Open Journal of Epidemiology, 2016, 6, 28-43
Published Online February 2016 in SciRes. http://www.scirp.org/journal/ojepi
http://dx.doi.org/10.4236/ojepi.2016.61005

Extracting Leading Joint Causes of Death and Mining Associations between Them

Keamogetse Setlhare, Ntonghanwah Forcheh

Department of Statistics, University of Botswana, Gaborone, Botswana
Email: setlhark@mopipi.ub.bw, Forchehn@mopipi.ub.bw

Received 21 November 2015; accepted 13 February 2016; published 16 February 2016

Abstract

The gains in analyzing death from a multiple cause perspective have been recognized for a very long time. Methods that have been adopted have sought to determine additional gains made by treating death as a multiple cause phenomenon as compared to analysis based on a single underlying cause. This paper shows how association rules mining methodology can be adapted to determine joint morbid causes with strong and interesting associations. Results show that some causes of death that do not appear among the leading causes show strong associations with other causes that would otherwise remain unknown without the use of association rules methodology. Overall, the study found that the leading joint pair of causes of death in South Africa was metabolic disorders and intestinal infectious diseases which accounted for 18.9 deaths per 1000 in 2008, followed by cerebrovascular and hypertensive diseases which accounted for 18.3 deaths per 1000.

Keywords

Association Rule, Confidence, Interestingness, Multiple Cause of Death, South Africa, Odds-Ratio, Prevalence, Support

1. Introduction

The analysis and reporting of causes of death continue to be dominated by the single underlying cause of death concept. As pointed out by [1], the philosophy behind the underlying cause concept is that if the starting point of the sequence of events that lead to death is known, then death can be prevented by preventing the initiating cause from operating. Hence, identifying leading underlying causes of death is very important in public health. Therefore, the underlying cause of death continues to be coded on death certificates. However, it has been recognised since the first attempts to develop international systems for classification of causes of death that death rarely arises from a single cause [2], especially death caused by chronic pathologies among the elderly [3]. It has

also noted that "the "lethality" of any one condition may be affected by the presence of other conditions" [4] and that it would be more informative to code the level of severity of each condition [5].

Changes and revisions in International Classification of Diseases (ICD) have been found to lead to differences in which conditions present at time of death are coded as underlying. This also leads to differences in the prevalence of certain underlying causes of death in the same country over time. Differences in the rules for selecting underlying causes have also been found to contribute to differences in prevalence between countries [6]-[9]. When death occurs outside the hospital or when the medical history for the deceased is unavailable and there is no qualified physician to determine the underlying cause, coding any condition present at the time of death as the underlying cause may be misleading. In some cultures, there is a strong belief in the need to "let the dead rest" meaning that upon death, there is often no desire to conduct autopsies or other follow-up to accurately ascertain the underlying cause of death.

Following a review of studies that have assessed the accuracy of causes of death coded using ICD systems, [10] found agreement rates of 90% or higher for various types of cancer. Even for this class of causes, the specific (underlying) cause of death is sometimes unknown—a situation referred to as masking in competing risk models.

In recognition of the importance of analysing multiple causes of death, one of the resolutions of the 1899 conference of the International Statistical Institute was a request to the United States government to continue "to undertake studies of the statistical treatment of joint causes of death" [11]. Some researchers [12], have gone as far as suggesting that the "underlying cause concept be replaced by a new class of rates that would define the relative frequency with which diagnosed diseases of sufficient importance appeared among those deceased" and that a "table indicating the frequency with which this group of diseases entered the death certificate should be used to describe the medical circumstances surrounding deaths, especially from chronic or degenerative diseases" [12]. It has been suggested that a matrix of primary or underlying causes (UC) of death by contributory causes (CCs) be used to give an overall picture of mortality patterns and that this matrix could serve as a scheme to identify disease combinations and facilitate the examination of proportional mortality ratios (PWR) by cause of death [13]. Other studies such as [7] have proposed that the average number of causes appearing on the death notification forms be used as a measure of additional diagnostic information gained over the use of single (underlying) cause of death.

Attempts to develop indicators of multiple cause of death are now being consolidated and two recent international workshops in Europe have been dedicated to these efforts (International workshop on the multiple cause-of-death analysis, Paris, 21-22 November 2012 and Rome, 12-13 June, 2014). These workshops have adopted 4 indicators for use in studying multiple causes of death, namely: 1) underlying and multiple cause mortality rates, 2) number of multiple causes per death certificate, 3) standardized ratios of multiple to underlying cause (SRMUs), and 4) cause of death association indicators (CDAIs). These indicators still essentially focus on modelling the additional gains in jointly analysing underlying and contributory causes to modelling underlying cause of death alone. Research is still lacking on how to extract and report leading joint causes of mortality in a database, and hence measure the associations between the leading joint causes as well as provide statistical models for their prediction.

Despite these prolonged calls to analyse mortality from a multiple cause framework, most attempts at analysing multiple causes of death especially from databases coded using the ICD systems have focused on comparing the prevalence of selected causes when coded as underlying cause to their prevalence when coded as multiple cause [1] [2] [4] [6] [8] [12] [13]-[21]. In these studies a cause is defined as a multiple cause if it is mentioned on the death certificate, whether as underlying or contributory. Some recent studies have sought to determine the associations between selected underlying causes of death and contributory causes [9] [20] and to develop models for relating multiple causes of death to background characteristics of the deceased [5] [9] [19] [21]-[23].

Some researchers have suggested that data mining be used to analyse mortality data because these methods are better suited to high dimensional, noisy and very large data sets than statistical techniques such as regression analysis [24]. Association rules can efficiently mine all rules of length l in a data set consisting of p binary attributes. Various statistical measures are available to determine if rules generated are interesting and whether they have statistical significance [25]-[34]. A summary of 38 measures of interestingness and statistical significance in use in data mining are presented in [35]. Among these are support, confidence, lift, chi-square, conviction, odds ratio, fisher's exact test, cosine, coverage, gini, hyper Confidence, hyper Lift, improvement and leverage. Various articles have discussed these measures in the context of market baskets ([36]-[39]).

In this paper, association rules mining (ARM) is presented as an adequate methodology to extract the leading joint morbidity conditions present at the time of death. Specifically, the paper shows how multiple cause of death data can be translated into transactional form suitable for ARM. It then demonstrates how the widely used apriori algorithm implemented in the arules package of R can be modified to extract the leading joint leading causes of death in the database. Finally, the paper shows how statistical measures of interestedness available in data mining literature such as support, confidence, lift, chi-squared and odds-ratio can be used to prune the discovered leading joint morbidity conditions and obtain statistically interesting and even medically surprising relationships among them.

The results of this paper should assist public health officers, epidemiologists, and other researchers to understand the joint morbid conditions often associated with death and enable them to move beyond focusing on single causes of death models. Medical experts should be able to determine which of the statistically significant associations are medically interesting and/or surprising.

2. Comparative Literature

A recent study compared the mortality from Alzheimer's disease, Parkinson's disease, and Dementias in Italy and France using the 2008 Multiple Cause of Death (MCOD) data from both countries [12]. Their analyses evaluated mortality levels when the selected causes were analyzed as MCOD and determined which other causes were listed as either UC or contributing cause (CC) along with the selected causes. The standardised ratio of multiple to underlying cause of death (SMRU) and cause of death association indicator (CDAI) were used as indicators of multiple cause of death. On another study, cancer related mortality from the 2003 MCOD data for Italy and France were analysed using SMRU and CDAI as MCOD indicators and various neoplasms as underlying and contributory causes [20].

A related study analysed MCOD data for deaths occurring in a hospital in Saudi Arabia from 1998 to 2007 and coded using ICD 9 [40]. The causes of death were grouped into 18 standard categories such as infectious and parasitic diseases (001.0 - 139.8) and diseases of the circulatory system (390 - 459). Each of the death certificates was found to have 2 or more causes listed, with an average of just over 6 causes. The authors reported the average number of causes, the frequency with which each cause was listed as main cause (UC), the frequency with which it was listed anywhere (MC) and the ratio of MC to UC by disease group and by age, sex and hospital ward.

In a study that analysed data from the Italian National Vital Statistics Death Registry for 2001 to determine groups of diseases listed together on death certificates of persons aged 70 years and older [11] presented a table of the ratio of the frequency that the disease occurred as underlying cause to the number of times the same disease occurs on death certificates (UC/MC) by selected diseases. Using cluster analysis, they determined which contributory causes were most associated with each of the selected underlying cause. They further used multiple logistic regression models to estimate adjusted and interaction odds ratios for ischaemic heart diseases, cerebrovascular diseases and other diseases of the circulatory system given the age, sex and other contributory causes identified from the cluster analysis.

