Enhancing understandings of social determinants of health in China: Linkage and analysis of a national multilevel population health surveillance with routinely collected mortality records for 98 058 people in 31 provinces of mainland China

Yunning Liu
National Center for Choronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
Thomas Astell-Burt
School of Public and Population Health, University of Wollongong
Xiaoqi Feng
School of Health and Society, University of Wollongong
Fan Mao
National Center for Choronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
Ruiming Liang
National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
Peng Yin
National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
Limin Wang
National Center for Choronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
Lijun Wang
National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
Maigeng Zhou (maigengzhou0906@163.com)

Research article

Keywords: Record linkage, Chronic disease, Death surveillance, Risk factor surveillance, Multilevel logistic regression

Posted Date: December 9th, 2019

DOI: https://doi.org/10.21203/rs.2.18425/v1

Abstract

Background: The aim of this study was to enhance capability in research on social determinants of health in China by linking and analyzing routinely-collected death records over 5 years with national population health surveillance.

Methods: Linkage of 98 058 participants in the 2010 China Chronic Disease and Risk Factor Surveillance (CCDRFS) to records in the national death surveillance data from 2011 to 2015 was conducted through a matching program involving identification numbers, name, gender and residential address, followed by a structured checking process. Multilevel regressions were used to investigate five-year odds of all-cause, non-communicable disease (NCD), infectious disease and injury mortality in relation to person- and county-level factors.

Results: A total of 3,365 deaths were observed in the linked mortality and population health surveillance. Cross-checks and comparisons with national mortality distributions provided assurance that the linkage was reasonable. Geographic variation in mortality was observed via age and gender adjusted median odds ratios for all-cause mortality (>1.30), infectious disease (>2.01), NCD (>1.24) and injury (>1.12). Increased odds of all-cause and all three cause-specific mortality outcomes were higher with age and among men. Low educational attainment was a predictor of all-cause, NCD and injury mortality. Longer mean years of education at the county-level was only associated with lower injury mortality. Divorcees had a higher odd of all-cause and NCD mortality than singletons. Rurality was a predictor of all-cause and NCD mortality.

Conclusion: The results of this study provide utility for future investigations of social determinants of health and mortality using linked data in China.

Background

Non-communicable diseases (NCDs) have become the leading causes of death in China brought about by decades of economic development, rapid urbanization and improvements in health care, living conditions and nutrition[1, 2]. Many studies have evaluated the impact of risk factors on NCDs around the world, such as unhealthy lifestyles[3–5], socioeconomic status[6–9], family history[10] and abnormal biological indicators[11–13]. In recent decades, there are also some studies on determinants of NCDs in China, such as the China Kadoorie Biobank (CKB) study[14], the Jinchang cohort on cancer[15, 16], and a 15-year and nationally representative prospective study[17, 18]. These studies have found meaningful results about the association between risk factors and NCDs, although they are not without some limitations.

For example, some studies were implemented in specific regional areas and were not nationally representative[14]. Some studies focused their attention on particular risk factors, certain groups of people or isolated diseases, providing a limited scope of results[15, 16]. Perhaps most importantly, given the abovementioned structural changes sweeping China, there has been insufficient focus on social determinants of health. There is a need for a more general, nationally representative and multilevel population health data that can enable researchers to investigate social and spatial determinants of NCD and related mortality in China.

Record linkage based on current data has proven to be a cost-effective means for integrating information from different sources[19]. In recent decades, it has been applied extensively to generate databases for epidemiological studies in higher income countries, such as United States of America[20, 21], Australia[22, 23], Canada[24] and the United Kingdom[25], but it is much less common in China. Obstacles include a lack of attention paid to record linkage, missing identification numbers, variation in the permanent or residence address coding, imprecision in the reporting information as well as some other issues related to data quality.

In this study, we link records from 98 058 participants in the 2010 Chronic Disease and Risk Factor Surveillance to the routinely-collected national mortality surveillance system in China from 2011 to 2015 inclusive. We also tested different matching strategies used to link records from both of the databases based on conventional personal identifiers (e.g., name, age, identification number, address). The linkage was evaluated by comparing the proportion of different causes of death based on the matched records and with data from the national mortality surveillance system. To test the utility of the data and to provide a demonstration of its potential, multilevel logistic regressions were used to investigate associations between the main cause of death categories and social determinants.

