

Vital Surveillances

National Cancer Data Linkage Platform of China: Design, Methods, and Application

Hongmei Zeng^{1,*}; Yunning Liu^{2,*}; Lijun Wang²; Peng Yin²; Baohua Wang²; Ruiying Fu¹; Xianhui Ran¹; Rongshou Zheng¹; Siwei Zhang¹; Jiangmei Liu²; Jinling You²; Kexin Sun¹; Shaoming Wang¹; Li Li¹; Ru Chen¹; Wenqiang Wei¹; Maigeng Zhou²; Jing Wu^{2,†}; Jie He^{1,†}

ABSTRACT

Background: The National Cancer Center (NCC) and China CDC cooperatively designed a National Cancer Data Linkage (NCDL) Platform to fulfill the task of sharing cancer outcome data through an automatic web-based system.

Methods: NCC and China CDC established a web-based NCDL Platform to link death information from China CDC with the cancer database from NCC. Overall, 76,708 cancer patients' data were analyzed to assess the feasibility and match rate of the NCDL Platform for 7 major cancers.

Results: The function of the platform includes a data application and approval system, data linkage module, and results visualization system. Through the platform, 38.9% cases were identified as deaths cases from the NCDL Platform in the first 3 years after cancer diagnosis. The linkage rate was highest in liver cancer and lowest in breast cancer.

Conclusions: The NCDL Platform provides a powerful and efficient way to link national vital statistics with national cancer programs' data. Expanding cancer outcome data linkage may not only improve data collection efficiency, but also improve data use.

INTRODUCTION

Cancer outcome data are important indicators to assess the magnitude of the cancer burden as well as monitor the effects of programs on cancer control. The National Cancer Center (NCC) is the Chinese government's principal agency for national cancer control programs, which regularly collects cancer-related data. Under the responsibility of the China CDC, the China Cause of Death Reporting System (CDRS) regularly collects death registration data from each county of the country based on an internet-based reporting system, which forms the National Mortality

Database (1). Strengthening data exchange and maximizing data use through informatics between NCC and China CDC have become important tasks in the Healthy China Program 2019–2030 (2). To fulfill this task, NCC and China CDC cooperatively established a web-based National Cancer Data Linkage (NCDL) Platform to retrieve the vital status for cancer patients. To develop the NCDL Platform and determine its efficacy among cancer patients, we used a multicenter hospital-based cancer database from NCC to link with National Mortality Database from China CDC.

METHODS

NCDL Platform Development and Architecture

Under a cooperative framework from NCC and China CDC, we first signed an agreement between two national bureaus, which described stepwise implementation regarding data linkage and sharing. We developed two methods for data linkage: deterministic linkage using individual participant identification cards and probabilistic linkage using identifiable information if the patient lacks identification card (Figure 1A). We developed a unique access portal to the webserver controlled by firewalls. The system requires timely servicing and monitoring to ensure there are no cyber security vulnerabilities. Real-time logs auditing aims to ensure the security of data transmission between two bureaus (Figure 1B).

Data Sources

The National Mortality Database was from CDRS (3). The CDRS includes data from the Vital Registration System, representative Disease Surveillance Points System, the expanded provincial and county registration system, and the in-hospital death reports. All deaths were reported online through

