Letter

A Case Study on Garbage Code Redistribution Methods for Heart Failure at City Level by Two Approaches

Liqun Liu[1,2,&], Zemin Cai[1,2,&], Xuewei Wang[1], Chunping Wang[3], Xiangyun Ma[4], Xianfeng Meng[5], Bofu Ning[4], Ning Li[5], and Xia Wan[1,2,#]

Cause of death surveillance data is most important 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 a 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, medical workers 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 actual practical work. These codes was regarded as "garbage code (GC)"[2]. 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 progression of the Global Burden of Disease (GBD) study, the identification and distinguishment of GCs kept getting further refined[4].

The percentage of GCs is the main index for the reliability of a death statistics dataset assessment. The existence of GCs affects the statistics on the 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, then unavoidably influence policy making and the population health promotion goals. In addition to the originally existing GCs, the results of GCs redistribution could also seriously influence the judgment on the trend of diseases.

Since the existence and the poor redistribution of GCs have adverse impacts on death data quality, scientists have explored reasonable approaches to redistribution[3]. GCs Redistribution could be understood as assigning a GC to a plausible correct ICD code for UCOD, based on the pathophysiological characteristics of certain diseases and so on. Nowadays, there are several ways for redistributing GCs, such as expert consultation[3], fixed proportional reassignment[3], computed proportional reassignment based on the information from the cause of death chain (including coarsened exact matching)[4], regression models (including linear regression)[5], and other complicated equations[6]. Among them, linear regression (LR) model and coarsened exact matching (CEM) are commonly used.

In previous studies[5], most of the approaches were used in global or national-level database scenarios. While, some causes of death at the city level with 1 to 10 million permanent residents would be impacted by GCs more sensitively than that of national or global level. It is unknown whether these methods are applied to the city level data or not. In China, administrative regions are divided into four levels: provincial, prefecture, county, and township. Normally, the number of populations from prefectures or counties is within 1 to 10 million, while the resident quantity of townships is smaller than that. Thus, taking a common GC - "heart failure"[7] as an example, this study aimed to test the application of CEM and LR at the city level, including a prefecture - Weifang city from Shandong Province and a county - Xuanwei city from Yunnan Province.

Chinese Cause-of-Death Reporting System (CDRS) had been established since 1990s, with only

doi: 10.3967/bes2024.173

1. Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (CAMS) and School of Basic Medicine, Peking Union Medical College (PUMC), Beijing 100005, China; 2. State Key Laboratory of Respiratory Health and Multimorbidity, Beijing 100005, China; 3. School of Public Health, Weifang Medical University, Weifang 261042, Shandong, China; 4. Xuanwei Center for Disease Control and Prevention, Xuanwei 655400, Yunnan, China; 5. Weifang Center for Disease Control and Prevention, Weifang 261061, Shandong, China

145 disease surveillance points across 31 provinces. After 2003, most of the cities/counties gradually established the cause of death surveillance system covering the whole administrative regions, as well as Weifang and Xuanwei. The cause of death report rules were gradually improved, forming a nest of criteria for cause of death reports and analysis procedures[8]. The causes of death data were from Weifang CDRS from 2010 to 2017, and from Xuanwei from 2010 to 2016, respectively. The average population of permanent residents in Weifang and Xuanwei during the study period was about 9.24 and 1.51 million, respectively. The data was desensitized before analyzed, and data cleaning was shown in previous paper[9].

The UCOD death records with ICD-10 codes I50, I50.0, I50.1 and I50.9 stated as heart failure, were extracted from the datasets. Firstly, these UCOD records were corrected based on the cause of death chain and the information on other diseases, following the ICD-10 rules and guidelines for morbidity coding established by WHO[1]. Then, those records with UCOD remaining to be heart failure entered the redistribution process.

Redistribution was carried out in Weifang and Xuanwei both by CEM[7] and LR method[10]. CEM combined with the death chain information based on fixed proportional reassignment[7]. The records with UCOD remaining to be heart failure were so-called “treatment records", while those UCOD records with a non-garbage code and non-injury disease, but with heart failure in the cause of death chain, were regarded as the “control records". Whether a UCOD was a GC or an injury was judged according to the disease classification by the Institute for Health Metrics and Evaluation (IHME) in their GBD 2017[3]. Notably, 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 they met the conditions. Five variables were selected to divide subgroups, including death year, urban or rural resident, gender, age, and the highest agency of diagnosis (Table 1). In each subgroup, we split the total number of heart failures (UCOD on the treatment records) into other diseases in accordance with 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 and Xuanwei, respectively.

