Vol 5, Issue 1, Jul 2018

Journal of Health Management & Informatics

jhmi.sums.ac.ir

Original Article

Data Quality Assessment and Recommendations to Improve the Quality of Hemodialysis Database

Neda Firouraghi[1], Shahrokh Ezzatzadegan Jahromi[2], Ashkan Sami[3], Mohamad Reza Morvaridi[4], Roxana Sharifian[5*]

Received 6 Jun 2016; Accepted 10 Dec 2016

ABSTRACT

Introduction: Since clinical data contain abnormalities, quality assessment and reporting of data errors are necessary. Data quality analysis consists of developing strategies, making recommendations to avoid future errors and improving the quality of data entry by identifying error types and their causes. Therefore, this approach can be extremely useful to improve the quality of the databases. The aim of this study was to analyze hemodialysis (HD) patients' data in order to improve the quality of data entry and avoid future errors.

Method: The study was done on Shiraz University of Medical Sciences HD database in 2015. The database consists of 2367 patients who had at least 12 months follow up (22.34±11.52 months) in 2012-2014. Duplicated data were removed; outliers were detected based on statistical methods, expert opinion and the relationship between variables; then, the missing values were handled in 72 variables by using IBM SPSS Statistics 22 in order to improve the quality of the database. According to the results, some recommendations were given to improve the data entry process.

Results: The variables had outliers in the range of 0-9.28 percent. Seven variables had missing values over 20 percent and in the others they were between 0 and 19.73 percent. The majority of missing values belong to serum alkaline phosphatase, uric acid, high and low density lipoprotein, total iron binding capacity, hepatitis B surface antibody titer, and parathyroid hormone. The variables with displacement (the values of two or more variables were recorded in the wrong attribute) were weight, serum creatinine, blood urea nitrogen, systolic and diastolic blood pressure. These variables may lead to decreased data quality.

Conclusion: According to the results and expert opinion, applying some data entry principles, such as defining ranges of values, using the relationship between hemodialysis features, developing alert systems about empty or duplicated data and entering directly HD data or lab results into the database can improve the data quality drastically. Experts' opinion in detecting outliers as a complement to statistical methods can have an effective role in detection of real outliers. For the analysis of HD databases, the relationship between the variables because of their effect on the quality should be focused more to improve the quality of the database.

Keywords: Database, Data entry, Hemodialysis, Data Quality, Outliers, Missing values

►*Please cite this paper as:*

Firouraghi N, Ezzatzadegan Jahromi SH, Sami A, Morvaridi M, Sharifian R. Data Quality Assessment and Recommendations to Improve the Quality of Hemodialysis Database. J Health Man & Info. 2018; 5(1): 21-26.

Introduction

Data gathering is an intrinsic part of the healthcare organizations. Analyzing and using these data is necessary. To achieve this objective, clinical data should be stored in well-designed databases (DB) and registries (1). Health care databases hold a valuable source of information needed to assess and improve the quality of health care services, assess the public health, perform epidemiological studies, evaluate the performance of health care providers, identify health needs, and extract knowledge (2, 3).

One of the large volume clinical databases is hemodialysis (HD) database. HD is the most widely used treatment modality for end-stage renal disease (ESRD) patients (4) that lose their kidney function and need renal replacement therapy (5). Because of Chronic kidney failure, HD patients received the treatment three times per week, so a large volume of data are being stored in the HD databases and patients' charts (6).

The values of health data depend on their quality and accuracy (2, 7). Since clinical data contain abnormalities (8),

1 *Faculty of Management and Medical Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran*

2 *Assistant Professor of Medicine, Shiraz Nephrology Urology Research Center, Nemazee Hospital, Shiraz University of Medical Sciences, Shiraz, Iran*

3 *Associate Professor of Computer Sciences, Shiraz University, Shiraz, Iran*

4 *Shiraz University of Medical Sciences Special Diseases Affairs, Shiraz, Iran*

5 *Associate Professor, Health Human Resources Research Center, Faculty of Management and Medical Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran*

Corresponding Author:** ***R Sharifian, *Associate Professor, Health Human Resources Research Center, Faculty of Management and Medical Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran, Email: Sharifianr@sums.ac.ir*

their quality assessment is necessary. There are no comprehensive guidelines for assessing and improving data quality in health care management systems and few studies have been done on reporting of clinical data errors. Therefore, reporting the data errors and quality assessment should be emphasized further. Data quality assessment consists of detecting errors, developing strategies, making recommendations to avoid future errors, and improving the quality of data entry.

