International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 13, Issue 7 [July. 2024] PP: 107-115

Studying medical data, using global data technologies (RWD), and extracting and analyzing big data (BDM) for cardiovascular diseases in diabetic patients

Fadwa Joma Abuhbail

Abstract:
Clinical research, which is typically focused on assessing the efficacy of medications and treatment interventions, serves as the foundation for medical practice. The utilization of Real World Data (RWD) and Big Data Mining (BDM) approaches is proving to be a beneficial tool for automated data analysis, thanks to advancements in technology and computational storage capacities. Clinics, hospitals, and other organizations that deal with medical practice on a daily basis have databases that include a wealth of information that should be examined to help physicians identify disease patterns, anticipate future trends, and develop therapeutic alliances. A nearly two-decade-old private clinic database was examined with the goal of evaluating the course of cardiovascular disease (CVD) in patients with diabetes. This paper's main objective was to assess the database's dependability in order to further the study of CVD.A manual review of the database's content identified inaccurate and missing fields, inconsistent instrumental data input, and field filling that depended on the user's actions over time, especially when it came to CV data. However, statistics generated from the records of 20222 diabetes patients with removed blanks showed that RWD results were consistent with those that had been reported.

Keywords — *Cardiology, Data Mining, Diabetes Mellitus, Real World Data*

Date of Submission: 12-07-2024 Date of acceptance: 26-07-2024

I.INTRODUCTION

In the modern period, 2.5 quintillion bytes of new data are produced daily. We now have access to vast amounts of data from numerous, diverse, and independent sources worldwide because to technological advancements. Large amounts of data can be summarised and cross-referenced in real time by creating processes. This is referred to as data mining (DM), big data processing, or simply big data mining (BDM). BDM's primary objective is to sift through massive amounts of data and find relevant knowledge or information for next actions [1]. Future primary research, especially in the biomedical field, is expected to be data-driven and produced entirely on its own with the aid of computer intelligence [2].

More generally, for data mining to work well, a conditioned sample of data that is associated with several other sources of information is necessary. This necessitates extremely effective database merging processes. Inaccurate frequency distributions, redundant data identification, and other aggregations can result in the production of erroneous or misleading statistics that create unreliable findings. Mistakes in data entry, inaccurate sensor readings, or malevolent actions result in a multitude of erroneous data sets, which in turn compound problems in every new generation of data [3]. Numerous research projects, examinations, and instruments have been developed in the medical domain to support physicians in their work. For many years, the most popular methods for creating and analysing data were Randomised Controlled Trials (RCTs) and Comparative Effectiveness Research (CER), which have now become the accepted methods for studies pertaining to medicine [4]. Randomised controlled trials (RCTs) are conducted with small populations randomised according to predetermined parameters in controlled settings. Because such a limited population was used in an artificially controlled setting, the data collection procedure may have produced results that are inaccurate due to the inability to appropriately reflect the medical problems under study. While CERs may be a little more dependable in that sense, they become a somewhat problematic instrument to use in order to get the best outcomes [5]. Because they can yield data that is either too or too lacking in parameters, these investigations are not the best option for case-specific diagnostics because they may not accurately reflect the medical condition under investigation. Noninterventional studies, whether prospective or retrospective, are regarded as naturalistic when they are carried out over an extended period of time in a heterogeneous population and provide an objective perspective on real-world outcomes [6].

Focusing on the connection between Diabetes Mellitus and Cardiovascular Diseases (CVD), a plethora of research in the area shows that both disorders share important factors that raise the risk of sickness, including age, gender, obesity, dietary habits, and others. [7]. There have been reports that diabetes is a risk factor for

cardiovascular morbidity and mortality overall. While congestive heart failure (CHF) and intermittent claudication (IC) have a bigger relative impact, heart disease (CHD) is often misunderstood due to its asymptomatic nature [8]. Diabetes also has a significant impact on a patient's likelihood of developing arterial hypertension (AH), which is another major risk factor for more serious and deadly CVDs [9]. The morbidity and mortality rates for each cardiovascular disease (CVD) are greater in those with diabetes than in those without the condition.

Data mining techniques have been applied in a few studies conducted in medical datasets, with a particular focus on populations with diabetes [10]. The majority of them work to aid in the diagnosis of diabetes and/or cardiovascular diseases (CVDs), with fewer blood tests and ultimately lower health expenses. They do this by supporting doctors and patients in obtaining early diagnoses and appropriate treatment [11]. Since correlation analysis and data visualisation are still in their infancy, the majority of these research use rather simplistic data visualisation.

