Establishing A Sorting Protocol for The Data Cleaning of Health Record Databases for Eco-Epidemiological Studies: The Case of Beirut

Elie Ghabi
University of Balamand Faculty of Medicine and Medical Sciences https://orcid.org/0000-0002-9026-9681
Wehbeh Farah
Saint Joseph University of Beirut
Maher Abboud
Saint Joseph University of Beirut
Elias Chalhoub
University of Balamand Faculty of Health Sciences
Nelly Ziade
Saint Joseph University of Beirut
Isabella Annesi-Maesano
Institut Pierre Louis d'Epidemiologie et de Sante Publique
Laurie Abi Habib
University of Balamand Faculty of Health Sciences
Myriam Mrad (myriam.mrad@balamand.edu.lb)
https://orcid.org/0000-0003-0030-0700

Research article

Keywords: Medical Record, Categorization, Sorting, Algorithm

Posted Date: September 19th, 2019

DOI: https://doi.org/10.21203/rs.2.14594/v1

Abstract

Background Health information records in many countries, especially developing countries, still rely on a paper-based system which, when compared to electronic systems, are disadvantageous in terms of storage and data extraction. Given the importance of health records as a data resource for epidemiological studies, guidelines for systematic data cleaning and sorting are essential, yet are largely absent in the literature. This paper discusses the process by which an electronic database was generated from emergency department registers in Lebanon and the data subsequently cleaned, sorted, and categorized.

Methods Demographic and health complaint-related information was extracted from emergency department registers of a convenience sample of seven hospitals in Beirut. Appropriate categories were selected for data categorization. For health-complaint related 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 appropriate categories and groups as required for future epidemiological studies. Analysis of the health information allowed for the creation of a sorting algorithm which then used to categorize and code the heath data. Several counts were then performed to represent and visualize the data numerically and graphically to aid in data interpretation.

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 and subsequent analysis. 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, following the historical 1952 London smog episode [1], and subsequent worldwide observations [1–7]. 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 [11]. 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, ranking ambient air particulate matter as the 6th leading cause of early death [12]. Internationally, air pollutant exposure, particularly chronic $PM_{2.5}$ exposure, has been linked to cardiorespiratory and neurocognitive disorders as well as diabetes [11] and future risk of chronic obstructive pulmonary disease [15], while in utero exposure to particulate matter may also predispose individuals to infectious and inflammatory disorders [13,14].

Similar associations have been found in Lebanon, where increases in $PM_{2.5}$ and PM_{10} are found to be positively correlated with increased emergency hospital admissions for respiratory diseases among children [8], and increased rates of emergency hospital admissions for respiratory and cardiovascular diseases among the adult and elderly populations [8]. Some studies in Lebanon have supported views of the association between air pollution and respiratory problems; one showed that geographic residence near to industrial factories, and therefore increased exposure to industrial air pollutants and particular matter, placed children at higher risk of developing respiratory problems than those living farther away (4–7km) [9], while another study investigating the factors associated with chronic bronchitis determined that living close to busy roads or local powerplants, and therefore exposure to outdoor air pollutants, are among the factors found to be associated with the development of chronic bronchitis [10].

It has become essential to understand the dynamics of pollutant levels and their relation to health events to provide a sound evidence base for policies to prevent and mitigate further deleterious health outcomes. In Lebanon, where socio-political circumstances do not always permit new research, there are abundant data sets available already, which, through the application of "Big Data" analysis, can lead to substantive evidence for preventive policies.

Many studies have suggested the utility of 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 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 [16]. While traditionally, pollutant levels have been determined by real-time or continuous monitoring, it has been possible to predict air pollution information comparable to the findings of traditional techniques by combining pollutant monitoring data with data obtained from meteorological monitoring, traffic flow, human mobility and road networks, [17,18]. Another study has showed that asthma related emergency department (ED) visits could be predicted from pollutant monitoring, social media posts and search engine queries with 70% precision [19].

Performing such studies first requires 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" [20]. Though the integrity of the dataset is essential, little discussion has occurred regarding data sorting and handling protocols and articles discussing data cleaning and handling have largely been subjects of grey literature [21]. Instead, the literature itself is saturated with discussions focusing on the roles of study design, protocol adherence and investigator experience in determining study validity [21]. Several protocols have been described [22–24] and recommendations have been outlined for database selection [24] and data management [25] but none have established guidelines for efficient and ethical data cleaning.

