A study on garbage code redistribution methods in small area: redistributing heart failure in two Chinese cities by two approaches

Liqun Liu
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College
https://orcid.org/0000-0002-6728-6088
Xuewei Wang
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College
Chunping Wang
School of Public Health, Weifang Medical University
Xiangyun Ma
Xuanwei Center for Disease Control and Prevention
Xianfeng Meng
Weifang Center for Disease Control and Prevention
Bofu Ning
Xuanwei Center for Disease Control and Prevention
Ning Li
Weifang Center for Disease Control and Prevention
Xia Wan (xiawan@ibms.pumc.edu.cn)
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College
https://orcid.org/0000-0002-0738-5631

Research Article

Keywords: garbage code, heart failure, small area, death data, China, coarsen exact matching, linear regression

Posted Date: February 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1242825/v1

Abstract

Background and Aim. Garbage code (GC) in death surveillance data may affect the statistics on causes of death, and the public health policies made based on it. Redistribution of GCs means to assign a GC to a plausible correct underlying cause of death (UCOD), ways of doing which include expert consultation, fixed proportional reassignment, computed proportional reassignment based on the information from cause of death chain, regression models, and so on. The Global Burden of Disease (GBD) study has used several ways to redistribute various GCs in the death data sets all over the world. In this study, using heart failure as an example, we aimed to discuss the GCs redistribution in relatively small city-level death data sets, and the suitability of certain redistribution methods.

Methods. We collected the cause of death surveillance data in two Chinese cities, Weifang and Xuanwei, checked and improved the data quality before analysis. We extracted the death records attributed to heart failure, then manually corrected their UCOD, based on the cause of death chain information and according to the rules and guidelines for morbidity coding established by World Health Organization (WHO). After this step, we conducted the redistribution process for the records with UCOD remaining to be heart failure, using two different approaches, coarsen exact matching and linear regression, respectively. At last, we calculated the cause-specific mortalities before and after to observe the changes.

Results. Death cases with UCOD stated as heart failure were 1556 (percentage: 0.33%) in Weifang and 226 (0.41%) in Xuanwei, respectively. After manual correction, in both cities UCOD remained the same in about 75% of the records. In Weifang, when using coarsen exact matching, heart failure was mainly redistributed to ischemic heart disease (IHD, 45.31%), hypertensive heart disease (HHD, 21.56%) and chronic obstructive pulmonary disease (COPD 8.98%), but the death counts due to HHD, rheumatic heart disease (RHD) and other cardiovascular diseases except for IHD, HHD, RHD and stroke (CD) increased the most, with increasing percentages being 3.288%, 2.451% and 1.619%. When using linear regression, heart failure was almost all redistributed to IHD (91.20%), the death counts due to CD and IHD increased 1.213% and 0.929%. In Xuanwei, when using coarsen exact matching, heart failure was mainly redistributed to IHD (24.70%), diabetes mellitus and chronic kidney disease (DMCKD 23.25%) and COPD (16.10%), but the death counts due to HHD, DMCKD and RHD increased the most, with increasing percentages being 7.786%, 4.107% and 2.156%. When using linear regression, heart failure was all redistributed to COPD (94.83%), and its death count increased 1.622%.

Conclusions. In cities with 1 to 10 million permanent residents, if the percentage of certain GC was quite low in their death data sets, the necessity of redistributing it would be worth discussed. If redistributing it, coarsen exact matching should probably be more suitable then linear regression, as linear regression may cause inappropriate centralization of redistribution target diseases. The fundamental way for improving the quality of death data is to improve the capacity of primary staff on UCOD identification and coding.

1. Background

Cause of death surveillance data is one of the most important material for developing effective health policies, whose quality is crucially affected by the accuracy of the underlying cause of death (UCOD) provided in death certificates. The World Health Organization (WHO) has clearly defined an UCOD as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstance of the accident or violence which produced the fatal injuries"[1], however, in actual practical work, hospital or public health staff may fill in some ambiguous or vague codes in the International Statistical Classification of Diseases and Related Health Problems (ICD) to be someone's UCOD. In 1996, Murray and Lopez[2] named these codes as "garbage code (GC)" for the first time. In 2010, Naghavi et al.[3] divided GCs into four categories, "causes that cannot or should not be considered as UCOD, intermediate causes in the cause of death chain, immediate causes in the cause of death chain, and unspecified causes within a larger cause grouping". With the progress of the Global Burden of Disease (GBD) study, the scope of GCs kept getting broader, and the distinction of GCs kept getting further refined[3–7].

GBD study has proposed a summary index containing six dimensions to objectively measure the performance of vital registration system in generating reliable mortality data[8], in other words, the quality of the data set. After assessment, they concluded that the percentage of GCs had a strong impact on the value of the index assigned to a death statistics data set[8]. The existence of GCs affects the statistics on composition and ranking of death causes in a population, which would lead to inaccurate or even incorrect cognition and judgment on the priorities of health issues. For example, in GBD study, after re-coding part of "senility (ICD10: R54)" to dementia, the mortalities of dementia in high- and middle-income countries increased significantly, which alerted a possible fact that the prevalence level of dementia worldwide might be much higher than we have recognized[4]. In South Africa, after re-coding "heart failure (I50)" to some diseases of the respiratory or circulatory systems, the burden resulted from chronic non-communicable diseases has exceeded that from communicable diseases. Inaccurate judgment will unavoidably influence the effects of the population health promotion policies made based on it. In addition to that, the results of GCs redistribution could also seriously influence the judgement on the trend of diseases. For example, GBD 2010 result showed that the age-standardized mortality rate (ASMR) of ischemic heart disease (IHD) in China increased from 1990 to 2010, while GBD 2013 reported that the ASMR of IHD in China remained stable from 1990 to 2013. The difference in the trend of IHD mortality between GBD 2010 and GBD 2013 was resulted from the uncorrected GC - chronic pulmonary heart disease (PHD) redistribution method in GBD 2013.[9]