However, a few studies have already taken things a little bit farther and used neural network analysis to help with data treatment, categorization, and conclusion drawing [12]. Instead of using fully developed databases, other studies opted to develop this idea using data from prior studies, demonstrating the significance of compiling all of the previous findings produced with the assistance of RCTs or CERs and transforming them into fresh and enhanced instruments for the corresponding medical fields [13].

Using data from a private database of a diabetic clinic in Lisbon, Portugal, a study was conducted regarding CVD in diabetic patients in light of BDM and RWD. Records of medical visits and cardiology exams were taken into consideration with the goal of developing a model that could be used to study the general population of the diabetic clinic with respect to cardiovascular conditions, in order to aid the cardiologists in treating diabetic patients.

The structure of this document is as follows. The methodologic technique used to determine which important elements inside the database would be searched for is described in Section II. The initial mining techniques are shown in part III, and the part after that, Section IV, contains the findings. Concluding thoughts and recommendations for further research are included in Section V.

II. METHODOLOGY

Understanding the type and structure of the database is a necessary first step. The size of the clinic database necessitates further refining the population set under investigation. Analysis of the database's contents can only be done later.

A. Database nature

The study's clinical database is from a private clinic that specialises in the treatment of diabetes. It was first designed in 1999 and has since been revised in response to suggestions from users for bettering the clinic's amenities and the expansion of the clinic's medical specialisations. Currently, the database has information from over a million appointment entries about the clinic. Procedures should result in the detection of data that is relevant for cardiovascular diagnosis, focusing on parameters that are appropriate for this type of study, given the size of the database and the fact that the primary focus of the current study is CVDs in a diabetic population. Vascular disorders were not taken into account in this early research, which allowed the study to concentrate on the field of cardiology.

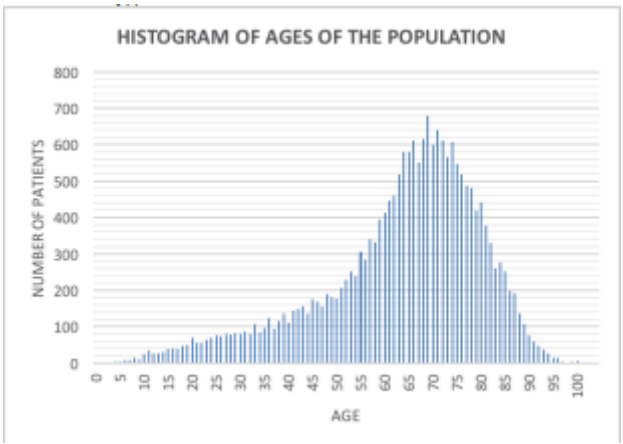

Fig. 1. Histogram of ages within the studied population

Given the sheer volume of appointments, defining the target demographic will be a crucial step after deciding to limit the database study area.

For the purposes of this study, it was determined that a patient must be alive and have had their most recent appointment within the last three years in order to be deemed active, as there is currently no system or set of regulations that determines whether a patient is active or not. A population of 20,222 patients with ages ranging from 4 to 103 years old was established for this set of rules.with the age distribution of the population under study illustrated in Fig. 1 and having a sample mean of 63.4, a median value of 50, a sample mode of 69, and a standard deviation of 32.7.

The fact that every medical record is made and handled differently has been noted, which only makes the situation worse. Therefore, the only option available was to carefully analyse the database by hand in order to determine how each table connects to the others, where the necessary information is located, and how the data is kept in the database.

B. Database structure

An outside business designed and maintains the database, handling both the front-end and back-end operations with a customised Java framework. The front office is connected to the clinician's software through a graphical user interface. The internal database, which houses the values and provides the structure for system manipulation, is correlated with the back office. As previously indicated, all of the SQL tables that were required to link the data were connected manually using pure logic, and values were retrieved using SQL queries.

A specific kind of SQL language/environment called PostgreSQL was used to create the database. While the SQL language is widely recognised as the industry standard for databases, it can be challenging to fully comprehend the hierarchy of individual tables, the import and export of values between tables, and the meaning of each field according to its original SQL name, which is assigned by the database management company.