In this article, the process by which a body of data was collected, handled and sorted is outlined. It is hoped that this could contribute towards establishing protocols by which entries of healthcare information in databases are allocated disease categories and codes following sequential analysis of existent information to allow for future data analysis.

Methods

Sample Selection

A convenience sample of 3 hospitals with emergency departments located in densely populated areas in the city of Beirut was chosen. The city of Beirut was chosen because of the availability of air pollutant level monitoring performed in previous research studies [26]. The selected hospitals, Hotel Dieu de France (HDF), Saint Georges University Medical Center (SGHUMC) and Makassed Hospital, were chosen based on their location close to where pollutant monitoring was performed, the presence of an ED and their reception of large volumes of patients. The sample was selected from a larger existing sample collected for the Beirut Air Pollution and Health Effects (BAPHE) study performed by Nakhle et al.

Data Collection

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 approvals to access patient records were obtained from the respective institutions. From each record, the following information was collected by trained hospital residents: patient's age, sex, date of presentation, chief complaint, differential diagnosis, final diagnosis, medications, and the name of the treating physician. This process is explained in detail in Nakhle et al.'s study entitled Beirut Air Pollution and Health Effects—BAPHE study protocol and objectives [26]. The data was then entered into Microsoft Excel, inspected by the principal investigator, and validated by two senior physicians/epidemiologists who were part of the research team.

Description of the Database

It was decided to sort the information into three major groups from hospital records identified as relevant by the study design. Logistic information contained the appropriate information required to identify the patient's admissions record. This included the patient's file number, the date of presentation to the ED and the hospital from which the record was obtained. Demographic information was also collected and included the patient's age and sex. Lastly, health information was extracted from the hospital record. The information extracted included the initial complaint, the differential diagnosis, the final diagnosis, admission/discharge history and the administered medications.

Table 1 summarizes the variables chosen for this study.

Description of Date of Presentation

The patient's date of presentation was entered using the format DD/MM/YYYY, given the versatility of this format in time-series analysis. Sorting through the dates present in the database revealed benign clerical errors that were easily corrected. For this study, the dates chosen range between 1/1/2012 and 31/12/2014. Patients who presented prior to 1/1/2012 or after 31/12/2014 were not included in 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 present health-related information, entries that lacked a date of presentation were not excluded The remainder of the information was studied and used to develop the coding and categorization algorithm.

To demonstrate when patients present themselves to the ED, counts were 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

(X being the appropriate integer to be inserted based on the relevant years and dates selected for the study)

Categorization by Age

As described in Table 2, age groups were defined according to the WHO recommended age groups for studying health, health services and nutrition [27]. When age was not available in the database, the code “NA” was entered.

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

Generally speaking, health-related information is recorded in each medical record sequentially starting with the chief complaint to finally reach a diagnosis and management plan. Similarly, to how a health record contains and categorizes existing data, the extracted health-related information is segregated into several categories. First, the chief complaint of the patient during is recorded as documented in the record. Second, the differential diagnosis based on the given presentation is recorded. When available, the final diagnosis is recorded as well as the medications used to manage the patient at the time of presentation and the name of the treating physician.

Several pathologies were found to be associated with increases in air pollutants, chiefly cardiovascular, cerebrovascular and respiratory pathologies as well as skin and cutaneous allergies. Therefore, based on the existing literature and associations, several disease categories are described relying on the International Classification of Diseases 10th edition (ICD10) [28]. Furthermore, to better communicate the data, disease codes are selected to reflect the various categories. Respiratory diseases, for example, are represented by their ICD10 code range of J00—J99.

Not only have general disease categories been found to be associated with rises in air pollutant levels but also specific pathologies such as asthma, bronchitis, emphysema, urticaria, etc. To better represent these pathologies, their codes were used instead of using the broader disease category and code range. For example, bronchitis, asthma and emphysema represent obstructive respiratory diseases, a subset of the more general respiratory disease category. To represent obstructive pathologies, the code range J40—J47 is used rather than J00—J99 which is the general code range for all respiratory pathologies.