Since both the existence and the poor redistribution of GCs have adverse impacts on death data quality, scientists around the world started studying them as the concept came up. Redistribution of GCs could be understood as to assign a GC to a plausible correct ICD code for UCOD, based on the pathophysiological characteristics of certain diseases and so on. It contains two aspects, determining the scope of the plausible UCOD and calculating the proportion of redistribution to each of them. At present, there are several ways for redistributing GCs, such as expert consultation,[3] fixed proportional reassignment,[3,6] computed proportional reassignment based on the information from cause of death chain,[6,7,10] and regression models.[11–14] In the GBD 2019 study[7], expert consultation,[3] fixed proportional reassignment,[3,6] the coarsen exact matching method[10] and the linear regression method[12] were all used to redistribute various GCs.

In this study, we took "heart failure" as an example from the death surveillance data sets in two Chinese cities to compare different GC redistribution methods. Through the differences in their redistribution results, we aimed to discuss the differences in the principles of the methods, and study whether certain methods would also be appropriate for redistributing GCs in much smaller, comparing to the global database or national level database, city-level data sets.

2. Methods

2.1 Death surveillance data

The cause of death surveillance data was from Disease Surveillance System in Weifang (prefecture-level city), Shandong province, from 2010 to 2017, and in Xuanwei (county-level city), Yunnan province, from 2010 to 2016, respectively. The average population of permanent residents in Weifang and Xuanwei during the study period was about 9.24 million and 1.51 million, respectively. The variables in the death surveillance data sets included demographic information of the deceased, identification (ID) card number, time and place of death, cause of death chain and the UCOD, the highest agency and basis of diagnosis, medical history, symptoms and signs of the deceased, and so on. Before analysis, we checked the data sets and improved the data quality as much as possible by five aspects: first, completeness, filling the missing values up if possible. Second, internal consistency, keeping the logical consistency between variables, the consistency of demographic information from different sources, and so on. Third, external consistency, modifying abnormal and extreme values. Fourth, removing duplicate records. Fifth, unifying the structure of the data set. The detailed information on data cleaning and quality improvement procedure has been shown in our previous papers[15,16].

2.2 Definition and correction of heart failure

We extracted the death records with UCOD stated as heart failure from the data sets. Heart failure contains ICD-10 codes I50, I50.0, I50.1 and I50.9. Firstly, we corrected the UCOD in these records, based on the cause of death chain and the information on other diseases, according to the ICD-10 rules and guidelines for morbidity coding established by WHO[1,17]. The UCODs on some records were changed to another disease or another GC. Then, those records with UCOD remaining to be heart failure entered the redistribution process.

2.3 Redistribution of heart failure

Redistribution was carried out in Weifang and Xuanwei respectively, using the coarsen exact matching method[10] and the linear regression method[12], respectively.

2.3.1 Coarsen exact matching

The records with UCOD remaining to be heart failure were so called "treatment records". The records with UCOD being a non-garbage code, non-injury disease, but at the same time having heart failure in the cause of death chain, were the "control records". Whether an UCOD was a GC or an injury was judged according to the diseases classification by the Institute for Health Metrics and Evaluation (IHME) in their GBD 2017 study[6]. Note that the records whose UCOD had been changed from heart failure to another ICD-10 code in the previous step would also be included in the control pool, if it satisfied the conditions.

We selected five variables to divide subgroups, death year, urban or rural resident, gender, age, and the highest agency of diagnosis, considering their influence on death, influence on the process of death recording, correlations, and so on. Because of different amount of data, the variables and the grading for each variable used in Weifang and Xuanwei at last were slightly different, as shown in Table 1.

Table 1
Variables for dividing subgroups in Weifang and Xuanwei

Weifang		**Xuanwei**	
death year	2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017	death year	two time periods: 2010-2013, 2014-2016
urban or rural resident	urban, rural, unknown		
gender	male, female	gender	male, female
age group	0-14, 15-44, 45-64, 65-74, 75-84, 85-	age group	0-14, 15-44, 45-64, 65-74, 75-84, 85-
the highest agency of diagnosis	primary hospital, secondary hospital, tertiary hospital, no hospital diagnosis or other or unknown	the highest agency of diagnosis	primary hospital, secondary hospital, tertiary hospital, no hospital diagnosis or other or unknown

Table 2 Results of the correction of heart failure according to the ICD-10 rules and guidelines for morbidity coding established by WHO

	Weifang (Total number of heart failure = 1556)			Xuanwei (Total number of heart failure = 226)		
		Count	Percent (%)		Count	Percent (%)
Garbage codes	Remaining heart failure	1154	74.16	Remaining heart failure	174	76.99
	Other garbage codes	200	12.85	Other garbage codes	23	10.18
Non-communicable diseases [#]	Ischemic heart disease (IHD)	126	8.10	Chronic obstructive pulmonary disease (COPD)	9	3.98
	Hypertensive heart disease (HHD)	34	2.19	Malignant neoplasms except for TBLC	6	2.65
	Stroke	3	0.19	Ischemic heart disease (IHD)	3	1.33
	Rheumatic heart disease (RHD)	2	0.13	Hypertensive heart disease (HHD)	3	1.33
	Cardiovascular diseases except for IHD, HHD, RHD and Stroke (CD)	2	0.13	Rheumatic heart disease (RHD)	1	0.44
	Chronic obstructive pulmonary disease (COPD)	10	0.64	Other non-communicable diseases	1	0.44
	Chronic respiratory diseases except for COPD	1	0.06	Digestive diseases	1	0.44
	Malignant neoplasms except for TBLC	5	0.32			
	Tracheal, bronchus, and lung cancer (TBLC)	3	0.19			
	Diabetes mellitus and Chronic kidney disease (DMCKD)	4	0.26			
	Other non-communicable diseases	2	0.13			
	Digestive diseases	1	0.06			
	Neurological disorders	1	0.06			
Injuries		7	0.45		4	1.77
Communicable, maternal, neonatal, and nutritional diseases		1	0.06		1	0.44

