BMJ Open

Outcomes and prognosis of non-small cell lung cancer patients who underwent curable surgery: a protocol for a real-world, retrospective, population-based and nationwide Chinese National Lung Cancer Cohort (CNLCC) study

Xin Wang,[1] Yicheng Liang,[2] Yuanzhuo Wang,[3] Xiangzhi Meng,[2] Boxuan Zhou,[2] Zhenyi Xu,[4] Hui Wang,[5] Wenjing Yang,[5] Ning Li ,[1] Yushun Gao ,[2] Jie He,[2] on behalf of the CNLCC -study group

To cite: Wang X, Liang Y, Wang Y, *et al*. Outcomes and prognosis of non-small cell lung cancer patients who underwent curable surgery: a protocol for a real-world, retrospective, population-based and nationwide Chinese National Lung Cancer Cohort (CNLCC) study. *BMJ Open* 2023;**13**:e070188. doi:10.1136/bmjopen-2022-070188

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-070188).

XW, YL and YW contributed equally.

Received 16 November 2022
Accepted 30 May 2023

For numbered affiliations see end of article.

Correspondence to
Yushun Gao;
ysgaopumc@163.com and
Dr Jie He;
hejie@cicams.ac.cn

ABSTRACT

Introduction Surgery is one of the main approaches for the comprehensive treatment of early and locally advanced non-small cell lung cancer (NSCLC). This study conducts a nationwide multicentre study to explore factors that could influence the outcomes of patients with I–IIIA NSCLC who underwent curable surgery in real-world scenarios.

Methods and analysis All patients diagnosed with NSCLC between January 2013 and December 2020 will be identified from 30 large public medical services centres in mainland China. The algorithm of natural language processing and artificial intelligence techniques were used to extract data from electronic health records of enrolled patients who fulfil the inclusion criteria. Six categories of parameters are collected and stored from the electronic records, then the parameters will be structured as a high-quality structured case report form. The code book will be compiled and each parameter will be classified and designated a code. In addition, the study retrieves the survival status and causes of death of patients from the Chinese Centre for Disease Control and Prevention. The primary endpoints are overall survival and the secondary endpoint is disease-free survival. Finally, an online platform is formed for data queries and the original records will be stored as secure electronic documents.

Ethics and dissemination The study has been approved by the Ethical Committee of the Chinese Academy of Medical Sciences. Study findings will be disseminated via presentations at conferences and publications in open-access journals. This study has been registered in the Chinese Trial Register (ChiCTR2100052773) on 11 May 2021, http://www.chictr.org.cn/showproj.aspx?proj=136659.

Trial registration number ChiCTR2100052773.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The main strength of the study is the complete and accurate ascertainment and follow-up of evenly distributed stage I–IIIA non-small cell lung cancer (NSCLC) patients in mainland China, which can counteract the selection bias.

⇒ The wide coverage of patients and prospective and efficient design contribute to the establishment and maintenance of a pool of patients with NSCLC in China.

⇒ The sample size will be large enough to perform reliable survival and regression analyses in subgroups of patients.

⇒ The inconsistency and incompleteness of data sources could result in data discrepancy and disparities.

⇒ Blinding is not performed in the retrospective study and attrition may be a source of bias that could not be ignored.

INTRODUCTION

Lung cancer is the first most common and lethal cancer worldwide with 2.2 million new diagnosed cancer cases and 1.8 million deaths in 2020 across the globe.[1–3] In China, lung cancer is also a primary public health problem, which is the leading cause of cancer death and accounts for approximately a quarter of all cancer-related deaths,[4] of which remain a primary public health problem. As the main classification of lung cancer, non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancers.[5] Since 2013, lung cancer survival had improved primarily due to the promotion of new drug development (eg, targeted therapies and immunotherapy) in the past decade.[6] However, the improvement in prognosis of patients with II–IIIA (N1–N2) was limited. The current standard of treatment for patients with stage II–IIIA (N1–N2) NSCLC is surgery followed by adjuvant cisplatin-based chemotherapy (vinorelbine plus cisplatin), irrespective of epidermal growth factor receptor mutation status.[7–11] The 5year overall survival (OS) in these patients is poor, which estimated to be between 36% and 49%, with a median (m) survival time of 35.0–58.9 months, mostly because of the high recurrence rates (30%).[12–14] However, currently most studies evaluating prognosis of postoperative NSCLC patients in China were retrospective analyses with small-scale patients in single centre, which may not represent real-world settings.