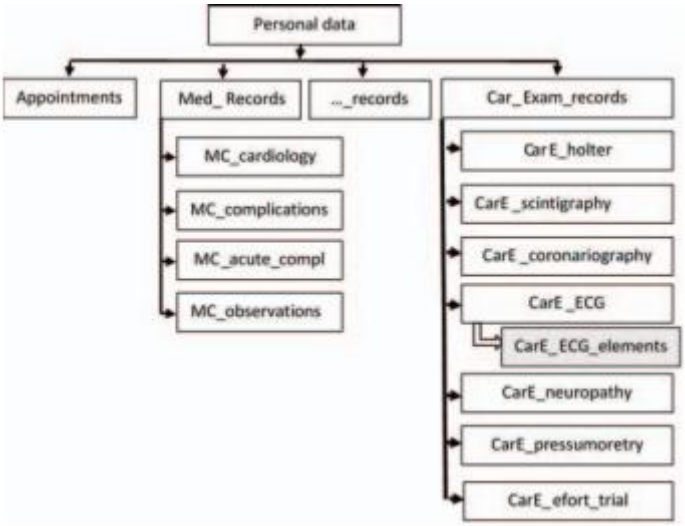

Fig. 2. Hierarchy of the tables considered for this study.

Data on routine appointments, medical records, and other general information were found by analysing the back-office database. Every pertinent piece of information related to the specialisation of cardiology was sought after. At this stage, the data could be linked together in terms of imported and/or exported values between tables, and the database's overall structure and pertinent information flow could be ascertained. The hierarchy of the tables that will be taken into consideration for the population of the entire clinic with reference to this specific study is depicted in Fig. 2. The names given to database tables correspond to the original SQL names assigned by the design firm, which is still in charge of maintenance.

The primary SQL database named Personal Data, which is created for each individual and contains the patient's personal information such as age, address, full name, health insurance number, and contacts, is shown in Fig. 2. It should be noted that only the personal data table displays private data pertaining to a certain patient. The only references in the other tables are to an identification number (ID), which is an integer that corresponds to a specific personal data item. This explains why each table in Fig. 2 eventually leads to the personal data table.

As a result, each sub-table has an integer ID that corresponds to its corresponding parent table. The use of anonymised data for study is made possible by this database design technique.

Connected to the Personal Data table, we examine three tables that are pertinent to this investigation: Appointments, Med_Records, and Car_Exam_records, which correspond to the patient's appointment registration, the clinician-introduced medical record, and the sub-table containing information related to cardiology exams, respectively. Fig. 2 features a module called..._records to symbolise all other tables pertaining to annotations pertaining to specialisations other than cardiology.

Tables Med_Records and Car_Exam_records have subtables; only those pertinent to this specific investigation are shown in Fig. 2. In relation to Med_Records, the primary subtables allow for the storage of data pertaining to cardiology (MC_cardiology), medical data pertaining to the identification of a patient's health complications (MC_complications), and, independent of each other, the description of the occurrence of acute complications episodes is stored in the sub-table MC_acute_complications.

There are multiple sub-tables in Table Car_Exam_records that hold records for particular cardiovascular exams. We focused solely on the sub-tables found in Figure 2 that were designated with the prefix CarE_ and corresponded to the Holter, scintigraphy, electrocardiogram (ECG), neuropathy, pressurometry, and effort trial examinations, respectively. Furthermore, it was seen that CarE_ECG is linked to the sub-table CarE_ECG_elements, which contains even more information on echocardiogram examinations.

III. ANALYSIS AND DISCUSSION OF IDENTIFIED DATASETS

A portion of the reliability assessment findings from our case-study database for mining cardiovascular clinical data will not be shown due to space constraints. The full examination of the database's overall structure as well as the datasets including the tables Appointments, Med_ Records, and Car_Exam_records shown in Fig. 2 are covered in the description that follows.

By first looking at the front-office, we were able to confirm that the physicians' interface is made up of four tabs for every table in a medical record: cardiology, complications, acute complications, and observations. There are several filling fields on these graphical tabs. A SQL table is generated for every graphical tab, including the topic's data, underscoring the importance of carefully reviewing each table and its contents separately. The cardiology tab would be empty in this instance because not all patients, for instance, have problems linked to cardiology. Because of this, data added to the corresponding tab is the only time an entry in a SQL table is created.

