



OPEN ACCESS

EDITED BY

Jacques Demongeot,
Université Grenoble Alpes, France

REVIEWED BY

Olumide Babatope Longe,
Academic City University
College, Ghana
Mustapha Rachdi,
Université Grenoble Alpes, France

*CORRESPONDENCE

Sana S. BuHamra
sana.buhamra@ku.edu.kw

SPECIALTY SECTION

This article was submitted to
Infectious Diseases: Epidemiology and
Prevention,
a section of the journal
Frontiers in Public Health

RECEIVED 15 October 2022

ACCEPTED 14 November 2022

PUBLISHED 01 December 2022

CITATION

BuHamra SS, Almutairi AN,
Buhamrah AK, Almadani SH and
Alibrahim YA (2022) An NLP tool for
data extraction from electronic health
records: COVID-19 mortalities and
comorbidities.

Front. Public Health 10:1070870.
doi: 10.3389/fpubh.2022.1070870

COPYRIGHT

© 2022 BuHamra, Almutairi,
Buhamrah, Almadani and Alibrahim.
This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

An NLP tool for data extraction from electronic health records: COVID-19 mortalities and comorbidities

Sana S. BuHamra^{1*}, Abdullah N. Almutairi¹,
Abdullah K. Buhamrah², Sabah H. Almadani¹ and
Yusuf A. Alibrahim³

¹Department of Information Science, Kuwait University, Kuwait City, Kuwait, ²Surgery Department, Al-Adan Hospital, Al Ahmadi, Kuwait, ³Department of Medical Imaging, Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Background: The high infection rate, severe symptoms, and evolving aspects of the COVID-19 pandemic provide challenges for a variety of medical systems around the world. Automatic information retrieval from unstructured text is greatly aided by Natural Language Processing (NLP), the primary approach taken in this field. This study addresses COVID-19 mortality data from the intensive care unit (ICU) in Kuwait during the first 18 months of the pandemic. A key goal is to extract and classify the primary and intermediate causes of death from electronic health records (EHRs) in a timely way. In addition, comorbid conditions or concurrent diseases were retrieved and analyzed in relation to a variety of causes of mortality.

Method: An NLP system using the Python programming language is constructed to automate the process of extracting primary and secondary causes of death, as well as comorbidities. The system is capable of handling inaccurate and messy data, this includes inadequate formats, spelling mistakes and mispositioned information. A machine learning decision trees method is used to classify the causes of death.

Results: For 54.8% of the 1691 ICU patients we studied, septic shock or sepsis-related multiorgan failure was the leading cause of mortality. About three-quarters of patients die from acute respiratory distress syndrome (ARDS), a common intermediate cause of death. An arrhythmia (AF) disorder was determined to be the strongest predictor of intermediate cause of death, whether caused by ARDS or other causes.

Conclusion: We created an NLP system to automate the extraction of causes of death and comorbidities from EHRs. Our method processes messy and erroneous data and classifies the primary and intermediate causes of death of COVID-19 patients. We advocate arranging the EHR with well-defined sections and menu-driven options to reduce incorrect forms.

KEYWORDS

natural language processing, text mining, information extraction, SARS-CoV-2, mortality, decision tree, prediction

Introduction

The COVID-19 pandemic has had a significant impact on how and where healthcare is delivered effectively and efficiently. During the pandemic, the need for novel and current technologies arise to assist in predicting clinical outcomes in critical time with the high overflow of patients. Clinical (text) notes constitute a major source of medical data and are rarely used to their full capacity, even though they include a wealth of subjective information. Prior to electronic health records (EHRs), practitioners had to manually collect data from clinical notes, which was costly and difficult to scale up. Despite the expanding volumes of healthcare data, Kong (1) claims that over 80% of text, image, signal, and other medical data collections remain unstructured and unused. One main goal in medical research is to use EHRs to extract and analyze well-structured data. Many methods were devised and evaluated using EHRs for detecting patients with known risk factors for consequences such as stroke and significant bleeding (2), as well as investigating the difficulties of decoding and comprehending clinical narratives (3). Natural language processing (NLP) can expedite diagnosis and care to patients who are most vulnerable during pandemics by using textual data from medical records. According to Zhou et al. (4), only NLP can extract information about a patient's family history from free-text clinical papers. The researchers employed word embeddings and a Convolutional Neural Network (CNN) to recognize International Classification of Diseases (ICD-10) diagnostic codes in discharge notes and outperformed current methods with little data preparation (5).

Artificial Intelligence (AI) and Machine Learning (ML) technologies including NLP can be used to aid in the diagnosis and treatment of individuals suffering from acute and chronic diseases during the COVID-19 pandemic. DeCapprio et al. (6) used medical records that had already been made public as COVID-19 proxies (pneumonia, influenza, acute bronchitis, and upper respiratory illnesses). Zoabi et al. (7) came up with a machine learning decision tree model that predicts a positive COVID-19 infection in an RT-PCR test during the first month of the pandemic. Izquierdo et al. (8) used a mix of traditional epidemiological methods, NLP, and ML predictive modeling to find out what symptoms COVID-19 patients have that make them likely to be admitted to the ICU. Guan et al. (9) employed simple-tree XGBoost to identify high-risk COVID-19 cases and assessed how much faster causes of death may be identified using minimally preprocessed notes.

This study intends to construct an NLP system to automate the extraction of primary and secondary causes of death, as well as comorbidities, from the mortality EHRs of COVID-19 patients admitted to the ICU in Kuwait during the pandemic. Since many of the free-text notes were inadequately formatted, contained spelling mistakes and were placed in the wrong field, acquiring sufficient and reliable data was the largest hurdle. In fact, the causes of death in most records in our data were not

expressed precisely nor was in the correct field although the EHRs file is mortality specific.

Other work in the literature used available clean EHRs for their analysis. However, EHRs may sometimes be inaccurate and noisy due to them being compiled under extreme pressures of time and manpower due to the large influx of patients with critical cases, such as the case during the pandemic. EHRs need to be first corrected and cleaned to be used for proper analysis or be used in medical systems such the Unified Medical Language System (UMLS) and SNOMED CT. Otherwise, a significant amount of information will be lost.

To correct the EHRs we used physicians as the domain knowledge experts to understand and extract the common mistakes in the EHRs that were done by their fellow physicians. Their knowledge and findings were converted to a Python language code to automate cleaning and fixing the data in the EHRs. Also, the Python code used the domain expert knowledge to distinguish between acute diseases and causes of death in some circumstances. In addition, the causes of death were classified to a direct cause or a related one. Comorbidities were used as an important factor in analyzing the cause of death. This will offer precise information on the casualty and spectrum of comorbidities in fatal instances, allowing for an accurate evaluation of COVID-19's hazardous nature. Finally, we have utilized a decision tree-based model to predict death due to ARDS or other complications. These findings can assist healthcare systems to plan for the spread of future pandemics and identify groups at risk.

Methods

The data

Data on COVID-19 mortalities were retrieved from Jaber Hospital's mortality Electronic Health Records (EHR) for all patients admitted to the ICU between March 7, 2020, and August 19, 2021, and death reported between March 7, 2020, and August 27, 2021. The data set contains 1691 cases after excluding 12 children (<17 years old) and 46 with no data entries. The monthly total death rate in Kuwait is depicted in Worldometer cite (10). On the final day of data collection for this study, the total number of COVID-19 deaths was reported to be 2415; thus, our sample size covers 70% (1691/2415) of the COVID-19 mortality population. We also covered all death peaks and pandemic main waves during this time.