Real-world analyses may identify subgroups of patients who responded to specific therapeutic regimens and contribute to the comprehension of disease staging and treatment choices.[15] This study intends to conduct a population-based, real-world nationwide multicentre study using original data from the Chinese National Lung Cancer Cohort (CNLCC), in combination with artificial intelligence (AI)-aided electronic medical records processing, and take the first step to retrospectively analyse the outcome and risk factors of patients with NSCLC who underwent curable surgery in China.

AIMS

Primary outcomes

1. To analyse the postoperative OS and disease-free survival (DFS) in real-world Chinese NSCLC patients who underwent curable surgery.
2. To explore the factors that estimate the prognosis of the patients after surgery, including demographic characteristics, baseline health status, tumour disease characteristics, clinical diagnosis, and treatment information. Then, the risk prediction models based on these factors will be established and validated.
3. To describe recurrence rate and outcomes in Chinese NSCLC patients after surgery.

Explorative outcomes

1. To identify the national and regional trends of the comprehensive demographic data in Chinese NSCLC patients over the years.
2. To describe the characteristics of genotype-guided precision medicine (proportion of genetic testing, distribution of genetic testing methods, results of genetic testing, and timing of testing) in Chinese NSCLC patients.
3. To estimate the changes and disparities in total costs as a function of changes in practice patterns for the treatment of NSCLC over years.

METHODS

The CNLCC retrospective cohort is established from patient registries at 30 large public medical services centres in 19 province-level administrative units of mainland China, which cover over 60% of Chinese people. The programme is initialised and maintained by the China National Cancer Centre and the Cancer Institute and Hospital, Chinese Academy of Medical Sciences (CAMS). A multicentre CNLCC-collaborative committee has been established and continuously provided consistently updated national and regional data on patients with lung cancers. The collaborative includes institution leaders, senior thoracic surgeons, oncologists, local principal investigators, pathologists and biostatisticians.

Study design and patient enrolment

This study is a population-based, real-world nationwide multicentre study using original data from postsurgery patients with NSCLC in CNLCC. The initial settled study period is from 1 January 2013 to 31 December 2020. The study period will be expanded every year to keep the cohort updated on the most recent data. The study is preliminarily settled to cease 5 years after the last patient registration.

The study population of NSCLC patients undergoing surgery in this study is obtained from the National Cancer Centre (NCC) Tumour Information Database, covering patients with stage I–IIIA primary NSCLC who undergo surgical resection from January 2013 to December 2020 in 30 tertiary hospitals in 19 provinces (figure 1). Figure 1 is jointly drawn by package ‘ggplot2’, ‘plyr’, ‘maptools’, ‘sp’, ‘cairo’, ‘RColorBrewer’, ‘openxlsx’ and ‘rgdal’ of R programme (V.4.2.2). Considering the average number of patients registered at the CNLCC per year, a total of over 10000 eligible patients will be enrolled in this study during the study period, after conducting the inclusion criteria.[16 17]

Baseline and endpoints

The baseline of follow-up for involved patients is defined as the date of surgery. The primary endpoint is OS after resection, defined as the time from baseline to death (from any cause) or last follow-up (whichever occurred first). The secondary endpoint is DFS, defined as the time between the patient’s baseline and the time of disease recurrence or death (from any cause) or the last follow-up (whichever occurs first).

Inclusion criteria and exclusion criteria

The patients fulfilling the following inclusion criteria are included:

1. Age of at least 18 years old.
2. Patients diagnosed with primary stage I–IIIA NSCLC with pathological confirmation (classified according to the American Joint Committee on Cancer eighth staging system for NSCLC)[18] who had undergone complete resection.
3. Patients had complete information and routine follow-up.
4. Patients without history of other malignancy within 5years.

