
Data quality issues in the GIS modelling of air pollution and cardiovascular mortality in Bangalore

Anitha K. Chinnaswamy*

Faculty of Engineering and Computing,
Coventry University, UK
Email: chinnasa@uni.coventry.ac.uk
*Corresponding author

Hewa Balisane

Faculty of Science,
Soran University,
Kurdistan Regional Government, Iraq
Email: hewa.balisane@soran.edu.iq

Quynh T. Nguyen and Raouf N.G. Naguib

Faculty of Science,
Soran University,
Kurdistan Regional Government, Iraq
and
BIOCORE Research and Consultancy International,
UK
Email: q.nguyen@biocoreinternational.com
Email: r.naguib@biocoreinternational.com

Nigel Trodd

Faculty of Business,
Environment and Society,
Coventry University, UK
Email: n.trodd@coventry.ac.uk

Ian M. Marshall and Norlaily Yaacob

Faculty of Engineering and Computing,
Coventry University, UK
Email: i.marshall@coventry.ac.uk
Email: n.yaacob@coventry.ac.uk

Gil Nonato Santos, Edgar A. Vallar and Maria Cecilia D. Galvez

Physics Department,
De La Salle University, Philippines
Email: gil.santos@dlsu.edu.ph
Email: edgar.vallar@dlsu.edu.ph
Email: maria.cecilia.galvez@dlsu.edu.ph

Mohyi H. Shaker

Ecology and Environment, Inc.,
USA
Email: mohyishaker@gmail.com

Nilmini Wickramasinghe

Epworth HealthCare
Deakin University, Australia
Email: n.wickramasinghe@deakin.edu.au

Tuan Nghia Ton

Environmental Health Unit,
World Health Organization Representative Office, Vietnam
Email: tont@wpro.who.int

Abstract: Cardiovascular disease (CVD) is the world's number one cause of mortality. Research in recent years has begun to illustrate a significant association between CVD and air pollution. As most of these studies employed traditional statistics, cross-sectional or meta-analysis methods, a study undertaken by the authors was designed to investigate how a geographical information system (GIS) could be used to develop a more efficient spatio-temporal method of analysis than the currently existing methods mainly based on statistical inference. Using Bangalore, India, as a case study, demographic, environmental and CVD mortality data was sought from the city. However, critical deficiencies in the quality of the environmental data and mortality records were identified and quantified. This paper discusses the shortcomings in the quality of mortality data, together with the development of a framework based on WHO guidelines to improve the defects, henceforth considerably improving data quality.

Keywords: data quality; information quality; air pollution; cardiovascular disease; cardiovascular mortality; Bangalore; India; geographical information systems; GIS.

Reference to this paper should be made as follows: Chinnaswamy, A.K., Balisane, H., Nguyen, Q.T., Naguib, R.N.G., Trodd, N., Marshall, I.M., Yaacob, N., Santos, G.N., Vallar, E.A., Galvez, M.C.D., Shaker, M.H., Wickramasinghe, N. and Ton, T.N. (2015) 'Data quality issues in the GIS modelling of air pollution and cardiovascular mortality in Bangalore', *Int. J. Information Quality*, Vol. 4, No. 1, pp.64–81.

Biographical notes: Anitha K. Chinnaswamy is an Assistant Lecturer in Information Systems at the Faculty of Engineering and Computing, Coventry University UK. She received her BEng (Hons) degree in Instrumentation and Electronics from Bangalore University, Bangalore, India. She received her MSc degree in Health Informatics from Coventry University. She is actively involved in Environmental health research and is currently on a PhD program, the subject of her research is 'The impact of air pollution on cardiovascular diseases in Bangalore, India'. Her other research interests include healthcare informatics, hospital information systems and knowledge management in healthcare.

Hewa Balisane is the Dean of the Faculty of Education and is also affiliated to the Faculty of Science at Soran University, Kurdistan, Iraq. He obtained his Master's degree from Salford University, UK, in 2008. He obtained his doctorate degree from Manchester Metropolitan University, UK, in 2011. Prior to his position at Soran University, he was an Associate Lecturer at Manchester Metropolitan University. His research interests lie in advanced information technology, multimedia technology, data telecommunications and networks, and biometrics.

Quynh T. Nguyen is a certified SPSS specialist and statistician, having obtained her Master's degree (with Distinction) and a PhD from Coventry University, UK, in 2007 and 2011, respectively. She is an experienced data analyst in structural equation modelling, regression analysis and data mining for a wide range of research topics, especially in environmental health. Over the last eight years, she worked on the integration of different datasets within geographical information systems, health informatics, environmental systems and website defacement detection. She was a former Post-Doctoral Research Assistant with the BIOCORE Applied Research group at Coventry University, UK, and she is currently affiliated to Soran University, KRG. She has published 28 peer reviewed journals and conferences papers and is an invited speaker at De La Salle University, Manila, Philippines. In addition to the above areas of work, her research interests include information retrieval, data extraction and mining, and environmental health in developing countries.