[#] Diseases of three systems, cardiovascular diseases, chronic respiratory diseases and malignant neoplasms, are further divided into detailed diseases.

Table 3 Total mortality and cause-specific mortalities in Weifang and Xuanwei after UCOD correction

Causes			Weifang 2010-2017	Xuanwei 2010-2016
			Crude mortality (/100000)	Crude mortality (/100000)
Total			645.41	524.56
Communicable, maternal, neonatal, and nutritional diseases	HIV/AIDS and sexually transmitted infections	HIV/AIDS	0.03	0.94
		Sexually transmitted diseases excluding HIV	0.01	0.04
	Respiratory infections and tuberculosis	Tuberculosis	0.41	1.20
		Lower respiratory infections	0.95	7.24
		Upper respiratory infections	0.04	0.18
	Enteric infections	Diarrheal diseases	0.07	0.72
		Invasive Non-typhoidal Salmonella (iNTS)	0.02	0.03
	Neglected tropical diseases and malaria	Malaria	<0.01	0.03
		Leprosy	0.01	0.08
		Chagas disease	<0.01	0.00
		Cysticercosis	<0.01	0.00
		Dengue	0.00	0.02
		Rabies	0.04	0.09
		Other neglected tropical diseases	0.06	0.00
	Other infectious diseases	Meningitis	0.19	0.63
		Encephalitis	0.18	0.11
		Tetanus	0.02	0.03
		Measles	<0.01	0.01
		Varicella and herpes zoster	0.01	0.09
		Acute hepatitis	0.17	0.49
		Other unspecified infectious diseases	0.05	0.26
	Maternal and neonatal disorders	Maternal disorders	0.08	0.40
		Neonatal disorders	2.09	5.70
	Nutritional deficiencies	Protein-energy malnutrition	0.12	0.56
		Iron-deficiency anemia	<0.01	0.01
		Other nutritional deficiencies	0.22	0.64
Non-communicable diseases	Neoplasms	Lip and oral cavity cancer	0.35	0.26
		Nasopharynx cancer	0.60	0.60
		Other pharynx cancer	0.21	0.09
		Esophageal cancer	10.07	2.72
		Stomach cancer	28.01	3.09
		Colon and rectum cancer	7.83	3.81
		Liver cancer	25.32	12.43
		Gallbladder and biliary tract cancer	1.97	0.68

Table 3 Total mortality and cause-specific mortalities in Weifang and Xuanwei after UCOD correction

	Pancreatic cancer	4.28	1.25
	Larynx cancer	0.92	0.75
	Tracheal, bronchus, and lung cancer	56.77	76.64
	Malignant skin melanoma	0.13	0.05
	Non-melanoma skin cancer	0.38	0.41
	Breast cancer	3.71	1.30
	Cervical cancer	1.21	0.93
	Uterine cancer	0.57	0.13
	Ovarian cancer	1.06	0.19
	Prostate cancer	1.05	0.63
	Testicular cancer	0.03	0.03
	Kidney cancer	0.93	0.19
	Bladder cancer	1.86	1.04
	Brain and nervous system cancer	3.86	2.94
	Thyroid cancer	0.29	0.19
	Mesothelioma	0.05	0.03
	Hodgkin lymphoma	0.24	0.04
	Non-Hodgkin lymphoma	2.01	1.03
	Multiple myeloma	0.52	0.23
	Leukemia	3.93	2.56
	Other malignant cancers	3.05	1.64
	Other neoplasms	0.63	0.21
Cardiovascular diseases	Rheumatic heart disease	1.99	3.70
	Ischemic heart disease	153.22	34.12
	Stroke	126.71	52.20
	Hypertensive heart disease	10.24	2.11
	Non-rheumatic valvular heart disease	0.10	0.03
	Cardiomyopathy and myocarditis	0.48	0.77
	Atrial fibrillation and flutter	0.04	0.01
	Aortic aneurysm	0.41	0.12
	Peripheral artery disease	0.08	0.03
	Endocarditis	0.04	0.07
	Other cardiovascular and circulatory diseases	0.26	0.21
Chronic respiratory diseases	Chronic obstructive pulmonary disease	61.18	96.10
	Pneumoconiosis	0.06	0.46
	Asthma	0.68	1.62
	Interstitial lung disease and pulmonary sarcoidosis	0.67	0.05
	Other chronic respiratory diseases	0.07	0.37
Digestive diseases	Cirrhosis and other chronic liver diseases	3.23	4.88

Table 3 Total mortality and cause-specific mortalities in Weifang and Xuanwei after UCOD correction