To continue the example, there won't be a cardiology table entry in the database for that patient if the clinician leaves the cardiology tab empty. As a result, a patient may have an entry in the medical record database but not in the cardiology table or any other table.

A. Patients appointments records

After analyzing the table derived from Appointments, we found that it generally better reflects a patient's current condition because, in order to produce a medical record or an exam, a prior appointment is necessary, during which the clinician obtains the necessary data to warrant the examination or to fill out a medical record. The Appointments table has numerous variables, the most of which are not relevant to our study. The date, appointment status, and specialty are the three most crucial fields.

There are 1646613 appointments in this table. In light of the enormous number of appointments and our ultimate research objective, it was decided that appointments with cardiologists would comprise a subgroup of this population. This limitation resulted in 64834 appointments in total, with the frequency of appointments for each patient displayed in Fig. 3.

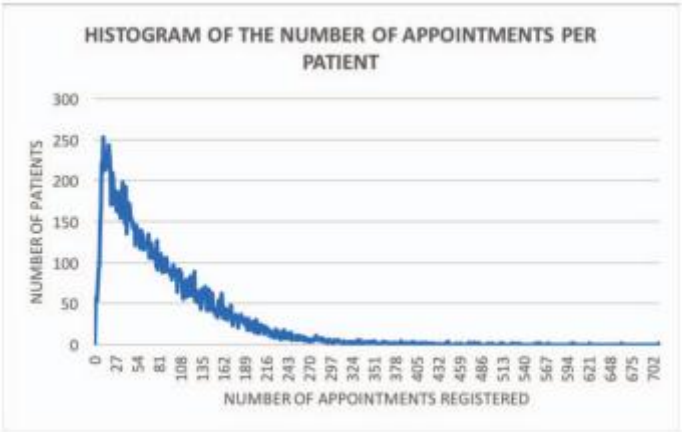

Fig. 3. Histogram of the number of cardiology appointment entries in database table Appointments per patient.

Fortunately, the number of patients with zero appointments is actually zero, which means that all patients possess at least one appointment, confirming that there are no significant data errors in this table. Fig. 3 shows that as the number of cardiology appointments per patient increases, the number of patients decreases exponentially. The most common number of cardiology appointments being 11 with 253 patients having that amount of entries.

B. Patients medical records

We discovered 707726 patient medical record entries in table Med_ Records (Fig. 2), which is slightly less than half of the total number of visits. This is where one would expect to find the medical record obtained during an appointment.

The distribution of the recorded data from these 707726 patients by sub-table is shown in Fig. 4. Figure 4 shows that only around one third of the population has entries in the cardiology table, but more than two thirds have observations table entries. Only 1.12% of patients have entries for acute complications, despite the fact that half of the population has entries in the complications table. Compared to the other tables, the Acute Complications table functions differently. The majority of the other table entries make reference to a sort of medical record about something that the technician or clinician has observed, said, or tested. On the other hand, the Acute Complications table is just a record of a particular patient's medical events.

Episodes of hypoglycemia, hyperglycemia, or other problems related to diabetes mellitus are reported by the patient to the clinician. Additionally, the text field for appointments is all that is shown in the bar named Observations. In order to facilitate a more straightforward data treatment process using more standardised datasets, Acute Complications and Observations will be eliminated for the time being.

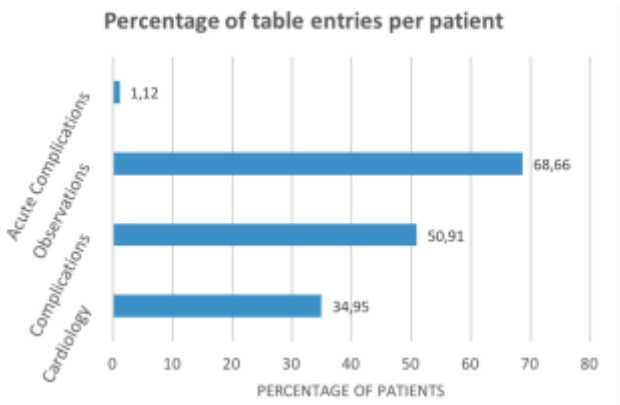

Fig. 4. Percentage of table entries for each sub-table linked to table Med_ Records

TABLE I DATABASE FIELDS INCLUDED IN SUB-TABLES MC_cardiology AND MC_complications