Analysis of Variance has been used to explore how the epidemiological and sociological factors (age, underlying cause, race, gender, education, year of death) correlated with the total number of medical conditions (TCs) reported on death certificates of adult residents of Michigan who died aged 25 or older in 1989-1991 [41]. Multiple logistic regression models were used to model the association between multiple cause of death as a function of factors available on the death certificate from all death certificates issued by the state of Minnesota between 1990 and 1998 [19]. The predictive factors included demographics of decedent, place of death, type of certifier, disposal method, whether an autopsy was performed, and year of death) [19]. The indicators of multiple cause of death used as dependent variables were 1) whether there were one or more than one cause listed, 2) whether ischemic heart disease was mention or not and 3) whether diabetes mellitus was mentioned or not.

Association rules methodology is being recognised as a potential method for studying multiple causes of death. For example, [42] applied association rules mining to identify new unexpected and interesting patterns in hospital infection control and Public Surveillance data. A general framework to understand the data mining approach in medical data containing the patient profile such as background information and medical history dates is provided in [43]. Another study by [44] applied association rules to multi-item Adverse Drug Effect (ADE) and demonstrated the feasibility of association rules mining to identify multi-item ADEs. Association rules

mining methodology was used by [24] to extract associations between socio-economic variables and mortality rates for 4 different types of cancer from 1988-1992 in the USA. The age-sex standardised mortality rates were categorised into quintiles. The rules were mined using classification based association program (CBA ver 2.0) with minimum support of 3% and minimum confidence of 40%. The lift was used to measure associations between the antecedent and consequent of each rule.

In further use of association rules to analyse mortality data, [45] applied ARM to find relations among activated brain areas in single photon emission computed tomography (SPECT) imaging, with the aim of using the determined patterns for early diagnosis of Alzheimer's disease (AD). The Apriori algorithm was used and the measures of support, confidence and lift used to mine, prune and assess the strength of associations. They noted that ARM is an innovative yet to be explored method and that the strength of association rules is based on the capability of operating with large data bases in an efficient way. Association rules mining have also been used to analyse the frequencies and associations among prescription patterns of Chinese medicinal formulae used for treating and preventing breast cancer [46]. In order to mine the rules the minimum support and confidence were set at 0.1 and 0.6 respectively, in keeping with current practice. For one set of patterns analysed, a total of 11 rules were analysed with 8 of them having the same pattern on the RHS. Another application of ARM to model mortality data was by [47], who analysed a database consisting of 10,000 diabetes patients' records gathered from General Hospital diabetes clinic in Sri Lanka. The attributes selected for analysis were age, gender, diabetes type, level of education, occupation, monthly income, FBS (Fasting Blood Sugar), BMI (Body Mass Index), Potassium level, Cholesterol level, Sodium level, Diastolic blood pressure, Systolic blood pressure, Edema, and Wheezes. They used ARM implement in the WEKA software to generate the rules. They then selected the top three rules based on confidence that had diabetes as the consequent and used decision trees to determine how the identified factors in the antecedent of each rule affected diabetes. The top rule was age = 57_75 & gender = F & diabetes Type = Type 2 & cholesterol =< 5.17 & wheezes = Yes => edema = Yes which had a confidence of 0.95, and the decision tree analysis revealed that wheezing was the most important determinant of edema. Thus association rule is increasingly being used to study multiple causes of death. However, our application of association rules in this paper differs from that of all previous researchers in that the primary intention is to mined all leading joint causes of mortality rather than looking for associations between selected causes of death. Thus our proposed methodology is different from existing methods as explained below.

3. Methodology

3.1. Data Description

The South African multiple causes of death data for each year are derived from notifications of deaths occurring in the specified year that reached Statistics South Africa by the end of processing phase [48]. The death notification form used provides space for reporting up to four possible causes, starting with the immediate condition leading to death. If present, the second, third, and fourth conditions leading to death are recorded sequentially. These are recorded in part 1 of the death notification form. The underlying cause, defined as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury" is selected from this group. Other "significant conditions contributing to death but not resulting in the underlying cause" are recorded in part 2 of the death notification. For the South African multiple causes of death data, only one such additional condition is coded if present, and is referred to as the "other cause" in the data base. France by comparison, provides space for up to 4 causes in part 1 and 2 other causes in part 2, while Italy provides space for up to 8 causes in part 1 and 5 other causes in part 2 [20]. Furthermore like France, South Africa uses the American Automated Classification of Medical Entities (ACME) to determine the underlying cause of death.

In terms of coverage, the data set does not include any deaths in 2008 that were not registered at home affairs or that were registered but had not reached Stats Office during the processing phase [48]. However, The 606699 deaths coded in 2008 represent 98% of the 618324 estimated to have occurred in South Africa in 2008 [49]. This high coverage in a country with high disparities in service availability validates an editorial in the Royal Statistical Society journal that registration of the fact of death must be uncoupled from registration of the cause of death (see Editorial [50]). As with mortality data collected through notifications, some conditions may have been present but undiagnosed, especially causes with no obvious symptoms or that are difficult to diagnose, leading to under reporting of such causes. Similarly certain causes of death may be over reported especially if

the conditions were present at time of death but played no role in the lethal process [20]. These data shall not be adjusted for any under or over reporting.

The causes of death are recorded using 3 digit codes such as A00, A01, B20, etc. However, in this paper, we are interested only in the broad causes of death such as TB, HIV, etc. A given broad condition is considered to be a cause of death if any one or more of its derivatives is coded on the death certificate, whether as underlying cause, contributory cause or both. For example, if B22, A16 and A18 are coded on a death certificate, then the broad causes of death are A15-A19 (TB) and B20-B24 (HIV). Hence each broad group is counted only once, and no distinction is made between underlying and contributory cause. This is consistent with the definition of a cause of death as a multiple cause of death that has been used in the literature. The resulting data file therefore, excludes duplicates and satisfies the transactional form of data required for data mining as shown on the extract in **Table 1**.

3.2. Application of Association Rules Mining to South Africa MCOD

Let t denote the set of causes coded on a typical death certificate, for example in **Table 1**. t = [W00-X59] for first death and t = [A15-A19, B25-B34, B20-B24, N25-N29] for the third death. The first death certificate represents a 1-item set, while the 3rd represents a 4-item set in ARM terminology. Suppose that X and Y denote two non-intersecting subsets of causes that can appear on a death certificate, then an association rule between X and Y involves $n(X)+n(Y)$ total causes of death where $n(X)$ and $n(Y)$ denote the number of causes in set X and Y respectively. For example if $X_1=[A15-A19]$, $X_2=[A15-A19, B25-B34]$ and $Y=[B20-B24]$ then the rule $X_1 => Y$ involves 2 causes of death while the rule $X_2 => Y$ involves 3 causes of death.

The support of the rule $X_1 => Y$ is the proportion of all certificates in which the pair: Tuberculosis (A15-A19) and human immunodeficiency virus [HIV] disease (B20-B24) are coded together. Similarly the support for the rule $X_2 => Y$ is the proportion of all certificates in which all three of tuberculosis (A15-A19), other viral diseases (B25-B34) and human immunodeficiency virus [HIV] disease (B20-B24) are coded together. Therefore, leading pairs of causes of death are those pairs with the highest support in the database among all possible pairs of causes, while leading triplets are those triplets with the highest support, among all possible triplets.

In order to use association rules method to mine leading pairs of causes of death, we set $n(X)=n(Y)=1$. In association rules algorithms such as apriori and arules, $n(Y)=1$, hence rules of length k correspond to rules with $n(X)=k-1$. Suppose that the maximum number of causes that can be coded on each certificate is K, then we can determine joint frequencies of occurrences of $k \le K$ conditions. For the South African MCOD data, $K=\max(k)=5$ hence $n(X)\le 4$.

Thus suppose that

$$C_X=\begin{cases}1 \text{ if all causes in } X \text{ are code on certificate}\\ 0 \text{ otherwise}\end{cases},\quad C_Y=\begin{cases}1 \text{ if cause } Y \text{ is code on certificate}\\ 0 \text{ otherwise}\end{cases}$$

The occurrence of causes in X and Y can be summarised in a 2 × 2 contingency table shown in **Table 2**. This table forms the basis of most measures of interestingness that are currently used to mine, prune and describe quality association rules. A comprehensive summary of these measures is given in [51].