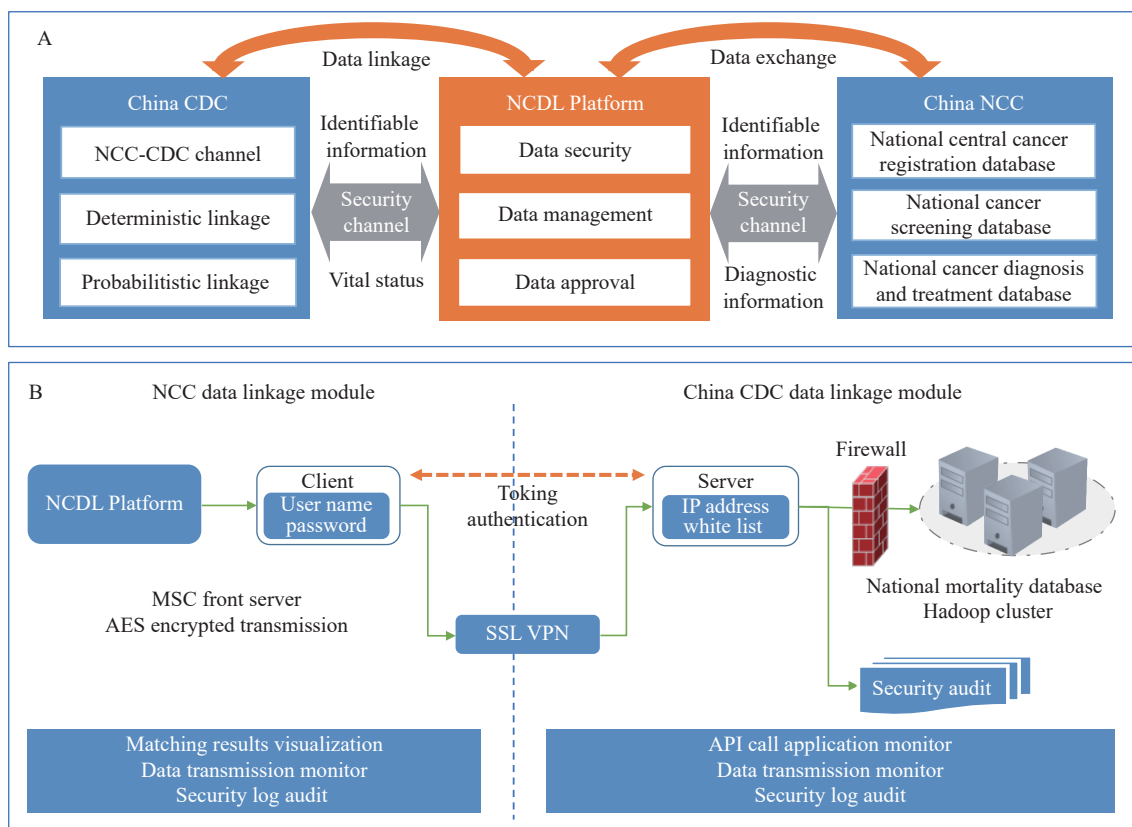


FIGURE 1. NCDL Platform architecture developed by NCC China and China CDC in 2021; (A) The framework of NCDL Platform; (B) Data security infrastructure of NCDL.

Abbreviations: NCDL=National Cancer Data Linkage; NCC=National Cancer Center.

China CDC's Death Information System with detailed information on the date of death and causes of death. To ensure data quality, CDC workers undertook routine data checks.

The multicenter hospital-based cancer database from NCC was used to test the feasibility of NCDL Platform, which included detailed, high-quality cancer data (4). We abstracted the information covering both urban and rural areas across six geographical regions of China. We identified all eligible cases diagnosed with first primary invasive cancer during 2016–2017 and whose home address was in the selected regions. We further linked the patients' information with the local population-based cancer registries, where registries' staff followed up the cancer patients by linking the local mortality surveillance system and/or actively contacting the patients or the next of kin to retrieve vital status (5–6).

Statistical Analysis

December 31, 2019 was used as the last date of contact in the study. The data match rate was calculated with the number of deaths identified by the

NCDL Platform divided by the corresponding number of cancer patients. We examined the match rate overall, by age at diagnosis, area of residence, and stage at diagnosis. We examined if the match rates were different in patients with different characteristics using chi-squared test. We analyzed all cancers combined and separately for each cancer type.

RESULTS

The function of the platform included three parts: a data application and approval system, data linkage module, and data visualization system. Through the platform, a multicenter hospital-based cancer database from NCC was successfully linked with National Mortality Database from China CDC securely and automatically.

Table 1 listed the selected characteristics for the linked dataset. A total of 76,708 cancer patients were included. With use of the NCDL Platform, 29,814 deaths were identified with an overall match rate of 38.9%. Patients with liver cancer had the highest match rate (56.1%), followed by lung cancer (50.0%),

TABLE 1. Baseline characteristics and results of the linked cancer dataset for patients diagnosed using National Cancer Data Linkage Platform, China, 2016–2017.