Raouf N.G. Naguib obtained both his Master's degree (with distinction) and doctorate degree from Imperial College, University of London, UK, in 1983 and 1986, respectively. He is the Director of BIOCORE Research and Consultancy International, UK and Professor at Soran University, KRG. He has published two books and over 360 journal and conference papers and reports in many aspects of health informatics, environmental health, social health, biomedical and digital signal/image processing, and the applications of artificial intelligence and evolutionary computation in cancer research. He was awarded the Fulbright Cancer Fellowship in 1995–1996 when he carried out research in the USA on the applications of artificial neural networks in breast cancer diagnosis and prognosis. He is a member of several national and international research committees, boards and review panels, an Adjunct Research Professor at the University of Carleton, Ottawa, Canada, and an Honorary Professor at De La Salle University, Manila, Philippines.

Nigel Trodd is the Associate Head of Environmental GIS and Remote Sensing at Coventry University, UK. He is an advocate for the intelligent use of location in the study of environmental problems. With over two decades of experience in research and technical development in geoinformation and geoinformatics, he provides expertise on the design and creation of interoperable geoinformation services. He is also an educator in GIS, satellite navigation systems and Earth observation by remote sensing with more than 13

years practice in distance/e-learning. His specialties include international and multidisciplinary research in the environmental and social sciences; image processing for the analysis of land cover dynamics; remote sensing of drylands; GIS for regeneration, development and disaster risk reduction; multidisciplinary fieldwork in the UK, Europe, China, West and Southern Africa.

Ian M. Marshall was Dean of Engineering and Computing until January 2007 and is currently the Deputy Vice-Chancellor (Academic) at Coventry University, UK. His applied research interests are focused around the use of games in education and training and in development effort estimation for multimedia and other interactive courseware. He has worked extensively as a computer and information systems consultant, flexible learning material developer and trainer. His client list includes BP Exploration, World Health Organisation, TSB plc, B&Q plc, Shell plc, Hydro Electric plc, Scottish Power plc, GEC Marconi 3SI, British Gas TransCo, WL Gore and Associates, NCR/AT&T Financial Services. In addition, he has extensive experience of working with small to medium sized enterprises as well as city councils and regional development agencies.

Norlaily Yaacob is a Senior Lecturer in the Faculty of Engineering and Computing at the Coventry University, UK. She obtained her PhD in Computer Science from the University of Exeter, UK. Her research interests include grid computing, concurrent programming, quality of service of grid applications and e-learning.

Gil Nonato C. Santos is a Full Professor and the Chairman of the Physics Department, De La Salle University, Manila, Philippines. He has a Master's degree in Physics and Doctorate degree in Materials Science and Engineering from the University of the Philippines. He is a Research Fellow at the University of Fukui, and a Visiting Professor at Osaka and Howard Universities. He is the Research Head of the Solid State Physics Research Laboratory and published local and international publications in nanomaterials. He is the author of several textbooks in science, math and computer technology. He is a member of the International Association of Engineers, Samahang Pisika ng Pilipinas, Philippine Physics Society and Ex-Officio Chair of the Microscopy Society of the Philippines. He was awarded the Most Outstanding Service in Physics Award in 2010 by the Philippine Physics Society.

Edgar A. Vallar is a faculty member of the Physics Department of De La Salle University, Manila, Philippines, where he currently co-chairs the Environment and Remote sensing research (EARTH) Group. He is the President of the Researchers for Clean Air (RESCUEAIR, Inc.), a non-governmental organisation composed of academics and researchers from the Philippines' top research and educational institutes. His current research focuses on clouds, aerosols and air quality (using LIDAR, Earth-Observing Satellites, Sunphotometry, DOAS, Air Sampling, SEM/EDX and Modeling) along with their correlations with human health. The EARTH group is also developing fluorescence setups for bioaerosols, water quality, and plant studies. Moreover, the group is building reliable, robust, and cost-effective weather and air quality instrumentation. The EARTH group is also part of a university-wide group implementing climate change and disaster preparedness projects for the Philippine Government.

Maria Cecilia D. Galvez is a Full Professor in the Physics Department at the De La Salle University, Manila, Philippines. Her current research interests include laser remote sensing (LIDAR), sun photometry, and air quality measurement and modelling using different methods.

Mohyi H. Shaker is a cardiologist with over 18-year professional background, which spans cardiology and internal medicine, emergency medical support and intensive care, health systems and administration, and medical facility management.

Nilmini Wickramasinghe is a Professor and the Director of Health Informatics Management at Epworth HealthCare. He is a Professor of Health Informatics Management at Deakin University, Australia. She actively researches and teaches with a particular focus on developing suitable models, strategies and techniques grounded in various management disciplines to facilitate more effective design, development and implementation of technology solutions to effect superior, patient centric healthcare delivery. She has collaborated with leading scholars at various premier healthcare organisations throughout Australasia, USA and Europe and is well published with more than 300 refereed scholarly articles, more than ten books, numerous book chapters, an encyclopaedia and a well established funded research track record. He is the Editor-in-Chief of two scholarly journals published by Inderscience: *Intl. J. Biomedical Engineering and Technology* and *Intl. J. Networking and Virtual Organisations*.

Tuan Nghia Ton earned his Master's degree in Environmental Science from Indiana University in Bloomington, Indiana, USA, under a Fulbright Scholarship program in 1995–1997. Since joining the WHO Representative Office in Vietnam in 2006, he has been a national professional officer in charge of environmental health covering water and sanitation, air pollution, climate change, chemical safety and healthcare waste management. One of his main recent interests is to successfully implement a water safety plan in the urban water supply sector in Vietnam. He currently also works closely with the Vietnamese National Institute for Occupational and Environmental Health (NIOEH) in conducting research on health impacts of indoor air pollutants and setting up standards on indoor air quality.