		Upper digestive system diseases	1.85	8.78
		Appendicitis	0.03	0.25
		Paralytic ileus and intestinal obstruction	0.49	1.29
		Inguinal, femoral, and abdominal hernia	0.02	0.01
		Inflammatory bowel disease	0.08	0.45
		Vascular intestinal disorders	0.02	0.00
		Gallbladder and biliary diseases	0.67	0.63
		Pancreatitis	0.38	0.42
		Other digestive diseases	0.11	0.53
	Neurological disorders	Alzheimer disease and other dementias	2.77	1.40
		Parkinson disease	0.28	0.07
		Epilepsy	0.58	1.17
		Multiple sclerosis	0.01	0.00
		Motor neuron disease	0.18	0.07
		Other neurological disorders	0.32	0.26
	Mental disorders	Eating disorders	0.24	0.00
	Substance use disorders	Alcohol use disorders	0.43	2.19
		Drug use disorders	0.03	0.17
	Diabetes and kidney diseases	Diabetes mellitus	3.70	3.70
		Chronic kidney disease	4.72	4.90
		Acute glomerulonephritis	0.20	0.70
	Skin and subcutaneous diseases	Bacterial skin diseases	0.02	0.11
		Decubitus ulcer	0.14	0.77
		Other skin and subcutaneous diseases	0.02	0.01
	Musculoskeletal disorders	Rheumatoid arthritis	0.19	0.83
		Other musculoskeletal disorders	0.39	0.53
	Other non-communicable diseases	Congenital birth defects	1.80	2.21
		Urinary diseases and male infertility	0.27	1.19
		Gynecological diseases	0.03	0.05
		Hemoglobinopathies and hemolytic anemias	0.63	0.38
		Endocrine, metabolic, blood, and immune disorders	0.38	0.65
		Sudden infant death syndrome	0.03	0.00
Injuries	Transport injuries	Road injuries	21.00	10.67
		Other transport injuries	0.38	0.60
	Unintentional injuries	Falls	3.11	18.32
		Drowning	2.14	4.37
		Fire, heat, and hot substances	0.55	0.74

Table 3 Total mortality and cause-specific mortalities in Weifang and Xuanwei after UCOD correction				
		Poisonings	2.00	3.05
		Exposure to mechanical forces	1.22	5.62
		Adverse effects of medical treatment	0.07	0.40
		Animal contact	0.02	0.83
		Foreign body	0.48	1.23
		Environmental heat and cold exposure	0.17	0.13
		Exposure to forces of nature	0.02	0.20
		Other unintentional injuries	0.79	3.64
	Self-harm and interpersonal violence	Self-harm	6.19	9.49
		Interpersonal violence	0.39	1.61
		Executions and police conflict	0.01	0.04
Garbage codes			59.62	97.48
Still birth or Missing			0.17	0.03

Table 4 Results of the redistribution of heart failure

	Weifang (Total number of heart failure = 1154)				Xuanwei (Total number of heart failure = 174)			
Approach 1 coarsened exact matching	**Redistribution target diseases**		**Count**	**Percent (%)**	**Redistribution target diseases**		**Count**	**Percent (%)**
	Remaining heart failure		61.00	5.29	Remaining heart failure		15.00	8.62
	Non-communicable diseases #	Ischemic heart disease (IHD)	522.90	45.31	Non-communicable diseases #	Ischemic heart disease (IHD)	42.97	24.70
		Hypertensive heart disease (HHD)	248.82	21.56		Hypertensive heart disease (HHD)	17.36	9.98
		Rheumatic heart disease (RHD)	36.13	3.13		Rheumatic heart disease (RHD)	8.45	4.86
		Stroke	31.84	2.76		Stroke	0.25	0.14
		Cardiovascular diseases except for IHD, HHD, RHD and Stroke (CD)	16.87	1.46		Cardiovascular diseases except for IHD, HHD, RHD and Stroke (CD)	2.23	1.28
		Chronic obstructive pulmonary disease (COPD)	103.60	8.98		Diabetes mellitus and Chronic kidney disease (DMCKD)	40.45	23.25
		Chronic respiratory diseases except for COPD	2.14	0.19		Chronic obstructive pulmonary disease (COPD)	28.02	16.10
		Malignant neoplasms except for TBLC	52.12	4.52		Chronic respiratory diseases except for COPD	0.25	0.14
		Tracheal, bronchus, and lung cancer (TBLC)	16.68	1.45		Other non-communicable diseases	4.57	2.62
		Diabetes mellitus and Chronic kidney disease (DMCKD)	32.53	2.82		Malignant neoplasms except for TBLC	3.17	1.82
		Digestive diseases	7.68	0.67		Tracheal, bronchus, and lung cancer (TBLC)	2.30	1.32
		Other non-communicable diseases	6.94	0.60		Digestive diseases	2.90	1.67
		Neurological disorders	4.51	0.39		Neurological disorders	0.50	0.29
		Mental disorders and Substance use disorders	3.00	0.26				
		Musculoskeletal disorders	1.79	0.16				
		Other neoplasms	0.13	0.01				
	Communicable, maternal, neonatal, and nutritional diseases		5.33	0.46	Communicable, maternal, neonatal, and nutritional diseases		5.58	3.21
Approach 2 linear regression	**Redistribution target diseases**		**Count**	**Percent (%)**	**Redistribution target diseases**		**Count**	**Percent (%)**
	Remaining heart failure		41.00	3.55	Remaining heart failure		9.00	5.17
	Non-communicable diseases #	Ischemic heart disease (IHD)	1052.42	91.20	Non-communicable diseases #	Chronic obstructive pulmonary disease (COPD)	165.00	94.83
		Cardiovascular diseases except for IHD, HHD, RHD and Stroke (CD)	12.64	1.10				
		Hypertensive heart disease (HHD)	1.33	0.12				
		Rheumatic heart disease (RHD)	0.58	0.05				
		Diabetes mellitus and Chronic kidney disease (DMCKD)	0.67	0.06				
	Other garbage codes		45.36	3.93				

Table 4 Results of the redistribution of heart failure
Diseases of three systems, cardiovascular diseases, chronic respiratory diseases and malignant neoplasms, are further divided into detailed diseases.

