pagepress

Journal of Public Health Research 2021; volume 10:1722

Article

Establishing a sorting protocol for healthcare databases

Elie Ghabi,[1] Wehbeh Farah,[2] Maher Abboud,[2] Elias Chalhoub,[3] Nelly Ziade,[4] Isabella Annesi-Maesano,[5] Laurie Abi Habib,[6] Myriam Mrad Nakhle[6]

[1]Faculty of Medicine, University of Balamand, Lebanon; [2]UEGP, Faculty of Sciences, Saint Joseph University of Beirut, Lebanon; [3]Medical Laboratory Sciences Department, Faculty of Health Sciences, University of Balamand, Lebanon; [4]Faculty of Medicine, Saint Joseph University of Beirut, Lebanon; [5]Institut Pierre Louis d'Epidémiologie et de Santé Publique, Equipe EPAR, Sorbonne Universités, Paris, France; [6]Public Health Department, Faculty of Health Sciences, University of Balamand, Lebanon

Abstract

Background: Health information records in many countries, especially developing countries, are still paper based. Compared to electronic systems, paper-based systems are disadvantageous in terms of data storage and data extraction. Given the importance of health records for epidemiological studies, guidelines for effective data cleaning and sorting are essential. They are, however, largely absent from the literature. The following paper discusses the process by which an algorithm was developed for the cleaning and sorting of a database generated from emergency department records in Lebanon.

Design and methods: Demographic and health related information were extracted from the emergency department records of three hospitals in Beirut. Appropriate categories were selected for data categorization. For health information, disease categories and codes were selected according to the International Classification of Disease 10th Edition.

Results: A total of 16,537 entries were collected. Demographic information was categorized into groups for future epidemiological studies. Analysis of the health information led to the creation of a sorting algorithm which was then used to categorize and code the health data. Several counts were then performed to represent and visualize the data numerically and graphically.

Conclusions: The article describes the current state of health information records in Lebanon and the associated disadvantages of a paper-based system in terms of storage and data extraction. Furthermore, the article describes the algorithm by which health information was sorted and categorized to allow for future data analysis using paper records.

Background

The association between increased air pollutant levels and increased mortality and hospital admissions for respiratory diseases has been well established for the past 70 years. It was first observed following the historical 1952 London smog episode.[1] Moreover, it continues to be observed.[1-11] Recent developments in public health have led to air pollution being declared the leading environmental cause of premature death globally, surpassing poor sanitation and the lack of drinking water.[12] According to the Health Effects Institute (HEI) State of Global Air report, 4.1 million deaths from heart disease, stroke, lung cancer, chronic lung disease and respiratory infections were due to exposure to $PM_{2.5}$ in 2016. This ranks ambient particulate matter concentrations as the 6th leading cause of early death.[13] In the UK, elevated levels of PM_{10} were correlated with higher hospital admissions for respiratory disorders and higher mortality from respiratory disorders.[14] Similar findings were observed in France for SO_2 and PM_{13}.[15] Internationally, chronic $PM_{2.5}$ exposure is linked to cardiorespiratory and neurocognitive disorders as well as diabetes[12,16] and the risk of developing chronic obstructive pulmonary disease[17] and severe lower respiratory tract infections.[18] In utero and childhood exposure to particulate matter may also predispose individuals to infectious and inflammatory disorders[19,20] as well as impaired lung function,[21] especially in asthmatic children.[22]

Similar associations have been found in Lebanon, where increasing concentrations of $PM_{2.5}$ and PM_{10} were found to be positively correlated with increased emergency hospital admissions for respiratory diseases among children.[23] They were also associated with increased rates of emergency hospital admissions for respiratory and cardiovascular diseases among the adult and elderly populations.[23] Moreover, one study showed that geographic residence near industrial factories led to increased exposure to indus-