Initially, the data was extracted as a pdf file and then converted to an Excel spreadsheet. Patients' demographics (age, gender, and residency), date of ICU admission, date of death, reasons for admission, admission diagnosis, final diagnosis, cause of death, brief history, brief summary, and contributing factors are all included in each record. To ensure confidentiality, all data was anonymized and all patient identifiers were

TABLE 2 Intermediate causes of death.

Intermediate COD (Abbrev.)	Alternative terms
Acute respiratory distress syndrome (ARDS)	Mechanical ventilation, acute respiratory failure, hypoxic respiratory failure, HRF
Acute kidney failure (AKI)	Acute kidney injury, renal impairment, anuric, hyperkalemia, dialysis
Pulmonary embolism (PE)	DVT collapse, thrombosis
Heart failure (HF)	Rescue PCI, cardiomyopathy, myocarditis
Stroke (ST)	CVA, cerebrovascular accident, failed thrombolysis, hemorrhagic cerebral, subdural, subarachnoid hemorrhage, hge
Pneumothorax (PN)	Tension pneumothorax, hemothorax, hemopneumothorax, hydropneumothorax, pneumoperitoneum, bilateral chest tubes, chest tube
Myocardial infarction (MI)	STEMI, PCI, CCU, ischemic changes, cardiac strain, st elevation, troponin elevated, NStemi
Arrhythmia (AR)	Ventricular fibrillation, VFib, ventricular tachycardia, vtach, rhythm, atrial fibrillation, AF, PAF
Bleeding (BL)	ICH, hematoma, AVM, intracerebral hemorrhage, epistaxis, PRBC, transfusion, melena, upper GI bleeds
Disseminated intravascular coagulation (DI)	DIC
Urinary tract infection (UT)	UTI, urinary tract infection, urosepsis, E.col

TABLE 3 General disease categories (GDC), comorbidities and other risk factors.

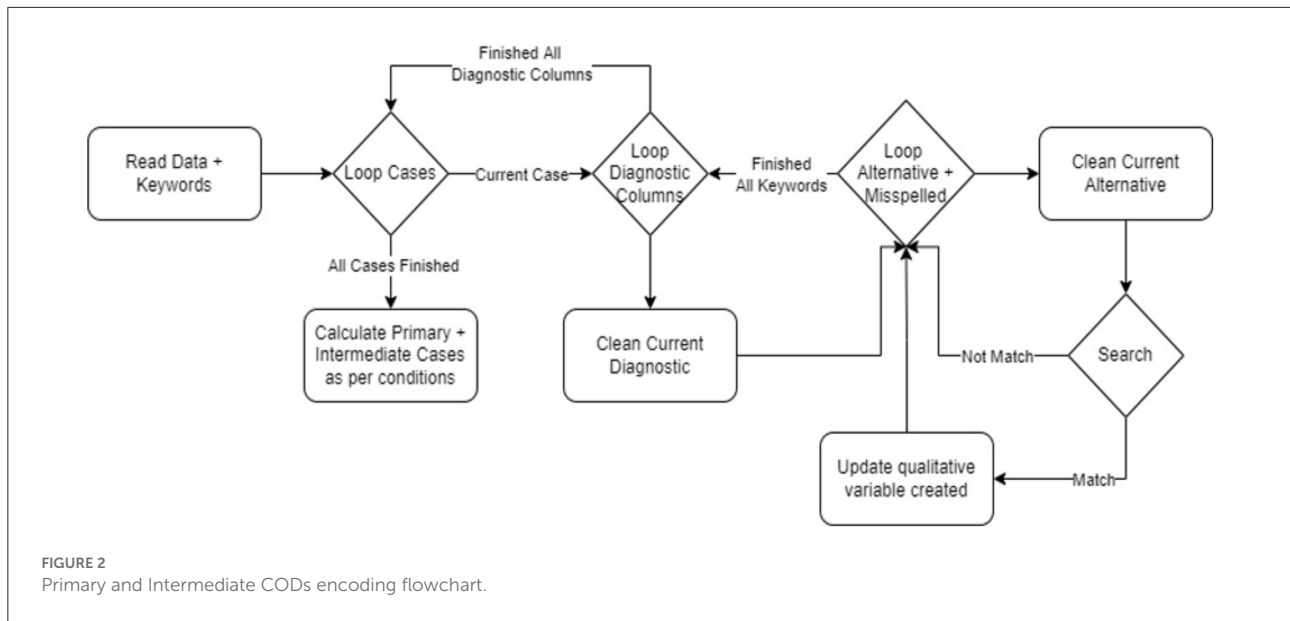
GDC (Abbrev.)	Comorbidity/risk factor (Abbrev.)
Endocrine, nutritional, and metabolic diseases (ENMs)	Diabetes mellitus (DM), thyroid disease (THY), dyslipidemia (DLP), obesity (OB), Addison disease (ADs)
Diseases of the nervous system (DNS)	Stroke (CVA), Parkinson's disease (PD), dementia (DEM), multiple sclerosis (MS), epilepsy (EP), psychiatric disorders (OCD)
Diseases of the circulatory system (DCS)	Hypertension (HTN), anemia (IDA), pulmonary embolism (PE), peripheral vascular disease (PVD), bleeding disorders (BDs)
Cardiovascular system diseases (CVD)	Coronary artery disease (CAD), cardiomyopathy (HCM), valvular heart disease (AVR), heart failure (HF), arrhythmia (AF)
Respiratory diseases (RDs)	Asthma (BA), chronic obstructive pulmonary disease (COPD), lung disease (LD)
GI disorders (GIDs)	Inflammatory bowel disease (IBD), gastroesophageal reflux (GERD), liver disease (LD)
Diseases of the genitourinary (DGS)	Chronic kidney disease (CKD), benign prostatic hyperplasia (BPH)
Autoimmune disorders (ADs)	Rheumatoid arthritis (RA), Immunocompromised (IC) (<i>a risk factor</i>)
Ortho disorders (ODs)	Bone disorders (OA)
Infectious diseases (IDs)	HIV-infection (HIV)
Neoplasms (CRC)	Cancer (CA) of any kind
Congenital disorders (CDs)	Down syndrome (DS)

patient's case and each column represents each medical term relating to a cause of death or comorbidity, all other word tokens will be omitted. The cells of the matrix will contain a 0 or 1 representing the occurrence or absence of the term from the case. The terms related to cause of death will be categorized to three stages similar to the fashion of death certificates. These stages are the primary, intermediate and the underlying cause (which led to the intermediate).

The list of primary causes of death, according to WHO guidelines, denotes the condition (injury, complication, or disease) that directly preceded death. WHO issued an updated International Classification of Diseases (ICD) and health-related problems to accommodate COVID-19-related death complications (11). The condition(s) that led to the primary COD are reflected in the intermediate COD. Multiple

complications contributing to the intermediate COD were identified in the majority of COVID-19 decedents in the ICU in this study. Additionally, COVID-19 pneumonia was the most frequently encountered underlying cause in those ICU cases, resulting in an intermediate stage of complication.