Entries were selected for inclusion based on their chief complaints. For the purposes of the study, a chief complaint was taken to mean a statement the patient reports to describe his symptoms, condition or reason to seek medical attention. Ultimately, a chief complaint is nonspecific and may or may not be reflective of the final diagnosis. Furthermore, since various pathologies may have common symptoms, the same chief complaint may be reflective of more than one distinct pathologic entity. Chest pain, for example, is a nonspecific complaint that may be present in cardiac, respiratory, gastrointestinal, musculoskeletal or psychiatric pathology. Given the non-specific nature of chief complaints, several entries could erroneously be deemed eligible. To represent these erroneously collected entries, the algorithm was developed to include error codes and categories.

Tables 3 and 4 summarize the various disease categories and codes used in this study.

Statistical Analysis

Descriptive analysis of the dataset was performed using Microsoft Excel. Counts were established for each variable and bar graphs plotted. Several graphs were also plotted to demonstrate certain variables versus others for subsequent analysis of possible correlations between the variables. 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 that addresses a specific complaint. Often, a final diagnosis is not reached, and definitive management is not provided. Rather, a differential diagnosis is described, and further investigation is recommended. An issue with paper-based records in Lebanon is the incompleteness of the record. A record may be missing the administered medications, the final diagnosis, or the differential diagnosis. Furthermore, due to institutional differences, classification by ICD10 code and category is not universally performed. To address the gaps caused by this, the collected health information was analyzed sequentially to allocate a disease code and category for each entry in the database. First, an entry was inspected to determine the completeness of the health information. If a final diagnosis was present, the appropriate code and category was allocated.

During the initial phases of the study, codes and categories were allocated to each entry based on the chief complaint. Given how a chief complaint may be common for several pathologies, the preliminary set of codes and categories may overrepresent certain pathologies in favor of others. Moreover, a final diagnosis may be documented as a general category of disease. This, compounded with the issue of record completeness, presents a challenge for code and category allocation. To overcome this challenge, the entirety of the health data present in each entry was analyzed to allocate the appropriate code and category and minimize errors, as outlined below.

The process of code and category allocation begins by determining if the entry had a documented chief complaint or not. If one was present, then a preliminary code and category was allocated during the process of data entry. To assess the adequacy and representability of the code and category, the remainder of the health information is analyzed beginning with the final diagnosis. If a single diagnosis is present and belongs to the disease categories chosen for the study, the appropriate code and category were selected. If these match the preliminary pair, no change is made. However, if a mismatch is present, the set pertaining to the final diagnosis is favored. If the final diagnosis does not belong to a relevant disease category, error codes and categories were chosen. To address the presence of multiple diagnoses, the relevance of each is assessed and irrelevant diagnoses are disregarded. Often, the diagnoses belong to the same system and 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 representative of 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 category is selected instead.

For entries that lacked a chief complaint, the analysis of the health data proceeded in a similar fashion. The difference is that a preliminary code and category were not present and so no validation was necessary. The diagnosis was assessed, and when a diagnosis was absent, the medications administered were assessed to determine which disease category was being addressed. Then the relevance of the category was assessed prior to allocating a disease code and category.

The sequence of steps used for code and category allocation are described in the following algorithm.

Figure 1: code and category allocation algorithm

Counts

Everything that did not fit into algorithm was classified under the error category. 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, 14940 entries were found to have relevant codes and categories, 240 entries were marked for revision and 1357 belonged to error groups.

All entries were allocated a disease category, however. The 1597 entries, which make up less than 1% of the sample population, that did not receive a disease code and were left blank are the entries that were given various error categories.

Several relative counts were performed and plotted to better visualize the data. The following graph, which represents the disease category distribution by patient age and sex, shows that pulmonary disease is the most common pathology across age groups and sex except in adult males where cardiac pathology is the most common. Cardiac pathology is the second most common pathology, followed by cardiorespiratory disease in general. These results are consistent with the general count performed for disease category and diseases code. Tables 5, 6, 7 and 8 as well as Figure 2 demonstrate these results.

Discussion

16,537 entries were gathered from the ED registers of the three participating hospitals. The records had varying levels of completeness which posed an initial challenge of code and category allocation. Sequentially analyzing 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 permitted the identification of issues that might arise during the data collection process, particularly the inclusion of entries based on their chief complaints rather than their final diagnosis, an issue that was addressed during data entry and code and category allocation. By following the steps outlined in the algorithm, researchers can categorize and code health information obtained from paper records in a setting where disease code and category allocation is not performed by the electronic health record automatically.

