of the deceased. The mortality dataset includes the following attributes: resident status, education, sex, age, place of death, date of death, injury at work, manner of death, method of disposition, autopsy, activity code, place of injury, underlying cause of death, ICD-10 code, and race. Each dataset is a collection of one year of data which is paired with JSON files and an ICD-10 code set. The ICD code description dataset generated by scrubbing the WHO website will be merged with the mortality dataset gathered from Kaggle in preparation for data mining analysis.

Dataset Preparation

The integrated dataset will need some more preparatory work before it can be used as an input for the classification algorithm. As the objective of the research is to classify healthcare devices based on mortality incidents, we need to generate a target class attribute. Diagnostic and monitoring healthcare devices like anesthesiology, cardiovascular, neurological, etc. will help us understand and classify the significant problems in healthcare due to associated device incidents. The classification of these device-related incidents will help us understand and classify the patient's death as high, medium and low risks. In order to prepare the dataset for classification algorithms, the number of adverse incidents that occurred due to a device would be aggregated across the dataset. Prior research work (Hougland et al. 2008) shows that positive predictive value for medical adverse events are in the range of 15% to 77%. Given that existing research provides a broad range, we generate the target class attribute based on following arbitrary categorization rules:

1. If the number of incidences is less than 30, the device will be categorized as low risk.
2. If the number of incidences is between 30 and 60, the device will be categorized as medium risk.
3. If the number of incidences is more than 60, the device it will be categorized as high risk.

Thus, the final dataset will contain the following in addition to existing attributes: ICD-10 code, ICD-10 description, name of the instrument used in the treatment of the patient, disease, treatment, number of the incidents, and associated risk.

Decision Tree Classification

We plan to use the decision tree classification technique to classify medical devices into risk-based categories. The decision tree technique is also known for its simplicity, comprehensibility, and robustness to handle missing values (Singh, Subramania, Holland, and Davis 2012). Decision tree classifiers are supervised learning algorithms. The goal of using decision tree algorithms is to create a model that can predict the values of target variables by learning simple decision rules inferred from the given training dataset. The decision tree algorithm produces a tree-like graph as an output (Tomar and Agarwal 2013). The decision tree is identical to the flowchart consisting of nodes, branches, and leaf nodes (terminal nodes). Each non-leaf node denotes a check on a particular attribute. Every branch denotes a result of that check and each leaf node has a class label (Tomar and Agarwal 2013). Traversal from root to leaf shows distinct class separation based on most statistics gain. The decision tree technique is commonly by many researchers within the health care field (Tomar and Agarwal 2013).

We will use RapidMiner to build decision trees using the integrated dataset. RapidMiner provides operators support for four different decision tree algorithms: Chi-squared Automatic Interaction Detector (CHAID), Interactive Dichotomizer version 3 (ID3), Random Forest, and Bagging (DT-RM 2019). CHAID algorithm builds a decision tree by starting with the entire dataset and repeatedly splitting the subset of the dataset into two or more child nodes (Ture, Tokatli, and Kurt 2009). It uses the chi-square test of association between independent and target variables, to determine the best node to split. The independent variable that has strongest association with the target variable becomes the first branch in the tree, and leaf nodes are created for each categorical value that is significantly different in relation to the target variable. This procedure is repeated until there is no statistically significant difference independent variable. ID3 algorithm builds a decision tree using a top-down greedy search through the input dataset at every tree node to determine the most useful variable for classification (Ture, Tokatli, and Kurt 2009). ID3 calculates information gain value for each variable and selects the variable with the most gain for branching the node. The algorithm performs the process recursively to construct the decision tree. Random Forest method generates trees based on a bootstrapped sample of the original training data and performs a random selection of a subset of variables from the training data to determine split for each node (Gislason, Benediktsson, and Sveinsson 2006). The random variable for the node split is determined by casting votes for the most popular class for the given input. The majority vote of the trees determines the output of the classifier. Bagging is a meta-algorithm that trains many classifiers on bootstrapped samples from the original training dataset and determines the final tree through a voting process (Gislason, Benediktsson, and Sveinsson 2006). Many researchers have shown that decision trees developed using C4.5 have produced better results. C4.5 is similar to the ID3 method but uses the gain ratio metric instead of information gain for determining variables to split nodes (Ture, Tokatli, and Kurt 2009). Since RapidMiner does not have an operator for C4.5, we will use Weka to build a decision tree using the C4.5 method. Decisions trees developed using different classifier algorithms will be evaluated using performance measures such as accuracy, sensitivity, and specificity. Accuracy is the ratio of the number of correct predictions over the total number of predictions. Sensitivity is the proportion of true positives correctly classified, whereas specificity is the proportion of true negatives correctly classified.

CONCLUSION

The purpose of this research-in-progress study is to explore the factors associated with adverse events caused due to medical devices reported in the CDC mortality dataset using decision tree analysis. To achieve the research objective, we have presented a data mining pipeline consisting of data scrubbing, preparation, and classification stages. The pipeline presented can also serve as a baseline for classifying various reasons for mortality rate other than the use of faulty instruments. We anticipate that the study results will serve as a reference for the hospitals and the medical instrument service providers to effectively manage the maintenance of these instruments. As an extension of this study, we intend to utilize the classification of risks associated with the number of adverse incidents to predict the risk of the medical instruments.