Significance for public health

In our protocol, we explain the process by which health information collected from paper-based records were analyzed to develop an algorithm to allocate appropriate ICD 10 disease codes and categories to each entry. The algorithm allows for the sequential analysis of the information present and serves to overcome the issue of incomplete records. This is especially relevant in developing countries where paper-based records are highly relied upon, noting that our protocol has already been applied to a study in the Ivory Coast. The protocol allows for a health information database to be cleaned and sorted for research in environmental, social and occupational health. Furthermore, since data cleaning and sorting protocols are sparse, it is crucial to establish a standardized tool to enhance the quality of the conducted research.

trial air pollutants and particular matter and placed children at higher risk of developing respiratory problems than those living farther away (4-7 km).[24] In another study investigating the factors associated with chronic bronchitis, the authors determined that living close to busy roads or local powerplants were among the factors associated with the development of chronic bronchitis.[25]

Understanding the dynamics of pollutant levels and their relationship to health events is essential to provide sound evidence for policies to prevent and mitigate further deleterious health outcomes. In Lebanon, where socio-political circumstances do not always permit new research, the existing abundance of data can lead to substantive evidence for preventive policies.

Many studies have suggested using a Big Data approach to generate evidence for policy making. The earliest application of Big Data analysis was performed by John Snow in the London cholera outbreak of 1854. This example was used by Khoury and Ionnides to argue that, despite numerous challenges, sifting through data to isolate true signals from massive amounts of noise must be performed to translate information into means by which societal health can be improved.[26] Pollutant levels are measured by real-time or continuous monitoring. Novel techniques for predicting air pollution information involves monitoring data obtained from meteorological monitoring, traffic flow, human mobility and road networks.[27,28] Moreover, asthma related emergency department (ED) visits could be predicted from pollutant monitoring, social media posts and search engine queries with 70% precision.[29]

To perform such searches, one would first require that the available data be properly handled to ensure its quality and reliability. As Huang *et al*. coin it, "the quality [of the dataset] determines the upper bound of the data product, *i.e*. garbage in garbage out".[30] Though the integrity of the dataset is essential, little exists regarding the management of data. Data sorting and handling protocols and articles discussing data cleaning and handling have largely been subjects of grey literature.[31] Instead, the literature is saturated with discussions focusing on the role of study design, protocol adherence and investigator experience in determining study validity.[31] Several protocols have been described[32-34] and recommendations have been outlined for database selection[34] and data management,[35] but none have established guidelines for efficient and ethical data cleaning.

In this article, the process of data collection, handling and sorting is outlined. We hope to contribute to the development of a standardized data cleaning and coding protocol that serves to improve the quality of a database generated from paper-based medical records.

Design and methods

Sample selection

Three hospitals with emergency departments located in densely populated areas in the city of Beirut were selected. The city of Beirut was chosen because of the availability of air pollution data.[36] The selected hospitals were chosen due to their proximity to pollutant measuring stations, the presence of an ED and their reception of large volumes of patients. The data sample was selected from a larger database collected for the Beirut Air Pollution and Health Effects (BAPHE) study performed by Nakhlé *et al*.[36]

Data collection

The data was collected from January 1, 2012 till December 31, 2014. The protocol, however, was developed in June 2016. At the time of data collection, Lebanese health institutions relied primarily on paper-based records. ED records for the years 2012, 2013 and 2014 were obtained from the hospitals' archives. IRB to access patient records was obtained from the respective institutions. From each record, the following information was collected by a trained professional: patient's age, sex, date of presentation, chief complaint, differential diagnosis, final diagnosis, medications, and the name of the treating physician. The data collection process is described in detail by Nakhlé *et al*. in another study.[36] The data was collected using Microsoft Excel, inspected by the principal investigator, and validated by two senior physicians/epidemiologists.

Description of the database

The information was sorted into three major categories. Logistic information included the patient's file number, the date of presentation to the ED and the hospital from which the record was obtained. Demographic information included the patient's age and sex. Lastly, health information included the initial complaint, the differential diagnosis, the final diagnosis, admission/discharge status and the administered medications.

The variables chosen for this study are summarized in Table 1.