Methods

Data source

Two separate, national representative databases were used in this study. The 2010 Chronic Disease and Risk Factor Surveillance database, consists of 98 058 individuals and was collected by the National Center for Chronic and Noncommunicable Disease Control and Prevention at China CDC between August and December in 2010. This data collection was based upon the National Disease Surveillance Points system (DSP), consisting of 161 districts/counties in all 31 provinces, autonomous regions and municipalities in mainland China. The population covered by the DSP system was approximately 73 million, or 6% of the Chinese population. This is a nationally representative population health survey with a multistage stratified cluster sampling strategy used to select participants among people 18 years and older[26]. The establishment, history and degree of representativeness of the DSP system are published elsewhere[27, 28]. The China CDC ethics committee approved the study and written informed consent was obtained from each participant before data collection.

Data was collected by gathering participants in certain central locations. For each participant, a face-to-face questionnaire was conducted to collect data including questions on socio-demographics, medical history, lifestyle-related factors and health service use. Identification numbers of all participants were collected at the beginning of 2011.

The second database consists of all the death records from the national mortality surveillance system from 2011 to 2015. In 2011 and 2012, the system was based upon the DSP system. The sampling strategy and the characteristics of this system have been described in detail elsewhere along with the quality control measures and the procedures for collecting data, coding the cause of death and determining the underlying cause of death[27, 28]. In 2013, the Ministry of Health combined the Chinese Center for Disease Control and Prevention's DSP system and the Ministry of Health's vital registration system into an integrated national mortality surveillance system to accelerate the development of a comprehensive vital registration and mortality surveillance for the whole country. This was expanded to include 605 surveillance points and population coverage increased to 300 million, or 24% of the Chinese population. Details on the development of the new surveillance have been published[29].

In the mortality surveillance, information on individual deaths in all population catchment areas are reported in real time based on an internet-based reporting system. Each death reported is systematically validated by local CDCs, which check the completeness, the underlying cause of death coding with the rules of the International Classification of Diseases[30] and internal logic of the items reported on death certificates. Causes-of-death are subsequently reported to the China CDC, where data are consolidated. Therefore, the national mortality surveillance system can provide data on total mortality, the broad cause-of-death distribution and the geographic distribution of deaths.

From 2013, the identification number of each death record was essential information and compulsory to fill in. The accuracy of the identification number was checked through comparing to the corresponding information from the police department from 2015. It is acknowledged that some death records do not have the identification number or an inaccurate identification number.

Record linkage

Two matching steps were adopted to link all the records from both of the two data sources. Before the match, some identifiers in the two data sources were collected including identification numbers, full name, gender, date of birth, address of permanent and residence and the corresponding codes. However, only partial coverage of the identification number and detailed address information can be gained for each record in each of the two data sources.

In the first step of matching the two data sources, a computer program was written including several strategies to match all possible records. The first program was to match records in population health surveillance with death records through identification number, which is unique identifier including 18 numbers. Those records without an identification number or where the identification number cannot be matched were progressed through a second program including four linkage strategies to find possible matched records one by one. If a record can be matched in any strategy, it will not be progressed to the next strategies. The four linkage strategies are:

i) Name, gender, permanent address coding;

ii) Name, gender and residence address coding;

iii) Name spelling, gender and permanent address coding;

iv) Name spelling, gender, residence address coding.

In the second step, program check and artificial check were adopted to get the confirmed matched records. Firstly, the name and identification numbers were combined together in the program check to confirm whether the records matched through identification number were correct. Secondly, all the other records matched through another four strategies were checked manually with two methods. If detailed village address were available, a check method that includes name or name spelling, gender and permanent or residence villages' address was used to confirm the linkage records. If the village address were not available, another check method that included name or name spelling, gender, permanent or residence town address and age differences less than 5 years would be used. Finally, the confirmed match database through program check and artificial check can be gained. The work flow of the whole matching process can be seen in Figure 1.

(Figure 1 The work flow of the matching process)

In order to guarantee the quality of matching, both of the two steps were preceded separately by two statisticians. All of the programs involved in program match and program check were run in SAS v9.4 (SAS Institute Inc. Cary, USA).

Statistical analysis

Two statistical analyses were performed. The first was to evaluate the matching effect and to assess the degree of completeness in the linked records. This included comparison of the mortality and proportion of different causes of death based on the linked database and from the national mortality surveillance system with chi-square test.

The second statistical analysis performed was to estimate and attempt to explain geographical variation in the odds of death over a discrete time period of 5-years using multilevel logistic regression[31]. A five-level model was fitted, with the individual as level 1, village as level 2, town as level 3, county as level 4, and province as level 5. An initial model included fixed effects for gender and age group which was divided by three subgroups. We then adjusted this model sequentially with variables describing education attainment, mean years of education, marital status and place of residence. Dummy variables for age, education attainment and marital status were added as fixed effects. The corresponding subcategories are shown in Table 3. All fixed effect parameters were expressed as odds ratios (OR) and 95% confidence intervals (95%CI). Geographic variation was expressed through the calculation of median odds ratios (MORs)[32]. Values of MORs equal to 1 mean no geographic variation in the outcome variable, whereas values above 1 indicate the necessity of taking context into account. After adding these variables the percentage change in variance (PCV) was calculated for different levels to illustrate that the proportion of the variation can be explained. All statistical analyses were conducted in MLwiN v3.01 (Centre for Multilevel Modelling, University of Bristol, Bristol, UK).