Let n be total number of deaths in the database, F_{ij} be the number of certificates in which $C_X=i$ and $C_Y=j$, $i,j=0,1$ and $M_{ij}=F_{ij}/n$. Hence F_{11} is the number of certificates on which X and Y are jointly coded, that is, the frequency of joint occurrence of all diseases listed in X and Y and M_{11} is the proportion of certificates on which the set of causes in X and Y are jointly coded. It is the estimated probability that all causes in X and Y will contribute to mortality—that is $P(X\cup Y)$ and could be expressed as a crude death rate or as age-sex standardised death rate. In data mining, M_{11} is the support of the rule $X => Y$.

The estimate of the probability $P(X)$, that all conditions listed in X will jointly contribute to death is given by $M_{1+}=M_{11}+M_{10}$. In ARM, it represents the coverage of the rule and it is also called the antecedent support. Similarly $M_{+1}=M_{11}+M_{01}$ is the support of the item set Y, referred to as the prevalence of the association rule [48].

Table 1. Sample of death certificates coded in transactional format for ARM.

	Broad causes listed				
Death ID	1	2	3	4	5 (other)
20080000001	W00-X59				
20080000002	R95-R99				
20080000003	A15-A19	B25-B34	B20-B24	N25-N29	
20080000004	A15-A19				
20080000005	I10-I15	E10-E14			
20080000006	V01-V99				
20080000007	J10-J18	R50-R69			

Table 2. Table of joint frequencies of two sets of causes of death.

	Y (rule consequent)		
X (rule antecedent)	Present $C_Y = 1$	Absent $C_Y = 0$	Total
Present: $C_X = 1$	$F_{11}(M_{11})$	$F_{10}(M_{10})$	$F_{1+}(M_{1+})$
Absent: $C_X = 0$	$F_{01}(M_{01})$	$F_{00}(M_{00})$	$F_{0+}(M_{0+})$
Total	$F_{+1}(M_{+1})$	$F_{0+}(M_{0+})$	$n = F_{0+}(M_{00} = 1)$

3.3. Data Processing and Transformation

The following approach was used to convert these transactional data into binary format useful for data mining:

$$\text{Let } c_{ij} = \begin{cases} 1 \text{ if broad group } j \text{ is coded on the } i\text{-}th \text{ certificate} \\ 0 \text{ if broad group } j \text{ is not coded on the } i\text{-}th \text{ certificate} \end{cases} \quad i = 1, \cdots, n;\ j = 1, \cdots, p$$

where n is the total number of deaths and p is the total number of broad causes of death in the data set. For example suppose $j = 1$ denotes the broad group A00-A09 (intestinal infectious diseases) then $c_{i1} = 1$ if and only if one or more of causes $\text{A00}, \text{A01}, \cdots, \text{A09}$ is coded on the $i\text{-}th$ death certificate. The $n \times p$ incidence matrix $\boldsymbol{C} = (c_{ij})$ is thus a binary incidence matrix for all causes of death in the data set. For the South African MCOD, there are 198 broad groups, hence $p = 198$.

The total number of broad causes listed on the $i\text{-}th$ certificate is $t_i = \sum_{j=1}^{p} c_{ij}$, while the total number of deaths due to cause j is $f_j = \sum_{i=1}^{n} c_{ij}$. The prevalence of broad cause j is $m_j = f_j / n$. This proportion has been referred to as the multiple cause of death ratio [20].

3.4. Extracting Leading Joint Causes of Mortality

In this paper, we are interested in the q leading MCOD, that is the sets of causes with the q largest values for m_j. The computation and ranking of m_j is accomplished using association rules mining methodology ([25] [29] [30] [51] and [52]). We use the R-package: arules in [38] to extract the interesting associations and compute relevant measures to determine the leading joint causes of mortality. The arules package uses the apriori algorithm [25] that requires the minimum values for support $(\mathrm{P}(X \cup Y))$ and confidence $(\mathrm{P}(Y/X))$ to be specified in order to mine the rules [52].

On application we found that the current default values in arules are 0.1 for support and 0.8 for confidence. This implies that only rules r: $X \rightarrow Y$ for which $\mathrm{P}(X \cup Y) \geq 0.1$ and $\mathrm{P}(Y/X) \geq 0.8$ are mined by default. Hence any rule with $\mathrm{P}(Y/X) < 0.8$ will not be generated even if (X, Y) is a set of leading joint causes. In order to use arules to generate the joint leading causes, we eliminate the effect of confidence on the rules generated by setting the minimum confidence to a very small value of to 10^{-7}. In order to mine leading pairs of causes of death, we set $n(X) = n(Y) = 1$. In arules, this is done by setting the rule length parameter to 2. Similarly, by setting rule length to 3 and then to 4, we generate the leading joint triplets and joint quartets of causes of death respectively.

3.5. Determining Interesting Associations

In order to mine interesting rules, we set confidence to 0.6 in line with common practice. We use the support, confidence, lift and odds-ratio as measures of interestingness and use chi-squared and fisher exact test to test for statistical significance of the associations between the rule antecedent (X) and rule consequence (Y).

We compare confidence, ($P(Y/X)$) to the rule prevalence, ($P(Y)$) and rule coverage, ($P(X)$) to determine the direction of association. If confidence is less than the prevalence, then the rule X => Y indicates that the mortality due to cause Y is lower in the general population than among those who also suffered from conditions in set X. In such a case, causes of death in set X are not background risk factors to cause Y. If confidence is greater than prevalence, more persons with conditions listed in set X are dying from cause Y than among the general population. Hence elements of set X are background risk factors to cause Y.

Lift values less than 1 indicate conditions in X and Y do not occur frequently together. This may happen if both sets tend to be underlying causes, or may be due to other medically known reasons or may reveal unexpected findings. A lift value close to 1 indicates that the two sets of causes are independent. The odds-ratio indicates how much more/less likely condition Y is likely to be a cause of death given that the set of conditions in X are present compared to when the set of conditions in X are not present.

Since the odds-ratio, chi-square, fisher exact tests are based on the 2 × 2 table shown in **Table 2**, each can be compared to a chi-squared distribution with 1 degree of freedom. Hence a value of 4 or more (for 5% level of significance) indicates that there is a significant association between causes X and Y. Very large chi-squared values should however, be interpreted with caution, since the chi-squared test suffers from large sample sizes.

3.6. Determining Predictors of a Cause of Interest

With the large number of interesting rules generated, several of the rules may have the same consequence (Y). For example, 8 of the top eleven rules reported by [42] had the same consequence. We group the rules generated by rule consequence and analyse the resulting antecedents (X) in order to determine the leading predictors of Y in the entire data set. For example, if 5 of the top rules each have TB as the consequence (Y), then the combination of causes in the 5 antecedents provide a list of causes of death associated with TB in the database.

4. Results

A total of 606,699 deaths were coded in South Africa in 2008. The deaths were attributed to a combined total of 954,401 natural causes and 54,446 unnatural causes, giving a total of 1,008,847 conditions that were coded. Hence, on average, 1.7 conditions were coded on each certificate of which averages of 1.6 coded conditions were due to natural causes. In total, 191 of the possible 198 broad causes contributed to one or more deaths in the 2008 database.

4.1. Leading Multiple Causes of Death

Figure 1 shows the contribution of the top 20 broad causes coded either as UC or associated cause on 2008 South Africa multiple cause of death data. Together, these conditions account for 71.9% of all coded natural causes and 74.6% of all natural causes coded as Cause A.