Description of date of presentation

A patient's date of presentation was entered using a DD/MM/YYYY format, given its versatility in time-series analyses. Sorting of the dates revealed benign clerical errors that were easily corrected. For this study, the dates ranged between 1st January 2012 and 31st December 2014. Patients who presented prior to 1st January 2012 or after 31st December 2014 were excluded from the study. For entries with missing dates, "NA" was entered. Given that the purpose of the study is to code and categorize entries based on health information, entries with missing dates of presentation were not excluded. The remaining information was studied and used to develop the coding and categorization algorithm.

Counts were then performed on a daily, monthly, seasonally, and yearly basis. For seasons, the following definitions were used:

- Winter: 1/1/201X - 20/3/201X
- Spring: 21/3/201X - 20/6/201X
- Summer: 21/6/201X - 20/9/201X
- Autumn: 21/9/201X - 31/12/201X

Categorization by age

Age groups were defined according to the United Nations recommended age groups for studying health, health services and nutrition.[37] When age was not available in the database, the code "NA" was entered.

Table 1. Variables chosen for the study.

File number	Date of presentation to the ER
Hospital	Sex
Age	Age category
Initial complaint	Differential diagnosis
ICD 10 code	Category
Diagnosis	Discharge/admission history
Administered medications	

Categorization by gender

Two genders were described and given appropriate codes for this study. Males were given the code "M" whereas females were given the code "F". Missing values were coded as "NA".

Categorization by disease

Health information in each medical record includes the chief complaint, the differential diagnosis, the final diagnosis and finally management plan. The collected health information was segregated into several categories. First, the chief complaint at presentation was recorded. Second, the differential diagnosis was recorded. When available, the final diagnosis was recorded as well as the medications used to manage the patient and the name of the treating physician. Several disease categories were described based on the WHO International Classification of Diseases 10th edition (ICD10).[38] Furthermore, to better demonstrate the data, disease codes were described for various disease categories. Respiratory diseases, for example, are represented by their ICD10 code range of J00-J99. Based on the literature, general disease categories with their corresponding code ranges as well as specific diseases with their specific ICD 10 codes were described. This was done to better represent these pathologies. For example, bronchitis, asthma and emphysema represent obstructive respiratory diseases, a subset of the general respiratory disease category. To represent obstructive pathologies, the code range J40 - J47 was described and used instead of the code range J00 – J99 which corresponds to the general respiratory disease category.

Entries were initially included based on chief complaints. Chief complaints, however, are nonspecific and may not be reflective of a final diagnosis. Furthermore, since various pathologies may have common presentations, the same chief complaint may be reflective of more than one distinct disorder. Chest pain, for example, is a nonspecific complaint that may be present in cardiac, respiratory, gastrointestinal, musculoskeletal or psychiatric disorders. Thus, entries that initially met the inclusion criteria may ultimately not fit the study objectives and would require reclassification or even exclusion. To represent these entries, the algorithm was developed to include error codes and categories.

Statistical analysis

Descriptive analysis of the data was performed using Microsoft Excel. Counts were established for each variable. Bar graphs were plotted as well. Lastly, code and category counts were performed by day, month, season and year to demonstrate the data over time for future analysis.

Results

Code and category allocation

An ED visit is a brief interview to address a specific complaint. Often, a final diagnosis is not reached. Paper-based medical records suffer from incomplete or missing data. This issue is especially prevalent in Lebanon. Furthermore, due to institutional preferences, classification using ICD10 codes and categories is not universally performed. To address this issue, the collected health information was sequentially analyzed to allocate an appropriate disease code and category for each entry. First, an entry was inspected to determine the completeness of the health information. If a final diagnosis was present, an appropriate code and category was then allocated.

During the initial phases of the study, codes and categories were allocated based on chief complaint. These served as preliminary set of codes and categories. To best represent the entry, the information present per entry was analyzed in its entirety to allocate the most appropriate code and category. This subsequently minimizes errors that may lead to over or underrepresentation or certain categories or codes.