Results

A total of 3365 records were linked to the death surveillance data in 2011 to 2015 using two matching steps, indicating 3365 participants surveyed in 2010 had died by the end of 2015. There were 726 records matched by identification number and checked by identification number plus name, which only accounted for 21.6% of the matched records. The remaining 2639 records were matched based on checking by the two artificial matching strategies from the 47102 possible matched-records, which supplemented almost four-fifths of the matched records in the absence of an available identification number. The matching efficiency of the identification number (726/816) was much higher than that of the Gender + Name + Address + Age matching (2639/47102). The matched number of participants increased year by year in this cohort, as expected due to ageing (Table 1).

(See Table 1)

Matching results

All-cause mortality based on linked participants in the cohort was 6.15%, 6.62%, 6.81%, 7.30% and 7.57% in 2011, 2012, 2013, 2014 and 2015, respectively. These mortality rates were close to the corresponding data from the national death surveillance system, at 7.04%, 7.26%, 7.87%, 7.93% and 7.97%, respectively (Table 2). Based on a chi-square test, there was no statistically significant difference in mortality measured in the cohort as compared with the national death surveillance (X^2=0.004, P=0.947).

(See Table 2)

According to the matched records, NCDs mortality rates in 2011 to 2015 inclusive were 84.41%, 85.43%, 88.92%, 91.34%, 89.76%, respectively. This was consistent with the NCDs mortality reported in the national death surveillance, which were 85.51%, 86.03%, 86.84%, 87.29%, 87.70%, respectively (Table 2). A chi-square test indicated no statistically significant differences in proportion of different death categories between the cohort and the national death surveillance (X^2=0.196, P=0.978).

Multilevel analyses

Table 3 illustrates the results of multilevel logistic regression models of the odds of mortality from all-causes, infectious diseases, NCDs and injury separately. After full adjustment, all-cause mortality varied between provinces (MOR=1.320), between counties or districts (MOR=1.346), between towns (MOR=1.299) and between villages (MOR=1.296). Infectious disease mortality also varied geographically, with the most substantial found between villages (MOR=7.650). Geographic variations in NCDs mortality varied between provinces (MOR=1.292), between counties or districts (MOR=1.362), between towns (MOR=1.312) and between villages (MOR=1.240), to a similar magnitude as was observed with all-cause mortality. There was no geographic variation observed in injury mortality between provinces, counties or districts or villages, but there was evidence of variation between towns (MOR=1.499).

Odds Ratio (ORs) in the multilevel logistic regression indicate that all-cause mortality was significantly lower among females (OR=0.745, 95%CI 0.694, 0.799). Compared to the people aged 18 to 44 years old, the odds of all-cause mortality were higher in people aged 45 to 59 years old (OR=3.300, 95%CI 2.917, 3.734) and people aged older than 60 years old (OR=14.296, 95%CI 12.735, 16.049). After adding education attainment, mean years of education, marital status and place of residence into fixed part of the models, 23.42% of the province-level variation, 36.60% of the variation between counties or districts, 10.71% of the variation between towns and 7.50% of the variation between villages were explained.

All-cause mortality was lower in participants with primary education or above compared to those with no educational qualifications. However, there were no significant differences in all-cause mortality among people with different mean years of education. Compared to single people, the odds of all-cause mortality was higher in divorced people (OR=1.462, 95%CI 1.158, 1.846). People living in rural areas had higher all-cause mortality than people living in urban areas (OR=1.461, 95%CI 1.196, 1.642).

(See Table 3)

Both all-cause mortality and NCDs mortality were significantly higher among male, elderly, lower education attainment, divorced and living in rural area. Infectious disease was only associated with gender and age, namely male and elderly having higher mortality. Injury mortality was higher among male, elderly, lower education attainment and fewer years of education. Full details are available in additional file 1.

Discussion

In recent years, the Chinese government has paid significant attention to the control and prevention of NCDs and allocated resources for cohort study data collections to help enhance understandings of the scale of the challenges at hand[33]. Linkage of survey data to other forms of data, such as routinely collected mortality records, as was done in this study with the 2010 Chronic Disease and Risk Factor Surveillance, can be a low-cost means of strengthening the capacity for evidence-based decision-making.