In order to create BoW, the COD and comorbidity terms were extracted from the EHR in several steps. Starting with a preliminary text analysis using the text mining package (*tm*) and the word cloud generator package (*wordcloud*) in R to extract the most common terms (Figure 1). To create glossary tables, our medical experts validated the extracted terms by reviewing 50–100 EHRs at random. The process was repeated four times to ensure that the majority of the terminologies were covered. This helped identify alternative terminologies and misspelled terms. Tables 1, 2 show the refined



```

For each case:
  If Total P = 1:
    Primary ← primary COD cell with 1
  If Total P > 1:
    If Septic shock = 1 and MOF = 1:
      Primary ← Septic shock + MOF
    If Septic shock = 1 and any other CODs = 1 and MOF = 0:
      Primary ← Septic shock
    If MOF=1 and any other CODs = 1 and Septic Shock = 0:
      Primary ← MOF
    If cardiac arrest = 1 and Respiratory failure=1:
      Primary ← Cardiopulmonary arrest
    If cardiopulmonary arrest=1 and (Cardiac arrest=1 or Respiratory failure=1):
      Primary ← cardiopulmonary arrest
    
```

FIGURE 3
Pseudocode for Primary COD.

list of primary COD and intermediate COD. In accordance with the International Classification of Diseases (11), Table 3 provides twelve general disease categories (GDC), 34 distinct comorbidities, and a risk factor associated with our data. Detailed versions of Tables 1–3, including all potential alternate terms and/or incorrect forms may be requested from the corresponding author.

Developing and applying NLP methods

We created an NLP method to identify, extract, and automatically encode natural language from mortality EHRs into structured clinical data. Tables 1, 2 are used as keywords to extract primary and intermediate CODs, while Table 3 presents

keywords to extract comorbidities. Method created in Python. Figure 2 shows our algorithm.

In this method, text is stripped of punctuation, special characters, capitalization, stop words, and tokenization. Used EHR variables include cause of death, final diagnosis, brief history, and brief summary. To create a case/COD term occurrence matrix, binary variables must be created for each primary/intermediate COD listed in Tables 1, 2. Initial occurrence matrix setting is zero. CODs or equivalents are compatible with tokens. The case/term occurrence matrix cell is set to 1 upon a match. Every case applies (rows). A COD abbreviation was not mistaken for a term, as PE is not present in hypertensive or hyperthyroid. Negation was also carefully handled; if a term is preceded by a negative or conditional word, it will not match. Exclusion words consist of (no, not, no sign of, non, no history of, no active, no previous medical, not known to have, no indications of, previous condition, old condition). Text format is used to list the final primary and intermediate CODs. The pseudocode used to extract the final primary COD is depicted in Figure 3.

Determining the actual intermediate CODs are handled differently. Multiple intermediate CODs are reported as a group. Our clinicians manually validated and separated the correct outcome to determine which disorders were terminal. A counter matching the extracted causes is also computed to help identify the terminal cause based on the most common causes to cross-check the accuracy of the findings.

The comorbidities for each case are identified using Table 3 in the same manner that CODs are identified. Preprocessed word tokens are extracted from the EHR reason for admission, contributing factors, admission diagnosis and brief summary.

Data manipulation and analysis

Original EHR mortality data had two sets of variables. First set included seven categorical and quantitative variables. Second set included eight free-text variables. The pdf data sheet was converted to an Excel sheet for data manipulation and cleaning. The second set of data was used to generate 70 variables using Python to determine death causes and comorbidities. During exploratory data analysis, we generated appropriate graphs (bar, pie, boxplots) and summary statistics (mean, median, SD, IQR). Hypothesis tests included Chi-square, TURF, ANOVA, and Kruskal Wallis. Finally, we built our prediction model

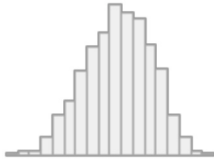


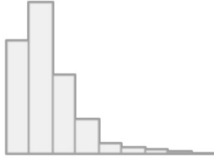
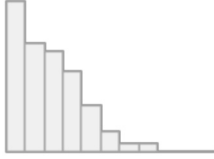

with a decision tree. SPSS V23 and R were used for the statistical analysis.

Results

Overall findings

The majority of the 1,691 anonymous COVID-19 decedents were male 963 (56.9%). The age at death ranges between 19.8 and 103.2 years with 63.8 years (SD 14.4). On the average the duration stay in ICU prior to death was 18.5

TABLE 4 Demographic and clinical characteristics.

Variable	Summary	Count (%)	Graph
Age	Mean (sd): 63.8 (14.4) min ≤ med ≤ max: 19.8 ≤ 64.5 ≤ 103.2 IQR (CV): 20.7 (0.2)	662 distinct values	
Age group	1. (< 50) years 2. (50–64) years 3. (65+) years	303 (17.9%) 573 (33.9%) 815 (48.2%)	
Gender	1. Female 2. Male	728 (43.1%) 963 (56.9%)	
ICU (days)	Mean (sd): 18.5 (12.8) min ≤ med ≤ max: 0 ≤ 16 ≤ 86 IQR (CV): 14 (0.7)	74 distinct values	
Total comorbidities	Mean (sd): 2.7 (1.9) min ≤ med ≤ max: 0 ≤ 3 ≤ 11 IQR (CV): 3 (0.7)	12 distinct values	
Total comorbidities group	Mean (sd): 2.6 (1.7) min ≤ med ≤ max: 0 ≤ 3 ≤ 6 IQR (CV): 3 (0.6)	0 : 172 (10.8%) 1 : 288 (18.1%) 2 : 333 (20.9%) 3 : 304 (19.1%) 4 : 245 (15.4%) 5 : 143 (9.0%) 6 : 110 (6.9%)	

days (SD 12.8). Two or more comorbidities were present (mean 2.5, SD 1.9) with hypertension and diabetes mellitus shared among more than half of them (Table 4). Since these patients died in the intensive care unit, COVID-19

pneumonia was mainly the underlying cause of death that resulted in intermediate and thus primary causes of death. COVID-19 pneumonia was detected in 94 percent of cases (1592/1691).

TABLE 5 Demographic and clinical characteristics by age group.

Variable	N	Overall N = 1,691 ^a	Age group (yrs.)			p-value ^b
			Age < 50 N = 303 ^a	Age [50-64] N = 573 ^a	Age =65+ N = 815 ^a	
Age	1,691	64 (54, 74)	43 (39, 47)	58 (54, 62)	75 (70, 81)	<0.001
Gender	1,691					0.003
Female		728 (43%)	114 (38%)	229 (40%)	385 (47%)	
Male		963 (57%)	189 (62%)	344 (60%)	430 (53%)	
ICU (days)	1,691	16 (10, 24)	14 (9, 23)	16 (10, 24)	16 (10, 24)	0.19
Total comorbidities	1,596	3 (1, 4)	1 (0, 2)	2 (1, 3)	3 (2, 5)	<0.001
Unknown		95	0	33	62	
Total group comorbidities	1,595					<0.001
0		172 (11%)	111 (37%)	61 (11%)	0 (0%)	
1		288 (18%)	85 (28%)	122 (23%)	81 (11%)	
2		333 (21%)	59 (19%)	136 (25%)	138 (18%)	
3		304 (19%)	30 (9.9%)	99 (18%)	175 (23%)	
4		245 (15%)	12 (4.0%)	72 (13%)	161 (21%)	
5		143 (9.0%)	2 (0.7%)	32 (5.9%)	109 (14%)	
6		110 (6.9%)	4 (1.3%)	18 (3.3%)	88 (12%)	
Unknown		96	0	33	63	

^aMedian (IQR) or Frequency (%).