Table 5 Changes in death counts and ranks of the target diseases

	Weifang							Xuanwei				
	original	after heart failure correction	further after heart failure redistribution		rank among all target diseases *	percentage of increase in death count (%, comparing to the count after correction)		original	after heart failure correction	further after heart failure redistribution		
			coarsened exact matching	*linear regression*		*coarsened exact matching*	*linear regression*			*coarsened exact matching*	*linear regression*	
Ischemic heart disease (IHD)	113147	113273	113795.90	114325.42	1	0.462	0.929	3609	3612	3654.97	3612.00	
Hypertensive heart disease (HHD)	7533	7567	7815.82	7568.33	6	3.288	0.018	220	223	240.36	223.00	
Rheumatic heart disease (RHD)	1472	1474	1510.13	1474.58	12	2.451	0.039	391	392	400.45	392.00	
Stroke	93671	93674	93705.84	93674.00	2	0.034	0	5527	5527	5527.25	5527.00	
Cardiovascular diseases except for IHD, HHD, RHD and Stroke (CD)	1040	1042	1058.87	1054.64	14	1.619	1.213	131	131	133.23	131.00	
Chronic obstructive pulmonary disease (COPD)	45219	45229	45332.60	45229.00	4	0.229	0	10165	10174	10202.02	10339.00	
Chronic respiratory diseases except for COPD	1101	1102	1104.14	1102.00	13	0.194	0	264	264	264.25	264.00	
Tracheal, bronchus, and lung cancer (TBLC)	41966	41969	41985.68	41969.00	5	0.040	0	8114	8114	8116.30	8114.00	
Malignant neoplasms except for TBLC	77195	77200	77252.12	77200.00	3	0.068	0	4146	4152	4155.17	4152.00	
Diabetes mellitus and Chronic kidney disease (DMCKD)	6370	6374	6406.53	6374.67	7	0.510	0.011	985	985	1025.45	985.00	
Digestive diseases	5083	5084	5091.68	5084.00	8	0.151	0	1825	1826	1828.90	1826.00	
Other non-communicable diseases	2324	2326	2332.94	2326.00	11	0.298	0	473	474	478.57	474.00	
Neurological disorders	3054	3055	3059.51	3055.00	10	0.148	0	314	314	314.50	314.00	
Mental disorders and Substance use disorders	513	513	516.00	513.00	15	0.585	0	250	250	250.00	250.00	
Musculoskeletal disorders	433	433	434.79	433.00	17	0.414	0	144	144	144.00	144.00	
Other neoplasms	464	464	464.13	464.00	16	0.028	0	22	22	22.00	22.00	
Communicable, maternal, neonatal, and nutritional diseases	3525	3526	3531.33	3526.00	9	0.151	0	2062	2063	2068.58	2063.00	

Table 5 Changes in death counts and ranks of the target diseases

* The ranks always kept the same.

In each subgroup, we split the total number of heart failures (UCOD on the treatment records) into other diseases according to the constituent ratio of UCOD on the control records. In some subgroups there was no control record, thus the redistribution could not be performed. At last, we added up the results of all subgroups to form the heart failure redistribution result of Weifang or Xuanwei.

2.3.2 Linear regression

The 12 target groups (TGs) for linear regression model establishment have been listed by Ahern and colleagues[12]. In each year, heart failure and the 12 TGs formed a "heart failure universe", and the percentage of each of them could be calculated by the following formula: *percentage (%) = death records with UCOD being heart failure or each of the 12 TGs / heart failure universe*. Using all the percentages data, we ran linear regression: $\%\ of\ a\ TG = \alpha + \beta * [\%\ of\ heart\ failure] + \varepsilon$, for 12 separate times, to estimate the relation across years between the proportion of heart failure-attributed deaths and the proportion of the deaths attributed to each TG. Then, if a TG / several TGs had statistically significant positive correlation with heart failure, we dropped it / them as Ahern and colleagues had instructed, formed a new "heart failure universe" consisting of heart failure and fewer TGs, re-calculated their percentages and re-ran the linear regression (the number of times was equal to the number of remaining TGs). The "forming universes – calculating percentages – running regression models" process should be repeated, until no TG was significantly positively associated with heart failure. We carried out the analysis in 8 subgroups (2 genders multiplied 4 age groups, see supplementary table 1), respectively.

In each subgroup, after all rounds of regression, we kept the TGs whose final regression coefficient (β) was negative and also significant ($p < 0.05$). We split the total number of heart failures into these TGs, according to the constituent ratio of their y-intercepts (α) yielded by the regression. We could further split each TG into the ICD-10 codes contained in it, according to the constituent ratio of the codes as well. At last, same as the first approach, we added up the results of all subgroups to form the heart failure redistribution result of Weifang or Xuanwei.

2.4 Change in cause-specific mortality

After data cleaning and quality improvement, we calculated original cause-specific mortalities in Weifang and Xuanwei (based on the diseases classification in GBD 2017 study[6], the following the same). Then, after heart failure correction and redistribution (by 2 approaches respectively), we aggregated the results to the level of diseases, and calculated again the cause-specific mortalities. We compared the cause-specific mortality before and after. All analyses were performed using SAS software (version 9.4).