TB (A15-A19) was the leading contributor to mortality in South Africa in 2008, appearing on 14.1% of all death certificates. The second and third leading causes were Ill-defined and unknown cause of mortality (R95-R99) and influenza and pneumonia (J10-J18), which appeared on 12.8% and 12.4% of death certificates respectively. HIV

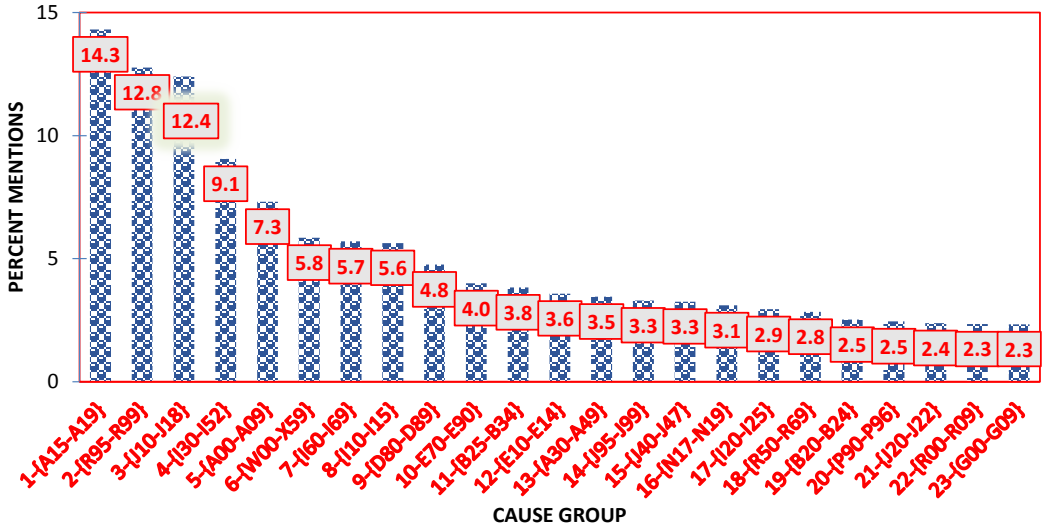

Figure 1. Leading broad causes of mortality in 2008 in South Africa.

(B20-B24) is not among the 10 leading causes of death in terms of mention. In fact HIV is mentioned only on 2.52% of all death certificates and is the 19th ranked leading cause of death mentioned.

4.2. Interesting Associations between Leading Joint Causes of Mortality

In this section, the interesting and statistically significant rules that were extracted using the support-confidence framework are presented in **Table 3**. As discussed in the methodology, the minimum confidence was 0.6, that is only rules in which the antecedence, *Y* occurs in at least 60% of deaths in which the precedence, *X* is a cause of death were mined. With this minimum confidence, 8 rules were generated with a minimum support of 0.0001 and 21 rules with a minimum support of 0.00005.

The top rule with a support of 356.4 deaths per 100,000 specifies that deaths with diabetes mellitus (E10-14) and cerebrovascular disease (I60-I69) listed tend to have hypertensive diseases (I10-I15) also listed. Rule #2 with support of 215.8 deaths per 100,000 specifies that deaths with other disorders of the skin and subcutaneous tissue (L80-L99) imply a high likelihood of having other bacterial diseases (A30-A49). The odds-ratio indicates that the likelihood that a death in which other bacterial diseases is a cause given that other disorders of the skin as well as subcutaneous tissue are present is 45.2 times the likelihood that such a death will not have other disorders of the skin as well as subcutaneous tissue.

From rule #3, the presence of symptoms and signs involving circulatory and respiratory systems (R00-R09) together with symptoms and signs involving digestive and abdominal systems (R10-R19) are predictive of the presence of ill-defined and unknown causes of mortality (R95-R99). However, the lift and odds-ratio is relatively small compared to values for the other rules, indicating that the association between the precedent (*X*) and antecedent may not be very strong.

Rule #4 specifies that cerebral palsy and other paralytic syndromes (G80-G83) together with other disorders of the skin and subcutaneous tissue (L80-L99) are predictors of the presence of other bacterial diseases (A30-A49). The support of 25.4 deaths per 100,000, the lift (20.4) and odds-ratio (67.0) are very high, indicating a strong association between the LHS and RHS causes. Rule #6 specifies that deaths with diabetes mellitus (E10-14) together with other disorders of the skin and subcutaneous tissue (L80-L99) imply other bacterial diseases (A30-A49) with a support of 17.6. This is sub-rule to rule #2. All the rules in **Table 3** have lift values and odds-ratios much greater than 1.0 indicating a highly positive association between the LHS and RHS causes.

The Pearson chi-squared values were all very large with a minimum value of 195. Correspondingly, the fisher exact test probabilities were all less than 0.00001. Thus the association between the precedent and antecedent in each of the rules was highly significant. The hyper confidence values were all equal to 1.0 and the difference in confidence (doc), gini, coverage, improvement, leverage, phi and RLD statistics were all computed but did provide any additional diagnostic information. Hence these measures were computed but excluded from **Table 3**.

Table 3. Rules with confidence of 0.6 or higher sorted by support.

Rule#	lhs	rhs	Support/100,000	Confidence	Lift	Odds-ratio	Hyper-lift	Conviction	cosine
1	{E10-E14, I60-I69}	{I10-I15}	356.4	0.69	12.2	38.9	10.4	3.0	0.21
2	{L80-L99}	{A30-A49}	215.8	0.60	17.5	45.2	13.8	2.4	0.19
3	{R00-R09, R10-R19}	{R95-R99}	32.1	0.61	4.8	10.7	3.5	2.2	0.04
4	{G80-G83, L80-L99}	{A30-A49}	25.4	0.70	20.4	67.0	11.0	3.3	0.07
5	{G80-G83, I10-I15}	{I60-I69}	22.6	0.71	12.7	42.1	7.2	3.3	0.05
6	{E10-E14, L80-L99}	{A30-A49}	17.6	0.66	19.1	53.9	8.9	2.8	0.06
7	{I10-I15, J60-J70}	{I60-I69}	16.8	0.63	11.2	28.2	6.0	2.5	0.04
8	{E10-E14, I20-I25, I60-I69}	{I10-I15}	11.5	0.64	11.4	30.1	5.8	2.6	0.04
9	{J10-J18, J95-J99, R50-R69}	{D80-D89}	9.4	0.66	13.8	38.1	6.3	2.8	0.04
10	{C81-C96, D70-D77}	{A30-A49}	9.1	0.72	21.0	73.7	7.9	3.5	0.04
11	{L80-L99, W00-X59}	{A30-A49}	8.9	0.70	20.4	66.1	7.7	3.2	0.04
12	{F10-F19, I26-I28}	{J40-J47}	8.4	0.71	22.7	75.6	8.5	3.3	0.04
13	{E10-E14, G80-G83}	{I10-I15}	8.2	0.65	11.5	31.0	5.0	2.7	0.03
14	{L80-L99, N17-N19}	{A30-A49}	7.6	0.61	17.8	44.6	6.6	2.5	0.04
15	{E40-E46, E70-E90, K50-K52}	{D80-D89}	7.1	0.61	12.9	31.9	5.4	2.5	0.03
16	{D80-D89, E40-E46, K50-K52}	{E70-E90}	7.1	0.61	16.0	39.9	6.1	2.5	0.03
17	{E10-E14, I60-I69, J60-J70}	{I10-I15}	5.9	0.71	12.5	40.2	5.1	3.2	0.03
18	{E10-E14, I10-I15, J60-J70}	{I60-I69}	5.9	0.65	11.7	31.9	4.5	2.7	0.03
19	{E10-E14, G80-G83, I60-I69}	{I10-I15}	5.1	0.76	13.4	51.9	5.2	3.9	0.03
20	{E10-E14, G80-G83, I10-I15}	{I60-I69}	5.1	0.62	11.1	27.5	4.4	2.5	0.02
21	{B25-B34, E70-E90, R50-R69}	{A00-A09}	5.1	0.62	8.5	20.7	3.9	2.4	0.02

4.3. Leading Pairs of Joint Causes of Death

A total of 19 pairs of leading joint causes were generated with a minimum support of 0.005 (*i.e.* 5 deaths per 1000) as shown in **Figure 2**.

As shown in **Figure 2**, the leading pair of joint causes of death was found to be metabolic disorders (E70-E90) and intestinal infectious diseases (A00-A09). Together, the pair contributed to 18.6 deaths per 1000. The second pair of joint causes, cerebrovascular diseases (I60-I69) and hypertensive diseases (I10-I15) together account for 18.3 deaths per 1000. The 3rd leading pair consists of certain disorders involving the immune mechanism (D80-D89) and tuberculosis (A15-A19) which appear together on 15.4 deaths certificates per 1000. The 4th leading pair of joint causes is diabetes mellitus (E10-E14) and hypertensive diseases (I10-I15) which account for 14.0 deaths per 1000. The 5^{th} and 6^{th} pairs consist respectively of other viral diseases (B25-B34) and TB (A15-A19) which are responsible for 13.6 deaths per 1000 and HIV (B20-B24) with TB (A15-A19) contributing to 9.5 deaths per 1000. All other joint pairs appear on less than 10 per 1000 deaths.