^bKruskal-Wallis rank sum test; Pearson's Chi-squared test.

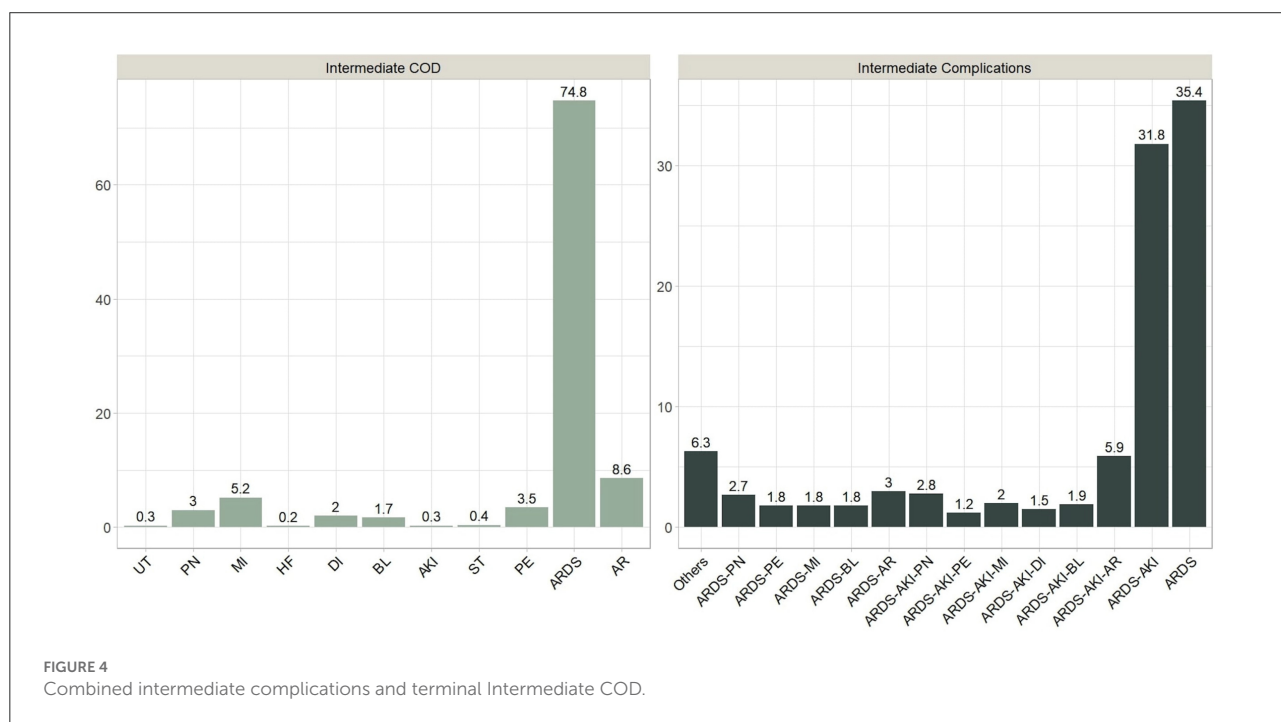


FIGURE 4 Combined intermediate complications and terminal Intermediate COD.

TABLE 6 Primary by intermediate causes of death.

Inter-mediate COD	Primary COD count (%)							Total
	SS	CPA	HRF	CA	SS + MOF	MOF	RF	
AR	53 (7.9)	26 (8.6)	11 (4.7)	26 (14.4)	14 (10.4)	12 (9.6)	3 (6.8)	145 (8.6)
ARDS	523 (78.4)	220 (72.4)	194 (82.6)	117 (65)	94 (69.6)	8 (64.8)	35 (79.5)	1,265 (74.8)
BL	12 (1.8)	3 (1)	5 (2.1)	3 (1.7)	1 (0.7)	5 (4)	0 (0)	29 (1.7)
DI	14 (2.1)	2 (0.7)	0 (0)	1 (0.6)	11 (8.1)	4 (3.2)	1 (2.3)	33 (2)
HF	0 (0)	2 (0.7)	0 (0)	1 (0.6)	0 (0)	1 (0.8)	0 (0)	4 (0.2)
MI	23 (3.4)	26 (8.6)	4 (1.7)	21 (11.7)	5 (3.7)	9 (7.2)	0 (0)	88 (5.2)
PE	14 (2.1)	16 (5.3)	12 (5.1)	7 (3.9)	5 (3.7)	5 (4)	1 (2.3)	60 (3.5)
PN	22 (3.3)	8 (2.6)	7 (3)	3 (1.7)	2 (1.5)	6 (4.8)	3 (6.8)	51 (3)
ST	2 (0.3)	1 (0.3)	1 (0.4)	1 (0.6)	1 (0.7)	0 (0)	0 (0)	6 (0.4)
UT	2 (0.3)	0 (0)	1 (0.4)	0(0)	1 (0.7)	1(0.8)	0 (0)	5 (0.3)
Total	667 (100)	304 (100)	235 (100)	180 (100)	135 (100)	125 (100)	44 (100)	1,691 (100)

When the mean ICU stay was compared across the three age groups of <50, 50–64, and 65 or more, no significant difference (Table 5) using the ANOVA F-test (p -value = 0.903). On the other hand, testing for mean total comorbidities across these three age groups was significant (p -value <0.0001), and the Tukey B multiple comparison test reveals significance with three means for groups in homogenous subsets of mean total comorbidities of 1.25, 2.19, and 3.28, respectively.

Clinical characteristics and common causes of death among COVID-19 patients

We identified primary and secondary causes of death. Septic shock was the primary COD in 667 patients (39.4%), followed by cardiopulmonary arrest 304 (18.0%), respiratory failure 235 (13.9%), and cardiac arrest 180 (10.6%). The percentages of cases with (septic shock & MOF), MOF, and renal failure were 135 (8.0%), 125 (7.4), and 44 (2.6%), respectively. Hepatic failure occurred in only one case and thus ignored from further analysis. On the other hand, ARDS was one of the main reasons for ICU admissions and was reported in all deaths. Numerous cases were reported in which a combination of intermediate death complications occurred. These cases were thoroughly examined by our physicians to determine which terminal complication is more likely to be classified as the intermediate COD. It was found that around 75% of these decedents had ARDS as an intermediate COD, while the remaining 25% had intermediate COD other than ARDS. Among the other causes are AKI, AR, BL, DI, HF, MI, PE, PN, ST, and UT. The frequency distribution of intermediate combined complications along with the frequency distribution of the terminal complication leading to intermediate COD are shown in Figure 4. Table 6 shows the

count and percentage of counts for primary and intermediate causes, as well as the column percentages for primary causes. While ARDS is the most prevalent intermediate COD regardless of primary cause, AR and MI disorders were significantly (7.2–14.4%) linked with cardiac arrest and MOF.

Age distribution appears to be similar by primary COD, with a median age at death of 64.5 years and an interquartile range (IQR = 20.7). However, a few young patients, approximately the age of 20, died because of MOF or renal failure (Figure 5). The median length of stay in the ICU prior to death was approximately 16 days overall but was significantly longer (~20 days) for those who died of septic shock or (septic shock + MOF). Patients who died because of MOF had an average of three or more comorbidities. Those who died of renal failure and (septic shock + MOF) died in a manner like that described above.