Code and category allocation begin by first determining if a chief complaint is documented. If one were present, then preliminary codes and categories were allocated during the data collection. To ensure the preliminary code and category adequately represents the entry, the remainder of the health information is analyzed beginning with the final diagnosis. If a single diagnosis is present and has been selected for the study, an appropriate code and category are allocated. If these match the preliminary set, no change is made. However, if a mismatch is present, the code and category allocated to the final diagnosis is favored. Error codes and categories are allocated when the final diagnosis is not selected for the study. When multiple diagnoses are present, each diagnosis is addressed individually. Often, the diagnoses belong to the same system, category and code range, thus a single pair is chosen for the entry and compared to the preliminary set as before. If, however, diagnoses belong to the same category but to different code ranges, the code range corresponding to the entire category is selected rather than that of a single diagnosis. If the diagnoses belong to different categories, then a single category or code range cannot represent the entry and so an error code and error category are allocated instead. When a chief complaint is absent, the analysis of the health data proceeds in a similar fashion. The difference is that a preliminary code and category are not present and so no validation is necessary. The diagnosis was assessed, and when a diagnosis was absent, the medications administered were assessed. The algorithm to sequentially assess the information present per entry is demonstrated in Figure 1.

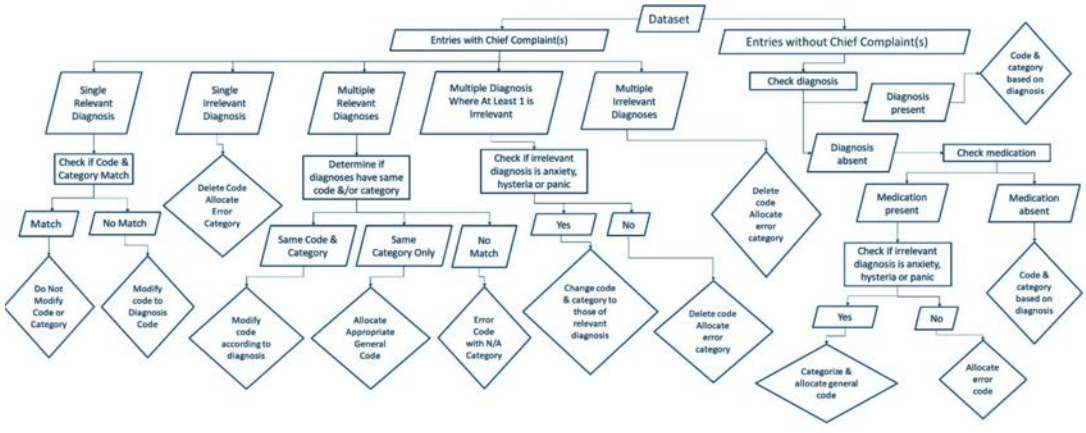

Figure 1. Code and category allocation algorithm.

Counts

After applying the appropriate formats and performing the code and category allocation, counts were performed for the various data categories present in the database. Of the 16,537 entries, 14,940 were found to have relevant codes and categories, 240 entries were marked for revision and 1357 belonged to error categories. However, all entries were allocated a disease category.

Several relative counts were performed and plotted to better visualize the data. Figure 2, which represents the disease category distribution by patient age and sex, shows that pulmonary diseases are the most common pathology encountered except in adult males where cardiac diseases are most common. Cardiac diseases are the second most common pathology, followed by cardiorespiratory diseases in general. These results are consistent with the general count performed for disease category and code (Tables 2 and 3).

Discussion

From the ED registers of the three participating hospitals 16,537 entries were gathered. The records had missing information to varying degrees which posed an initial challenge for proper code and category allocation. The sequential analysis of the available information with a similar framework to that of the ICD10 allowed for the creation of a code and category allocation protocol. Furthermore, inclusion of error values and missing data allowed the identification of issues that might have been encountered during the data collection process. By following the steps outlined in the algorithm, researchers can allocate categories and codes to health information obtained from paper records. This is particularly useful in settings where disease category and code allocation is not performed routinely or automatically.

Pulmonary diseases followed by cardiac diseases were the most encountered disorders in an ED between the years 2012 and 2014. Furthermore, the most prevalent age group was that of persons older than 75 years of age, accounting for 14.87% of the sample population. Pulmonary and cardiac diseases were more prevalent among males than females. Age group 2 which corresponds to ages 1 through 4, showed the highest prevalence of pulmonary diseases. The prevalence of cardiac diseases was found to progressively increase with age, which is well-established finding. Progressively, both cardiac and pulmonary disorders become more

Table 2. Disease and error category counts.