1 Introduction

CVD is the single leading cause of death globally (WHO, 2013). Although hereditary factors and age play a vital role in causing CVD, an individual's diet, exercise and lifestyle choices increase the likelihood of acquiring the disease. Environment also plays an insidious role in the onset and exacerbation of CVD (Bhatnagar, 2011). Numerous studies have reported positive associations between long- and short-term exposure to pollution and CVD (Brook, R.D. et al. (2004; Dockery et al., 19936; Dominici et al., 2003; Katsouyanni et al., 2001; Pope et al., 2014). The striking similarity of these studies is that they all emerge from developed nations. There is a noticeable gap of the research output from developing countries; in Su et al. (2011) highlighted this lack of research knowledge from South Asia. The authors report that less than 10% of references to

studies on air pollution and CVD are from the Asian sub-continent; they reiterate the need for more research output from Asia to fill this knowledge gap. South Asia especially has witnessed the burden due to mortality and morbidity from CVD. Whereas developed countries have robust systems for data recording, developing nations lack systems that methodically record health data. The lack of research output may be attributed to a number of factors, such as the insufficiency of data, availability of data in unstable formats, lack of research culture leading to non-sharing of data, or stakeholders being unaware of the prominence or advantages of evidence-based research.

India houses the cities with the highest air pollution in the world, with Delhi being classified as the one of the worst polluted cities in the world (The Indian Express, 2014). Also, ironically, the statistics for India show that it has the highest number of CVD patients globally (Ghaffar et al., 2004) and is projected to have increasingly more deaths from cardiovascular disease than any other country in the world over the next decade (Alwan, 2011). There is an urgent need for investigations in India and other fast developing nations to assess the extent to which the pollutants affect health and CVD in particular. Considering this need for studies in South Asia, this research focuses on Bangalore as a case study. Bangalore is one of India's booming economies and due to the immense employment opportunities it offers, the city has witnessed an exponential growth in population with an increase of 38% growth in just under a decade. The city is reportedly witnessing high levels of air pollution with levels being critical in some of its areas (Clean Air Asia, 2013).

To investigate the levels of pollutants and their effects on CVD mortality and to determine if there are patterns of the disease in certain areas over other areas in the city, a spatio-temporal methodology of analysis, using a geographic information system (GIS), is employed. Goodchild defines GIS as "a computer system for performing virtually any conceivable operation on geographical data, from acquisition and compilation through display, analysis and modelling, to sharing and archiving" (Goodchild, 2011). GIS integrates hardware, software and data for capturing, managing, analysing and displaying all forms of geographically referenced information (ESRI, 2011). GIS have the capability to store spatial data such as zones in the city and the related non-spatial attributes such as population within these zones.

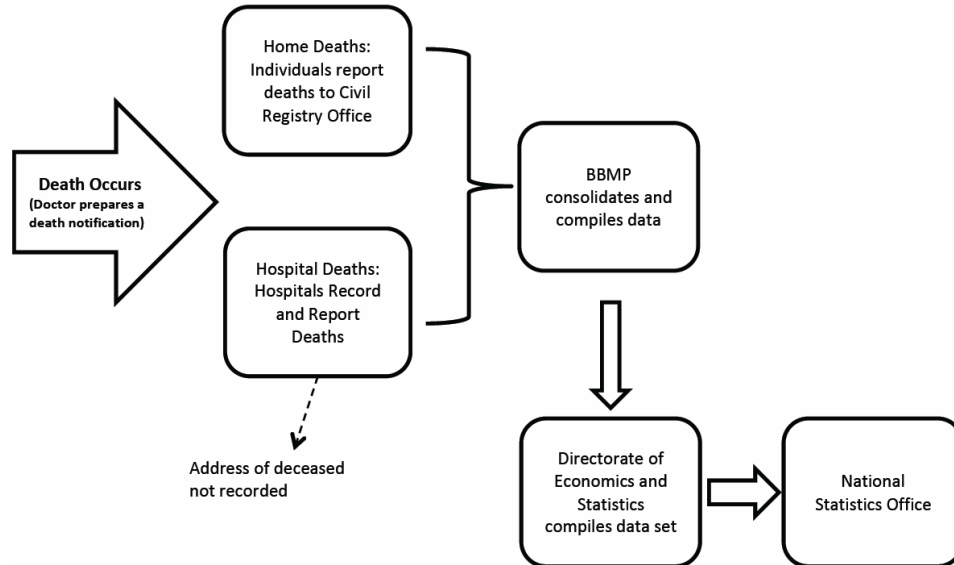
Data for the analysis has been obtained from various secondary sources. The data for the research such as zonal population data, levels of air pollution and mortality are stored in the relational database of the GIS and are arranged in files of related information, each file being called a layer. GIS has the capability to provide the relationships of these spatial and non-spatial data, going beyond just providing maps. It is also being increasingly used in health research and applications due to its strong ability to analyse any patterns or trends. Kistemann et al (2002) state that within the environmental health and public health domains, GIS are an indispensable tool that can efficiently process, analyse and visualise spatial data. It is vital however that data is encoded in proper format to aid in its manipulation, storage and analysis. Data quality has been defined as 'the fitness of data for all purposes that require it and measuring data quality requires an understanding of all intended purposes for that data' (Turner, 2004). Therefore, the datasets acquired should be subjected to testing for data quality parameters such as completeness, accuracy and other parameters as described in Table 1.