Those who died because of AR, DI, or MI had the highest average age (70 years) and total comorbidities (3 or more), as well as the shortest average stay in the ICU. Patients who died of HF were younger (average age 50 years) and had more than two comorbidities, with an ICU stay of < 20 days. We also noticed the sequences of (MI → septic shock) and (PE → respiratory failure) were associate with 4 or above comorbidities on the average.

Exploring the relationship between comorbidities and causes of death

The following is a list of the comorbidities of dead patients in this study. Hypertension (57%) is the most common condition, followed by diabetes (52%), coronary artery disease (23%), and chronic renal disease (14%). 12% for each arrhythmia and dyslipidemia, cancer (11%), Rheumatoid arthritis (10%),

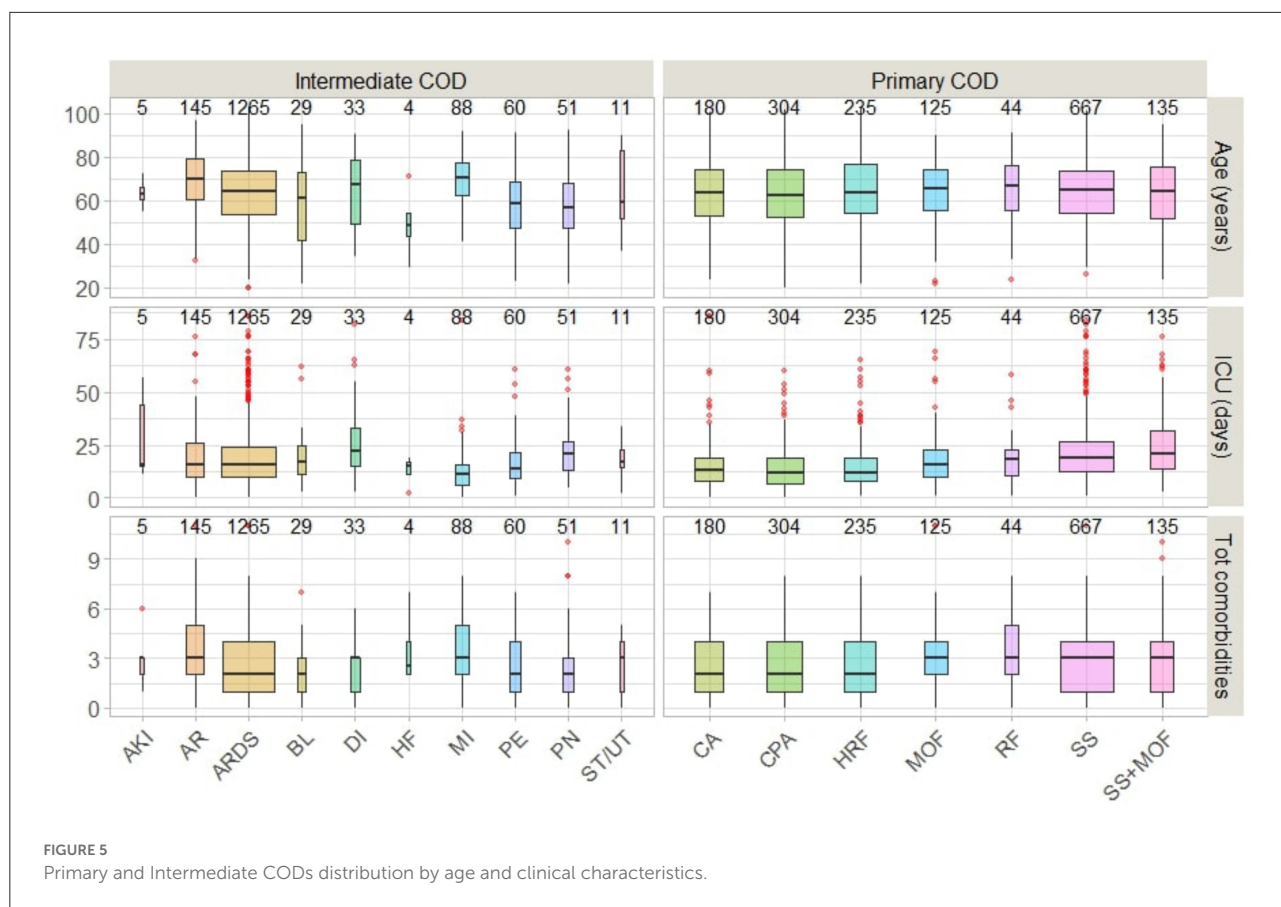


FIGURE 5 Primary and Intermediate CODs distribution by age and clinical characteristics.

TABLE 7 Best reach and frequency by group size.

	Size	Reach	Cases %	Count	Responses %
ADDED: HTN	1	962	56.9	962	27.7
ADDED: DM	2	1,139	67.4	1,834	52.9
KEPT: HTN					
ADDED: Cancer	3	1,195	70.7	2,026	58.4
KEPT: DM, HTN					
ADDED: RA	4	1,234	73.0	2,196	63.3
KEPT: Cancer, DM, HTN					
ADDED: AF	5	1,266	74.9	2,402	69.3
KEPT: Cancer, DM, HTN, RA					
ADDED: Obesity	6	1,296	76.6	2,550	73.5
KEPT: AF, Cancer, DM, HTN, RA					

obesity (9%), thyroid disease (8%), stroke (7%), pulmonary embolism (5%), asthma (4%), valvular heart disease (4%), bleeding disorders (4%), and 3% for each COPD and dementia. The remaining comorbidities with < 3% reported incidence include Anemia, heart failure, prostate hyperplasia, liver disease, epilepsy, cardiomyopathy, peripheral vascular disease, lung disease, psychiatric disorders, osteoporosis, multiple sclerosis, down syndrome, Parkinson’s disease, inflammatory bowel

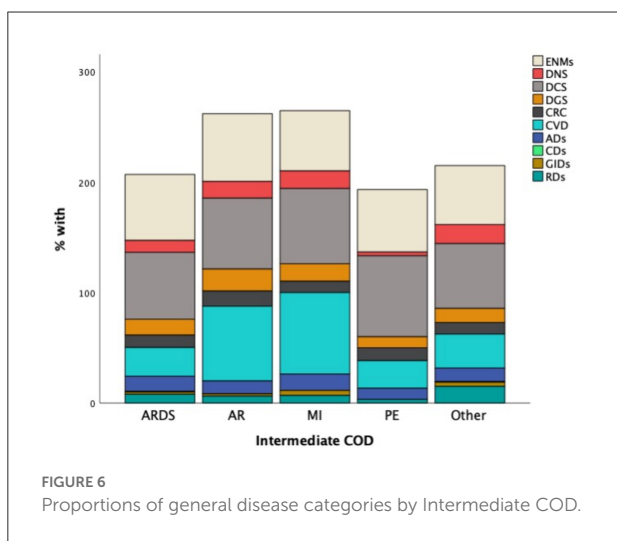
disease, gastroesophageal reflux disease, Addison disease, and HIV infection.

Next, we present the results of the Total Unduplicated Reach and Frequency (TURF) method. TURF is a popular statistical technique in market research that ranks product combinations according to the number of customers who favor them (12). In this study, we applied the method in a clinical setting, treating comorbidities and patients as products and people.