Items	All cancers	Lung	Stomach	Colorectum	Liver	Female breast	Esophagus	Ovary
No. of cases	76,708	22,820	12,807	11,338	6,519	11,975	9,471	1,778
Mean age at diagnosis (SD) (years)	61.4 (11.5)	63.0 (10.1)	63.6 (10.9)	63.2 (11.9)	58.2 (11.9)	53.5 (11.3)	66.1 (9.16)	55.6 (12.4)
Sex (%)								
Male	43,449/76,708 (56.6)	15,134/22,820 (66.3)	9,330/12,807 (72.9)	6,695/11,338 (59.0)	5,274/6,519 (80.9)	0/11,975 (0)	7,016/9,471 (74.1)	0/1,778 (0)
Female	33,259/76,708 (43.4)	7,686/22,820 (33.7)	3,477/12,807 (27.1)	4,643/11,338 (41.0)	1,245/6,519 (19.1)	11,975/11,975 (100)	2,455/9,471 (25.9)	1,778/1,778 (100)
Area (%)								
Urban	56,065/76,708 (73.1)	16,738/22,820 (73.3)	8,925/12,807 (69.7)	8,773/11,338 (77.4)	4,530/6,519 (69.5)	9,562/11,975 (79.8)	6,195/9,471 (65.4)	1,342/1,778 (75.5)
Rural	20,643/76,708 (26.9)	6,082/22,820 (26.7)	3,882/12,807 (30.3)	2,565/11,338 (22.6)	1,989/6,519 (30.5)	2,413/11,975 (20.2)	3,276/9,471 (34.6)	436/1,778 (24.5)
Total deaths (%)	29,814/76,708 (38.9)	11,411/22,820 (50.0)	5,458/12,807 (42.6)	3,041/11,338 (26.8)	3,656/6,519 (56.1)	1,016/11,975 (8.5)	4,632/9,471 (48.9)	600/1,778 (33.7)
Death from China CDC (%)	27,747/29,814 (93.1)	10,766/11,411 (94.3)	5,109/5,458 (93.6)	2,791/3,041 (91.8)	3,456/3,656 (94.5)	761/1,016 (74.9)	4,311/4,632 (93.1)	553/600 (92.2)
Death from cancer	24,691/27,747 (89.0)	9,473/10,766 (88.0)	4,571/5,109 (89.5)	2,489/2,791 (89.2)	3,086/3,456 (89.3)	692/761 (90.9)	3,881/4,311 (90.0)	499/553 (90.2)
Death from non-cancer	3,056/27,747 (11.0)	1,293/10,766 (12.0)	538/5,109 (10.5)	302/2,791 (10.8)	370/3,456 (10.7)	69/761 (9.1)	430/4,311 (10.0)	54/553 (9.8)
Death supplemented from NCC (%)	2,067/29,814 (6.9)	645/11,411 (5.7)	349/5,458 (6.4)	250/3,041 (8.2)	200/3,656 (5.5)	255/1,016 (25.1)	321/4,632 (6.9)	47/600 (7.8)

Abbreviation: NCC=National Cancer Center; SD=standard deviation.

esophageal cancer (48.9%), stomach cancer (42.6%), ovarian cancer (33.7%), colorectal cancer (26.8%), and breast cancer (8.5%). Because some registries actively tracked the patients' vital status, we tracked the vital status information from the hospital-based cancer database and added another 2,067 (6.9% of all death cases) deaths from the NCC database only.

Figure 2 showed the data match rates for cancer patients by sex, area, year of diagnosis and stage. We found the data match rates in patients who were 60 years and above were significantly higher than those who were less than 60 years (44.2% *vs.* 30.9%). Male patients generally had a higher match rate than females (47.8% *vs.* 27.2%). The match rate was higher in patients with stage III/IV than those with stage I/II (53.7% *vs.* 14.3%).

DISCUSSION

In the present study, we described the development and implementation of the NCDL Platform. This is the first nationwide cancer outcome data linkage system that enables a highly efficient data linkage and bilateral data sharing to the best of our knowledge.

Our study results demonstrated the feasibility of NCDL Platform as well as the advantages of data linkage and sharing. There is important public health significance of the NCDL Platform. First, through the complementation of the two systems, the data integrity of the cancer registration system and CDRS can be improved. Second, through the integration and linking of the two systems, indicators related to cancer outcomes such as mortality, survival time, and disease burden of cancer can be calculated more accurately.