Table 1 Dimensions of data quality

<i>Amount of data</i>	<i>Sufficiency/Insufficiency of information</i>
Timeliness	Freshness and up-to-date state of information
Objectivity	Whether the information was objectively collected
Relevancy	Whether the information is useful/relevant/appropriate/applicable to intended purpose
Completeness	The expectation that certain attributes are expected to have assigned values in a dataset
Accuracy	The level to which stored data agrees with accepted sources of correct information
Access security	Whether the information is protected against unauthorised access

2 Mortality recording

The study of mortality is vital to determine a community's current demographic conditions and also aids in analysing the health status of that community. The statistics from the analysis of death data provide an understanding of disease and death in populations and aid in understanding any patterns or trends in deaths. However, it is the quality of the data that determines the quality of the subsequent information and decisions (Jacke et al., 2012). It is therefore vital that the data on age- and cause-specific mortality is reliable, continuous and timely to inform proper evaluation of policies and research activities.

Figure 1 Mortality data recording and collection in Bangalore

Source: Directorate of Economics and Statistics, Government of Karnataka (2014)

According to the Registration of Births and Deaths Act, 1969 in India (Directorate of Economics and Statistics, Government of Karnataka, 2014), all births and deaths have to be reported, recorded and registered. Registration takes place in registration centres or hospitals. A death certificate is mandatory for deaths occurring outside of hospitals for issuance of a crematorium certificate. The head of the household or the nearest relative has to report the occurrence of deaths in the house, by obtaining a local doctor to certify the cause of death. In case of deaths in a hospital, the Medical Officer in charge or any authorised person is responsible for registering the death and reporting it to the local registration centre. Figure 1 provides a schematic representation of the death registration process in Bangalore.

3 Methods

The data register containing individual data are entered into a central database by the Bruhat Bangalore Mahanagara Palike (BBMP) and the death registers are sent to the Department of Economics and Statistics (DES) which maintains a database used for statistical reporting (Figure 1). Data was obtained from BBMP from the central repository data register that contained the mortality data integrated from all the registration centres in Bangalore. A total of $n = 1,090,899$ historic records were obtained for the years 1930–2013. Data was recorded according to date of death, address, age, gender and cause of death. Data management tools (Excel and Access) were used for the management of the data, and the statistical software package, IBM SPSS20 was used for its analysis.

4 Analysis and results

Data collected was input into SPSS, the variables being zone, age groups, year of death, place of death, gender and the cause of death. Cause of death was coded according to the WHO's ICD-10 classification for diseases. Codes start from A00-Z99, with CVD mortality codes ranging between 100–199. This data was subjected to descriptive statistical analysis to determine its quality in terms of the parameters described in Table 1.

4.1 Amount of data

Of the $n = 1,090,899$ records, initial analysis determined that not all records had the cause of death recorded. The absence of cause of death does not assist in the analysis of death rates attributed to specific conditions such as CVD. Hence, those records were deemed of no use for the purposes of this research. Eliminating these records from the dataset reduced it to 250,000 between the years 2007–2013. Bangalore has only recently started to digitise the paper records into digital formats. Records of 2007–2009 were not completely input into the database and hence deemed unfit for this analysis. Hence, the years 2010–2013 were selected for the purposes of the study which resulted in a dataset with $n = 183,893$. This dataset was thus subjected to the assessment of data quality using the data quality parameters specified in Table 1.

4.2 *Timeliness*

Although the records obtained spanned over the years (since 1930), it is only since 2010, as mentioned previously, that Bangalore has started to robustly store it in a digital database. As the processes are in place now with people trained to enter data into the designed database for purposes of recording mortality data, it can be inferred that Bangalore will now be in a position to adequately collate it.

4.3 *Objectivity*

Objectivity is a measure to check if information was objectively collected. It should be noted that Bangalore has two different procedures for recording deaths, home and hospital deaths. Its laws clearly state that every death has to be reported and recorded, and that all home deaths have to be certified by a doctor. All hospital deaths are reported by the hospital where the patient died. In the case of hospital deaths, a doctor also attaches the Medically Certified Cause of Death (MCCD) but they do not include the residential address of the patient. As a consequence, the spatial component of these deaths is then immediately lost. Thus, prior to commencing the data analysis in this study, deaths occurring in hospitals were classified as ‘Institutional Deaths’ and home deaths as ‘Non-Institutional Deaths’ in SPSS. The institutional Deaths accounted for 52.8% of the deaths – implying that the spatial component for over half of the records was immediately lost (Table 2). Although the recorded information will still assist in providing overall statistics for Bangalore, death patterns due to locational influence, if any, cannot be determined. This is valuable information which can guide environmentalists and public health officials in generating robust public health policies. Hence, the objectivity of collecting the data in usable formats has to be emphasised to all the stakeholders involved.