The database reveals that for the years 2012 through 2014, ED visits in the sample were chiefly for pulmonary disorders followed by cardiac disorders. Furthermore, the most prevalent age of the sample was that of older persons over the age of 75 making up 14.87% of the sample population. Pulmonary and cardiac pathologies 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, well-established finding. Progressively, both cardiac and pulmonary disorders become more prevalent with age, with cardiac disease more so among men than women. The significance of these differences and trends, however, will be explored in further 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. Validation of the algorithm needs to occur prior to recommending its use for future research. The process by which the data was cleaned is influenced by the general guidelines proposed by Van der Broeck, Cunningham, Eeckels and Herbst [21]. To our knowledge, no study describes a protocol by which health information is manually analyzed for code and category allocation. However, various tools that use pattern recognition and machine learning have been developed to automatically allocate ICD codes and categories [29]. When compared to our protocol, a similar sequence of data analysis is realized, particularly similar to the protocol described by Crammer, Dredze, Ganchev and Talukdar [30]. In their protocols, information is sequentially analyzed in a similar fashion in order to code and categorize the obtained information.

Conclusion

Data handling is a crucial step in a research study. Few guidelines exist regarding how a database should be created and how the data should be handled in preparation for data analysis. Sorting of the information is a critical step that, when performed properly, increases the validity and reliability of the dataset and subsequently the analysis. By describing the steps by which a database was generated from ED records, and the algorithm used to sort health information through disease code and category allocation, we hope to add to the efforts of establishing guidelines by which a database is created and by which health data is sorted. The steps used and the algorithm that was generated are tools that could be sued for future researchers faced with the similar task of extracting data from paper health records. 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, thereby leading to better evidence based public health policies concerning air pollution.

Abbreviations

HEIHeath Effects Institute

EDEmergency Department

HDFHotel Dieu de France

SGHUMCSaint Georges Hospital University Medical Center

BAPHEBeirut Air Pollution and Health Effects

ICD10International Classification of Diseases 10th Edition

Declarations

Ethics Approval and Consent to Participate

An IRB approval from the Saint George University Medical Center research committee was obtained to access the required medical health records to collect the data present in the database.

Consent for Publication

Not Applicable

Availability of Data and Material

The dataset used and/or analyzed during the current study is available from the corresponding author on reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Funding

No funding was provided to carry out this project.

Authors' Contributions

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

Acknowledgements

Not applicable.

References

1. Anderson HR, de Leon AP, Bland JM, Bower JS, Strachan DP. Air pollution and daily mortality in London: 1987-92. Bmj. 1996 Mar 16;312(7032):665-9. 2. Schwartz J, Marcus A. Mortality and air