MC_cardiology fields	MC_complications fields
Myocardial infarction	Ophthalmologic
Surgery	Cardiologic
Angioplasty	Podological
CVA/TIA	Nephrological
Coronary	Neurological
Cardiac insufficiency	Peripheral vascular disease
Arterial hypertension	Others (text)
Carotid stenosis	-
Observations (text)	--

In Table I, the sub-tables MC_cardiology and MC_complications are essentially completed by a yes/no selection within each field. Additionally, both sub-tables have a text typing window for observational information on the front office. The field in Table I denoted by CVA/TIA represents cerebrovascular attack, also referred to as stroke, and transient ischemic attack.

The aetiology of the acute complication that occurred is included in the sub-table MC_acute_compl (see Fig. 2). A text field for medical annotations, the date of the incident, and a code corresponding to the kind of complication—0 for hyperglycemia, 1 for hypoglycemia, 2 for other, and 3 for ketosis—must be entered.

We may obtain a rudimentary grasp of the incidence rate of each type of complication by analysing the values inside the MC_complications and MC_cardiology sub-tables, starting with the first one (the Complications bar on Fig. 4). To do this, the relationship between positive and negative values was evaluated and empty values were eliminated.

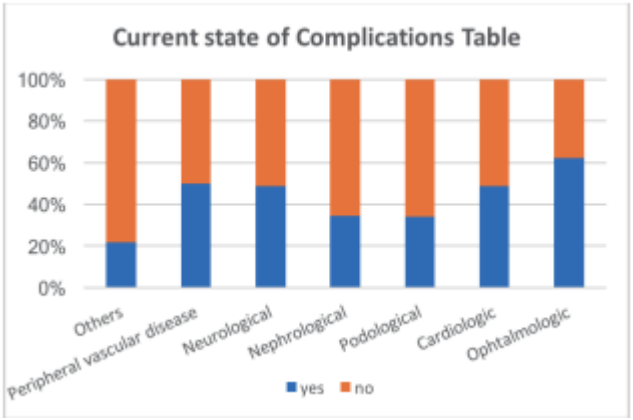

Fig. 5. Modified MC_complications table fields disregarding empty entries

As would be expected, we can see that if we add the incidence rates of peripheral vascular diseases and cardiovascular diseases (referred to as "doença_vasc_perif" and "cardiovascular" in Fig. 5) that CVD could be the most common pathology among the diabetic population.With our attention now focused on the MC_cardiology sub-table (the Cardiology bar on Fig. 4) and the rate at which yes, no, or null selections were made, we discovered that, on average, just 4.9% of the fields have values that are positively or negatively filled, making the percentage of unfilled fields extremely high. It is appropriate to look through this table for any fully empty items in light of the results. Out of the 7067 entries in the cardiology table, it was discovered that 6033 were completely blank—that is, not a single field had a yes or no entered—making these entries meaningless.

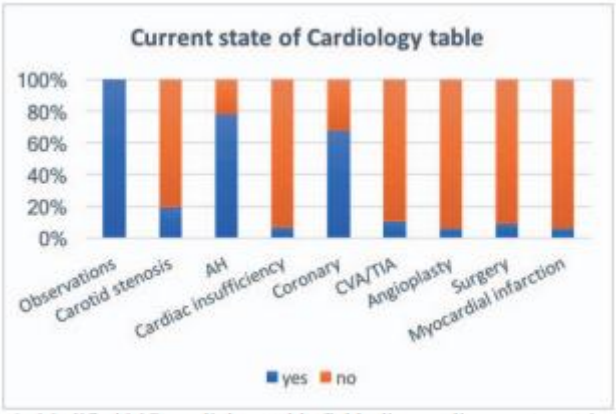

Fig. 6. Modified MC_cardiology table fields disregarding empty entries

In order to assess the usefulness of this tactic, we chose to examine a subsample of the cardiology table that was produced by eliminating the blank items from every field. With this step, the table's overall filling increased from 4.9% of filled values to 33.4%. However, two thirds of the values across all entries remain blank. As a result, in every field, all empty values were eliminated to display only the relationship between values that were positively and negatively filled. We can determine the potential proportion of occurrence of specific disorders in the population under consideration using this relationship.A bar graph representing the percentages of values in each field of the Car_Exam_records table that are filled in positively and negatively is displayed in Fig. 6. One of two conclusions can be drawn from the values in Fig. 6: either there is a serious inconsistency in the filling of the cardiology table, or the values indicate that AH ("HTA" in Fig. 6) and CHD ("coronaria" in Fig. 6) are the most common conditions among diabetics, with the rest of the fields being rarely used. It is impossible to know for sure that these numbers are accurate given the startlingly low percentage of filled submissions.In fact, this demonstrates the shortcomings of the Front and Back Office with regard to BDM; that is, a significant improvement in the quality of our dataset is necessary to move further with this study.