Follow-up

The study period is set from the baseline to 5 years after the last patient registration. All patients who undergo

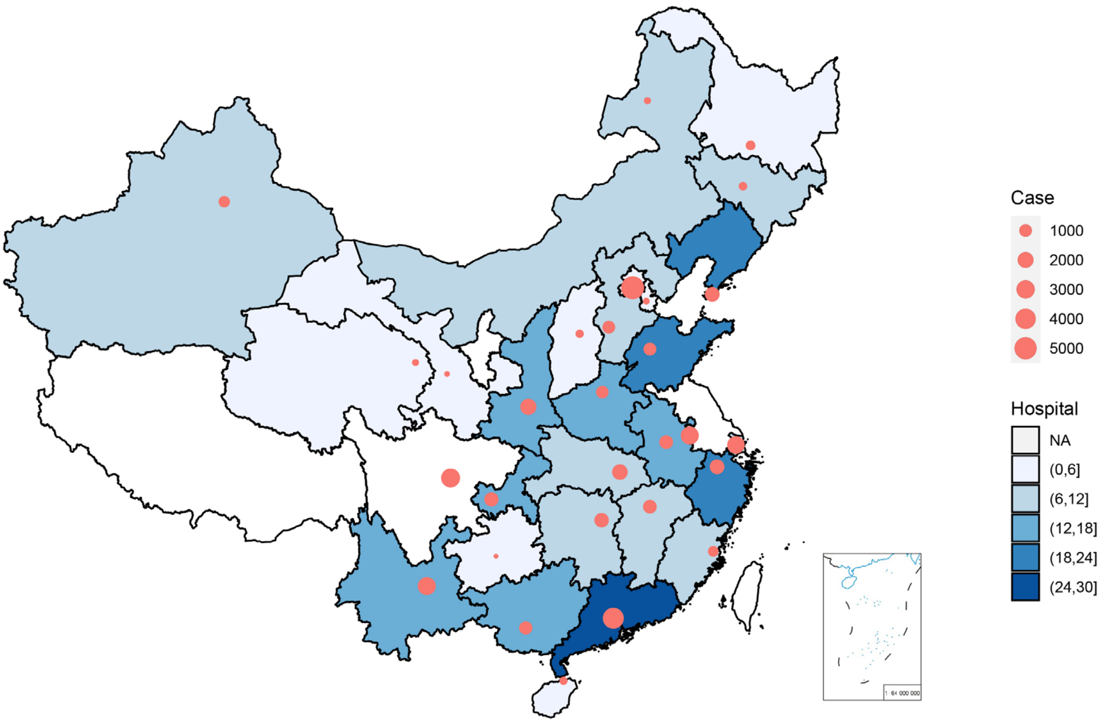

Figure 1 Map of distributions of 30 tertiary hospitals in 19 provinces. NA, not applicable.

surgical treatment will be asked to enter the follow-up schedule of the study site. In general, the study aims to make a follow-up for at least 5 years and continues as usual for early detection of disease progression or metastasis indefinitely, if requested by the patients. Routine follow-up and treatment are asked for all committee members to be in line with the Chinese Society of Clinical Oncology (CSCO) and the European Society for Medical Oncology (ESMO) Clinical Practice Guidelines.[19]

The follow-up mainly relies on the evaluation when patients come back to clinics or are readmission. The procedures include findings on clinical examination, results of laboratory investigations and imaging evaluation, to update the disease status and the tumour information. Basic examinations include the tumour markers (eg, the squamous cell carcinoma antigen and cytokeratin 19 fragment (CYFRA21-1), CT at 3 month or 6 month intervals are routinely evaluated in the first-year postoperation. Further intervals will refer to the guidelines, patient conditions and the opinion of the attending doctor. Bone scintigraphy, brain CT or MRI, and positron emission tomography-CT examinations are performed on patients in all necessary situations.

Data collection and document

A standard data collection and quality control procedure are designed before conducting the study. Six categories of parameters are collected and stored from the electronic records of the patients, to facilitate further AI-aided management. The parameters will be structured as a high-quality structured case report form (CRF), aiming to describe the features of the enrolled patients comprehensively, answer the clinical questions and reach the study objectives. The code book will be compiled and each parameter will be classified and designated a code, convenience to indexing and screening. In addition, the study retrieves the survival status and causes of death of patients from the China Centres for Disease Control and Prevention (CDC), an organisation dedicated to addressing diseases such as cancer and infectious diseases and public health issues.[20] Previously, NCC and CDC established a cooperation framework, namely National Cancer Data Linkage (NCDL) Platform of China, which was described in the previous study.[21]In short, we developed two methods of data linking: deterministic linking using individual participant ID, and probabilistic linking using identifiable information in the absence of patient ID information. Then, we developed a unique web server access portal controlled by a firewall, which needs to be maintained and monitored in a timely manner to ensure that there are no network security vulnerabilities. Using the NCDL platform, the data matching rate of lung cancer patients reached 50.0%.[21] The detail of the parameters is shown in table 1.