The goal is to determine the most likely disease combinations that these patients share. The analysis traverses all possible combinations of comorbidities and records two statistics for each: reach and frequency. The reach is the percentage of individuals who exhibit at least one comorbidity in a given combination, and the frequency is the total number of times comorbidities are exhibited in a given combination. We tested the method for all comorbidities listed in Table 3 and a range of reach values. Table 7 provides a summary of the ideal choices according to the number of diseases (Size). For instance, the optimal combination of four comorbidities has a 73 percent

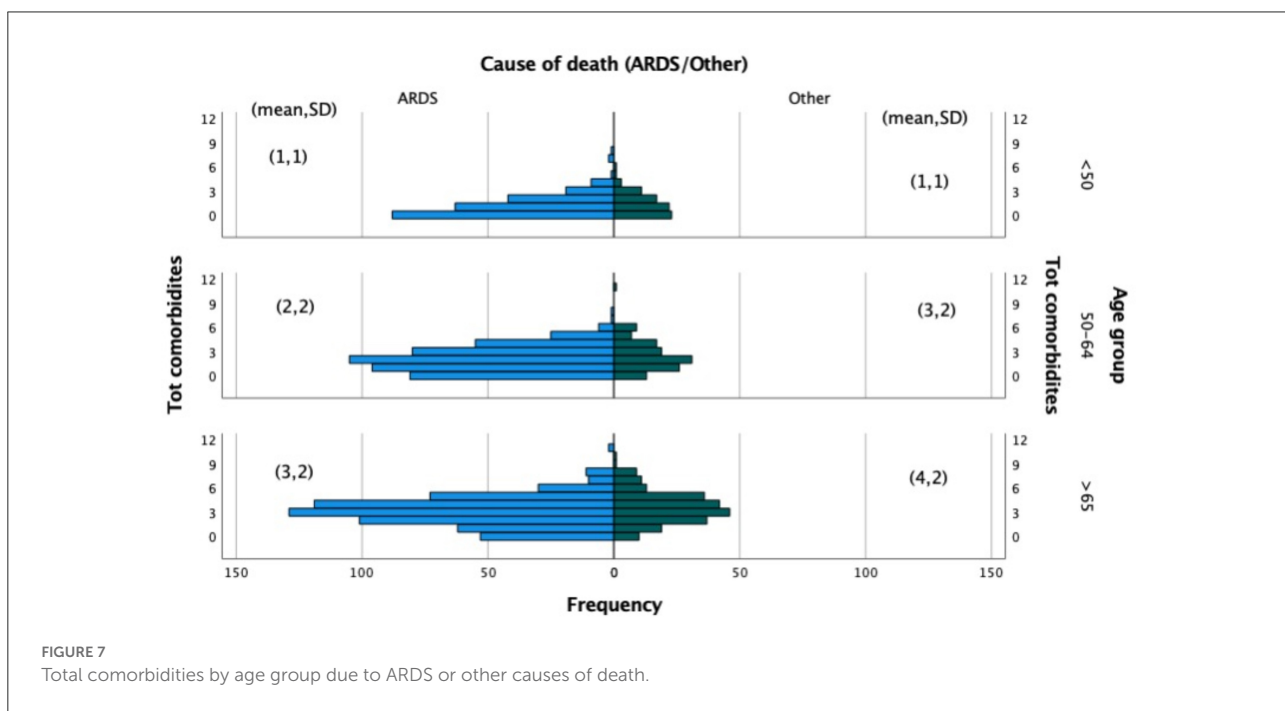
success rate with RA, cancer, DM, and HTN. This indicates that seventy-three percent of the patients had at least one of the conditions (rheumatological disorders, cancer, DM, HTN). If Diabetes and High Blood Pressure were eliminated from the analysis due to their high prevalence and we wanted to evaluate other possible combinations of diseases, the one with the highest prevalence was (obesity, CAD, Cancer, RA) with 43.6%.

When we looked at the general disease classification frequencies, we found that over 60% of the patients had circulatory (DCS) and endocrine (ENMS) disorders, one-third had cardiovascular diseases (CVD), and the remaining categories (RDs, CRC, DNS, ADs, DGS) varied from 8 to 15%. In compared to patients who died of ARDS/PE/Other, approximately 65 percent of patients who died of MI or AR had cardiovascular illnesses (Figure 6). Those who die from ARDS, on the other hand, usually have endocrine or circulatory system problems. Nervous system diseases were the least common among the PE dead. With chi-square test findings of (175.5, *p*-value 0.001) and (12.2, *p*-value = 0.016), the circulatory and nervous systems had the most significant association with intermediate COD.



Predicting death due to ARDS or other causes

The total comorbidities distribution by age group of COVID-19 deaths due to ARDS or other cause is displayed in



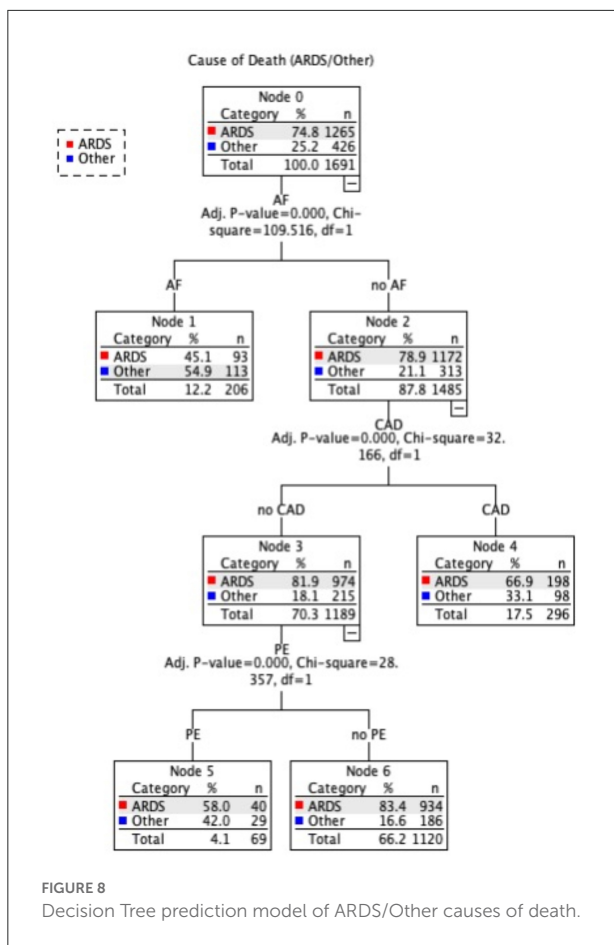


Figure 7. Patients under the age of 50 have a similar comorbidity distribution, with an average of one disease. Two comorbidities were found on average per age group (50–64) with more variation among those who died from causes other than ARDS. In contrast, older patients (age >65) who died from causes other than ARDS have an average of four comorbidities, compared to three for the other group who died mainly from ARDS.