Categories			
Pulmonary	8770	Misc. Musk	387
Cardiac	4159	Misc. Neuro	1
Cutaneous allergy	369	Misc. OBGYN	3
Cardiorespiratory	729	Misc. Psych	595
ENT	83	Misc. Subst. Abuse	12
Cerebrovascular	165	Misc. Surgical	3
N/A	240	Misc. Unidentified	69
Blank	0	Misc. Uro	10
Urticaria	665	Misc. Endo	12
Misc. Accident	26	Misc. Gastro	171
Misc. Allergic	30	Misc. Hemato-Onco	28
Misc. Andro	2	Misc. Immuno	1
Misc. Combined	2	Misc. Infectious	5

Table 3. Disease and error code counts.

Codes					
G00 - G09	0	I63 - I64	155	J90 – J94	69
G45 - G46	10	J00 - J98	393	L50 - L54	362
H60 - H95	83	J00 - J99	3775	T886	7
I00 - I52	953	J12 - J18	1442	N/A	240
I00 - I99	2109	J20 - J22	1125	BLANK	1357
I20 - I24	1779	J40 - J47	597	L50	665
I60 - I62	0	J45 - J46	1416		

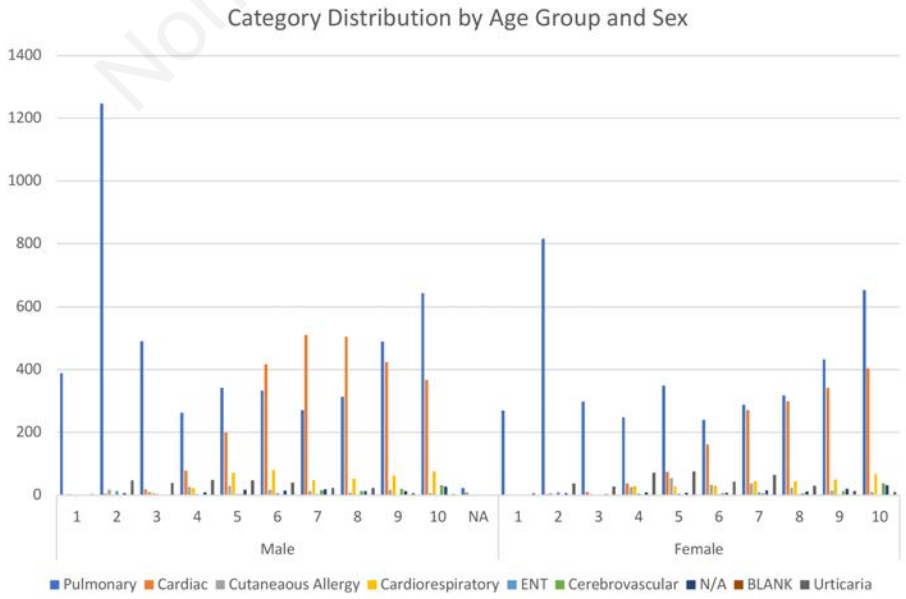

Figure 2. Category distribution by age group and sex.

prevalent with age, with cardiac diseases more so among men than women. The significance of these trends, however, will be explored in future research.

The potential weaknesses of this study chiefly stem from the integrity and completeness of the health records. The more incomplete the records, the less representative the code and category and the less reliable the allocation process becomes. Furthermore, the allocation was performed manually after entries were individually analyzed. The large sample size, however, serves to reduce bias and minimize the impact of errors. The algorithm should also be validated prior to being used for future research. The process of data cleaning was influenced by the general guidelines proposed by Van der Broeck *et al.*[31] Besides this, no guidelines or algorithms for cleaning and sorting health data has been described in the literature. However, various tools that use pattern recognition and machine learning have been devised to automatically allocate ICD codes and categories.[39] When compared to the electronic protocol described by Crammer *et al.*, our protocol is particularly similar to theirs with regards to the sequence of data analysis.[40]