In our study the results from the linked cohort and mortality data were consistent with the wider routinely collected mortality surveillance. This provided some assurance towards its reliability for future use[34], even while accounting for the multiple approaches used to perform the data linkage in the absence of an identification number for every person. Matching by a combination of gender, address and age helped in the absence of an identification number.

Convergence in mortality among people in the cohort and the routine mortality surveillance was observed in subsequent years after baseline. The initial difference may be attributable to subject bias wherein healthier people tend to participate in face-to-face surveys[35]. This convergence can be confirmed in following studies based on the two databases.

Further assurance of the data quality was provided by the results of the multilevel logistic regressions. These models revealed geographic variation and associations in the main cause of death with several social determinants. These results are in agreement with previous reports[36], confirming the utility of the linkage database. The multilevel logistic regression model afforded insights into geographic variations across multiple spatial scales. These models further demonstrate the importance of multilevel modelling of health data in large countries with varied populations, socioeconomics and topography, such as China. The results showed geographic variation in all-cause mortality and NCDs mortality varied among different area levels. These results are aligned with findings from previous studies[37–39]. Enhancing understandings of social determinants and inequities in all-cause and NCD-related mortality in China is critical to give a more complete, in-depth picture of the public health challenges that decision-makers face, while also providing data that can be used to evaluate specific policies and interventions.

To our knowledge, this is the first study to link the national chronic disease and risk factor surveillance with routinely collected cause-specific mortality data in China. This linked data will provide opportunities and possibilities for researchers and policy makers. There are, however, some limitations to acknowledge. First, the matching process of the detailed Chinese characters took a long time to implement and confirm. This work could be potentially conducted via machine learning to improve efficiency[40]. Second, the linked social data was cross-sectional and so common changes in social determinants over time in positive (e.g. educational attainment) and negative (e.g. job loss) directions could not be investigated in this particular data. Therefore, collection of longitudinal data and linkage to mortality records in future work is needed to build a more comprehensive understanding of the social determinants of mortality in China.

Conclusions

In this study, we linked 98 058 participants in the 2010 Chronic Disease and Risk Factor Surveillance to records in the national death surveillance data from 2011 to 2015. Program match, program check and artificial check were adopted to link all the records from both of the two data sources. Cross-checks and comparisons with national mortality distributions provided assurance that the linkage was reasonable. Multilevel logistic regression models revealed geographic variation and associations in the main cause of death with several social determinants. The results of this study provide utility for future investigations of social determinants of health and mortality using linked data in China.

Declarations

Ethics approval and consent to participate

The ethical review committee of Chinese Center for Disease Control and Prevention approved the 2010 CCDRFS and written informed consent was obtained from each participant before data collection.

The records in the national death surveillance data from 2011 to 2015 are obtained from the national mortality surveillance system. The information on individual deaths in all population catchment areas has been reported to the national mortality surveillance system according to the national guidelines since 1992.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the study will be made available by the corresponding author following a reasonable request.

Comflict of Interest

The authors declare that they have no conflict of interest.

Funding

This work was supported by The Australia-China Science and Research Fund[ACSRF17120].

Author's Contributions

Yunning Liu and Thomas Astell-Burt contributed to the conception and design, analysis and interpretation of data, drafted and wrote the paper. Xiaoqi Feng and Fan Mao contributed to analysis and interpretation of data and drafting of this paper. Ruiming Liang and Peng Yin contributed to analysis and interpretation of data. Limin Wang and Lijun Wang contributed to the acquisition of data. Maigeng Zhou contributed to the obtaining funding and supervision and critical revision of this paper for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Abbreviations