In this section, we used decision tree (DT) to determine the most parsimonious predictors of intermediate COD among COVID-19 patients in intensive care units. Decision trees learn to divide data into smaller and smaller categories to forecast the goal. The test is represented by a node, while the numerous outcomes are represented by edges. The dividing process is repeated until no further gains can be obtained or a preset rule is reached. Three common decision tree techniques include classification and regression tree (CART), chi-squared automatic interaction detection (CHAID), and quick unbiased efficient statistical tree (QUIEST). For mathematical explanations and performance comparisons of these DT approaches, see Lin et al. (13). Figure 8 illustrates the results of the QUEST model, which demonstrate that the existence of an arrhythmia (AF) was the best indicator of the intermediate cause (ARDS/Other). Patients

with AF are more likely to have a cause other than ARDS (54.9%). Node 1 is considered a terminal node for predicting a cause of death other than ARDS since no child nodes was found below it. In patients without AF, on the other hand, CAD was the second-best predictor of (ARDS/Other). In patients without AF but with CAD, the terminal Node 3 predicted 66.9 ARDS vs. 33.1% for other causes. PE is an additional predictor in the model for patients who do not have AF or CAD. ARDS is the main intermediate COD in this group, accounting for over 83% of patients without PE and 58% of patients with PE who died from ARDS. The risk and classification tables allow for a quick evaluation of the model’s performance. The risk of misclassifying the cause of death is estimated to be 0.272 (or 27.2%), which is consistent with the results of the classification table, which show that 76% of causes of death are correctly classified.

Discussion

We used Machine learning NLP to extract clinical data and causes of death from EHRs for COVID-19 patients at Jaber Hospital in Kuwait. Consistency and completeness issues with the text data in these records made extraction difficult. During the pandemic, Jaber hospital was restricted to COVID-19 admissions, with most critical cases transferred from other hospitals. Many patient records were incomplete due to patients being transferred from district hospitals where their original medical records were kept. Machine learning and big data analytics have been used to investigate disease-related prognostic factors (14).

Several clinical characteristics have been linked to COVID-19 mortality. Age, gender, comorbidities, ICU stay, and disease severity are all factors. Increased proportions of 65-year-olds or older led to a significant age-mortality association (15, 16). Males were more likely to die from COVID-19 (17, 18). More than double the number of death patients had two or more comorbidities, according to Ayed et al. (17). Combining old age and comorbidities was also a factor in death (19) and survival time (20). On the other hand, Zhou et al. (21) reported a median (IQR) time of 18.5 (15–22) days from onset of symptoms to death. In our study, 815 (48%) of 1691 deceased ICU COVID-19 patients were over 65, men were more prevalent (56.9 vs. 43.1%), patients with two or more comorbidities accounted for 52% of cases, and the mean (SD) survival time to death was 18.5 (12.8) days. Hypertension and diabetes accounted for more than half of all cases in this study. This confirms prior research (17, 22–24). In COVID-19 patients, cardiovascular disease and secondary infections increase disease severity and mortality (15, 25, 26). Circulatory and cardiovascular diseases account for 61.6 and 32.5% of these patients, respectively; HIV-infections are rare. COVID-19 patients had a higher incidence of kidney and heart disease, and myocardium damage reduced survival (16, 27, 28).

Previous research on comorbidities and death causes has linked dysfunction to mortality (17, 29). In this study, decedents with MOF and renal failure averaged three or more comorbidities. Septic shock was the leading primary cause, accounting for 667 deaths (39.4%), followed by cardiopulmonary arrest (304 deaths, 18%), respiratory failure (235 deaths, 13.9%), and cardiac arrest (180 deaths, 10.6%). The most common intermediate COD, on the other hand, was ARDS (1265, 74.8%). We also found 849 (50.2%) cases of sepsis. Other findings (21) revealed that sepsis was the leading cause of death (59%) among the 54 pandemic deaths, followed by respiratory failure (54%), ARDS (31%), heart failure (23%), and septic shock (20%).

Acute respiratory distress syndrome (ARDS) is a severe COVID-19 consequence. Patients with moderate-to-severe ARDS require invasive mechanical ventilation and intensive medical therapy (30, 31). ARDS was one of the most common reasons for ICU hospitalizations, as it was recorded in 81.8% of ICU survivors and all fatalities (32). This is also demonstrated in our data, as all patients were admitted to the intensive care unit, and ARDS was a common morbid consequence. However, complications other than ARDS were deemed the predominant intermediate COD in 25% of the cases (Figure 4). As a result, we employed decision trees to forecast the most significant contributing factors to intermediate COD, namely ARDS or Other cause. "Other" denotes a complication associated with AKI, AR, BL, DI, HF, MI, PE, PN, ST, or UT. We encountered only three significant predictors, namely arrhythmia (AF), coronary artery disease (CAD), and pulmonary embolism (PE). Patients with AF were more likely to have an etiology other than ARDS. According to Elezkurtaj et al. (33), the majority of decedents died from COVID-19, with preexisting health conditions and comorbidities only contributing to the mechanism of death. We agree because, among the many variables examined in this study, only a few contributing factors were found to be significant with intermediate COD.

Strengths, limitations, and future work

The dynamic nature of the method, its usability, and its potential to maintain self-control all contribute to its strength. In addition, the sampled data span both significant pandemic waves and death peaks, accounting for 70% of the total reported COVID-19 fatality cases in Kuwait. The death rate drastically decreased after then. Therefore, our sample represents the population under consideration to a high degree of accuracy. Nevertheless, our study has several limitations. First, there is a chance of selection or referral bias as the research was conducted at a single location, i.e., Jaber Hospital. Second, the lack of information extracted from the inadequate documentation of the patient records. The absence of a symptom (such as obesity, smoking, etc.) does not necessarily suggest that a patient is

symptom-free. Thirdly, patients were typically transferred late in the course of their disease, and their medical records lacked vital medical history information. Such discrepancies in clinical data may result in information bias that contributes to a decrease in model precision.

Future studies could potentially investigate the impact of vaccines on the time to death, provide survival time estimates by cause of death, and perform spatiotemporal analyses of transferable patients. Knowing the COVID-19 death rate and patient survival rate can help risk management experts. COVID-19 or its evolving variants can be avoided, and strategies can be used to slow their spread.

Conclusion

We employ self-developed natural language processing (NLP) to automate the extraction of causes of death and comorbidities from the EHRs of COVID-19 decedents from the beginning of the pandemic through all major pandemic waves in this study. We structured the acquired text data and used it to conduct additional research.

We analyzed the demographic, clinical, and causes of death data for 1,691 ICU patients and discovered that the most common primary causes of death, which were documented in 54.8% of cases, were infection-related and included septic shock or sepsis-related multi-organ failure. The second most common cause of death was respiratory failure or cardiopulmonary arrest, which were documented in 32.2% of cases. Furthermore, cardiac arrest and renal failure account for 10.6 and 2.6% of all deaths, respectively. ARDS, on the other hand, was the most common cause of mortality in the intermediate stage. Arrhythmia (AF) was revealed to be the strongest predictor of intermediate cause (ARDS/Other) using machine learning decision tree analysis.

We recommend structuring the EHR with well-defined sections and providing menu-driven options for reporting causes of death and comorbidities to minimize misspellings or incorrect forms. Comprehensive assessment and user guidance are required for standards to be effectively integrated into EHR systems.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethical Review Committee (ERC) at Kuwait Ministry of Health (No. 1529/2020). Written informed consent for participation was not required for this study in

accordance with the national legislation and the institutional requirements.