Table 2 Distribution of deaths – place of death

<i>Place of death</i>	<i>Frequency</i>	<i>%</i>
Institutional	97,075	52.8
Non-institutional	86,818	47.2
Total	183,893	100

4.4 *Relevancy*

The data obtained is useful for analysis and interpretation, and applicable for the intended purpose. The data includes characteristics such as age, gender, cause of death and address (only in cases of home death) that are appropriate for the purposes of this research. However, the absence of address of the deceased dying in hospitals and the short-comings in the cause-of-death recording which is further discussed in this section, need to be improved.

4.5 *Completeness*

Completeness of mortality recording, according to the World Health Organization (2013), is a measure of the extent to which the deaths that occur in a region in a given year are registered by the civil registration system. This is calculated by the following equation:

$$YD = \frac{(RD)}{(CDR * P)} * 100$$

where

YD estimated death registration completeness (%)

RD actual number of registered deaths

CDR crude death rate (per 1,000)

P total population size (% by 1,000).

Applying this equation to determine completeness of death registrations in Bangalore for the year 2010 yields:

$$YD = \frac{(49,000)}{(8 * 8,400)} * 100$$

$$YD = 73\%$$

While all hospital deaths (MCCD deaths) are registered in Bangalore, home deaths require one family representative to take the death notification to the registration centre and register it. It was acknowledged by the City Council authorities that some people are not aware of the process to register and assume the death notification as a registration process. Wider awareness of the requirements for death registration is likely to improve the coverage and completeness of death registrations.

4.6 *Accuracy*

Accuracy of the data is the level to which the acquired data agrees with accepted sources of correct information. To assess for accuracy, the dataset was subjected to detailed checking for the following parameters:

4.6.1 *Unknown cause of death*

The dataset comprised of causes of deaths recorded as 'Unknown' and there are missing values. This amounted to 4.1% of total deaths (Table 3). The World Health Organization (2013) states that this is an incorrect procedure for the recording of deaths as such practice does not favour disease control and prevention programs. It is a huge challenge for developing countries to ensure that all deaths are medically certified.

Table 3 Distribution of ‘unknown’, and ‘missing values’

<i>Death causes</i>	<i>Frequency</i>	<i>%</i>
Unknown	5,089	2.8
Missing values	2,429	1.3
Total	7,518	4.1

4.6.2 *Death on arrival*

The procedure of certifying death on arrival (DOA) cases differs from country to country. DOA cases are sometimes reported to the police and hence at times not included in the cause-of-death statistics. Only an autopsy can ascertain the correct cause of death and this is not always carried out. As a result, a large number of DOA cases are assigned to ill-defined or non-specific causes, which do not serve any public health purposes. Unnatural deaths that require further investigations in Bangalore are recorded as ‘Awaiting PM Report’ (PM referring to post-mortem); this accounted for 2% of the deaths in the dataset (Table 4).

4.6.3 *Natural and old age deaths*

The most challenging deaths to certify are those of infants and of older people. People suffer from various conditions as they age and if reliable medical records and lab findings are unavailable, which is typical of developing countries, ascribing the exact cause of death requires significant judgement. Sometimes the deaths of people above a certain age simply attributed to ‘natural’ or due to ‘old age’ causes. In this dataset, the Natural deaths (4.3%) and Old age deaths (2.6%) accounted for almost 7% of the total deaths (Table 4). The WHO states that this is incorrect practice and should not be used by certifiers as the actual cause of death has to be determined for accurate death statistics.

4.6.4 *Spelling errors*

The local language in Bangalore is Kannada and the forms recording deaths are predominantly in this local language. However, the database is in English and the deaths are translated from Kannada into English by the database operators. It is presumed that some information is lost during the translation, resulting in errors in spellings and incomprehensible abbreviations. Such spelling errors and abbreviations accounted for 6.4% of the total deaths in the dataset (Table 4).

Table 4 Distribution of causes of death

<i>Death causes</i>	<i>Frequency</i>	<i>%</i>
Awaiting PM report	3,663	2
Spelling errors and abbreviations	11,775	6.4
Natural death	7,959	4.3
Old age	4,856	2.6
Total	28,253	13.5

4.6.5 Ill-defined causes of deaths

There is a list of ill-defined causes of deaths such as symptoms and signs, R00-R99, J81, etc. Although these are used for coding deaths, ideally they should be frequently monitored so as to reduce their vague connotations. In general it is also recommended that the proportion of deaths coded according to these ill-defined categories should not exceed 10% of all deaths in the age group of 65 years and above. Also for those under the age of 65 years old, it should account for only 5% of deaths. On inspection, for Bangalore, these did not exceed the recommended percentage, but much can be achieved by improving the coding practice.

Following testing for accuracy and identifying all the short-comings, 19.4% of the overall records were deemed incomplete for use in this research (Table 5). Table 5 also provides the distribution of deaths according to the Institutional and Non-institutional deaths. It can be concluded that the MCCD deaths recorded by hospitals (Institutional), although appear to be better recorded than home deaths, still contribute to 10.3% of overall errors. The percentage of Unknown, Missing values, Old age, and Natural death individually contribute to less than 1%. But, the spelling errors and abbreviations had the major contribution to errors. These errors can be minimised through the appropriate training of coders on cause-of-death recording and training of database operators to input data accurately.