However, to guarantee their validity for further study investigations, the results of Fig. 6 were cross-checked with those of earlier studies. According to a Portuguese population survey, the percentage of stroke patients increases from 2.1% in a normal population to 5% in those with diabetes, and 78.3% of diabetics have hypertension [14]. Comparable rates of diabetics with cardiovascular problems linked to AH are found in other worldwide investigations [15, 16].

By examining the corresponding bars in Figure 6, we can quickly verify that the method created to process the current database validates values previously found in international studies. As a matter of fact, Fig. 6 indicates that around 78% of the population has AT, 68% has CHD, and that roughly 10% and 5% of the population, respectively, have records of stroke and heart attack ("ACV-AIT" and "enf_miocardio" in Fig. 6). Additionally, it is known that the magnitude of positively filled CHD-related values correlates with an increase in the likelihood of CHD of two fold in men and three fold in women, making this the most positively filled field, second only to AH [16].

C. Patients cardiology exams records

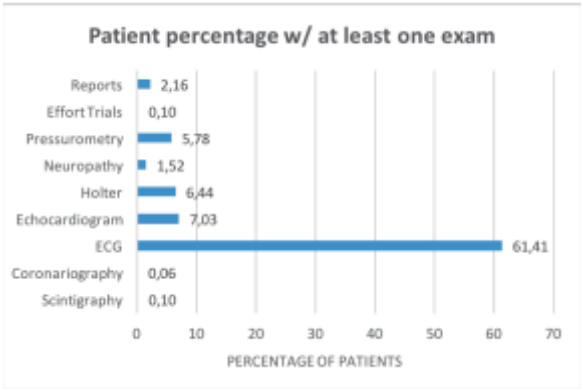

Fig. 7. Percentage of patients in the population with one or more entries per type of cardiology exam

The maximum number of Cardiology Exam entries a patient can have in the population is 43, and 63.8% of patients have at least one entry. When only patients with entries are taken into account, 72.2% of patients have only one exam, 14.8% have two exams, and the remaining 13.0% have three or more exams. The ECG exam is by far the most popular form of exam, as Fig. 7 shows. We decided not to go into greater detail on cardiology tests for the purpose of simplicity and conciseness, as these revealed superior filling procedures than the earlier sections of the database. However, more research on this subtable and the fields' reliance on instrumental data for completeness will be done in the future.

IV. CONCLUSION

The number of patients in the database and, more specifically, Fig. 2, which shows only the number of appointments per diabetes patient in the MC_cardiology table, indicate that a sizable clinical database was employed for this investigation. The abundance of information in the database is beneficial for both statistical analysis and BDM processes. Its capacity to access records dating back to 1999 raised some questions, nevertheless, regarding its suitability for RWD and BDM research. Further evidence for the necessity of a thorough study came from database updates, which occasionally included new tables to consider appointments for other specialties when the database's proprietary clinic supplied those services.

Another factor taken into consideration was the evolution of medical instrumentation since 1999 (the year the database was created), necessitating the adaptation of clinical exam recording processes to new communication protocols. Lack of information in some database entries, or at the very least, differential inputting strategies, indicates a consequence. Figures 5 and 6, which were produced by excluding null entries from the data analysis, attest to the presence of false information. In these situations, it is unknown whether the lack of anticipated data is ascribed to human error, instrumental neglect, or even the belief that it is clinically irrelevant.

Such data refutes the need for human involvement. When it comes to storing parameters, which can originate digitally or on paper, we frequently rely on a clinician's or technician's experience with the database. Furthermore, anytime data is entered by humans, as is the case in many other fields, a number of errors could happen as a result of misreading the information that is accessible, improperly filing or corrupting digital records, or even unintentionally erasing datasets. This typically occurs when a large number of users interact with a database simultaneously. To sum up, after staff changes or even when individuals develop different filling habits over time, information may be lost from person to person.