LR is established between the garbage code and the target code. The negative correlation and linear regression model were used to find the target coding range and allocation ratio method. The 12 target groups (TGs) for LR model establishment have been listed by Ahern et al[10]. In each year, heart failure and the 12 TGs formed a “heart failure universe". Then using all the percentages data, a LR formula: % TG = $\alpha + \beta \times$[% heart failure] + ε, was run 12 times, to estimate the relation across years between the proportion of heart failure attributed to deaths and the proportion of the deaths attributed to TG. If the TGs had statistically significant positive correlation with heart failure, they were dropped and then formed a new “heart failure universe”. After that, the procedure was repeated until no TG was significantly positively associated with heart failure. We carried out the analysis in 8 subgroups (2 genders multiplied by 4 age groups). After all rounds of calculation, the TGs whose final regression coefficient (β) was negative and statistically significant ($P < 0.05$) in each subgroup were kept. According to the constituent ratio of their y-intercepts (α) yielded by the regression, we split the total number of heart failures into TGs. We could further split each TG into the ICD-10 codes contained in it, in accordance with the constituent ratio of the codes as well.

After heart failure correction by using WHO guidelines and redistribution by two approaches, we aggregated the results to the categories of diseases, and compared the cause-specific mortalities before and after redistribution. All analyses were performed by SAS 9.4 (SAS Institute Inc., Cary, North Carolina, USA).

All methods of this study were carried out in accordance with relevant guidelines and regulations. All study protocols were approved by ethics committee of Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (CAMS) (authorization number: 037-2014).

UCOD Correction before the Redistribution

In total, Weifang and Xuanwei had 477,136 and 55,536 deaths, with 56.30% and 61.80% of males proportion, respectively. The proportion of the cause of death without a diagnosis from hospitals in Weifang (3.54%) was much lower than that of Xuanwei (23.34%). The percentages of category 1 to category 4 GCs in the death dataset were 4.61%, 1.62%, 1.39% and 1.67% in Weifang, and 7.73%, 4.69%, 0.83% and 5.38% in Xuanwei, respectively (Table 1), which indicated that some problems still existed in the cause of death coding process,

especially in county-level cities. The percentage of GCs in Weifang and Xuanwei was 9.29% and 18.63%, respectively (Supplementary Figure S1, available in www.besjournal.com). The percentage of category 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 encounter.

Within those, death cases with UCOD stated as heart failure were 1,556 (0.33%) in Weifang and 226 (0.41%) in Xuanwei, respectively. Following the WHO

Table 1. Variables for dividing subgroups in Weifang and Xuanwei

Characteristics		Weifang		Xuanwei	
		n	Percent (%)	*n*	Percent (%)
Gender	Male	268,639	56.30	34,320	61.80
	Female	208,425	43.68	21,216	38.20
	Unknown*	72	0.02	–	–
Age group (years)	0-	5,513	1.16	2,120	3.82
	15-	24,090	5.05	6,584	11.86
	45-	103,215	21.63	12,472	22.46
	65-	94,719	19.85	10,959	19.73
	75-	150,839	31.61	15,322	27.59
	85-	98,760	20.70	8,079	14.55
Death year	2010	56,264	11.79	7,054	12.70
	2011	56,446	11.83	7,993	14.39
	2012	56,333	11.81	7,737	13.93
	2013	59,725	12.52	7,924	14.27
	2014	58,323	12.22	8,335	15.01
	2015	60,645	12.71	7,869	14.17
	2016	63,638	13.34	8,610	15.50
	2017	65,762	13.78	–	–
	Between 2014—2016 but unknown*	–	–	14	0.03
The highest agency of diagnosis	Tertiary hospital	106,308	22.28	10,018	18.04
	Secondary hospital	270,190	56.63	22,579	40.66
	Primary hospital	83,771	17.56	9,976	17.96
	No hospital diagnosis or other or unknown*	16,867	3.54	12,963	23.34
Urban or rural resident	Rural	263,743	55.28	–	–
	Urban	213,166	44.68	–	–
	Unknown*	227	0.05	–	–
Garbage code#	Category 1	21,975	4.61	4,291	7.73
	Category 2	7,706	1.62	2,606	4.69
	Category 3	6,629	1.39	462	0.83
	Category 4	7,967	1.67	2,990	5.38
Total		477,136		55,536	