Table 5 Distribution of causes of death according to place of death

Death causes	<i>Institutional</i>	<i>Non-institutional</i>	<i>Total</i>
Unknown	1,050 1.1%	4,039 .6%	5,089 2.8%
Awaiting PM report	1,292 1.3%	2,371 2.7%	3,663 2.0%
Spelling errors and Abbreviations	6,268 6.5%	5,507 6.3%	11,775 6. %
Missing values	743 0.8%	1,686 1.9%	2,429 1.3%
Natural death	249 0.3%	7,710 8.9%	7,959 .3%
Old age	289 0.3%	4,567 5.3%	4,856 2.6%
Total percentage	10.3%	29.7%	19.4%

4.7 Access security

The mortality database is protected against unauthorised access by appropriate security measures. The database is only accessible to authorised personnel in the department and is secure with password controls. For the purposes of this research the BBMP provided data where the rights of the individuals were protected by removing all identifiable variables, such as name of the individual being replaced by a unique ID, Date of Birth

removed and replaced with age, and full address of the deceased replaced with only with the corresponding ward/zone.

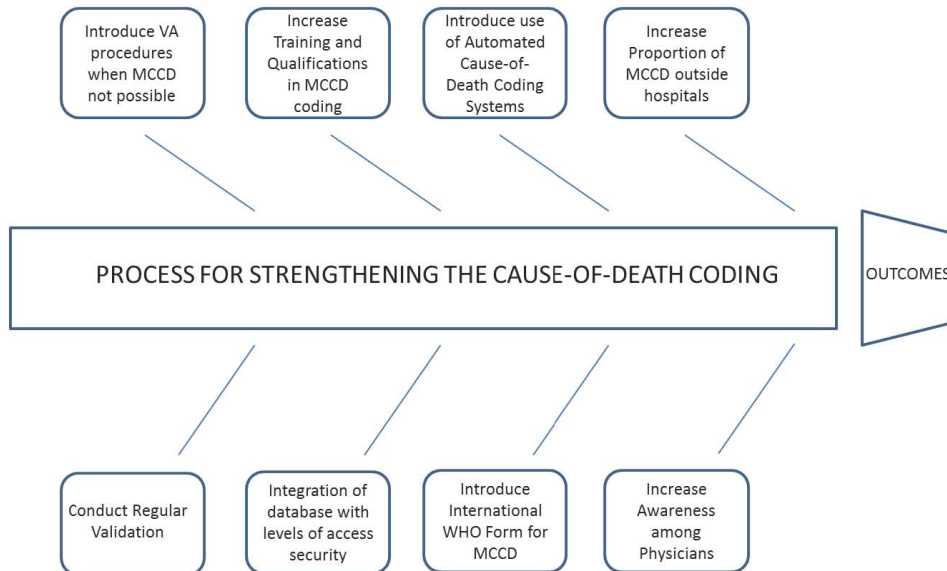
5 Framework for improvement

It is concluded from this assessment that Bangalore has a system that is functional but could be improved. Having identified the data quality issues, it is determined that if there is a focus on improving the elements that can be easily rectified, Bangalore will see an improved mortality database which will provide credible and trustworthy information that will assist in determining any patterns or trends in the deaths.

Certifying deaths by the ‘cause-of-death’ is important as it describes and explains levels and trends. Emerging diseases and conditions can be identified and the scale of how different groups are affected by the burden of specific diseases can be tracked. This will then guide the priorities for intervention programs and the decisions on directing and/or allocating resources according to areas of priority. According to the WHO, due to the many important public health uses to which the cause of death statistics are put, the accurate diagnosis of the cause of death should be ensured and coded according to international standards.

Hence, a framework (Figure 2) based on WHO guidelines for improving cause of death coding is recommended. The implementation of the proposed policies will aid Bangalore in improving data collection that will in turn result in recording unambiguous and correct cause-of-death information.

Figure 2 Roadmap to strengthen cause-of-death recording (see online version for colours)



5.1 Introduce verbal autopsy procedures when MCCD is not possible

The number of Unknown cases in mortality coding is considerably high in the dataset. A number of factors could lead to recording deaths in this manner, the most prominent ones being the coding practice and lack of training and qualifications. Where the cause of death is not known, verbal autopsy (VA) procedures can be used to determine the cause of death; VA is a procedure where carers or family members are asked using a standard questionnaire about signs and symptoms that led to the death of the person (Centre for Global Health Research, 2014). The sequence of causes that led to the death can then be determined by a MCCD trained practitioner and entered into the death certificate. This process will reduce the number of deaths recorded as 'Unknown' or 'Missing Values' that contribute to almost 4.1% of deaths in this dataset.

5.2 Introduce use of automated cause of death recording

Many countries, such as the UK and USA, use automated coding programs such as Automated Classification of Medical Entities (ACME), IRIS and Mortality Indexing, Classification and Retrieval (MICAR) primarily to

- 1 to introduce consistent and rapid assignment of underlying cause-of-death coding with reduced needs for manpower training
- 2 to allow better utilisation of medical information on death certificates for multiple cause-of-death analyses (Israel, 1990; O'Carroll et al., 2003).

These computer programs enable coders to select an accurate underlying cause of death, henceforth minimising the number of subjective decisions a coder normally makes. Introduction of this tool in Bangalore will enhance the quality of the dataset by tackling the problem of inconsistency among coders and standardise the production of mortality statistics.