NCD: Non-communicable disease **CKB**: the China Kadoorie Biobank study

OR: Odds ratios **CI**: Confidence intervals **MOR**: Median odds ratios

PCV: Percentage change in variance **DSP**: Disease Surveillance Points

CCDRFS: China Chronic Disease and Risk Factor Surveillance

ID: Identification numbers **N**: Name **G**: Gender **NS**: Name spelling

PAC: Permanent address coding **RAC**: Residence address coding

PVA: Permanent villages address **RVA**: Residence villages address

PTA: Permanent town address **RTA**: Residence town address

AD: Age differences

References

1. Huang C, Yu H, Koplan JP: **Can China diminish its burden of non-communicable diseases and injuries by promoting health in its policies, practices, and incentives?** *The Lancet* 2014, **384**(9945):783-792.
2. Zhou M, Wang H, Zhu J, Chen W, Wang L, Liu S, Li Y, Wang L, Liu Y, Yin P *et al*: **Cause-specific mortality for 240 causes in China during 1990–2013: a systematic subnational analysis for the Global Burden of Disease Study 2013**. *The Lancet* 2016, **387**(10015):251-272.
3. Okayama A, Okuda N, Miura K, Okamura T, Hayakawa T, Akasaka H, Ohnishi H, Saitoh S, Arai Y, Kiyohara Y *et al*: **Dietary sodium-to-potassium ratio as a risk factor for stroke, cardiovascular disease and all-cause mortality in Japan: the NIPPON DATA80 cohort study**. *BMJ Open* 2016, **6**(7):e011632.
4. Vorster HH, Kruger A, Wentzel-Viljoen E, Kruger HS, Margetts BM: **Added sugar intake in South Africa: findings from the Adult Prospective Urban and Rural Epidemiology cohort study**. *The American journal of clinical nutrition* 2014, **99**(6):1479-1486.
5. Wakabayashi M, McKetin R, Banwell C, Yiengprugsawan V, Kelly M, Seubsman SA, Iso H, Sleigh A, Thai Cohort Study T: **Alcohol consumption patterns in Thailand and their relationship with non-communicable disease**. *BMC public health* 2015, **15**:1297.
6. Hallal PC, Clark VL, Assuncao MC, Araujo CL, Goncalves H, Menezes AM, Barros FC: **Socioeconomic trajectories from birth to adolescence and risk factors for noncommunicable disease: prospective analyses**. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 2012, **51**(6 Suppl):S32-37.
7. Khan FS, Lotia-Farrukh I, Khan AJ, Siddiqui ST, Sajun SZ, Malik AA, Burfat A, Arshad MH, Codlin AJ, Reininger BM *et al*: **The burden of non-communicable disease in transition communities in an Asian megacity: baseline findings from a cohort study in Karachi, Pakistan**. *PloS one* 2013, **8**(2):e56008.
8. Mirelman AJ, Rose S, Khan JAM, Ahmed S, Peters DH, Niessen LW, Trujillo AJ: **The relationship between non-communicable disease occurrence and poverty—evidence from demographic surveillance in Matlab, Bangladesh**. *Health Policy and Planning* 2016, **31**(6):785-792.
9. Van Minh H, Lan Huong D, Wall S, Byass P, Thi Kim Chuc N: **Cardiovascular Disease Mortality and Its Association With Socioeconomic Status: Findings From a Population-based Cohort Study in Rural Vietnam, 1999–2003**. *Preventing Chronic Disease* 2006, **3**(3):A89.
10. Liu J, Sekine M, Tatsuse T, Hamanishi S, Fujimura Y, Zheng X: **Family History of Hypertension and the Risk of Overweight in Japanese Children: Results From the Toyama Birth Cohort Study**. *Journal of Epidemiology* 2014, **24**(4):304-311.