In this study, we used data cleaning method in order to assess the quality of hemodialysis DB. Data cleaning is an important part of pre-processing step of data mining for knowledge discovery of databases. This method was done by identifying and refining noisy data and errors or removing them, by using statistical methods like mean ±3×standard deviation (SD),and handling missing values by mean substitution in order to improve data quality (4, 8-12). Data cleaning can lead to identification of the fields and processes that cause the most common error; then, by giving feedback from these studies' results to data collecting centers, it might be possible to improve the data gathering quality and reduce errors. Identifying the types and causes of errors can decrease the health data anomalies. This procedure can be extremely useful and effective in improving the quality of the studies (13).

There are some studies about the clinical data quality that have focused on accuracy (correctness of data) and completeness (completely recorded necessary data), as the data quality characteristics (7, 13, 14). Identifying and refining the errors and noises can improve the accuracy; also, handling the missing values can lead to improvement of the completeness of data (13, 14). In HD databases, data cleaning was used to improve the quality, as well (8, 11, 12, 15). Several studies have shown the importance of data quality in clinical registries (7, 13, 14). Data quality is more important in adopting powerful systems like clinical decision support systems; moreover, poor data quality can affect medical decisions strongly (16).

There was no information about the quality of Shiraz University of Medical Sciences (SUMS) HD database, so this study was conducted to assess the quality of this database during 2012-2014, in order to detect the outliers and refine them, the proportion of the outliers and missing data percentage, handle missing values, identify the types of errors and their causes and finally give some recommendations on data entry process in order to improve the data quality for avoiding future errors.

Method

This was a descriptive study conducted on Shiraz University of Medical Sciences HD database in 2015. This DB contained hemodialysis data from 34 centers since 2012, so all demographic data, hemodialysis sessions indices, and all regular biochemical and hematological investigations had been recorded in the database. The study included data of 2367 adult HD patients who initiated their treatment during 2012-2014 and had at least 12 months of follow up (mean 22.34±11.52 months). The database contains 72 variables in 3 tables (Demographic: 6 fields, hemodialysis sessions: 29 fields, 266135 records, and regular investigations: 37 fields, 24919 records).

Quality analysis

Data quality analysis was done in three steps: removing the duplicated data, detecting outliers and correcting or removing them, and handling the missing values.

1-Removing the duplicated data

At this step, duplicated records in the dialysis and test variables were detected and omitted. This occurred when a patient had more than one record in the same date because of repeated tests or repeated data entry mistakes.

2-Detecting and removing outliers

Outliers are out of the expected and acceptable range values for each variable; these values are impossible data for that variable (13). Clinical data may have outliers caused by clinical instruments or human errors in data collection or data entry stage (7, 13), so the outlier detection is a very critical step in improving the quality of clinical databases. There are some techniques that could be used to define the outliers (17, 18). In this study, to detect and measure the percentage of the outliers per variable, we used three methods:

– Variance and Standard Deviation (Mean±3SD): It is a common method used in outlier detection.
– Quartiles (Q1-1.5IQR, Q3+1.5IQR): This method, which is calculated based on the first and third quarters, is another way to detect the outliers by defining acceptable value ranges.
– Frequency Histograms: It is used to visualize the outliers.

Some values may be considered as outliers in statistical methods, but they may be an acceptable or possible value for a HD patient, so the final decision about outliers was made by the clinical expert who approved the acceptable ranges. Minimum, maximum, average, histograms and bins obtained by statistical methods for each variable were evaluated by a nephrologist (Clinical expert). Eventually, based on the relationship between the variables, acceptable ranges for each variable were determined by an expert.

Some of the mistakenly recorded values were tagged as outliers by our methods where we found they were only incorrect recorded values (noisy data). Some validation rules based on the relationship between variables were applied to detect noisy value from the real outliers and correct the errors including:

Pre-dialysis weight should be equal or greater than post-dialysis weight. By comparing the outliers of the patient's weight with other recorded values of these variables for that patient and also comparing the pre-dialysis weight with the post- dialysis weight, we modified some noises.