Table 4 shows the association between the leading pairs of joint causes ranked with respect to support. The measures of interestingness presented include the support, which estimates the probability that a death would have both conditions listed as joint causes with or without additional causes; the prevalence which estimates probability that the rule will have the RHS condition as one of the causes and the coverage which gives an estimate of probability that a death would have the LHS condition listed as one of the causes. These are all expressed per 1000 deaths for ease of interpretation.

Among the top 20 rules, the maximum coverage value for the LHS conditions is 123.8 corresponding to rule 10,

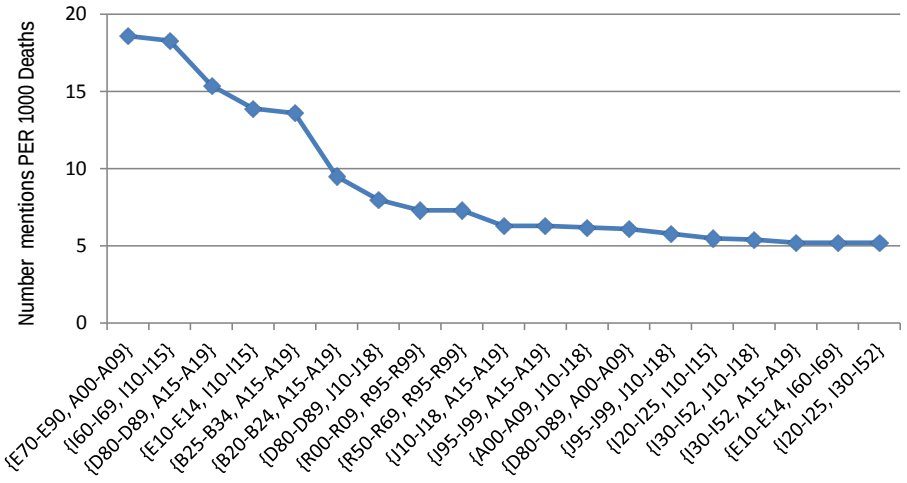

Figure 2. Leading pairs of joint MCOD in South Africa—2008.

Table 4. Leading pairs of joint causes of death.

Rank	lhs (X)	rhs (Y)	Support P(X, Y) /1000	Coverage P(X)/1000	Prevalence P(Y)/1000	Confidence P(Y/X)	Conviction P(X/Y)	Lift P(X, Y)/ [P(X)P(Y)]	Odds ratio	Chi-square	Fishers-exact test
1	{E70-E90}	{A00-A09}	18.6	38.4	73.1	0.48	1.80	6.6	15.6	60539.0	0.000
2	{I60-I69}	{I10-I15}	18.3	56.0	56.4	0.33	1.40	5.8	11.5	49287.8	0.000
3	{D80-D89}	{A15-A19}	15.4	47.6	141.3	0.32	1.27	2.3	3.1	8247.5	0.000
4	{E10-E14}	{I10-I15}	13.9	35.7	56.4	0.39	1.55	6.9	13.9	46986.9	0.000
5	{B25-B34}	{A15-A19}	13.6	38.2	141.3	0.35	1.33	2.5	3.6	9067.5	0.000
6	{B20-B24}	{A15-A19}	9.5	25.2	141.3	0.38	1.38	2.7	3.9	7218.2	0.000
7	{D80-D89}	{J10-J18}	8.0	47.6	123.8	0.17	1.05	1.4	1.5	534.1	0.000
8	{R00-R09}	{R95-R99}	7.3	22.4	127.6	0.33	1.30	2.6	3.5	4979.4	0.000
9	{R50-R69}	{R95-R99}	7.3	27.5	127.6	0.26	1.19	2.1	2.6	2906.4	0.000
10	{J10-J18}	{A15-A19}	6.3	123.8	141.3	0.05	0.90	0.4	0.3	5739.9	0.000
11	{J95-J99}	{A15-A19}	6.3	32.2	141.3	0.20	1.07	1.4	1.5	489.7	0.000
12	{A00-A09}	{J10-J18}	6.2	73.1	123.8	0.09	0.96	0.7	0.6	662.7	0.000
13	{D80-D89}	{A00-A09}	6.1	47.6	73.1	0.13	1.06	1.7	1.9	1317.9	0.000
14	{J95-J99}	{J10-J18}	5.8	32.2	123.8	0.18	1.07	1.5	1.6	608.6	0.000
15	{I20-I25}	{I10-I15}	5.5	27.1	56.4	0.20	1.18	3.6	4.6	6680.5	0.000
16	{I30-I52}	{J10-J18}	5.4	83.7	123.8	0.06	0.94	0.5	0.5	1769.2	0.000
17	{I30-I52}	{A15-A19}	5.2	83.7	141.3	0.06	0.92	0.4	0.4	2843.9	0.000
18	{E10-E14}	{I60-I69}	5.2	35.7	56.0	0.15	1.10	2.6	3.1	3393.8	0.000
19	{I20-I25}	{I30-I52}	5.2	27.1	83.7	0.19	1.13	2.3	2.7	2557.6	0.000
20	{B25-B34}	{J10-J18}	5.0	38.2	123.8	0.13	1.01	1.0	1.1	7.2	0.004
21	{B25-B34}	{A00-A09}	4.9	38.2	73.1	0.13	1.06	1.8	1.9	1087.5	0.000
22	{A00-A09}	{A15-A19}	4.6	73.1	141.3	0.06	0.92	0.4	0.4	2402.8	0.000
23	{P05-P08}	{P20-P29}	4.4	10.5	13.8	0.41	1.69	30.0	73.5	76127.3	0.000
24	{A30-A49}	{J10-J18}	4.2	34.4	123.8	0.12	1.00	1.0	1.0	1.3	0.875
25	{E10-E14}	{I30-I52}	3.8	35.7	83.7	0.11	1.03	1.3	1.3	159.6	0.000
26	{I10-I15}	{I30-I52}	3.8	56.4	83.7	0.07	0.98	0.8	0.8	134.9	0.000
27	{I20-I25}	{E10-E14}	3.8	27.1	35.7	0.14	1.12	3.9	4.7	5188.5	0.000

while the highest value for prevalence is 141.3, corresponding to rule number 3, 5, 6, 10, 11 and 17. All of these rules have TB (A15-19) as the antecedent of the rule, hence the common prevalence value. The diseases mention in these rules are the best set of associated causes of death to TB in the data set.

The chi-squared statistics for the top 20 pairs are all very large and the corresponding p-values for the fisher exact test are all quite small $(p < 0.005)$, suggesting highly significant associations between the pairs of causes. However, both of these statistics are influence by the large sample sizes involved. The odds-ratio and lift values suggest that not all of the associations may be interesting, since in some cases both statistics are close to 1.

For the leading joint pair, the odds-ratio of 15.6 and lift value of 6.6 indicate a strong positive association. Similarly the lift for the next 5 leading pairs is 2.3 or greater, while the corresponding odds-ratios are 3.1 or greater. Thus these associations are all highly significant and positive. The lift of 1.4 and odds ratio of 1.5 for the 7th pair: certain disorders involving the immune mechanism (D80-D89) with influenza and pneumonia (J10-J18) suggest that the association is weak. Some of the associations have lift and odds ratios less than 1, such as the 10th, 12th, 16th and 17th pairs. These lift and odds ratios indicate negative associations between the pairs.

4.4. Leading Triplets, Quartets and Quintets of Joint Causes of Death

Table 5 presents the leading triplets, quartets and quintets of joint causes of death. The leading triplets of joint causes consist of diabetes mellitus (E10-E14), cerebrovascular diseases (I60-I69) and hypertensive diseases (I10-I15) which together are responsible for 356 deaths in 100,000. All measures of interestingness are high, indicating a strong association among the causes involved.

The second leading triplet consists of diabetes mellitus (E10-E14), Ischaemic heart diseases (I20-I25) and hypertensive diseases (I10-I15) which together accounted for 186 deaths in 100,000 while Certain disorders involving the immune mechanism (D80-D89), Metabolic disorders (E70-E90) and Intestinal infectious diseases (A00-A09) are the 3rd leading triplets responsible for 143 deaths per 100,000.

Table 5. Leading triplets, quartets and quintets of joint causes of death in South Africa in 2008.