11. Inohara T, Kohsaka S, Okamura T, Watanabe M, Nakamura Y, Higashiyama A, Kadota A, Okuda N, Ohkubo T, Miura K *et al*: **Long-Term Outcome of Healthy Participants with Atrial Premature Complex: A 15-Year Follow-Up of the NIPPON DATA 90 Cohort**. *PloS one* 2013, **8**(11):e80853.
12. Okamura T: **Dyslipidemia and Cardiovascular Disease: A Series of Epidemiologic Studies in Japanese Populations**. *Journal of Epidemiology* 2010, **20**(4):259-265.
13. Okamura T, Hayakawa T, Hozawa A, Kadowaki T, Murakami Y, Kita Y, Abbott RD, Okayama A, Ueshima H, Group NDR: **Lower levels of serum albumin and total cholesterol associated with decline in activities of daily living and excess mortality in a 12-year cohort study of elderly Japanese**. *Journal of the American Geriatrics Society* 2008, **56**(3):529-535.
14. Du H, Bennett D, Li L, Whitlock G, Guo Y, Collins R, Chen J, Bian Z, Hong LS, Feng S *et al*: **Physical activity and sedentary leisure time and their associations with BMI, waist circumference, and percentage body fat in 0.5 million adults: the China Kadoorie Biobank study**. *The American journal of clinical nutrition* 2013, **97**(3):487-496.
15. Ma L, Bai YN, Pu HQ, He J, Bassig BA, Dai M, Zhang YW, Zheng TZ, Cheng N: **A retrospective cohort mortality study in Jinchang, the largest nickel production enterprise in China**. *Biomed Environ Sci* 2014, **27**(7):567-571.
16. Qu HM, Bai YN, Cheng N, Dai M, Zheng TZ, Wang D, Li HY, Hu XB, Li JS, Ren XW *et al*: **Trend Analysis of Cancer Mortality in the Jinchang Cohort, China, 2001-2010**. *Biomedical and Environmental Sciences* 2015, **28**(5):364-369.
17. Yang L, Zhou M, Smith M, Yang G, Peto R, Wang J, Boreham J, Hu Y, Chen Z: **Body mass index and chronic obstructive pulmonary disease-related mortality: a nationally representative prospective study of 220,000 men in China**. *International journal of epidemiology* 2010, **39**(4):1027-1036.
18. Yin P, Brauer M, Cohen A, Burnett RT, Liu J, Liu Y, Liang R, Wang W, Qi J, Wang L *et al*: **Long-term Fine Particulate Matter Exposure and Nonaccidental and Cause-specific Mortality in a Large National Cohort of Chinese Men**. *Environmental Health Perspectives* 2017, **125**(11).
19. Jutte DP, Roos LL, Brownell MD: **Administrative record linkage as a tool for public health research**. *Annu Rev Public Health* 2011, **32**.
20. Holian J: **Live birth and infant death record linkage**. *J Health Soc Policy* 2000, **12**.
21. Holian J, Mallick MJ, Zaremba CM: **Maternity and infant care, race and birth outcomes**. *J Health Soc Policy* 2004, **18**.
22. Amin J, Law MG, Bartlett M, Kaldor JM, Dore GJ: **Causes of death after diagnosis of hepatitis B or hepatitis C infection: a large community-based linkage study**. *Lancet* 2006, **368**.
23. Holman CAJ, Bass AJ, Rouse IL, Hobbs MS: **Population-based linkage of health records in Western Australia: development of a health services research linked database**. *Aust N Z J Public Health* 2008, **23**.
24. Chen J, Fair M, Wilkins R, Cyr M: **Maternal education and fetal and infant mortality in Quebec. Fetal and Infant Mortality Study Group of the Canadian Perinatal Surveillance System**. *Health Reports/Statistics Canada, Canadian Centre for Health Information* 1998, **10**.
25. Chard T, Penney G, Chalmers J: **The risk of neonatal death in relation to birth weight and maternal hypertensive disease in infants born at 24–32 weeks**. *Eur J Obstet Gynecol Reprod Biol* 2001, **95**.
26. Li Y, Wang L, Jiang Y, Zhang M, Wang L: **Risk factors for noncommunicable chronic diseases in women in China: surveillance efforts**. *Bulletin of the World Health Organization* 2013, **91**(9):650-660.
27. Yang G, Hu J, Rao KQ, Ma J, Rao C, Lopez AD: **Mortality registration and surveillance in China: History, current situation and challenges**. *Population health metrics* 2005, **3**(1):3.
28. Zhou MG, Jiang Y, Huang ZJ, Wu F: **Adjustment and representativeness evaluation of national disease surveillance points system**. *Disease Surveillance* 2010, **13**(3):6295-6378.
29. Liu S, Wu X, Lopez AD, Wang L, Cai Y, Page A, Yin P, Liu Y, Li Y, Liu J *et al*: **An integrated national mortality surveillance system for death registration and mortality surveillance, China**. *Bulletin of the World Health Organization* 2016, **94**(1):46-57.
30. **International statistical classification of diseases and related health problems, 10th revision**. Geneva: World Health Organization; 1992.
31. Vaughn BK: **Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J**. *Journal of Educational Measurement* 2008, **45**(1):94-97.
32. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Råstam L, Larsen K: **A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena**. *Journal of Epidemiology and Community Health* 2006, **60**(4):290-297.
33. Wang H, Chen P, Zhang Z, Dong E: **The present situation, opportunity and challenge of cohort study in China**. *Chinese Journal of Preventive Medicine* 2014, **48**(11):1016-1021.
34. Shkolnikov VM, Jasilionis D, Andreev EM, Jdanov DA, Stankuniene V, Ambrozaitiene D: **Linked versus unlinked estimates of mortality and length of life by education and marital status: Evidence from the first record linkage study in Lithuania**. *Social Science & Medicine* 2007, **64**(7):1392-1406.
35. Megan B, Damien J, Vijaya S, Sue E, David P, Ian S, Caroline B: **Data Linkage: A powerful research tool with potential problems**. *BMC Health Serv Res* 2010, **10**.
36. Guilmoto CZ: **Mortality in China, India and Indonesia: an Overview**. In: *Contemporary Demographic Transformations in China, India and Indonesia.* edn.: Springer; 2016: 89-94.
37. Zhou M, Astell-Burt T, Yin P, Feng X, Page A, Liu Y, Liu J, Li Y, Liu S, Wang L *et al*: **Spatiotemporal variation in diabetes mortality in China: multilevel evidence from 2006 and 2012**. *BMC public health* 2015, **15**:633.