Study records with identifiable numbers of patients will be stored as secure electronic documents in the Datacentre of the Cancer Institute and Hospital, CAMS and retained in line with the CAMS Data Confidential Regulations. A professional confidentiality consent is asked to be signed by all committee members in patient enrolment, data collection and data processing. All cooperation units must be approved by the collaborative committee supervisors before accessing the data. For data sharing, anonymised electronic data will be reorganised and made public to enable secondary research, educational use and

Table 1 Collected data throughout the Chinese National Lung Cancer Cohort (CNLCC) study

Parameters	Details
Sociodemographic characteristics	Factors included sex, date of birth; educational level, identified number; registered permanent residence; age at the first diagnosis; date at operation; hospital name and hospital class where the operation is performed; body mass index (BMI); treatment of costs and inpatient length of hospital stay health-related quality of life from the perspective of patients.
Baseline Health Status	Eastern Cooperative Oncology Group performance status (ECOG PS);[27] physical examination information; blood routine; blood clotting; liver function; renal function; etiological examinations; important previous medical history; operation history; concurrent conditions.
Tumour Disease Characteristics	TNM classification of stages, clinical grades, CT or other imaging reports; tumour diameter, numbers, location and other information; histological type; morphology and pathology characteristics; invasiveness and lymph node metastases; tumour markers.
Treatment information	Type and timing of surgery; surgical procedures; margins of tumour resection; adverse events; preoperative and postoperative neoadjuvant/adjuvant therapy (specifically including the existence of neoadjuvant/adjuvant therapy, drug class, timing, cycles, start-end interval and additional detailed data).
Genetic testing variables	Cancer-associated genotyping results; testing methods; timing and variation of results during preoperative, perioperative, postoperative and after disease advance or recurrence (if exist).
Disease progression, recurrence and survival information.	Time of disease progress or recurrence; tumour recurrence sequence; corresponding recurrence site; survival status and death time; cause of death.

TNM, Tumor-Node-Metastasis.

scholarly communication after passing the data masking period.

Data cleansing and data management

As stated above, the basic data are collected from patients' Electronic Health Record (EHR) of the registered hospital. The clinical data in each centre were main exported from the Electronic Medical Record (EMR), Hospital Information System (HIS), Laboratory Information Management System (LIS), Picture Archiving and Communication System (PACS) and Pathology Information System (PIS). Then, it will be encrypted, and divided into structured data, such as laboratory values, or unstructured data from photographs, charts or paragraphs (eg, smoking or drinking history, family history of cancer). The data will be transferred to the National Cancer Institute via the virtual private network tunnel which is highly secured.

With the help of the senior computer experts in the committee, the algorithm of natural language processing (NLP) will cooperate with other AI techniques to transform unstructured data into structured variables to facilitate further analysis. This is enabled by text mining on keywords and computer identification. Regardless of the source or category, all data will not be docked into the analysis database unless they undergo data cleansing and standardising. Data from the CDC death databases will be linked to the analysable database by research identify number to form a generic research database, as they had already been screened and validated. The detail of the data cleansing is shown in figure 2, which was generated by Adobe Illustrator software.

An online platform available to users for public data sharing is being developed. The EHR data will be provided by an application programming interface, to enable structured search using the search engine by users. During the development phase, the summarised data are routinely reported back to the committee to facilitate feedback from experts and subsequent users. In addition, the EHR context will be shown to enable manual validation of results for users during the phase. Specific codes for each parameter are also applicated in the system to index and increase search speed. Those codes will be classified into six categories, which had one key parameter of the patient identification number to connect them as an ensemble. The advantages of the decentralised operation are increasing the speed of queries, facilitating comparisons, and decreasing the data storage spaces.

Quality assurance

Data quality assurance is mainly guaranteed in the data collecting and cleansing process. To create a gold standard, manual chart review will be performed by a specialist in a local hospital centre, who is experienced in dealing with the EHR. The original manually retracted data selected randomly will run a comparative test with the data extracted by the computer programmes. The consistency of data will be verified through random sampling at the National Cancer Institute as well, which examined the data quality after data cleansing. A discrepancy lower than 10% is considered acceptable for both the data collecting and cleansing process. This is in line with the thresholds set by Hernandez Boussard *et al*.[22]

Another concern for the data quality is the missing values. For each patient, when the CRF has generated automatically from parameters provided by programmes, the quality of the data will be evaluated, and CRF with more than 10% of selected values missing will not be