pollution in London: a time series analysis. American journal of epidemiology. 1990 Jan 1;131(1):185-94. 3. Schwartz J, Dockery DW. Increased mortality in Philadelphia associated with daily air pollution concentrations. American review of respiratory disease. 1992 Mar;145(3):600-4. 4. Zanobetti A, Schwartz J. The effect of fine and coarse particulate air pollution on mortality: a national analysis. Environmental health perspectives. 2009 Feb 13;117(6):898-903. 5. Filleul L, Rondeau V, Vandentorren S, Le Moual N, Cantagrel A, Annesi-Maesano I, Charpin D, Declercq C, Neukirch F, Paris C, Vervloet D. Twenty five year mortality and air pollution: results from the French PAARC survey. Occupational and Environmental Medicine. 2005 Jul 1;62(7):453-60. 6. Hoek G, Brunekreef B, Goldbohm S, Fischer P, van den Brandt PA. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. The lancet. 2002 Oct 19;360(9341):1203-9. 7. Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, Monopolis Y, Rossi G, Zmirou D, Ballester F, Boumghar A, Anderson HR. 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 Sep 1;12(5):521-31. 8. Nakhlé MM, Farah W, Ziade N, Abboud M, Salameh D, Annesi-Maesano I. Short-term relationships between emergency hospital admissions for respiratory and cardiovascular diseases and fine particulate air pollution in Beirut, Lebanon. Environmental monitoring and assessment. 2015 Apr 1;187(4):196. 9. Kobrossi R, Nuwayhid I, Sibai AM, El-Fadel M, Khogali M. Respiratory health effects of industrial air pollution on children in North Lebanon. International journal of environmental health research. 2002 Sep 1;12(3):205-20. 10. Salameh P, Salame J, Khayat G, Akhdar A, Ziadeh C, Azizi S, Khoury F, Akiki Z, Nasser Z, Abbass LA, Saadeh D. Exposure to outdoor air pollution and chronic bronchitis in adults: a case-control study. Int J Occup Environ Med (The IJOEM). 2012 Aug 22;3(4 October). 11. Kelly FJ, Fussell JC. Air pollution and public health: emerging hazards and improved understanding of risk. Environmental geochemistry and health. 2015 Aug 1;37(4):631-49. 12. Health Effects Institute. State of Global Air 2018 [Internet]. stateofglobalair.org. [cited 2018May21]. Available from: https://www.stateofglobalair.org/sites/default/files/soga-2018-report.pdf 13. Latzin P, Röösli M, Huss A, Kuehni CE, Frey U. Air pollution during pregnancy and lung function in newborns: a birth cohort study. European Respiratory Journal. 2009 Mar 1;33(3):594-603. 14. Jedrychowski WA, Perera FP, Spengler JD, Mroz E, Stigter L, Flak E, Majewska R, Klimaszewska-Rembiasz M, Jacek R. Intrauterine exposure to fine particulate matter as a risk factor for increased susceptibility to acute broncho-pulmonary infections in early childhood. International journal of hygiene and environmental health. 2013 Jul 1;216(4):395-401. 15. Grigg J. Particulate matter exposure in children: relevance to chronic obstructive pulmonary disease. Proceedings of the American Thoracic Society. 2009 Dec 1;6(7):564-9. 16. Khoury MJ, Ioannidis JP. Big data meets public health. Science. 2014 Nov 28;346(6213):1054-5. 17. Zheng Y, Liu F, Hsieh HP. U-air: When urban air quality inference meets big data. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 2013 Aug 11 (pp. 1436-1444). ACM. 18. Zheng Y, Chen X, Jin Q, Chen Y, Qu X, Liu X, Chang E, Ma WY, Rui Y, Sun W. A cloud-based knowledge discovery system for monitoring fine-grained air quality. MSR-TR-2014–40, Tech. Rep.. 2014. 19. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE journal of biomedical and health informatics. 2015 Jul;19(4):1216-23. 20. Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and challenges of big data computing in health sciences. Big Data Research.

2015 Mar 1;2(1):2-11. 21. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine. 2005 Sep 6;2(10):e267. 22. Winkler WE. Data cleaning methods. InProc ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation 2003 Aug. 23. Loureiro A, Torgo L, Soares C. Outlier detection using clustering methods: a data cleaning application. InProceedings of KDNet Symposium on Knowledge-based systems for the Public Sector 2004. Springer. 24. Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, Casale RL. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiology and drug safety. 2012 Jan;21(1):1-0. 25. Borer ET, Seabloom EW, Jones MB, Schildhauer M. Some simple guidelines for effective data management. The Bulletin of the Ecological Society of America. 2009 Apr;90(2):205-14. 26. Nakhlé MM, Farah W, Ziade N, Abboud M, Coussa-Koniski ML, Annesi-Maesano I. Beirut Air Pollution and Health Effects-BAPHE study protocol and objectives. Multidisciplinary respiratory medicine. 2015 Dec;10(1):21. 27. Department of International Economic and Social Affairs. Provisional Guidelines on Standard International Age Classifications. 28. World Health Organization. International statistical classification of diseases and related health problems. World Health Organization; 2004. 29. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics. 2008;17(01):128-44. 30. Crammer K, Dredze M, Ganchev K, Talukdar PP, Carroll S. Automatic code assignment to medical text. InProceedings of the workshop on bionlp 2007: Biological, translational, and clinical language processing 2007 Jun 29 (pp. 129-136). Association for Computational Linguistics.