Note. *If the missing values could not be filled, 'unknown' was used as a category. #Garbage Code categories: category 1: causes that cannot or should not be considered as underlying cause of death (UCOD); category 2: intermediate causes in the cause of death chain; category 3: immediate causes in the cause of death chain; category 4: unspecified causes within a larger cause grouping; — means no data.

guidelines, in Weifang, 74.16% of records remained heart failure, 12.85% of records were changed to other GCs, and 12.98% of records were assigned to a plausible correct UCOD. While in Xuan, these three proportions were 76.99%, 12.39% and 10.62%, respectively. These plausible correct UCODs were highly overlapping with the original top 20 causes of death in both cities (Table 2).

The average total cause of death rates was 645.41/100,000 in Weifang, which was higher than that of Xuanwei (524.56/100,000). After corrected the causes of death by WHO guidelines, the top two causes of death were ischemic heart disease (IHD) (153.22/100,000) and stroke (126.71/100,000) in Weifang, and chronic obstructive pulmonary disease (COPD) (96.10/100,000) and tracheal, bronchus, and lung cancer (TBLC) (76.64/100,000) in Xuanwei (Supplementary Table S1, available in www.besjournal.com).

Heart Failure Redistribution

Overall, after using CEM and LR approaches, the proportions of heart failure deaths unable to be changed to any other UCOD were 3.55% to 8.62%, respectively (Table 3).

When using CEM, the numbers of "treatment

Table 2. The correction of heart failure by WHO guidelines

Item	Weifang (*n* = 1,556)			Xuanwei (*n* = 226)		
	Corrected UCODs	*n*	Percent (%)	Corrected UCODs	*n*	Percent (%)
Garbage codes	Remaining heart failure	1,154	74.16	Remaining heart failure	174	76.99
	Other garbage codes	200	12.85	Other garbage codes	28	12.39
	Correct UCOD	202	12.98	Correct UCOD	24	10.62
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

Note. [#]Diseases of three systems, cardiovascular diseases, chronic respiratory diseases and malignant neoplasms, are further divided into detailed diseases. ICD: International Statistical Classification of Diseases and Related Health Problems; UCOD: underlying cause of death; — means no data.

records" and "control records" were 1,154 and 6,506 in Weifang, and 174 and 163 in Xuanwei, respectively. After redistribution by CEM, in Weifang, the deaths due to hypertensive heart disease (HHD), rheumatic heart disease (RHD) and cardiovascular diseases except for IHD, HHD, RHD and stroke (CD) most increased by 3.29%, 2.45% and 1.62%, respectively. In Xuanwei, the deaths due to HHD, diabetes mellitus and chronic kidney disease (DMCKD), RHD, and CD increased the most, with 7.79%, 4.11%, 2.16% and 1.70% of the increasing proportion, respectively (Supplementary Table S2,

Table 3. Results of the redistribution of heart failure

Weifang (*n* = 1,154)			**Xuanwei (*n* = 174)**		
Redistribution target diseases	Count, *n*	Percent (%)	Redistribution target diseases	Count, *n*	Percent (%)
Approach 1 of coarsened exact matching					
Remaining heart failure	61.00	5.29	Remaining heart failure	15.00	8.62
Non-communicable diseases[#]	1,093.00	94.71	Non-communicable diseases[#]	159.00	91.38
Ischemic heart disease (IHD)	522.90	45.31	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					
Remaining heart failure	41.00	3.55	Remaining heart failure	9.00	5.17
Non-communicable diseases [#]	1,113.00	96.45	Non-communicable diseases [#]	165.00	94.83
Ischemic heart disease (IHD)	1,052.42	91.20	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	–	–	–

Note. [#]Diseases of three systems, cardiovascular diseases, chronic respiratory diseases and malignant neoplasms, are further divided into detailed diseases; — means no data.

available in www.besjournal.com).

After redistribution by LR, in Weifang, the deaths due to CD and IHD increased by 1.21% and 0.93%, respectively. In Xuanwei, the deaths due to COPD increased by 1.62% (Supplementary Table S2, available in www.besjournal.com). 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, those were estimated for heart failure deaths of 4 subgroups, 0–14 or 15–44 male and female (Supplementary Table S3, available in www.besjournal.com).