3. Results

In total, Weifang had 477136 deaths from 2010 to 2017, and Xuanwei had 55536 deaths from 2010 to 2016. Within those, death cases with UCOD stated as heart failure were 1556 (0.33%) and 226 (0.41%), respectively. Table 2 shows the results of UCOD correction by using rules from WHO before the redistribution. We could see that only 12.98% records in Weifang and 10.62% records in Xuanwei could be assigned to a plausible correct UCOD directly in this step, respectively. These plausible correct UCOD were highly overlapping with the original top 20 causes of death in both cities. In about 10% of the records, UCOD was changed from heart failure to another GC. And in about 75% of the records, UCOD was not changed. Table 3 shows the different cause-specific and GCs mortalities in Weifang and Xuanwei after the correction step. In Weifang, during 2010-2017, the average total mortality was 645.41/100,000, IHD mortality was 153.22/100,000 and stroke mortality was 126.71/100,000. In Xuanwei, during 2010-2016, the average total mortality was 524.56/100,000, chronic obstructive pulmonary disease (COPD) mortality was 96.10/100,000 and tracheal, bronchus, and lung cancer (TBLC) mortality was 76.64/100,000. The percentage of GCs in Weifang and Xuanwei was 9.24% and 18.58%, respectively. The constituent ratios of the four levels of GCs were shown in Figure 1. Table 4 shows the results of heart failure redistribution. Overall, in 2 cities, by 2 approaches, the proportions of heart failure deaths not able to be changed to any other UCOD after the redistribution were 3.55–8.62%. When using coarsen exact matching, the numbers of "treatment records" and "control records" were 1154 and 6506 in Weifang, 174 and 163 in Xuanwei, respectively. Finally, in Weifang, 1093 heart failure deaths (94.71%) were matched to 4940 deaths with other non-garbage, non-injury UCOD, mainly redistributed to IHD (45.31%), hypertensive heart disease (HHD, 21.56%) and COPD (8.98%). In Xuanwei, 159 heart failure deaths (91.38%) were matched to 128 deaths with other non-garbage, non-injury UCOD, mainly redistributed to IHD (24.70%), diabetes mellitus and chronic kidney disease (DMCKD 23.25%) and COPD (16.10%). The unsuccessfully matched heart failure deaths mostly happened before 2014, and had no or undefined diagnostic agency.

When using linear regression, in Weifang and Xuanwei the data set for redistribution work contained 164217 and 11886 records, respectively. Finally, in Weifang, 1113 heart failure deaths (96.45%) were successfully redistributed, almost all to IHD (91.20%). In Xuanwei, 165 heart failure deaths (94.83%) were successfully redistributed, all to COPD. In Weifang, no redistribution proportions were estimated for heart failure deaths of 3 subgroups, 0-14 or 15-44 male and 45-64 female; in Xuanwei, for heart failure deaths of 4 subgroups, 0-14 or 15-44 male and female.

Table 5 shows the changes in death counts of the target diseases. We displayed the change after heart failure correction, and the further change after heart failure redistribution. Moreover, we calculated the percentages of increase in the death counts after redistribution, comparing with the counts after correction.

In Weifang, after redistribution by coarsen exact matching, the deaths due to HHD, rheumatic heart disease (RHD) and cardiovascular diseases except for IHD, HHD, RHD and stroke (CD) increased the most, the increasing percentages were 3.288%, 2.451% and 1.619%, respectively. The deaths due to mental disorders and substance use disorders, DMCKD and IHD increased the next most, the percentages were around 0.5%. After redistribution by linear regression, the deaths due to CD increased 1.213%, and the deaths due to IHD increased 0.929%.

In Xuanwei, after redistribution by coarsen exact matching, the deaths due to HHD and DMCKD increased 7.786% and 4.107%, the deaths due to RHD and CD increased 2.156% and 1.702%, while the deaths due to IHD and other non-communicable diseases increased 1.190% and 0.963%, respectively. After redistribution by linear regression, the deaths due to COPD increased 1.622%.

4. Discussion

4.1 The quality of UCOD determination

Previous international studies[18–24] have reported the misjudgment of UCOD on death certificate (not limited to GCs). For some causes such as malignant neoplasms or injuries, the agreement between the initial coders and the reviewers or autopsy results was relatively good, while for other causes such as nephrotic diseases, circulatory diseases or pneumonia, the agreement was poor. Although owing to the "compensatory effect of errors", mortality statistics of a certain area may not be affected significantly, it is still necessary for the administration to undertake routine internal studies on the quality of UCOD coding practices[19].

In China, primary medical and public health units started reporting the information on death cases directly over the Internet via "Cause of Death Registration Management System" since 2005.[25] Over the years, Chinese Center for Disease Control and Prevention (CDC) has regularly conducted certain work to control and improve the quality of death data, such as surveys on under-reported deaths, improvement of regulations and protocols, training for the staff carrying out the registration and report, and so on.[26] These efforts have produced remarkable effect, however, in this study our group still detected some problems existing in the cause of death coding process.

According to the classification by IHME[6], the percentages of level 1 to level 4 GCs in the death data set were 4.61%, 1.62%, 1.39% and 1.67% in Weifang, and 7.73%, 4.69%, 0.83% and 5.38% in Xuanwei. The much higher latter percentages indicated that in county-level cities, the cause of death coding was likely to be less accurate. One counter proof was that in this study, in Xuanwei, our manual UCOD correction process impacted the change of cause of death ranks as much as the heart failure redistribution process, while in Weifang the redistribution had a bigger impact (see supplementary table 2). This reminded us again that the fundamental way for improving the quality of death certificate is to improve the capacity of relevant primary staff on UCOD identification and coding, which may be especially true in small areas where current quality of cause of death registration is not so high. Some new developed software tools such as "ANACONDA"[27] could help assess the accuracy of the mortality and cause of death data, and help find out where the main problems lie on.