The match rates revealed the proportion of death across cancers in different patients (5). The validated results were consistent with the intrinsic characteristics of the death surveillance data, such as cancer sites with poor prognosis, or poor prognosis with late cancer stage being more likely to get death outcome in a shorter period. The linked dataset from the NCDL Platform is a potentially valuable resource that allows for further cross-sectional and longitudinal studies. Given that NCC actively followed-up cancer patients through Cancer Registration and Follow-up Program, it may also provide a channel to improve data completeness of death registration through the NCDL Platform (3,7).

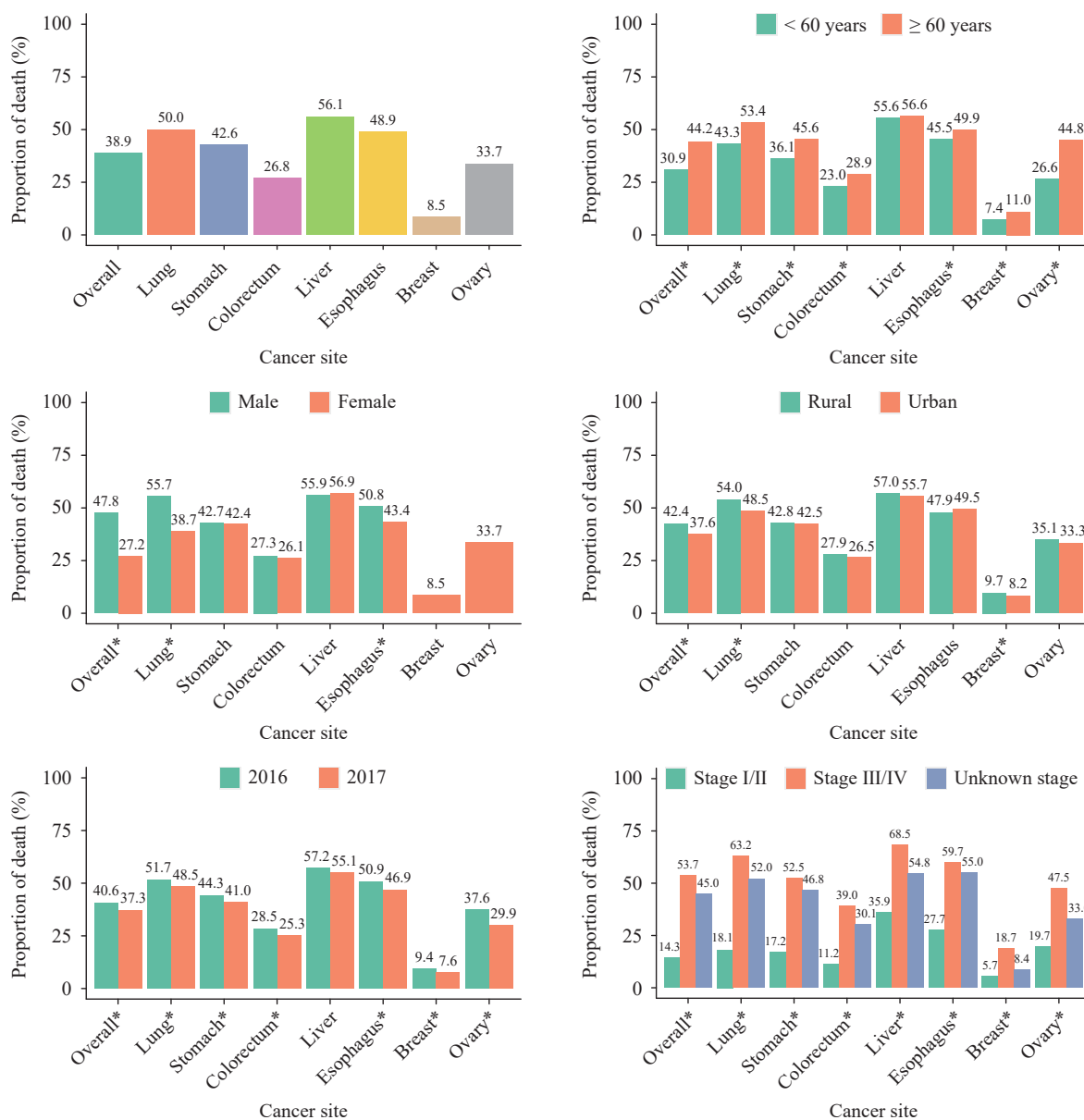


FIGURE 2. Data match rates (proportion of death) for cancer patients diagnosed during 2016–2017 and followed up to 2019 using NCDL Platform in China.