Companies should be ready to provide detailed instructions on improved database filling techniques that aim to produce consistent data that can be used in BDM processes in order to prevent such human errors.
Beyond these drawbacks of employing an outdated database that was not created with BDM in mind, there is no denying the wealth of clinical data that can be utilised for RWD after some pre-processing and the definition of targeted sub-samples of the population.

ACKNOWLDGEMENT

The authors express their gratitude to the private clinic's doctors and technical team for their helpful conversations. Sponsor of the study: LINK: 692023 H2020, (2016-2019).

REFERENCES

[1]. X. Wu, X. Zhu, G.-q. Wu, and W. Ding, “Data Mining with Big Data,” IEEE Transactions on Knowledge and Data Engineering, Vol. 26, no. 1, pp. 97-107, 2014.
[2]. A. Kusiak, J. A. Kern, K. H. Kernstine, and B. T. Tseng, “Autonomous decision-making: a data mining approach”, IEEE Transactions on Information Technology in Biomedicine, vol. 4, no. 4, pp. 274–284, 2000.
[3]. M. Hernandez and S. Stolfo, “Real-World Data Is Dirty-Data Cleansing and the Merge or Purge Problem”, Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 9–37, 1998.
[4]. M. A. Laken, R. Dawson, O. Engelman, et al, “Comparative effectiveness research in the "real" world: Lessons learned in a study of treatment-resistant hypertension”, J. of the American Society of Hypertension, vol. 7, no. 1, pp. 95–101, 2013.
[5]. G. Y. H. Lip, T. Potpara, G. Boriani, and C. Blomström-Lundqvist, “A tailored treatment strategy: A modern approach for stroke prevention in patients with atrial fibrillation”, Journal of Internal Medicine, vol. 279, no. 5, pp. 467–476, 2016.
[6]. G. Bonnelye, A. Miniuks, and A. Goncalves, “The importance of realworld data to the pharma industry”, http://www.pmlive.com/pharma_thought_leadership/the_importa nce_of_real-world_data_to_the_pharma_industry_740092, 2015.
[7]. A. Neil, “Diabetes and cardiovascular disease,” Diabetes, Obesity and Metabolism, vol. 5, no. SUPPL. 1, pp. S11–S18, 2003
[8]. O. Kittnar, “Electrocardiographic changes in Diabetes Mellitus”, Physiol. Res. 64 (Suppl. 5): S559-S566, 2015.
[9]. J. R. Sowers, M. Epstein, and E. D. Frohlich, “Diabetes, Hypertension, and Cardiovascular Disease: An Update”, Hypertension, vol. 37, no. 4, pp. 1053–1059, 2001.
[10]. K. Rajesh and V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”, International Journal of Engineering Research and Innovative Technology (IJEIT), vol. 2, no. 3, pp. 224–229, 2012.
[11]. A. Singh and S. Kumari, “A Data Mining Approach for the Diagnosis of Diabetes Mellitus”, In Procedings of the 7th International Conference on IEEE In Intelligent Systems and Control (ISCO), pp. 373–375, 2013.
[12]. I.-N. Lee, S.-C. Liao, and M. Embrechts, “Data mining techniques applied to medical information”, Medical Informatics and the Internet in Medicine, vol. 25, no. 2, pp. 81–102, 2000.

[13]. M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, "Data- Mining Technologies for Diabetes: A Systematic Review", J. of Diabetes Science and Technology, vol. 5, no. 6, pp. 1549–1556, 2011.
[14]. N. Cortez-Dias, S. Martins, A. Belo, and M. Fiuza, "Prevalence, management and control of diabetes mellitus and associated risk factors in primary health care in Portugal," 2010.
[15]. D. Wentworth, J. Stamler, F. O. R. The, et al, "Diabetes, Other Risk Factors, and 12- Yr Cardiovascular Mortality for Men Screened in the Multiple Risk Factor Intervention Trial," vol. 16, no. 2, pp. 434–444, 1993.
[16]. J. A. E. Manson, G. A. Colditz, M. J. Stampfer, et al, "A prospective study of maturity-onset diabetes mellitus and risk of coronary heart disease and stroke in women," Archives of Internal Medicine, vol. 151, no. 6, pp. 1141–1147, 1991