38. Liu Y, Astell-Burt T, Liu J, Yin P, Feng X, You J, Page A, Zhou M, Wang L: **Spatiotemporal variations in lung cancer mortality in China between 2006 and 2012: A multilevel analysis**. *International journal of environmental research and public health* 2016, **13**(12):1252.

39. Yin P, Feng X, Astell-Burt T, Qi F, Liu Y, Liu J, Page A, Wang L, Liu S, Wang L: **Spatiotemporal variations in chronic obstructive pulmonary disease mortality in China: multilevel evidence from 2006 to 2012**. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 2016, **13**(3):339-344.

40. Yao L, Liu Y, li X, Liu H: **Chinese named entity recognition via word boundary based character embedding**. *Chinese named entity recognition via word boundary based character embedding* 2016, **1**(1):37-42.

Tables

Table 1 The number of matching records with different matching strategies

Matching Strategy	2011		2012		2013		2014		2015		Total	
	N	%	N	%	N	%	N	%	N	%	N	%
ID + Name	15	2.49	123	19.07	138	20.94	206	28.77	244	32.88	726	21.58
Name/Name Spelling + Gender + Permanent/Residence Villages Address	360	59.70	316	48.99	320	48.56	300	41.90	292	39.35	1588	47.19
Name/Name Spelling + Gender + Permanent/Residence Town Address + Age Differences less than 5 years	228	37.81	206	31.94	201	30.50	210	29.33	206	27.76	1051	31.23
Total	603	100.00	645	100.00	659	100.00	716	100.00	742	100.00	3365	100.00

Table 2 Comparison of mortality and the proportion of different causes among 18 years and older in 2011-2015

Year	All cause		Infectious disease		NCDs		Injury		Others	
	Matched records	Death surveillance	Matched records	Death surveillance	Matched records	Death surveillance	Matched records	Death surveillance	Matched records	Death surveillance
Mortality（‰）										
2011	6.15	7.04	0.22	0.24	5.19	6.12	0.68	0.58	0.06	0.10
2012	6.62	7.26	0.30	0.26	5.65	6.33	0.60	0.57	0.07	0.10
2013	6.81	7.87	0.18	0.24	6.05	6.91	0.52	0.59	0.06	0.13
2014	7.30	7.93	0.15	0.25	6.67	7.00	0.46	0.58	0.02	0.10
2015	7.57	7.97	0.23	0.25	6.79	7.05	0.48	0.56	0.07	0.11
Proportion（%）										
2011	100.00	100.00	3.65	4.27	84.41	85.51	11.11	8.86	0.83	1.36
2012	100.00	100.00	4.50	4.27	85.43	86.03	8.99	8.42	1.08	1.28
2013	100.00	100.00	2.58	3.61	88.92	86.84	7.59	8.00	0.91	1.55
2014	100.00	100.00	2.09	3.63	91.34	87.29	6.28	7.67	0.29	1.41
2015	100.00	100.00	3.10	3.52	89.76	87.70	6.33	7.45	0.81	1.33

Table 3 Odds ratios and 95% confidence intervals for the mortality in different cause in the linked database