Rule No	LHS (X) triplets	RHS (Y)	Support P(X,Y) /100,000	Coverage P(X) /100,000	Confidence P(Y/X)	Conviction	Lift P(X,Y)/ [P(X)P(Y)]
1	{E10-E14, I60-I69}	{I10-I15}	356.4	519.7	0.69	3.00	12.17
2	{E10-E14, I20-I25}	{I10-I15}	186.1	375.6	0.50	1.87	8.79
3	{D80-D89, E70-E90}	{A00-A09}	143.2	352.1	0.41	1.56	5.57
4	{E10-E14, I10-I15}	{I30-I52}	101.0	1392.9	0.07	0.99	0.87
5	{A00-A09, E70-E90}	{J10-J18}	99.7	1860.4	0.05	0.93	0.43
6	{B25-B34, E70-E90}	{A00-A09}	92.8	217.2	0.43	1.62	5.85
	Quartets						
1	{E10-E14, I10-I15, I20-I25}	{I30-I52}	18.8	186.1	0.10	1.02	1.21
2	{E10-E14, I10-I15, I60-I69}	{I30-I52}	11.9	356.4	0.03	0.95	0.40
3	{E10-E14, I20-I25, I60-I69}	{I10-I15}	11.5	18.0	0.64	2.64	11.39
4	{A00-A09, B25-B34, E70-E90}	{A15-A19}	10.1	92.8	0.11	0.96	0.77
	Quintets						
1	{I10-I15, I20-I25, I30-I52, I60-I69}	{E10-E14}	1.0	1.5	0.3	1.3	8.0
2	{A30-A49, I10-I15, I70-I79, Y40-Y84}	{E10-E14}	0.8	1.6	0.6	2.6	17.5
3	{A30-A49, E10-E14, I60-I69, L80-L99}	{I10-I15}	0.8	2.0	0.4	1.5	6.3
4	{A00-A09, A15-A19, E70-E90, I30-I52}	{B25-B34}	0.8	2.3	0.2	1.2	5.4
5	{A00-A09, A15-A19, B25-B34, E70-E90}	{I30-I52}	0.8	2.5	0.1	1.0	1.0

The leading quartets of joint causes are shown in part 2 of **Table 5**. The first of these quartets contributed to 18.8 deaths per 100,000 and consisted of diabetes mellitus (E10-E14), hypertensive diseases (I10-I15), ischaemic heart diseases (I20-I25) and other forms of heart disease (I30-I52). The second leading quartet of joint causes comprises diabetes mellitus (E10-E14), hypertensive diseases (I10-I15), cerebrovascular diseases (I60-I69), and other forms of heart disease (I30-I52).

The precedent of these joint causes comprises of the 1st leading triplets of joint causes which had a support of 35.6 deaths per 10,000 (*i.e.* 356.4 per 100,000) as shown under the coverage for rule 2 of the joint quartets.

When ischaemic heart diseases (I20-I25) is replaced with cerebrovascular diseases (I60-I69) to get the second leading quartet, the association between other forms of heart disease (I30-I52) and the LHS set becomes negative (lift = 0.40). The chi-squared test is found to 72.0 on 1df indicating a strong negative association ($p < 0.001$). Thus, a medical explanation is required for the relationship between these causes.

The 3rd leading quartet is similarly to the first 2. Indeed, the first 3 leading quartets all include diabetes mellitus (E10-E14) and hypertensive diseases (I10-I15) along with a combination of pairs of causes from: other forms of heart disease (I30-I52), cerebrovascular diseases (I60-I69) and ischaemic heart diseases (I20-I25). The fourth leading quartet is entirely different, comprising of intestinal infectious diseases (A00-A09), tuberculosis (A15-A19), other viral diseases (B25-B34) and metabolic disorders (E70-E90). This quartet is responsible for 10 in every 100,000 deaths. The lift value of 0.77 indicates that although these are leading joint causes, the presence of the other members (in the LHS) is inversely related to the presence of TB.

While considering the quintet of joint causes of death, it should be recalled that only a small proportion (5.2%) of deaths in the database involved 5 causes. The leading quintet of joint causes accounts for just 1 in 100,000 deaths as shown in the third part of **Table 4**. This comprises of the causes in the 3 leading quartets ({E10-E14, I10-I15, I20-I25, I30-I52, I60-I69}), which is therefore, not statistically surprising except for the fact that some of the causes are negatively associated as discussed in the preceding paragraph.

5. Discussion

This paper found that the leading single multiple cause of death in South Africa was TB (A15-A19) followed by ill-defined and unknown causes of mortality (R95-R99), influenza and pneumonia (J10-J18), other forms of heart disease (I30-I52) and intestinal infectious disease (A00-A09). This is consistent with the leading MCOD reported in [53], Table 4.15. HIV is mentioned on 2.5% of all death certificates and is the 19th ranked leading multiple cause of death. A study by [21] showed that when it is mentioned, HIV is almost always coded as the underlying cause of death. Hence it is not surprising that it is the 9th leading underlying cause of death in this data set as reported in [53], Table 4.16.

The rules shown in **Table 3** were mined using traditional implementation of association rules techniques that is based on the support-confidence framework with minimum confidence set at 0.6. Hence these rules, ranked by support, represent the list of leading interesting associations among the causes of death in association rules terminology.

The leading rule in terms of support involves 3 causes: cebrovascular diseases (I60-I69), diabetes mellitus (I10-I15) and hypertensive diseases (E10-E14). All three causes are ranked outside the top 5 leading single multiple causes of death, indicating that their appearance in the most interesting rule is not a consequence of high prevalence in the population and should be of interest to researchers, public health and medical diagnosis. As reported in [53], these causes are ranked 4th, 5th and 10th underlying leading causes of death. The association indicates that they tend to occur together.

The association between other disorders of the skin and subcutaneous tissues (L80-L99) with bacteria diseases (A30-A49) is ranked as the second leading interesting rule even though neither is ranked in the top 10 leading underlying causes nor top 10 leading multiple causes of death. Only bacterial diseases, ranked #13, is among the top 20 leading MCOD.

Among the three causes in rule #3, ill-defined and unknown causes of mortality (R95-R99) is ranked as the second leading cause of death, while symptoms of signs involving circulatory and respiratory systems (R00-R09) and symptoms of signs involving digestive and abdominal (R10-R19) are not ranked among the top 20. Similar analysis suggests that most of the leading interesting rules ranked by support are not a consequence of high a prevalence of the individual causes that constitute the rules. For example, although TB is by far the leading single MCOD as well as underlying cause, it featured for the first time in rule #39: {A15-A19, R47-R49} => {J95-J99}

(not shown) of the most interesting rules in **Table 3**. An analysis of this rule revealed that if a death certificate has TB (A15-A19) and symptoms & signs involving speech & voice (R47-R49) as a joint cause of death, then the odds that the certificate will also have other diseases of the respiratory systems (J95-J99) listed, are 56.8 times the odds that other diseases of the respiratory systems will not be listed.

Leading Joint Causes of Mortality

The leading joint causes of death in South Africa in 2008 have been determined and tabulated (**Table 4** and **Table 5**) in this study. The first leading pair of joint causes is found to be metabolic disorders (E70-E90) and intestinal infectious diseases (A00-A09). Curiously, metabolic disorders do not appear along with any other cause among the top 20 pairs of leading causes. In fact, further analysis reveals that it next occurs with certain disorders involving the immune mechanism (D80-D89) as the 30th leading pair that cause 3.5 per 1000 deaths, and as the 44th leading pair with influenza and pneumonia which as pair cause 2.9 deaths per 1000. Its association with certain disorders involving the immune mechanism is positive and moderate—with a lift value of 1.9 and an odds ratio of 2.1 while its association with pneumonia is also moderate but negative, with a lift of 0.6 and an odds ratio of 0.6. Its association with TB is also found to be negative, with a odds ratio and lift of 0.4.

The second leading pair of joint causes consists of cerebrovascular diseases (I60-I69) and hypertensive diseases (I10-I15) contributed 18.3 deaths per 1000. The 3rd leading pair consists of certain disorders involving the immune mechanism (D80-D89) and tuberculosis (A15-A19) contributing to 15.4 deaths per 1000. Hypertensive diseases together with diabetes mellitus are the fourth leading pair of joint causes of death, while hypertensive diseases with ischaemic heart diseases are the 15th leading pair of joint causes of death. Cerebrovascular diseases next appears with Diabetes mellitus (E10-E14) as the 18th leading pair.