REFERENCES

1. Bowman, S. E. (2008). Why ICD-10 is Worth the Trouble. *Journal of AHIMA*, 79(3), 24-29.
2. Cartwright, D. J. (2013). ICD-9-CM to ICD-10-CM Codes: What? Why? how? *Advances in Wound Care*, 2(10), 588-592.
3. CDC-IDC. (2015). International Classification of Diseases, (ICD-10-CM/PCS) Transition - Background, https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm. Retrieved on December 30, 2019.

4. DT-RM. (2019). Decision Tree operators in RapidMiner, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html. Retrieved on January 9, 2020.
5. Escorpizo, R., N. Kostanjsek, C. Kennedy, M. M. R. Nicol, G. Stucki, and T. B. Ustün. (2013). Harmonizing WHO's International Classification of Diseases (ICD) and International Classification of Functioning, Disability and Health (ICF): Importance and Methods to Link Disease and Functioning. *BMC Public Health*, 13(1), 742.
6. Gislason, P. O., J. A. Benediktsson, and J. R. Sveinsson. (2006). Random Forests for Land Cover Classification. *Pattern Recognition Letters*, 27(4), 294-300.
7. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75-105.
8. Hougland, P., Nebeker, J., Pickard, S., Van Tuinen, M., Masheter, C., Elder, S., Williams, S., and Xu, W. (2008). Using ICD-9-CM codes in hospital claims data to detect adverse events in patient safety surveillance. In *Advances in patient safety: new directions and alternative approaches* (Vol. 1: Assessment). Agency for Healthcare Research and Quality.
9. Jay, N., G. Nuemi, M. Gadreau, and C. Quantin. (2013). A Data Mining Approach for Grouping and Analyzing Trajectories of Care using Claim Data: The Example of Breast Cancer. *BMC Medical Informatics and Decision Making*, 13(1), 130.
10. Kaggle-Mortality. (2017). Death in the United States - Learn more about the leading causes of death from 2005-2015, https://www.kaggle.com/cdc/mortality. Retrieved on January 8, 2020.
11. Kojic, E. M., and R. O. Darouiche. (2004). Candida Infections of Medical Devices. *Clinical Microbiology Reviews*, 17(2), 255-267.
12. Koopman, B., S. Karimi, A. Nguyen, R. McGuire, D. Muscatello, M. Kemp, D. Truran, M. Zhang, and S. Thackway. (2015a). Automatic Classification of Diseases from Free-Text Death Certificates for Real-Time Surveillance. *BMC Medical Informatics and Decision Making*, 15(1), 53.
13. Koopman, B., G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson. (2015b). Automatic ICD-10 Classification of Cancers from Free-Text Death Certificates. *International Journal of Medical Informatics*, 84(11), 956-965.
14. Lavergne, T., A. Névéol, A. Robert, C. Grouin, G. Rey, and P. Zweigenbaum. (2016). A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage. *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, December 12, Osaka, Japan, 60–69.
15. Mierswa, I., and Klinkenberg, R. (2020). RapidMiner Studio, RapidMiner, Inc., https://rapidminer.com. Retrieved on January 7, 2020.
16. Mujtaba, G., L. Shuib, R. G. Raj, R. Rajandram, K. Shaikh, and M. A. Al-Garadi. (2017). Automatic ICD-10 Multi-Class Classification of Cause of Death from Plaintext Autopsy Reports through Expert-Driven Feature Selection. *Plos One*, 12(2).
17. Naik, A., and L. Samant. (2016). Correlation Review of Classification Algorithm using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*, 85, 662-668.
18. NVSS - Mortality. (2019). National Vital Statistics System (NVSS) - Mortality Data, https://www.cdc.gov/nchs/nvss/deaths.htm. Retrieved on December 30, 2019.
19. Python. (2020). Python: A dynamic, open-source programming language. https://www.python.org/, Retrieved on January 8, 2020.
20. Singh, S., H. S. Subramania, S. W. Holland, and J. T. Davis. (2012). Decision Forest for Root Cause Analysis of Intermittent Faults. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6), 1818-1827.
21. Subotin, M., and A. R. Davis. (2016). A Method for Modeling Co-Occurrence Propensity of Clinical Codes with Application to ICD-10-PCS Auto-Coding. *Journal of the American Medical Informatics Association*, 23(5), 866-871.
22. Sundararajan, V., T. Henderson, C. Perry, A. Muggivan, H. Quan, and W. A. Ghali. (2004). New ICD-10 Version of the Charlson Comorbidity Index Predicted in-Hospital Mortality. *Journal of Clinical Epidemiology*, 57(12), 1288-1294.

23. Tomar, D., and S. Agarwal. (2013). A Survey on Data Mining Approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5), 241-266.
24. Ture, M., F. Tokatli, and I. Kurt. (2009). Using Kaplan–Meier Analysis Together with Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) in Determining Recurrence-Free Survival of Breast Cancer Patients. *Expert Systems with Applications*, 36 (2), Part 1, 2017-2026.
25. Weiner, M. G. (2018). POINT: Is ICD-10 Diagnosis Coding Important in the Era of Big Data? Yes. *Chest*, 153(5), 1093-1095.
26. WHO-ICD. (2019). International Classification of Diseases (ICD), https://www.who.int/classifications/icd/en/. Retrieved on December 29, 2019.
27. Zweigenbaum, P., and T. Lavergne. (2016). Hybrid Methods for ICD-10 Coding of Death Certificates. *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI)*, November 5, Austin, TX, USA, 96-105.