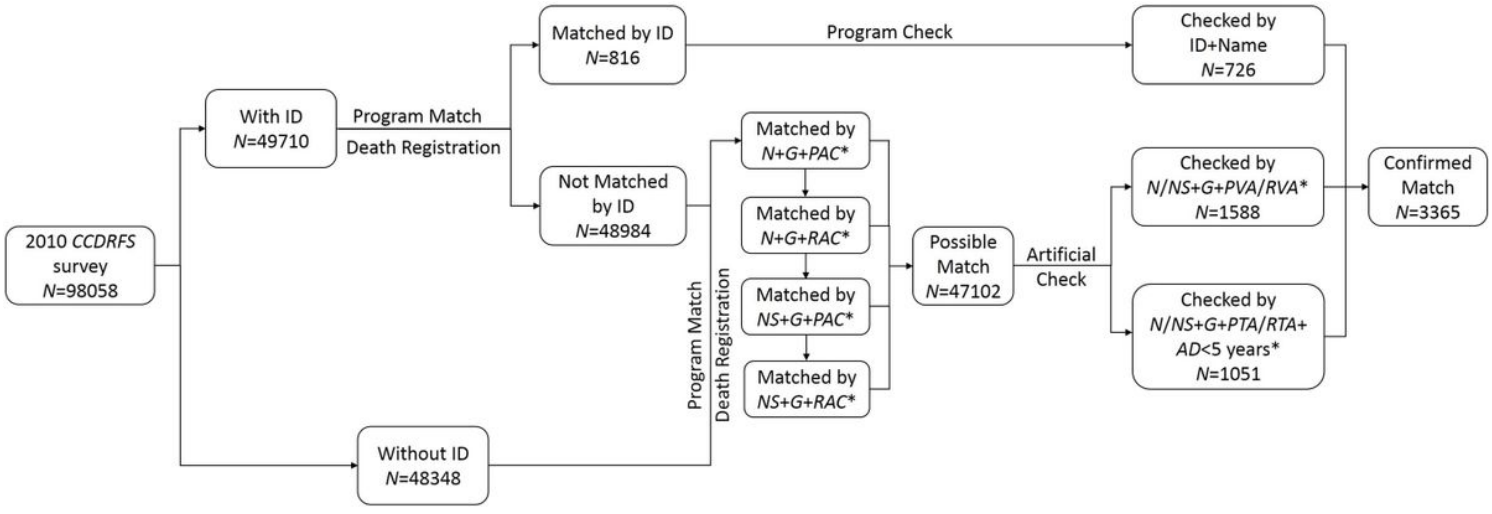
	All cause	Infectious disease	NCDs	Injury
Fixed effect	Incidence Odds Ratio (95% Confidence Interval)			
Gender (ref: male)				
female	0.660(0.602,0.723)*	0.578(0.364,0.918)*	0.700(0.636,0.771)*	0.345(0.260,0.458)*
Age(ref: 18-44)				
45-59	2.951(2.593,3.358)*	1.623(0.899,2.927)	3.600(3.096,4.187)*	1.418(1.042,1.928)*
60-	10.762(9.456,12.248)*	4.745(2.635,8.542)*	14.112(12.159,16.378)*	1.960(1.377,2.789)*
Education attainment				
(ref: none)				
Primary	0.780(0.710,0.856)*	0.908(0.534,1.544)	0.775(0.703,0.855)*	0.900(0.625,1.296)
Secondary	0.608(0.542,0.681)*	0.775(0.415,1.445)	0.580(0.514,0.653)*	0.890(0.599,1.323)
University	0.417(0.325,0.535)*	0.580(0.161,2.085)	0.417(0.319,0.545)*	0.411(0.176,0.9 61)*
Mean years of education	1.024(0.956,1.097)	0.803(0.603,1.069)	1.047(0.974,1.126)	0.813(0.713,0.927)*
Marital Status				
(ref: single)				
Married	1.051(0.849,1.302)	1.587(0.492,5.125)	1.065(0.835,1.358)	0.882(0.544,1.432)
Divorced	1.462(1.158,1.846)*	1.914(0.521,7.032)	1.465(1.129,1.902)*	1.146(0.598,2.196)
Widowed	1.229(0.919,1.642)	2.237(0.502,9.959)	1.138(0.822,1.575)	1.706(0.874,3.328)
Place of residence				
(ref: urban)				
Rural	1.461(1.196,1.784)*	0.759(0.346,1.662)	1.498(1.217,1.844)*	1.106(0.761,1.609)
Random effects				
Provinces				
Variance (standard error)	0.085(0.031)*	0.778(0.352)*	0.072(0.029)*	0(0)
MOR	1.320	2.320	1.292	1
PCV	23.42%	60.96%	7.69%	100.00%
Counties/Districts				
Variance (standard error)	0.097(0.022)*	0.706(0.371)	0.105(0.024)*	0(0)
MOR	1.346	2.229	1.362	1
PCV	36.60%	-3.67%	36.36%	-
Towns				
Variance (standard error)	0.075(0.022)*	0.519(0.608)	0.081(0.024)*	0.180(0.138)
MOR	1.299	1.988	1.312	1.499
PCV	10.71%	3.35%	11.96%	43.93%
Villages				
Variance (standard error)	0.074(0.027)*	4.550(0.896)*	0.051(0.028)	0(0)
MOR	1.296	7.650	1.240	1
PCV	7.50%	-37.01%	5.56%	-

* $p < 0.05$

MOR = Median Odds Ratio

PCV = proportional change in variance in the final model compared to the initial model.

Figures



* *N+G+PAC*: name, gender, permanent address coding; *N+G+RAC*: name, gender, residence address coding
NS+G+PAC: name spelling, gender, permanent address coding; *NS+G+RAC*: name spelling, gender, residence address coding
N/NS+G+PVA/RVA: name/name spelling, gender, permanent/residence villages address
N/NS+G+PTA/RTA+AD<5 years: name/name spelling, gender, permanent/residence town address, age differences less than 5 years

Figure 1

The work flow of the matching process