Ultrafiltration (UF) was calculated by using the formula (pre-dialysis weight minus post- dialysis weight), and then by comparing the calculated UF with existing UF values in the DB, pre- dialysis weight and post-dialysis weight in out of accepted ranges; we found the values of weight variables in some cases which were recorded mistakenly.

During four hours of each dialysis session, systolic blood pressure (SBP) and diastolic blood pressure (DBP) were measured hourly. SBP should be greater than DBP in each measurement, so SBP was compared with DBP. It was found that in some cases these two variables were displaced. The rate of displacement was calculated for systolic and diastolic blood pressure.

Since serum creatinine (Cr) and blood urea nitrogen (BUN) drop during dialysis, post- dialysis serum Cr and BUN should be less than pre-dialysis values. We fixed the displacements and modified some errors.

Direct bilirubin is a part of total bilirubin, so total bilirubin levels must be greater than direct bilirubin. Serum albumin (Alb) is a part of serum total protein, so the total protein must be greater than Alb. In addition, the blood hematocrit (HCT) must be greater than hemoglobin (Hb). With these principles, noisy data were detected. Some of them were corrected by expert opinion, if the correction was possible, and the others were removed from the data. Then, the real outliers based on approved bins were detected and the percentages of them were calculated and then excluded. In other variables, after comparing the out of range values with others of that variable, some errors were corrected and the outliers were removed based on expert approved ranges.

3-Handling missing values

After removing the outliers, the percentage of missing values (the not recorded necessary values) for each variable was calculated. In this study, the mean (for continues variables) and the mode (for categorical variables) of each patient's available values were used for handling that patient's missing values. This method was a common way in the studies that apply all available values effects for missing observation (8, 11, 12).

Results

The mean age of the patients was 58.37± 16.30 years, and 58.2% of them were male.

Baseline demographic variables, hemodialysis sessions indices, and all regular biochemical and hematological investigations are listed in Tables 1-4.

Outliers and missing percentages for variables are shown in the Tables. Blood group type and etiology of ESRD were the variables with missing values from demographic features (Table 1). Outliers' percentage ranged between 0-8 in hemodialysis sessions indices and maximum percentage was for erythropoietin dose (8%). The percentage of missing values in these variables ranged between 0.95-19.73; the minimum was for SBP in the first hour and maximum percentage was for DBP in the third hour (Table 2). The outliers' percentage in monthly test variables ranged between 0% and 5.63% and the maximum percentage was for serum uric acid. The percentage of missing values was between 0.44% and 51.83%; the minimum was for Hb and the maximum percentage was for serum uric acid. Serum uric acid and alkaline phosphatase (AlkPh) had missing values more than 20 percentage (>50%) (Table 3). The percentage of the outliers ranged between 0.16-9.28 in three-month test variables; the minimum was for total and direct bilirubin and maximum percentage was for hepatitis B surface antibody (HBs Ab) titer. The percentage of missing values ranged between 3.56% and 56.8%; the minimum was for serum glutamic pyruvic transaminase (SGPT) and the maximum percentage was for HBs Ab titer. Low density lipoprotein (LDL), high density lipoprotein (HDL), serum total iron-binding capacity (TIBC), HBs Ab titer and parathyroid hormone (PTH) had missing values above 20 percent (Table 4).

Table 1. The Percentages of Outliers and Missing Values of Demographic Variables

Variable Name	%Outliers	%Missing	Variable Name	%Outliers	%Missing
The Dialysis Start Date	0	0	The Date of Dialysis Termination	0	0
Blood Group Type	0	5.1	Etiology of ESRD	0	5.4
Gender	0	0	Date of birth	0	0

Table 2. The Percentages of Outliers and Missing Values of Hemodialysis Sessions Indices

Variable Name	%Outliers	%Missing	Variable Name	%Outliers	%Missing
Pre-dialysis weight	0.52	5.72	Diastolic Blood Pressure in the First Hour	0	1.36
Post- dialysis weight	0.57	8.28	Diastolic Blood Pressure in the Second Hour	0	16.93
Dry weight	0.12	7.18	Diastolic Blood Pressure in the Third Hour	0	19.73
Height	0.92	6.23	Diastolic Blood Pressure in the Fourth Hour	0	6.85
Body mass index (BMI)	0	10.6	Erythropoietin Dose	8	14.9
Dialysate temperature	0.23	6.22	Ultrafiltration (UF) volume	0	1.7
Blood flow rate in the First Hour	0.07	1.04	Pulse rate in the First Hour	0	1.66