TB which is the leading single multiple cause of death and pneumonia which is the third leading cause of death both appear among 6 of the 20 leading pairs, but metabolic disorders which is the second leading cause, appears in only 2 of the 20 leading pairs of joint causes. Interestingly, while TB and HIV appear as the 6th leading pair and TB and pneumonia appear as the 10th leading pair, pneumonia and HIV do not appear among the 20 leading joint pairs of causes. In fact, further analysis reveals that pneumonia and HIV appear as the 36th leading pair of multiple causes of death, and that their joint occurrence is purely due to chance.

Among the leading quartets the first 3 conditions (diabetes mellitus, ischemic heart disease and hypertensive disease) are the same as the second leading joint triplets shown in part 1 of **Table 5**. The appearance of other forms of heart disease (I30-I52) on the RHS of the fourth leading triplet and the RHS of the first and second leading quartets is statistically surprising given that it only previously appeared among the 16th (with Influenza and pneumonia (J10-J18)) and 17th (with TB (A15-A19)) leading pairs of joint causes. Influenza and pneumonia does not appear with any forms of heart disease among the leading multiple causes in **Table 5**. TB appears on three of the 15 rules in **Table 5**, and in two of these rules (4th and 5th quintets) it is in association with other forms of heart disease.

Traditional association rules mining results (**Table 3**) also failed to pick up this association because of the high minimum confidence threshold. The second and third quintets are quite interesting as each includes a cause that has not featured among any of the previous leading causes or joint causes. Further analysis revealed that one of the members of the second leading quintet, complications of medical and surgical care (Y40-Y84) appears as the 47th leading single MCOD in the database. Similarly, other disorder of the skin and subcutaneous tissue (L80-L99) which appears among the 3rd leading quintets of causes of death is the 54th single leading MCOD.

6. Conclusions

This study confirms the medical knowledge that TB and HIV are associated causes of death and as a pair are among the leading joint causes of death in South Africa. However, there are a number of more prevalent single causes, joint pairs as well multiple combinations of causes of death with higher prevalence and stronger associations than TB and HIV. Furthermore, although pneumonia has long been seen as an opportunistic infection with HIV, and indeed has been used as a case defining condition for HIV, no interesting rules involving influenza and pneumonia and HIV as joint causes where found. Indeed HIV appears among the leading joint causes with TB only.

A possible explanation is that influenza and pneumonia are highly prevalent in this population, with bacterial pneumonia much more treatable than TB. It is possible that HIV infected persons tend to use medical facilities

more than the general population, and hence are more likely to seek treatment for other illnesses including opportunistic infections like pneumonia. As a result the prevalence of curable infections among HIV+ patients at the time of death may be lower than the prevalence of such conditions among the general population.

The methodology advanced in this study has unveiled some interesting leading joint associations, leading pairs and multiple combinations of causes of death in South Africa in 2008. The strength of these associations suggests that they can be expected to be present in other years. A striking finding is that the leading joint triplets are mainly consisting of combinations of diabetes mellitus, hypertensive diseases, metabolic disorders and intestinal infectious diseases occurring together with other causes such as cerebrovascular disease, ischaemic heart diseases and disorders involving the immune systems.

Acknowledgements

The authors wish to thank Statistics South Africa for making data available in a very useful friendly format and free of charge for researchers to use.

Disclosure

Authors report no conflicts of interest in this work.

References

[1] Moriyama, I.M. (1956) Development of the Present Concept of Cause of Death. *American Journal of Public Health*: *Mortality Analysis*, **46**, 437-441. http://dx.doi.org/10.2105/ajph.46.4.436

[2] Weiner, L., Bellows, M.T., McAvoy, G.H. and Cohen, E.V. (1955) Use of Multiple Causes in the Classification of Deaths from Cardivascular-Renal Disease. *American Journal of Public Health*, **45**.

[3] D'Amico, M., Agozzino, E., Biagino, A., Simonetti, A. and Marinelli, P. (1999) Ill-Defined and Multiple Causes on Death Certificates—A Study of Misclassification in Mortality Statistics. *European Journal of Epidemiology*, **15**, 141-148. http://dx.doi.org/10.1023/A:1007570405888

[4] Nam, C.B., Eberstein, I.W., Deeb, L.C. and Terrie, E.W. (1989) Infant Mortality by Cause: A Comparison of Underlying and Multiple Cause Designations. *European Journal of Population*, **5**, 45-70. http://dx.doi.org/10.1007/BF01796788

[5] Forcheh, N., Setlhare, K. and Amey, A.K.A. (2014) Modeling Severity of Tuberculosis as a Multiple Cause of Death in South Africa. *Journal of Tuberculosis Research*, **2**, 16-29. http://dx.doi.org/10.4236/jtr.2014.21003

[6] Erhardt, C.L. (1958) What Is "the Cause of Death"? *JAMA*, **168**, 161-168. http://dx.doi.org/10.1001/jama.1958.03000020023005

[7] Dorn, H.F. and Moriyama, I.M. (1964) Uses and Significance of Multiple Cause Tabulations for Mortality Statistics. *American Journal of Public Health*, **54**, 400-406. http://dx.doi.org/10.2105/AJPH.54.3.400

[8] Frova, L., Salvatore, M.A., Pappagallo, M. and Egidi, V. (2009) Multiple Cause of Death Approach to Analyze Mortality Patterns. *Genus*, **65**.

[9] Désesquelles, A., Demuru, E., Salvatore, M.A., Pappagallo, M., Frova, L., Meslé, F. and Egidi, V. (2014) Mortality from Alzheimer's Disease, Parkinson's Disease, and Dementias in France and Italy: A Comparison Using the Multiple Cause-of-Death Approach. *Journal of Aging and Health*, **26**, 283-315. http://dx.doi.org/10.1177/0898264313514443

[10] Basu, S. (2010) Breast Cancer Survival, Competing Risks and Mixture Cure Model: A Bayesian Analysis. *Journal of the Royal Statistical Society*: *Series A* (*Statistics in Society*), **173**, 307-329. http://dx.doi.org/10.1111/j.1467-985X.2009.00618.x

[11] WHO (1990) History of the Development of the ICD. http://www.who.int/classifications/icd/en/HistoryOfICD.pdf

[12] Treloar, A.E. (1956) The Enigma of Cause of Death. *JAMA*, **162**, 1376-1379. http://dx.doi.org/10.1001/jama.1956.02970320024007

[13] Wong, O., Rockette, H.E., Redmond, C.K. and Heid, M. (1978) Evaluation of Multiple Causes of Death in Occupational Mortality Studies. *Journal of Chronic Diseases*, **31**, 83-193. http://dx.doi.org/10.1016/0021-9681(78)90033-4

[14] Guralnick, L. (1966) Some Problems in the Use of Multiple Cause of Death. *Journal of Chronic Diseases*, **19**, 979-990. http://dx.doi.org/10.1016/0021-9681(66)90031-2

[15] Speizer, F.E., Trey, C. and Parker, P. (1977) The Uses of Multiple Causes of Death Data to Clarify Changing Patterns of Cirrhosis Mortality in Massachusetts. *American Journal of Public Health*, **67**, 333-336. http://dx.doi.org/10.2105/AJPH.67.4.333

[16] Goodman, R.A., Manton, K.G., Nolan Jr., T.F., Bregman, D.J. and Hinman, A.R. (1982) Mortality Data Analysis Using a Multiple-Cause Approach. *JAMA*, **247**, 793-796. http://dx.doi.org/10.1001/jama.1982.03320310041026

[17] Israel, R.A., Rosenberg, H.M. and Curtin, L.R. (1986) Analytic Potential for Multiple Cause-of-Death Data. *American Journal of Epidemiology*, **124**, 161-179.