Tables

Table 1: variables chosen for the study

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	

Table 2: age groups

Group	Age Range	Code
Group 1	< 1	1
Group 2	1 - 4	2
Group 3	5 - 14	3
Group 4	15 - 24	4
Group 5	25 - 34	5
Group 6	35 - 44	6
Group 7	45 – 54	7
Group 8	55 – 64	8
Group 9	65 – 74	9
Group 10	75+	10
Group 11	NA	NA

Table 3: disease codes

Codes		
G00 - G09	G45 - G46	J90 - J94
H60 - H95	I00 - I52	L50 - L54
I00 - I99	I20 - I24	T886
I60 - I62	I63 - I64	0
J00 - J98	J00 - J99	L50
J12 - J18	J20 - J22	N/A
J40 - J47	J45 - J46	BLANK

Table 4: disease categories

Categories				
Pulmonary	Cardiac	Cutaneous Allergy	ENT	Cardiorespiratory
Cerebrovascular	Urticaria	N/A	Blank	Misc. Accident
Misc. Allergic	Misc. Andro	Misc. Combined	Misc. Endo	Misc. Gastro
Misc. Hemato-Onco	Misc. Infectious	Misc. Musk	Misc. Neuro	Misc. OBGYN
Misc. Psych	Misc. Subst. Abuse	Misc. Surgery	Misc. Unidentified	Misc. Uro

Table 5: 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 6: 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		

Table 7: gender counts

Genders	
M	9175
F	7320
Missing	0
NA	40

Table 8: age group counts

Age Groups	
1	681
2	2239
3	948
4	1080
5	1622
6	1728
7	1872
8	1793
9	2013
10	2460
NA	101
Total	16537