Blood flow rate in the Second Hour	0.08	16.48	Pulse rate in the Second Hour	0	17.10
Blood flow rate in the Third Hour	0.15	19.49	Pulse rate in the Third Hour	0	19.70
Blood flow rate in the Fourth Hour	0.57	6.33	Pulse rate in the Fourth Hour	0	6.79
Dialysate Sodium concentration	1.82	4.34	Heparin-Bolus dose	7.34	11.45
Systolic blood pressure in the first Hour	0	0.95	Systolic Blood Pressure in the Third Hour	0	19.68
Systolic Blood Pressure in the Second Hour	0	16.57	Systolic Blood Pressure in the Fourth Hour	0	6.46
High filters	0	1.50	Type of vascular access	0	1.3
Duration of dialysis	1.62	3.40			

Table 3. The percentages of outliers and missing values of regular monthly biochemical and hematologic investigations

Variable Name	%Outliers	%Missing	Variable Name	%Outliers	%Missing
Pre-dialysis BUN	0.2	1.6	Serum Calcium (Ca)	0.2	8.82
Post-dialysis BUN	0.4	9.1	Serum Phosphorus (P)	0.15	12.8
Pre-dialysis -Creatinine	0.27	5.73	Hemoglobin (Hb)	0.28	0.44
Post-dialysis Creatinine	0.22	14.7	Serum Sodium (Na)	0.4	10.31
Fasting Blood Sugar	0.38	12.01	Serum Potassium (K)	0.31	10.2
Mean Corpuscular Hemoglobin (MCH)	0.49	8.87	Hematocrit (HCT)	0.32	0.68
Mean Corpuscular Hemoglobin Concentration (MCHC)	1.49	4.45	Platelets (Plt)	0.15	1.11
Mean Corpuscular Volume (MCV)	0.77	7.18	Red Blood Cell (RBC) count	0.57	7.67
Urea Reduction Rate (URR)	0	8.51	White Blood Cell count (WBC)	0.61	5.15
The Normalized Protein Catabolic Rate(En-PCR)	0	1.93	Alkaline Phosphatase AlkPh))	0.13	50.75
Dialysis Adequacy(Kt/v)	0.2	7.16	Serum Uric Acid	5.63	51.83

Table 4. The Percentages of Outliers and Missing Values of regular Three-monthly Biochemical and Hematologic Investigations (9, 10)

Variable Name	%Outliers	%Missing	Variable Name	%Outliers	%Missing
Serum Albumin (Alb)	0.6	3.71	Serum Glutamic Oxaloacetic Transaminase (SGOT)	1.35	4.14
Serum Ferritin	2.26	3.81	Serum Glutamic Pyruvic Transaminase (SGPT)	0.71	3.56
Triglyceride (TG)	0.35	3.94	Total Protein	0.67	13.15
Direct Bilirubin	0.16	8.47	Serum Cholesterol	0.5	4.95
Total Bilirubin	0.16	6.75	Serum Iron (Fe)	3.04	12.6
High Density Lipoprotein (HDL)	0.67	35.21	Low Density Lipoprotein (LDL)	1.52	34.66
Total Iron-Binding Capacity (TIBC)	5.02	21.5	Parathyroid Hormone (PTH)	1.92	40.13
Hepatitis B Surface antibody (HBS ab) Titer	9.28	56.8			

No outlier was detected in demographic variables; the percentage of the outliers in hemodialysis session variables was 0 to 8, in monthly test variables 0-5.63 and in three-month variables 0.16-9.28. Overall, some variables had no outliers and the rest had less than 9.28 percent. The variables the dose of bouls heparin erythropoietin dose, uric acid, Fe, TIBC and HBS ab titer showed a percentage of outlier above 3 percent. Seven variables had missing values over 20 percent. These variables were AlkPh, uric acid, HDL, LDL, TIBC, HBS ab Titer, and PTH.

Regardless of the variables in which the missing values were above 20 percent, the percentage of missing values in other variables was as follows: demographic variables 0-5.4, dialysis session variables 0.95-19.73, monthly test variables 0.44-14.7, and three-month variables 3.56-13.15. Totally, the missing value of all variables was 0-19.73 percent.