5.3 Increase training and qualifications in MCCD coding

Mortality coding is a highly specialised task that requires thorough understanding of the rules for determining the underlying cause of death and for assigning codes. The level of knowledge and understanding of the coder is a key determinant of the accuracy of the resulting data. It is essential for coders (working with these statistics) to be qualified professionals. An understanding of medical terminologies and medical sciences will ensure effective coding and will aid the coders in selecting the appropriate codes for the underlying cause of death as per the requirements of the ICD-10. Where coders do not have the opportunity to undergo such training, the WHO has developed an electronic self-training tool (World Health Organisation, 2014). As the highest percentage of errors in the dataset is a result of spelling errors and abbreviations, some investment in cause-of-death training would result in considerable improvement of the quality of the dataset.

5.4 Introduce WHO form for MCCD

The WHO recommends the use of their standardised cause-of-death recording form for wide use of death recording. As Bangalore already has their own standard form for recording deaths, amendments that need to be incurred to improve the recording of data is that forms need to compulsorily include the characteristics described in Table 6.

Table 6 Characteristics of cause of death recording

<i>Characteristics of the event</i>	<i>Characteristics of the deceased</i>
Date of occurrence	Date of birth and age (derived)
Date of registration	Sex
Place of occurrence	Marital status
Place of registration	Place of usual residence
Locality of occurrence (derived)	Locality of residence (derived)
Urban or rural occurrence (derived)	Urban or rural residence (derived)
Cause(s) of death	
Certifier and type of certification (derived)	

If these fields are made mandatory at the time of recording deaths, the risk of losing the spatial location of the deceased will be minimised. This will aid in determining the spatial pattern of disease, if any, to guide policies and interventions.

5.5 Conduct regular evaluation

It is good practice that the work of the coders should be systematically evaluated on a regular basis to ensure good quality coding. Additional training or guidance can be provided when inconsistencies (or errors) are identified. Routine internal reviews to check the quality of coding will highlight any weaknesses and, with adequate training, will improve the overall quality of deaths recording.

5.6 Increase awareness of MCCD

One of the important factors is the awareness of the importance of MCCD. It is important for the different stakeholders, coders, practitioners, physicians and related personnel to understand the purposes of collecting and using cause-of-death information. Appropriate training and awareness will influence the quality of the coded data with a possible substantial impact on its quality.

5.7 Database integration with levels of access/security

It has been noted that there is considerable duplication of data in Bangalore with both BBMP and DES having databases of mortality data. As databases have powerful capabilities of integration, security and access, it is recommended that one central database be utilised to digitally record all deaths. Appropriate access controls can then ensure who gets access to the necessary fields of the data.

6 Discussion

An assessment of the current data quality issues in recording mortality events has determined that Bangalore has a system that has a completeness of 73%. The accuracy of the dataset has resulted in 19.4% of the records being unusable as the cause-of-death recording is either absent or inaccurate. As the geographic distribution of deaths is vital to determine the trends and patterns of deaths, it is necessary that all records have the address of the deceased accurately recorded; the current MCCD recording practice in Bangalore although focuses on determining the cause-of-death, does not prioritise the address of the patient, hence losing valuable information for over 52% of the records.

Bangalore could address these inadequacies by developing an improvement plan; by adopting the components of the framework described in this paper, a drastic improvement in data quality can be achieved. As the resources are already available, with very minimal costs, a robust health information management system can be advocated and maintained. Avoiding duplication of the data will result in cost-savings which can then be channelled into more public health interventions.

Countries such as Thailand (Tangcharoensathien et al., 2006), Sri Lanka (Rampatige et al., 2009) and Iran (Khosravi et al., 2008) have in recent years conducted an assessment of their systems that highlighted their shortcomings such as absence of a strong legislation for registration of births and deaths, decentralised systems with wide differences in practice, absence of usage of the International Form for MCCD, incorrect coding of cause of death, paper-based recording of deaths, lack of awareness among citizens and training for coding cause of death accurately. These countries reviewed their policies to strengthen their Health Information Management System as a priority. The areas for improvement such as hospital death certification, deaths occurring outside the hospitals, coverage and completeness of registration, ICD-10 compliance practices for death certification, review of forms used for registration were recommended. The countries developed a set of recommendations and a feasible set of actions that was likely to have a significant impact on the data quality. The countries learned a lot from the assessments which became building blocks from which they developed strategic and long-term improvement plans for their systems.

Simple measures to improve the mortality recording system in Bangalore will result in reliable and accurate mortality statistics. The resulting database can then be imported into a GIS system which will enable mapping of the deaths in the city and highlight if there are any hotspots due to specific diseases or certain causes of death. This model can then also be duplicated in other major cities in India to enable the government to better regulate health policies and interventions.

Acknowledgements

The authors wish to thank the BBMP and DES for providing the data used for the purposes of this research.