Conclusion

Data handling is crucial in any research study. Few guidelines exist on how a database should be created and how the data should be handled prior to data analysis. Proper sorting of collected information increases the validity and reliability of the data and, subsequently, the data analysis. By describing the steps followed to generate a database from ED records, as well as the algorithm used to sort the collected data, we hope to contribute to the development of heath data sorting guidelines. Furthermore, given the increasing influence of air pollution on health, the rising popularity of big data research and the efficacy of big data research in public health and air pollution studies, proper care should be given to data handling and database creation to increase the validity and reliability of the data analysis, thus leading to better evidence-based public health policies concerning air pollution.

Correspondence: Myriam Mrad, Public Health Department, Faculty of Health Sciences, University of Balamand, P.O. Box 166378, Achrafieh, Beirut 1100 2807, Lebanon.
Tel. +961.1562108 Ext: 5127-5202.
E-mail: myriam.mrad@balamand.edu.lb

Key words: Sorting protocol; data cleaning; database; health record.

Conflict of interest: The authors declare that they have no competing interests.

Contributions: EG, MM, EC, designed the project; WF, MA, MM, NZ, collected the data from emergency department medical records; IAM, EG, analyzed the collected data to develop the sorting protocol presented in the article; MM, EG, LAH, interpreted the results obtained after the protocol was applied to the database. The manuscript has been read, reviewed, and approved by all the involved authors.

Received for publication: 16 December 2020.
Accepted for publication: 15 January 2021.

Journal of Public Health Research 2021;10:1722
doi:10.4081/jphr.2021.1722

References

1. Anderson HR, de Leon AP, Bland JM, et al. Air pollution and daily mortality in London: 1987-92. BMJ 1996;312:665-9.
2. Schwartz J, Marcus A. Mortality and air pollution in London: a time series analysis. Am J Epidemiol 1990;131:185-94.
3. Schwartz J, Dockery DW. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am Rev Respir Dis 1992;145:600-4.
4. Zanobetti A, Schwartz J. The effect of fine and coarse particulate air pollution on mortality: a national analysis. Environ Health Perspect 2009;117:898-903.
5. Filleul L, Rondeau V, Vandentorren S, et al. Twenty five year mortality and air pollution: results from the French PAARC survey. Occup Environ Med 2005;62:453-60.
6. Verhoeff AP, Hoek G, Schwartz J, van Wijnen JH. Air pollution and daily mortality in Amsterdam. Epidemiology 1996:225-30.
7. Hoek G, Brunekreef B, Goldbohm S, et al. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. Lancet 2002;360:1203-9.
8. Katsouyanni K, Touloumi G, Samoli E, et al. Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project. Epidemiology 2001;12:521-31.
9. Touloumi G, Samoli E, Katsouyanni K. Daily mortality and "winter type" air pollution in Athens, Greece--a time series analysis within the APHEA project. J Epidemiol Commun Health 1996;50:s47-51.
10. Sunyer J, Castellsagué J, Sáez M, et al. Air pollution and mortality in Barcelona. J Epidemiol Commun Health 1996;50:s76-80.
11. Spix C, Heinrich J, Dockery D, Schwartz J, et al. Air pollution and daily mortality in Erfurt, east Germany, 1980-1989. Environ Health Perspect 1993;101:518-26.
12. Kelly FJ, Fussell JC. Air pollution and public health: emerging hazards and improved understanding of risk. Environ Geochem Health 2015;37:631-49.
13. Health Effects Institute [Internet]. State of Global Air 2018. Accessed: 2018 May 21]. Available from: https://www.stateofglobalair.org/sites/default/files/soga-2018-report.pdf
14. Wordley J, Walters S, Ayres JG. Short term variations in hospital admissions and mortality and particulate air pollution. Occup Environ Med 1997;54:108-16.
15. Dab W, Medina S, Quenel P, et al. Short term respiratory health effects of ambient air pollution: results of the APHEA project in Paris. J Epidemiol Commun Health 1996;50:s42-6.
16. Kelly FJ, Fussell JC. Health effects of airborne particles in relation to composition, size and source. In: FJ Kelly, JC Fussell, editors. Airborne Particulate Matter: Sources, atmospheric processes and health. London: Royal Society of Chemistry; 2016. p. 344-82.
17. Grigg J. Particulate matter exposure in children: relevance to chronic obstructive pulmonary disease. Proc Am Thorac Soc 2009;6:564-9.
18. Grigg J. Air pollution and children's respiratory health–gaps in the global evidence. Clin Experiment Allergy 2011;41:1072-5.
19. Latzin P, Röösli M, Huss A, et al. Air pollution during pregnancy and lung function in newborns: a birth cohort study. Eur Respir J 2009;33:594-603.
20. Jedrychowski WA, Perera FP, Spengler JD, et al. Intrauterine exposure to fine particulate matter as a risk factor for increased susceptibility to acute broncho-pulmonary infections in early childhood. Int J Hygiene Environ Health 2013;216:395-401.