Figures

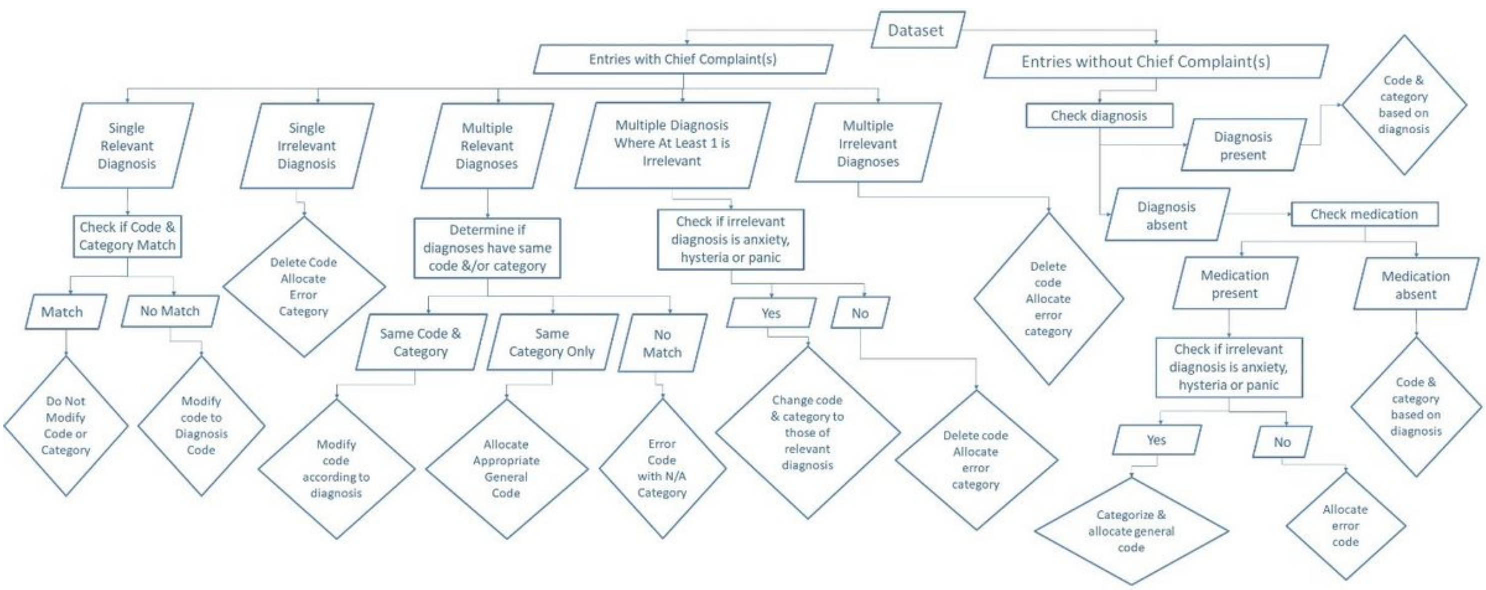

Figure 1

code and category allocation algorithm

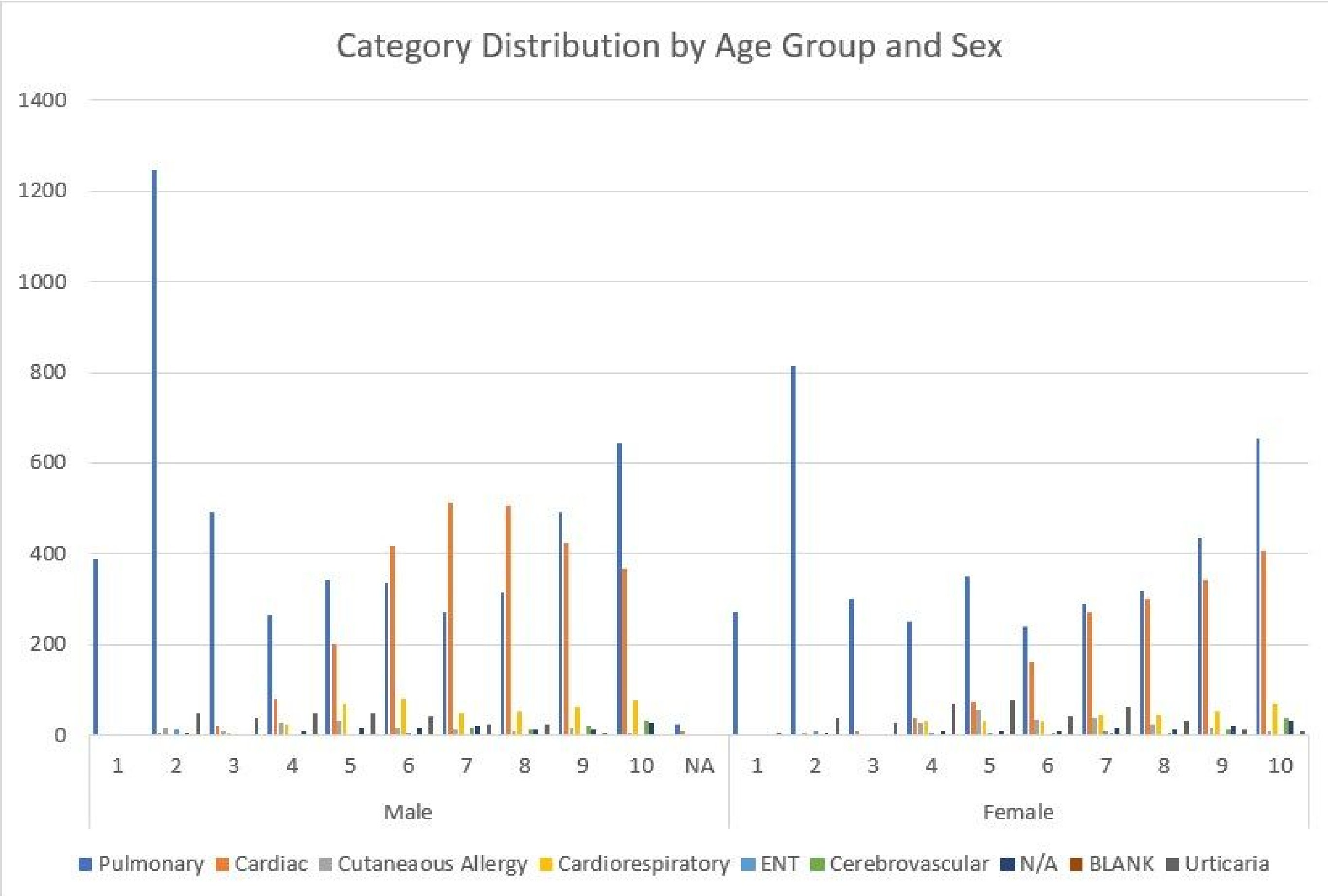

Figure 2

category distribution by age group and sex