Based on the above results, CEM led to a much more diverse spectrum of redistribution target diseases compared to LR. CEM determines the target diseases by considering the possible diseases occurred before the deceased die, while LR chose its TGs in advance based on the pathophysiology of heart failure. The LR redistribution results seemed to be strongly driven by the variation of the percentages of one or two TGs over the years, which may be the dominant ones in the dataset with large percentages. While in the publication of Ahern[10], the central tendency of redistribution to target diseases was not obvious. In some subgroups, heart failure was redistributed to 4 or more TGs out of 12. This may be because they used super region level databases (developed or developing countries). However, from the different age groups results, if the percentage of certain disease was high enough, we inferred that 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 (Supplementary Table S4, available in www.besjournal.com).

Finally, the mortalities and ranks of all the causes originally after the redistribution of heart failure by two approaches were shown in Supplementary Table S5 (available in www.besjournal.com).

Our study illustrated that the CEM might be better than the LR at city level. However, another study[6] conducted in Taiwan showed that multinomial logistic regression was the suitable model for GC redistribution, compared to naive Bayes classifier and CEM, which might be because Taiwan has a larger population quantity (23 million population) than that of Weifang or Xuanwei, indicating that the application of different models might be related to the scale of the population.

Several suggestions for the applicability of the two methods are proposed. LR is more recommended when the target codes are very clear, or the data sample size is large enough for modeling. While, CEM is recommended if the data sample size is small, or primary health workers are not requested for complex technology, without establishing statistical model.

Funding This study was supported by State Key Laboratory Special Fund (2060204); Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2023-I2M-2-001); 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; 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).

Competing Interests The authors declare that there are no conflict of interest.

Authors' Contributions Organizing and analyzing the data, interpreting the results and drafting the manuscript, including the tables and figures: Liqun Liu and Zemin Cai. Supporting the data acquisition process, and helping explain the questions relating the operation of the death surveillance system: Xuewei Wang, Chunping Wang, Xiangyun Ma, Xianfeng Meng, Bofu Ning and Ning Li. Designing the study, guiding its implementation, helping with the interpretation of the results and revising the manuscript: Xia Wan. Reading and approving the final manuscript: All authors.

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

[&]These authors contributed equally to this work.

[#]Correspondence should be addressed to Xia Wan, Professor, Ph.D, Tel: 13621024640, E-mail: xiawan@ibms.pumc.edu.cn

Biographical notes of the first authors: Liqun Liu, female, born in 1983, associate professor, majoring in burden of disease and tobacco control; Zemin Cai, female, born in 1994, doctoral candidate, majoring in burden of disease and tobacco control.

Received: July 1, 2024;

Accepted: October 15, 2024

REFERENCES

1. World Health Organization. International statistical classification of diseases and related health problems (10th revision): volume 2 Instruction manual. 5th ed. World Health

Organization. 2015, 592-600.
2. Murray CJL, Lopez AD. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary. World Health Organization. 1996.
3. Naghavi M, Makela S, Foreman K, et al. Algorithms for enhancing public health utility of national causes-of-death data. Popul Health Metr, 2010; 8, 9.
4. 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.
5. Ellingsen CL, Ebbing M, Alfsen GC, et al. Injury death certificates without specification of the circumstances leading to the fatal injury - the Norwegian Cause of Death Registry 2005-2014. Popul Health Metr, 2018; 16, 20.
6. Ng TC, Lo WC, Ku CC, et al. Improving the use of mortality data in public health: a comparison of garbage code redistribution models. Am J Public Health, 2020; 110, 222–9.
7. Stevens GA, King G, Shibuya K. Deaths from heart failure: using coarsened exact matching to correct cause-of-death statistics. Popul Health Metr, 2010; 8, 6.
8. The National Center for Chronic and Noncommunicable Disease Control and Prevention in China Center for Disease Control and Prevention. Chinese cause-of-death surveillance dataset 2015. China Science and Technology Press. 2016. (In Chinese).
9. Liu LQ, Liu XY, Liu Y, et al. 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. Dis Surveill, 2021; 36, 261–9. (In Chinese)
10. Ahern RM, Lozano R, Naghavi M, et al. Improving the public health utility of global cardiovascular mortality data: the rise of ischemic heart disease. Popul Health Metr, 2011; 9, 8.