References

- Alwan, A. (2011) 'Global status report on noncommunicable diseases 2010', *World Health Organization*.
- Bhatnagar, A. (Ed.) (2011) *Environmental Cardiology: Pollution and Heart Disease*, Royal Society of Chemistry, Cambridge.
- Brook, R.D., Franklin, B., Cascio, W., et al. (2004) 'Air pollution and cardiovascular disease: a statement for healthcare professionals from the expert panel on population and prevention science of the American Heart Association', *Circulation*, Vol. 109, No. 21, pp.2655–2671.
- Centre for Global Health Research (2014) *What is Verbal Autopsy* [online] <http://www.cghr.org/index.php/projects/million-death-study-project/what-is-verbal-autopsy/> (accessed 18 June 2015).
- Clean Air Asia (2013) *Bangalore's Air Pollution Can Reach Critical Levels* [online] <http://cleanairinitiative.org/portal/node/11742> (accessed 28 December 2014).
- Directorate of Economics and Statistics, Government of Karnataka (2014) *Registration of Births and Deaths* [online] http://des.kar.nic.in/crs_acts_rules.asp (accessed 15 December 2014).
- Dockery, D.W., Pope, C.A. and Xu, X. et al. (1993) 'An association between air pollution and mortality in six US cities', *N. Engl. J. Med.*, Vol. 329, pp.1753–1759.
- Dominici, F., McDermott, A. and Daniels, D. et al. (2003) 'Mortality among residents of 90 cities', *Special Report: Revised Analyses of Time-Series Studies of Air Pollution and Health*, pp.9–24, Health Effects Institute, Boston, Mass.
- ESRI (2011) *What is GIS?* [online] <http://www.esri.com/what-is-gis> (accessed 20 June 2015).
- Ghaffar, A., Reddy, K.S. and Singhi, M. (2004) 'Burden of non-communicable diseases in South Asia', *BMJ*, Vol. 328, No. 7443, pp.807–810.
- Goodchild, M.F. (2011) 'Challenges in geographical information science', *Proceedings of the Royal Society (A)* [online] <http://rspa.royalsocietypublishing.org/content/early/2011/04/16/rspa.2011.0114.short> (accessed 16 December 2014).
- Israel, R.A. (1990) 'Automation of mortality data coding and processing in the United States of America', *World Health Statistics Quarterly*, Vol. 43, No. 4, pp.259–262.
- Jacke, C.O., Kalder, M., Wagner, U. and Albert, U-S. (2012) 'Valid comparisons and decisions based on clinical registers and population based cohort studies: assessing the accuracy, completeness and epidemiological relevance of a breast cancer query database', *BMC Research Notes*, Vol. 5, p.700.
- Katsouyanni, K., Touloumi, G. and Samoli, E. et al. (2001) 'Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 Project', *Epidemiology*, Vol. 12, No. 5, pp.521–531.
- Khosravi, A., Rao, C., Naghavi, M., Taylor, R., Jafari, N. and Lopez, A.D. (2008) 'Impact of misclassification on measures of cardiovascular disease mortality in the Islamic Republic of Iran: a cross-sectional study', *Bulletin of the World Health Organization*, Vol. 86, No. 9, pp.688–696 [online] <http://www.who.int/bulletin/volumes/86/9/07-046532.pdf> (accessed 20 November 2014).
- Kistemann, T., Dangendorf, F. and Schweikart, J. (2002) 'New perspectives on the use of geographical information systems (GIS) in environmental health sciences', *International Journal of Hygiene and Environmental Health*, Vol. 205, No. 3, pp.169–181.
- O'Carroll, P.W., Yasnoff, W.A., Ward, E.M., Ripp, L.H. and Martin, E.L. (2003) *Public Health Informatics and Information Systems*, Springer, New York.
- Pope, C.A., Burnett, R.T. and Thurston, G.D. et al. (2014) 'Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease', *Circulation*, Vol. 109, No. 1, pp.71–77.

- Rampatige, R., Mikkelsen, L., Gamage, S. and Peiris, S. (2009) *Promoting the Periodic Assessment of the Quality of Medical Records and Cause-of-Death Data: Lessons Learned from a Medical Records Study in Sri Lanka*, Working paper No. 14, University of Queensland School of Population Health, Health Information Systems Knowledge Hub, Brisbane, Australia [online] <http://www.uq.edu.au/hishub/wp14> (accessed 14 December 2014).
- Su, T., Chen, S. and Chan, C. (2011) 'Progress of ambient air pollution and cardiovascular disease research in Asia', *Progress in Cardiovascular Disease*, Vol. 53, No. 5, pp.369–378.
- Tangcharoensathien, V., Faramnuayphol, P., Teokul, W., Bundhamcharoen, K. and Wibulpholprasert, S. (2006) 'A critical assessment of mortality statistics in Thailand: potential for improvements', *Bulletin of the World Health Organization*, Vol. 84, No. 3, pp.233–238 [online] <http://www.who.int/bulletin/volumes/84/3/233.pdf> (accessed 17 January 2015).
- The Indian Express (2014) *Delhi Most Polluted City in the World: WHO* [online] <http://indianexpress.com/article/cities/delhi/delhi-most-polluted-city-in-the-world-who/> (accessed 16 December 2014).
- Turner, S. (2004) *Defining and Measuring Traffic Data Quality*, White paper on recommended approaches, Transportation Research Record No. 1870, 2004.
- WHO (2013) *Cardiovascular Diseases* [online] <http://www.who.int/mediacentre/factsheets/fs317/en/> (accessed 18 January 2015).
- World Health Organisation (2014) *ICD-10 Interactive Self Learning Tool* [online] <http://apps.who.int/classifications/apps/icd/icd10training> (accessed 23 November 2014).
- World Health Organization (2013) *Strengthening Civil Registration and Vital Statistics for Births, Deaths and Causes of Death: Resource Kit*, World Health Organization, Luxembourg.