Displacement percentage based on the relationship between variables was in the range of 0.26% and 0.5%, and the maximum percentage was for serum creatinine displacement (Table 5).

Table 5. The Percentage of Displacement in some Variables

Feature name	Feature name	Percent of displacement
Pre-hemodialysis weight	Post- hemodialysis weight	0.3
Pre- hemodialysis serum creatinine	Post- hemodialysis serum creatinine	0.5
Systolic blood pressure1[a]	Diastolic blood pressure1	0.33
Systolic blood pressure2[b]	Diastolic blood pressure2	0.3
Systolic blood pressure3[c]	Diastolic blood pressure3	0.29
Systolic blood pressure4[d]	Diastolic blood pressure4	0.33
Pre- hemodialysis serum BUN	Post- hemodialysis serum BUN	0.26

[a, b, c, d] Systolic blood pressure measured in the first, second, third, and fourth hour of hemodialysis accordingly.

Recommendations

Outlier detection and quality assessment of databases can lead to identification of the variables with most errors. In this study, the variables with the most outliers (heparin-bolus dose, erythropoietin dose, uric acid, Fe, TIBC, HBS ab Titer), the most missing values (AlkPh, uric acid, HDL, LDL, TIBC, HBS ab Titer, PTH), and with displacement such as weight, creatinine, SBP, DBP and BUN may lead to reduced quality, so entering values in these features should be done more carefully.

By applying some data entry principles in HD database, these errors can be prevented; these principles include limiting data entry by defining acceptable and valid value ranges for each continuous variable, not allowing to re-enter the existing data, not permitting to empty the field, setting the length of features specifically, developing alert systems about empty or duplicated data, and entering of HD data or lab results into the database directly.

Relationship between HD features can be used to develop data entry rules to restrict data input such as:

- Pre_HD_Weight should be ≥ Post_HD_Weight
- SBP (of each hour) should be >DBP (of each hour)
- Pre_HD_Serum Cr should be ≥ Post_HD_Serum Cr
- Pre_HD_BUN should be ≥ Post_HD_ BUN
- Total Bilirubin should be ≥ Direct Bilirubin
- Total Protein should be ≥ Serum Alb

By using the mentioned principles in all HD databases, the percentage of noises can be decreased drastically.

Discussion

Real data have usually have abnormalities. It may be caused by human or machine errors, so quality analysis of data is essential to have accurate and cleaned data for research, especially in health databases. Data quality is more important in adopting powerful systems, and poor data quality can affect medical decisions adversely (16). The studies showed that missing handling and improving data quality lead to increased accuracy in the prediction model as compared with real data (9) and poor data quality leads to decreased accuracy in clinical decision support systems (19).

Totally, some of the variables in this study had no outliers and the rest had less than 9.28 percent. The variables' missing percentage was 0-19.73. In Wagner's study (2011), BMI, blood pressure, cholesterol and PTH had missing values over 20 percent(20). In Titapiccolo's study (2012), just lab results had missing values (1-22 percent) and total protein and C-reactive protein had missing values over 20 percent in this study (8). The missing value percentage of variables in the van Diepen's study was an average of 1.9 percent (21). In Floege's study, the missing value of HD variables was 0 to 44 percent (15). All variables in Rhee's study except Creatinine, residual urea clearance and glucose, had missing values less than one percent (12 Studies show that having missing values is an inevitable fact, but reducing these values leads to improved quality of data in healthcare databases.

In the studies, only algorithms and statistical methods were used to determine the outliers and noisy data and handle missing values like using mean substitution for missing handling and mean ±3×SD to detect the outliers of HD patients' data and exclude the variables with missing values over 15 percent to improve the data quality (4). Wagner analyzed the United Kingdom's renal registry for predicting mortality in dialysis patients. They detected the outliers, removed the variables with missing values over 20 percent, and imputed other missing values by the Markov Chain Monte Carlo method (20). In the Somasundaram and Nedunchezhian's study, after comparing 3 methods in missing value imputation, using attribute mean substitution for missing value imputations was recommended (11). Titapiccolo handled missing values by mean substitution in HD patients' database, too (8). Rhee handled missing values by mean or medians to assess the effect of serum sodium on the mortality of HD patients (12). In this study, in addition to popular and recommended methods, the clinical expert opinion, as well as the specific relationships among the variables, has been used to detect the noises, outliers and acceptable ranges for each variable. This hybrid method can increase the accuracy of the study. Also, using the variables' principles can help distinguish the noisy data from the outliers and correct many cases instead of removing them.