Abbreviations: NCDL=National Cancer Data Linkage.

* statistical significance between groups.

Automatic data linkage, data security and data confidentiality were among the highest priorities of the NCDL Platform design. The application of innovative informatics ensures the security of bilateral data transmission. Through the NCDL Platform, National Mortality Database and cancer control programs' database could be easily connected, which is more time-efficient for data exchange and sharing. Through this feasibility study, NCC and China CDC have established a standardized procedure for future data exchange.

Records linkage improves data completeness and quality. However, when unique identifiers are unavailable, successful record linkage cannot be assessed using deterministic linkage methods. The algorithm of probabilistic linkage is still under validation and optimization. Further research in this area will help to improve the successful data match rate. Considering the security issue, the NCDL Platform is not currently assessable to the public. We only issued institutional account with strict rules to ensure data transmission safety. The development and

fulfillment of the NCDL Platform had fulfilled the goal of efficient collection of cancer outcome data and maximized cancer data use between institutions.

In conclusion, the study demonstrated the feasibility of using NCDL Platform to bring together information on cancer diagnosis and treatment with information on vital status. Continued use of the NCDL platform will increase cancer outcome data collection efficiency and boost cancer data use.

Funding: Science and Technology Innovation 2030 Program (2020AAA0109500); The National Key Research and Development Program of China (2018YFC1311704).

doi: 10.46234/ccdcw2022.068

Corresponding authors: Jing Wu, wujing@chinacdc.cn; Jie He, hejie@cicams.ac.cn.

¹ National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; ² National Center for Chronic and Non-communicable Disease Control and Prevention, China CDC, Beijing, China.

⊗ Joint first authors.

Submitted: December 23, 2021; Accepted: March 30, 2022

REFERENCES

1. Liu SW, Wu XL, Lopez AD, Wang LJ, Cai Y, Page A, et al. An integrated national mortality surveillance system for death registration and mortality surveillance, China. *Bull World Health Organ* 2016;94(1):46 – 57. <http://dx.doi.org/10.2471/BLT.15.153148>.
2. Wei WQ, Zeng HM, Zheng RS, Zhang SW, An L, Chen R, et al. Cancer registration in China and its role in cancer prevention and control. *Lancet Oncol* 2020;21(7):e342 – 9. [http://dx.doi.org/10.1016/S1470-2045\(20\)30073-5](http://dx.doi.org/10.1016/S1470-2045(20)30073-5).
3. Zeng XY, Adair T, Wang LJ, Yin P, Qi JL, Liu YN, et al. Measuring the completeness of death registration in 2844 Chinese counties in 2018. *BMC Med* 2020;18(1):176. <http://dx.doi.org/10.1186/s12916-020-01632-8>.
4. Zeng HM, Ran XH, An L, Zheng RS, Zhang SW, Ji JS, et al. Disparities in stage at diagnosis for five common cancers in China: a multicentre, hospital-based, observational study. *Lancet Public Health* 2021;6(12):e877 – 87. [http://dx.doi.org/10.1016/S2468-2667\(21\)00157-2](http://dx.doi.org/10.1016/S2468-2667(21)00157-2).
5. Zeng HM, Chen WQ, Zheng RS, Zhang SW, Ji JS, Zou XN, et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. *Lancet Glob Health* 2018;6(5):e555 – 67. [http://dx.doi.org/10.1016/S2214-109X\(18\)30127-X](http://dx.doi.org/10.1016/S2214-109X(18)30127-X).
6. Zeng HM, Zheng RS, Guo YM, Zhang SW, Zou XN, Wang N, et al. Cancer survival in China, 2003-2005: a population-based study. *Int J Cancer* 2015;136(8):1921 – 30. <http://dx.doi.org/10.1002/ijc.29227>.
7. Wang L, Wang LJ, Cai Y, Ma LM, Zhou MG. Analysis of under-reporting of mortality surveillance from 2006 to 2008 in China. *Chin J Prev Med* 2011;45(12):1061 – 4. <http://dx.doi.org/10.3760/cma.j.issn.0253-9624.2011.12.002>. (In Chinese).