Author contributions

SB conceived and gained ethical approval for the project. SB, AB, and YA participated in the retrieval, processing, and purification of data. AB and YA developed the clinical concepts, played a key role in establishing the data extraction clinical criteria, and validations. SB and AA created both the method and the programming. SB and SA carried out statistical analysis and produce visuals. SB, AA, and SA contributed to the paper's drafting. All authors have reviewed, offered comments, and approved the submission of the work.

Acknowledgments

We would like to express our gratitude to the administration of Jaber Al-Ahmad Hospital for their cooperation and support.

References

- Kong HJ. Managing unstructured big data in healthcare system. *Health Inform Res.* (2019) 25:1. doi: 10.4258/hir.2019.25.1.1
- Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. *J Am Med Inform Assoc JAMIA.* (2017) 24:339–44. doi: 10.1093/jamia/ocw082
- Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform.* (2019) 7:e12239. doi: 10.2196/12239
- Zhou L, Lu Y, Vitale CJ, Mar PL, Chang F, Dhopeswarkar N, et al. Representation of information about family relatives as structured data in electronic health records. *Appl Clin Inform.* (2014) 5:349–67. doi: 10.4338/ACI-2013-10-RA-0080
- John Lin CC, Yu K, Hatcher A, Huang TW, Lee HK, Carlson J, et al. Identification of diverse astrocyte populations and their malignant analogs. *Nat Neurosci.* (2017) 20:396–405. doi: 10.1038/nn.4493
- DeCapprio D, Gartner J, McCall CJ, Burgess T, Kothari S, Sayed S. Building a COVID-19 Vulnerability Index. *MedRxiv.* (2020). doi: 10.1101/2020.03.16.20036723
- Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med.* (2021) 4:3. doi: 10.1038/s41746-020-00372-6
- Izquierdo JL, Ancochea J, Soriano JB. Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. *J Med Internet Res.* (2020) 22:e21801. doi: 10.2196/21801
- Guan X, Zhang B, Fu M, Li M, Yuan X, Zhu Y, et al. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. *Ann Med.* (2021) 53:257–66. doi: 10.1080/07853890.2020.1868564
- Coronavirus. *Worldometer.* Available online at: <https://www.worldometers.info/coronavirus/country/kuwait/> (accessed April 12, 2022).
- (ICD-10). *International Classification of Diseases, Tenth Revision (ICD-10).* (2021). Available online at: <https://www.cdc.gov/nchs/icd/icd10.htm> (accessed April 11, 2022).

We would like to thank Eng. Naser Alibrahim for his support with Python coding.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Data Scientist. *Reflections of a Data Scientist.* (2018). Available online at: <https://www.reflectionsofadatascientist.com/2018/05/r-turf-analysis-spss.html> (accessed November 11, 2022).
- Lin CL, Fan CL. Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *J Asian Archit Build Eng.* 18:539–53. doi: 10.1080/13467581.2019.1696203
- Darabi H, Tsinis D, Zecchini K, Whitcomb W, Liss A. "Forecasting mortality risk for patients admitted to intensive care units using machine learning." In: *Procedia Computer Science, vol. 140.* Chicago, IL: Elsevier (2018). p. 306–313. doi: 10.1016/J.PROCS.2018.10.313
- Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med.* (2020) 46:846–8. doi: 10.1007/s00134-020-05991-x
- Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. *J Med Virol.* (2020) 92:441–7. doi: 10.1002/jmv.25689
- Ayed M, Borahmah AA, Yazdani A, Sultan A, Mossad A, Rawdhan H. Assessment of clinical characteristics and mortality-associated factors in COVID-19 critical cases in Kuwait. *Med Princ Pract.* (2021) 30:185–92. doi: 10.1159/000513047
- Galbadage T, Peterson BM, Awada J, Buck AS, Ramirez DA, Wilson J, et al. Systematic review and meta-analysis of sex-specific COVID-19 clinical outcomes. *Front Med.* (2020) 7:348. doi: 10.3389/fmed.2020.00348
- Moon SS, Lee K, Park J, Yun S, Lee YS, Lee DS. Clinical characteristics and mortality predictors of COVID-19 patients hospitalized at nationally-designated treatment hospitals. *J Korean Med Sci.* (2020) 35:e328. doi: 10.3346/jkms.2020.35.e328
- Sousa GJB, Garces TS, Cestari VRE, Florêncio RS, Moreira TMM, Pereira MLD. Mortality and survival of COVID-19. *Epidemiol Infect.* (2020) 148:e123. doi: 10.1017/S0950268820001405
- Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet Lond Engl.* (2020) 395:1054–62. doi: 10.1016/S0140-6736(20)30566-3

22. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, et al. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy. *JAMA*. (2020) 323:1574–81. doi: 10.1001/jama.2020.5394
23. Nada KM, Hsu E shuo, Seashore J, Zaidan M, Nishi SP, Duarte A, et al. Determining cause of death during Coronavirus Disease 2019 pandemic. *Crit Care Explor*. (2021) 3:e0419. doi: 10.1097/CCE.0000000000000419
24. Yan Y, Yang Y, Wang F, Ren H, Zhang S, Shi X, et al. Clinical characteristics and outcomes of patients with severe COVID-19 with diabetes. *BMJ Open Diabetes Res Care*. (2020) 8:e001343. doi: 10.1136/bmjdr-2020-001343
25. Li B, Yang J, Zhao F, Zhi L, Wang X, Liu L, et al. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. *Clin Res Cardiol Off J Ger Card Soc*. (2020) 109:531–8. doi: 10.1007/s00392-020-01626-9
26. Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *Int J Infect Dis IJID Off Publ Int Soc Infect Dis*. (2020) 94:91–5. doi: 10.1016/j.ijid.2020.03.017
27. Arentz M, Yim E, Klaff L, Lokhandwala S, Riedo FX, Chong M, et al. Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State. *JAMA*. (2020) 323:1612–4. doi: 10.1001/jama.2020.4326
28. Rodriguez-Morales AJ, Cardona-Ospina JA, Gutiérrez-Ocampo E, Villamizar-Peña R, Holguin-Rivera Y, Escalera-Antezana JP, et al. Clinical, laboratory and imaging features of COVID-19: a systematic review and meta-analysis. *Travel Med Infect Dis*. (2020) 34:101623. doi: 10.1016/j.tmaid.2020.101623
29. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA*. (2001) 286:1754–8. doi: 10.1001/jama.286.14.1754
30. Gibson PG, Qin L, Pua SH. COVID-19 acute respiratory distress syndrome (ARDS): clinical features and differences from typical pre-COVID-19 ARDS. *Med J Aust*. (2020) 213:54–6.e1. doi: 10.5694/mja2.50674
31. Tzotzos SJ, Fischer B, Fischer H, Zeitlinger M. Incidence of ARDS and outcomes in hospitalized patients with COVID-19: a global literature survey. *Crit Care Lond Engl*. (2020) 24:516. doi: 10.1186/s13054-020-03240-7
32. Alshukry A, Ali H, Ali Y, Al-Taweel T, Abu-Farha M, AbuBaker J, et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) patients in Kuwait. *PLoS ONE*. (2020) 15:e0242768. doi: 10.1371/journal.pone.0242768
33. Elezkurtaj S, Greuel S, Ihlow J, Michaelis EG, Bischoff P, Kunze CA, et al. Causes of death and comorbidities in hospitalized patients with COVID-19. *Sci Rep*. (2021) 11:4263. doi: 10.1038/s41598-021-82862-5