Moreover, among the four levels, the percentage of level 1 GCs (causes that cannot or should not be considered as UCOD) was the highest in both cities, which indicated that we should further enhance certain training on the difference between UCOD and diagnoses in clinical or health service encounters.

4.2 Difference between the two redistribution approaches

Coarsen exact matching and linear regression are both recommended internationally, however, their redistribution results showed obvious difference.

Coarsen exact matching led to a much more diverse spectrum of redistribution target diseases, including malignant neoplasms, diabetes or chronic kidney disease, although cardiovascular diseases, cerebrovascular diseases, COPD, etc. were still dominant. Since the original counts of deaths from various causes could be quite different, after redistribution by coarsen exact matching, for some causes such as HHD, RHD, CD or DMCKD, the number of heart failure deaths redistributed to which was not the largest, or even relatively small, but the extent of increase of which was quite high, or even the biggest. This happened in both cities.

When using linear regression, the heart failure deaths were almost totally redistributed to IHD in Weifang, and to COPD in Xuanwei. Moreover, in Xuanwei, the most important target disease would change from IHD to COPD, if the redistribution approach changed from coarsen exact matching to linear regression.

The difference in redistribution results reflected the difference in the core principles of the two methods. Coarsen exact matching determined the target diseases by certain records and their constituent ratio in the data set itself. The deceased on these "control records" had heart failure before death, and also his or her UCOD was possible to be misjudged as "heart failure" by doctors or public health staff. Therefore, in reverse, their UCOD may all be possible target causes to which heart failure deaths could be re-assigned. Linear regression chose its TGs in advance based on the pathophysiology of heart failure, which was also reasonable. However, the redistribution results seemed to be strongly driven by the variation of the percentages of one or two TGs among years. These TGs may be the dominant ones in the data set used for redistribution work, or on the contrary, the ones with quite small percentages.

We observed the percentages of heart failure and the 12 TGs in each year in Weifang, as shown in supplementary table 3. The percentage of IHD was always the highest, however, in 45-64 male and 15-44 female, year by year the percentage was around or higher than 75%, while in 65 or older male and female, the percentage was around or lower than 70%. And to the redistribution results, in the latter two subgroups, heart failure was redistributed to IHD, while in the first two groups, heart failure was redistributed to other TGs, whose percentages were lower than 1% in every year. Therefore, we inferred that when the percentage of IHD was high enough, the variation of it between years may become difficult to identify, and this "stationary trend" could not drive the redistribution of heart failure towards it.

In the publication of Ahern[12], the centralization of redistribution target diseases was not obvious. In some subgroups, heart failure was redistributed to 4 or more TGs out of 12. This maybe because that they used super region level databases (developed and developing countries).

4.3 Strengths and limitations

In this study, we chose two internationally certified methods to redistribute heart failure in two Chinese cities. Through these processes, we explored their suitability of usage on relatively small death data sets, and also compared and discussed the difference in their principles. We believe that results from this

study could provide some references for redistributing other GCs in small areas.

As a first attempt, we only conducted the study in two cities and for one GC, this would limit the extrapolation of the results. Further study on more GCs and in more cities would be in need.

5. Conclusions

In cities with 1 to 10 million permanent residents, if the percentage of certain GC was quite low in their death data sets, its potential adverse impact on death statistics and subsequent policy making, or in other words, the necessity of its redistribution, would be worth discussed. If redistributing, coarsen exact matching approach should probably be more suitable, comparing to linear regression. The fundamental way for improving the quality of death data is to improve the capacity of primary staff on UCOD identification and coding.

Declarations

-Ethics approval and consent to participate

Not applicable

-Consent for publication

Not applicable

-Availability of data and materials

Data and materials used in this study may be obtained after consultation with the corresponding author.

-Competing interests

The authors declare that they have no competing interests.

-Funding

1) the Collaborative Innovation Team Project: Health Effect of Environmental Factors and Gut Microbiome on Digestive Tract-Related Diseases: Population-Based Cohort Studies (2016-12M-3-001) supported by CAMS Innovation Fund for Medical Sciences;

2) Strengthen Capacity of Study and Application on the Burden of Disease in Health Care Systems in China: Establishment and Development of Chinese Burden of Disease Research and Dissemination Center (15-208) supported by the China Medical Board (CMB).

The funding bodies played no role in the study design and data collection, analysis, and interpretation and writing the manuscript.

-Authors' contributions

Liqun Liu organized and analyzed the data, interpreted the results and drafted the manuscript, including the tables and figures.

Xuewei Wang, Chunping Wang, Xiangyun Ma, Xianfeng Meng, Bofu Ning and Ning Li supported the data acquisition process, and helped explain the questions relating the operation of the death surveillance system.

Xia Wan designed the study, guided its implementation, helped with the interpretation of the results and revised the manuscript.

-Acknowledgements

We thank Prof. Gonghuan Yang for her valuable advice during the study design and the data analysis.

References

1. WHO. International statistical classification of diseases and related health problems (10th revision): Volume 2 Instruction manual (fifth edition); 2016.
2. Murray C, Lopez A. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Cambridge: Harvard University Press; 1996.
3. Naghavi M, Makela S, Foreman K, O'Brien J, Pourmalek F, Lozano R. Algorithms for enhancing public health utility of national causes-of-death data. Population Health Metrics. 2010;8:9.
4. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;385(9963):117–71.
5. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specifi c mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388:1459–544.
6. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980 - 2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392:1736–88.

7. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396:1204–22.
8. Phillips DE, Lozano R, Naghavi M, et al. A composite metric for assessing data on mortality and causes of death: the vital statistics performance index. *Population Health Metrics* 2014; 12(14).
9. Wan X, Yang G. Is the Mortality Trend of Ischemic Heart Disease by the GBD2013 Study in China Real? Biomed Environ Sci. 2017;30(3):204–9.
10. Stevens GA, King G, Shibuya K. Deaths from heart failure: using coarsened exact matching to correct cause-of-death statistics. Population Health Metrics. 2010;8:6.
11. Lozano R, Murray CJ, Lopez AD, Satoh T. Miscoding and misclassification of ischaemic heart disease mortality. Global Programme on Evidence for Health Policy Working Paper No 12: World Health Organization; 2001.
12. Ahern RM, Lozano R, Naghavi M, Foreman K, Gakidou E, Murray CJ. Improving the public health utility of global cardiovascular mortality data: the rise of ischemic heart disease. Population Health Metrics. 2011;9:8.
13. Ellingsen CL, Ebbing M, Alfsen GC, Vollset SE. Injury death certificates without specification of the circumstances leading to the fatal injury - the Norwegian Cause of Death Registry 2005 - 2014. Population Health Metrics. 2018;16:20.
14. Foreman KJ, Naghavi M, Ezzati M. Improving the usefulness of US mortality data: new methods for reclassification of underlying cause of death. Population Health Metrics. 2016;14:14.
15. Wan X, Liu L, Yang G. Discussion on data cleaning of vital registration system in Xuanwei, Yunnan, 2014-2016. Disease Surveillance. 2018;33(6):520–4.
16. Liu L, Liu X, Liu Y, Wan X, Yang G. Building the standard operating procedure for improving health insurance data quality: Quality evaluation and improvement on the reimbursement records data of new rural cooperative medical system of a county in Henan province, 2013-2015. Disease Surveillance. 2021;36(3):261–9.
17. WHO. ICD-10 Interactive Self Learning Tool. 2010. https://apps.who.int/classifications/apps/icd/icd10training/ICD-10%20training/Start/index.html.
18. Naka Mieno M, Tanaka N, Arai T, et al. Accuracy of Death Certificates and Assessment of Factors for Misclassification of Underlying Cause of Death. J Epidemiol. 2016;26(4):191–8.
19. Lu T, Lee M, Chou M. Accuracy of cause-of-death coding in Taiwan: types of miscoding and effects on mortality statistics. Int J Epidemiol. 2000;29(2):336–43.
20. Lahti R, Penttilä A. The validity of death certificates: routine validation of death certification and its effects on mortality statistics. Forensic Sci Int. 2001;115(1-2):15–32.
21. Rampatige R, Mikkelsen L, Hernandez B, Riley I, Lopez AD. Systematic review of statistics on causes of deaths in hospitals: strengthening the evidence for policy-makers. Bull World Health Organ. 2014;92(11):807–16.
22. Hazard RH, Rahman Chowdhury H, Adair T, et al. The quality of medical death certification of cause of death in hospitals in rural Bangladesh: impact of introducing the International Form of Medical Certificate of Cause of Death. BMC Health Serv Res. 2017;17(1):688.
23. McGivern L, Shulman L, Carney JK, Shapiro S, Bundock E. Death Certification Errors and the Effect on Mortality Statistics. Public Health Rep. 2017;132(6):669–75.
24. Falci L, Lee Argov EJ, Van Wye G, Plitt M, Soto A, Huynh M. Examination of Cause-of-Death Data Quality Among New York City Deaths Due to Cancer, Pneumonia, or Diabetes From 2010 to 2014. Am J Epidemiol. 2018;187(1):144–52.
25. Wan X, Liu L, Yang G. Discussion on data cleaning of vital registration system in Xuanwei, Yunnan, 2014-2016. DISEASE SURVEILLANCE. 2018;33(5):441–5.
26. The National Center for Chronic and Noncommunicable Disease Control and Prevention in China CDC. Work Instruction Handbook on Population and Cause of Death Surveillance. Beijing: China Population Publishing House; 2017.
27. Mikkelsen L, Moesgaard K, Hegnauer M, Lopez AD. ANACONDA: a new tool to improve mortality and cause of death data. BMC Med. 2020;18:61.

Figures

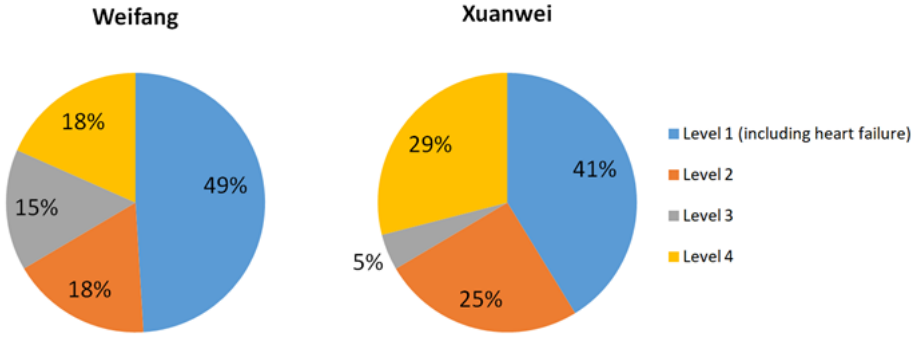

Figure 1

The constituent ratio of GCs in Weifang and Xuanwei

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 3Supplementarytables2022.1.9.xlsx