The quality problems with SUMS hemodialysis database could be due to new implementation of the database and lack of routine data quality check on the entered data, so quality assessment of this database and reporting the errors can be an effective approach in finding data abnormality types and improving data quality.

Conclusion

Experts' opinion in detecting the outliers as a complement to statistical methods can have an effective role in detection of real outliers. For the analysis of HD databases, the relationship between variables, because of their effect on the quality, could be focused to improve all HD databases' quality. To achieve the practical and clinical principles, all steps in data quality assessment should be done by using clinical expert's idea and data quality should be more focused by administrative and clinical staff.

Acknowledgements

The present article was extracted from Firouraghi’s MSc. thesis in Medical Informatics, Faculty of Management and Medical Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran. The authors would like to thank the Research Vice-Chancellor of Shiraz University of Medical Sciences for financially supporting the research (Contract No. 94-7482), Center for Development of Clinical Research of Nemazee Hospital and Dr. Nasrin Shokrpour for editorial assistance. The present study could not be done without the cooperation of Shiraz University of Medical Sciences Special Diseases Affairs.

Conflict of Interest

None declared.

References

1. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE). 2010;2(02):250-5.
2. Martin GS. The essential nature of healthcare databases in critical care medicine. Critical Care. 2008;12(5):176.
3. Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine: Springer Science & Business Media; 2013.
4. Yeh J-Y, Wu T-H, Tsao C-W. Using data mining techniques to predict hospitalization of hemodialysis patients. Decision Support Systems. 2011;50(2):439-48.
5. Goldman L, Schafer AI. Goldman's Cecil Medicine E-Book: Elsevier Health Sciences; 2011.
6. Titapiccolo JI, Ferrario M, Cerutti S, Barbieri C, Mari F, Gatti E, et al. Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. Expert Systems with Applications. 2013;40(11):4679-86.
7. Arts DG, De Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. Journal of the American Medical Informatics Association. 2002;9(6):600-11.
8. Titapiccolo JI, Ferrario M, Cerutti S, Signorini MG, Barbieri C, Mari F, et al., editors. Mining medical data to develop clinical decision making tools in hemodialysis. Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on; 2012: IEEE.
9. Razavi A, Gill H, Åhlfeldt H, Shahsavar N. A data pre-processing method to increase efficiency and accuracy in data mining. Artificial Intelligence in Medicine. 2005:434-43.
10. Kantardzic M. Data mining: concepts, models, methods, and algorithms: John Wiley & Sons; 2011.
11. Somasundaram R, Nedunchezhian R. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications, Vol21. 2011;21(10).
12. Rhee CM, Ravel VA, Ayus JC, Sim JJ, Streja E, Mehrotra R, et al. Pre-dialysis serum sodium and mortality in a national incident hemodialysis cohort. Nephrology Dialysis Transplantation. 2015;31(6):992-1001.
13. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine. 2005;2(10):e267.
14. Wagner MM, Hogan WR. The accuracy of medication data in an outpatient electronic medical record. Journal of the American Medical Informatics Association. 1996;3(3):234-44.
15. Floege J, Gillespie IA, Kronenberg F, Anker SD, Gioni I, Richards S, et al. Development and validation of a predictive mortality risk score from a European hemodialysis cohort. Kidney international. 2015;87(5):996-1008.
16. Hasan S, Padman R, editors. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. AMIA annual symposium proceedings; 2006: American Medical Informatics Association.
17. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.
18. Cody R, Johnson R. Data cleaning 101. Robert Wood Johnson Medical School, Piscataway, NJ. 2008.
19. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK, editors. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annual Symposium Proceedings; 2005: American Medical Informatics Association.
20. Wagner M, Ansell D, Kent DM, Griffith JL, Naimark D, Wanner C, et al. Predicting mortality in incident dialysis patients: an analysis of the United Kingdom Renal Registry. American Journal of Kidney Diseases. 2011;57(6):894-902.
21. van Diepen M, Schroijen MA, Dekkers OM, Rotmans JI, Krediet RT, Boeschoten EW, et al. Predicting mortality in patients with diabetes starting dialysis. PloS one. 2014;9(3):e89744.