[18] Rushton, L. (1994) Use of Multiple Causes of Death in the Analysis of Occupational Cohorts—An Example from the Oil Industry. *Occupational and Environmental Medicine*, **51**, 722-729. http://dx.doi.org/10.1136/oem.51.11.722

[19] Wall, M.M., Huang, J.Z., Oswald, J. and McCullen, D. (2005) Factors Associated with Reporting Multiple Causes of Death. *BMC Medical Research Methodology*, **5**, 4. http://dx.doi.org/10.1186/1471-2288-5-4

[20] Désesquelles, A., Salvatore, M.A., Pappagallo, M., Frova, L., Pace, M., Meslé, M. and Egidi, V. (2012) Analysing Multiple Causes of Death: Which Methods for *Désesquelles et al.* 313 Which Data? An Application to the Cancer-Related Mortality in France and Italy. *European Journal of Population/Revue Européenne de Démographie*, **28**, 467-498. http://dx.doi.org/10.1007/s10680-012-9272-3

[21] Amey, A.K.A., Forcheh, N. and Setlhare, K. (2012) Multiple Causes of Death Models for Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome and Related Mortality in South Africa in 2006 and 2007. *Open Access Medical Statistics*, **2**, 1-13. http://dx.doi.org/10.2147/OAMS.S23627

[22] Johnson, N.E. and Christenson, B.A. (1998) Socio-Demographic Correlates of Multiple Causes of Death: Real or Artifactual? *Population Research and Policy Review*, **17**, 261-274. http://dx.doi.org/10.1023/A:1005985730638

[23] Jung, R.S., Bennion, J.R., Sorvillo, F. and Bellomy, A. (2010) Trends in Tuberculosis Mortality in the United States, 1990-2006. A Population-Based Case-Control Study. *Public Health Reports*, **125**, 389-397.

[24] Vinnakota, S. and Lam, N.S.N. (2006) Socioeconomic Inequality of Cancer Mortality in the United States: A Spatial Data Mining Approach. *International Journal of Health Geographics*, **5**, 9. http://dx.doi.org/10.1186/1476-072X-5-9 http://www.ij-healthgeographics.com/content/5/1/9

[25] Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the* 1993 *ACM SIGMOD International Conference on Management of Data*, Washington DC, 26-28 May 1993, 207-216. http://dx.doi.org/10.1145/170035.170072

[26] Hong, T., Kuo, C. and Chi, S. (2001) Trade-Off between Computation Time and Number of Rules for Fuzzy Mining from Quantitative Data. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **9**, 587-604. http://dx.doi.org/10.1142/S0218488501001071

[27] Liu, F., Lu, Z.D. and Lu, S.F. (2001) Mining Association Rules Using Clustering. *Intelligent Data Analysis*, **5**, 309-326.

[28] Melab, N. (2001) Data Mining: A Key Contribution to E-Business. *Information & Communications Technology Law*, **10**, 309-318. http://dx.doi.org/10.1080/13600830120081935

[29] Berzal, F., Blanco, I., Sanchez, D. and Vila, M.A. (2002) Measuring the Accuracy and Interest of Association Rules: A New Framework. *Intelligent Data Analysis*, **6**, 221-235.

[30] Hahsler, M.A. (2005) A Comparison of Commonly Used Interesting Measures of Association Rules. http://wwwai.wu.wien.ac.at/~hahsler/research/assosciation_rules/measures.html

[31] Li, Q., Feng, L. and Wong, A. (2005) From Intra-Transaction to Generalized Inter-Transaction: Landscaping Multidimensional Contexts in Association Rule Mining. *Information Sciences*, **172**, 361-395. http://dx.doi.org/10.1016/j.ins.2004.07.006

[32] Yan, P. and Cheng, G.Q. (2005) Discovering a Cover Set of ARsi with Hierarchy from Quantitative Databases. *Information Sciences*, **173**, 319-336. http://dx.doi.org/10.1016/j.ins.2005.03.003

[33] Richards, G. and Rayward-Smith, V.J. (2005) The Discovery of Association Rules from Tabular Databases Comprising Nominal and Ordinal Attributes. *Intelligent Data Analysis*, **9**, 289-307.

[34] Rodríguez, A., Carazo, J.M. and Trelles, O. (2005) Mining Association Rules from Biological Databases. *Journal of the American Society for Information Science and Technology*, **56**, 493-504. http://dx.doi.org/10.1002/asi.20138

[35] Geng, L. and Hamilton, H.J. (2006) Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, **38**, 5-31. http://dx.doi.org/10.1145/1132960.1132963

[36] Ekosse, G.I.E. and Forcheh, N. (2007) Mining Association Rules Applied to Goethite and Haematite Abundances in Manganese-Contaminated Soils. *Polish Journal of Environmental Studies*, **16**, 531-538.

[37] ABS, Australian Bureau of Statistics (2003) Multiple Cause of Death Analysis 1997-2001. http://www.abs.gov.au/ausstats/abs@.nsf/956c382b0b05ba7d4a2568010004e173/0971aa07f3c12518ca256d6b0003b678! Open Document

[38] Hahsler, M., Chelluboina, S., Hornik, K. and Buchtac, C. (2011) The Arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Datasets. *Journal of Machine Learning Research*, **12**, 1977-1981.

[39] Hahsler, M. (2006) A Model-Based Frequency Constraint for Mining Associations from Transaction Data. *Data Mining and Knowledge Discovery*, **13**, 137-166. http://dx.doi.org/10.1007/s10618-005-0026-2

[40] Bah, S. and Qutub, H. (2010) Insights into Data on Multiple Causes of Death Obtained from the Information System of a University Teaching Hospital, Al-Khobar, Saudi Arabia. *Journal of Health Informatics in Developing Countries*, **4**, 18-26. http://www.jhidc.org/index.php/jhidc/issue/view/9

[41] Johnson, N.E. and Christenson, B.A. (1998) Socio-Demographic Correlates of Multiple Causes of Death: Real or Artifactual? *Population Research and Policy Review*, **17**, 261-274. http://dx.doi.org/10.1023/A:1005985730638

[42] Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T. and Moser, S.A. (1998) Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, **5**, 373-381. http://dx.doi.org/10.1136/jamia.1998.0050373

[43] Ordonez, C., Santana, C.A. and Braal, L.D. (2000) Discovering Interesting Association Rules in Medical Data. *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Dallas, 14 May 2000, 78-85.

[44] Harpaz, R., Chase, H.S. and Friedman, C. (2010) Mining Multi-Item Drug Adverse Effect Associations in Spontaneous Reporting Systems. *BMC Bioinformatics*, **11**, S7. http://dx.doi.org/10.1186/1471-2105-11-S9-S7

[45] Chaves, R., Gorriz, J.M., Ramırez, J., Illan, I.A., Salas-Gonzalez, D. and Gomez-Rio, M. (2011) Efficient Mining of Association Rules for the Early Diagnosis of Alzheimer's Disease. *Physics in Medicine and Biology*, **56**, 6047-6063. http://dx.doi.org/10.1088/0031-9155/56/18/017

[46] He, Y., Zheng, X., Sit, C., Loo, W.T.Y., Wang, Z., Xie, T., Jia, B., Ye, Q., Tsui, K., Chow, L.W.C. and Chen, J. (2012) Using Association Rules Mining to Explore Pattern of Chinese Medicinal Formulae (Prescription) in Treating and Preventing Breast Cancer Recurrence and Metastasis. *Journal of Translational Medicine*, **10**, S12. http://www.translationalmedicine.com/content/10/S1/S12 http://dx.doi.org/10.1186/1479-5876-10-s1-s12

[47] Nuwangi, S.M., Oruthotaarachchi, C.R., Tilakaratna, J.M.P.P. and Calder, H.A. (2010) Usage of Association Rules and Classification Techniques in Knowledge Extraction of Diabetes. *Proceedings of the* 6*th International Conference on Advanced Information Management and Service*, Seoul, 30 November-2 December 2010, 372-377.

[48] Stats SA (2010) Mid-Year Population Estimates 2010. (Statistical Release P0302). Pretoria Statistics South Africa.

[49] Stats SA (2013) Mid-Year Population Estimates 2013. (Statistical Release P0302). Pretoria Statistics South Africa.

[50] Bird, S.M. (2013) Editorial: Counting the Dead Properly and Promptly. *Journal of the Royal Statistical Society*: *Series A*, **176**, 815-817. http://dx.doi.org/10.1111/rssa.12035

[51] Stallard, E. (2002) Underlying and Multiple Cause Mortality at Advanced Ages: United States 1980-1998. *North American Actuarial Journal*, **6**, 64-87. http://dx.doi.org/10.1080/10920277.2002.11073999

[52] Hahsler, M. and Hornik, K. (2006) Building on the Arules Infrastructure for Analyzing Transaction Data with R. *Proceedings of the* 30*th Annual Conference of the Gesellschaftfür Klassifikatione*. *V*., Berlin, 8-10 March 2006, 449-456.

[53] Statistics South Africa (2010) Mortality and Causes of Death in South Africa, 2008: Findings from Death Notification. www.statssa.gov.za