21. Mortimer K, Neugebauer R, Lurmann F, et al. Air pollution and pulmonary function in asthmatic children: effects of prenatal and lifetime exposures. Epidemiology 2008:550-7.
22. Morales E, Garcia-Esteban R, de la Cruz OA, et al. Intrauterine and early postnatal exposure to outdoor air pollution and lung function at preschool age. Thorax 2015;70:64-73.
23. Nakhlé MM, Farah W, Ziade N, et al. Short-term relationships between emergency hospital admissions for respiratory and cardiovascular diseases and fine particulate air pollution in Beirut, Lebanon. Environ Monitor Assess 2015;187:196.
24. Kobrossi R, Nuwayhid I, Sibai AM, et al. Respiratory health effects of industrial air pollution on children in North Lebanon. Int J Environ Health Res 2002;12:205-20.
25. Salameh P, Salame J, Khayat G, et al. Exposure to outdoor air pollution and chronic bronchitis in adults: a case-control study. Int J Occup Environ Med 2012;3:165-77.
26. Khoury MJ, Ioannidis JP. Big data meets public health. Science 2014;346:1054-5.
27. Zheng Y, Liu F, Hsieh HP. U-air: When urban air quality inference meets big data. In: Proceedings 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago; 2013. p. 1436-44.
28. Zheng Y, Chen X, Jin Q, et al. A cloud-based knowledge discovery system for monitoring fine-grained air quality. MSR-TR-2014-40, |Microsoft Research. 2014.
29. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 2015;19:1216-23.
30. Huang T, Lan L, Fang X, et al. Promises and challenges of big data computing in health sciences. Big Data Res 2015;2:2-11.
31. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2005;2:e267.
32. Winkler WE. Data cleaning methods. Proceedings ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington DC, 2003. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.2066&rep=rep1&type=pdf
33. Loureiro A, Torgo L, Soares C. Outlier detection using clustering methods: a data cleaning application. Proceedings of KDNet Symposium on Knowledge-based systems for the Public Sector, 2004.
34. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf 2012;21:1-0.
35. Borer ET, Seabloom EW, Jones MB, Schildhauer M. Some simple guidelines for effective data management. Bull Ecol Soc Am 2009;90:205-14.
36. Nakhlé MM, Farah W, Ziade N, et al. Beirut air pollution and health effects-BAPHE study protocol and objectives. Multidiscip Respir Med 2015;10:21.
37. United Nations, Department of International Economic and Social Affairs. Provisional Guidelines on Standard International Age Classifications, Statistical Paper Series M, No. 74. 1982- Available from: https://unstats.un.org/unsd/publication/SeriesM/SeriesM_74e.pdf
38. WHO. International statistical classification of diseases and related health problems. Geneva: World Health Organization; 2004.
39. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;47:128-44.
40. Crammer K, Dredze M, Ganchev K, et al. Automatic code assignment to medical text. Proceedings of the Workshop on BioNLP 2007: Biological, translational, and clinical language processing. Stroudsburg: Association for Computational Linguistics. p. 129-36.