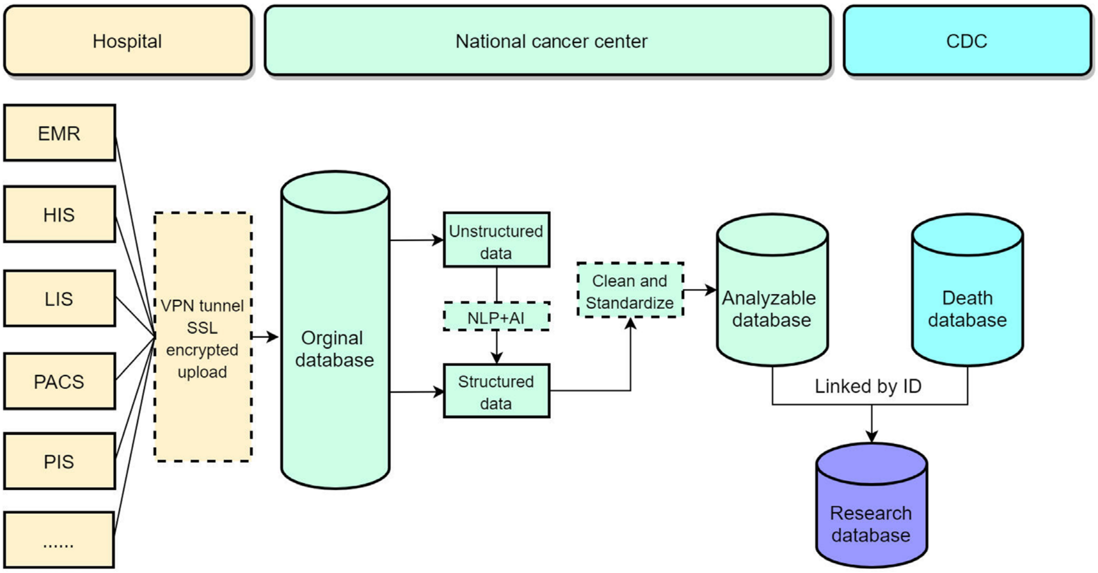

Figure 2 Flow chart of the study procedure. AI, artificial intelligence; CDC, Centre for Disease Control and Prevention; EMR, Electronic Medical Record; HIS, Hospital Informations System; LIS, Laboratory Information Management System; NLP, natural language processing; PACS, Picture Archiving and Communication System; VPN, virtual private network.

included for analysis. For those variables with less than 10% of values missing, the median values are estimated from the imputation of previous and subsequent values of respective variables.

Statistical analysis plan

Data management and analysis in this study will be conducted under the guidance of biostatisticians (PO). After data collection, a cohort profile will be published. For the cohort profile, the number of new cancer cases per year and the number of cancer cases operated per year will be calculated from the registry data. The baseline characteristics of cohort will be reported in the table. All-cause mortality will be reported according to the life table method and described using Kaplan-Meier curves. Each substudies will be planned in detail, including a prior study protocol, describing the rationale, aims, hypothesis and statistical analysis plan, for example, how to reduce potential confounding and bias for specific research questions.

Patient and public engagement

The development of the CNLCC study involved representatives from institution leaders, senior thoracic surgeons, oncologists, local principal investigators, pathologists, biostatisticians and most importantly, the patients. The representatives are grouped and continuously provided constructive feedback on the research design, study protocol, involved population and data analysis in the whole process. The committee ensures that the patient and public are actively interviewed and involved through activities of conferences, forums, lectures and social media dissemination.

Ethics and dissemination

Results of the study will be disseminated publicly through annual reports on a study-specific website, presentations at conferences and publications in open-access journals. The final reports will be turned into policy briefs for decision-making by the relevant government departments responsible for cancer prevention, control and treatment. An educational campaign for patients will be held once a year using the results of the study. Other stakeholders could be targeted by presentations on traditional social media, mobile phone applications, lectures and open-access publications.

DISCUSSION

Currently, there are some population-based nationwide multicentre lung cancer databases in other countries. For example, The Society of Thoracic Surgeons General Thoracic Surgery Database (STS GTSD) is the largest and most robust thoracic surgical database in the world. Since its inception in 2002, the GTSD has provided multiple mechanisms for high-quality clinical research using data from 274 participating sites and 781 000 procedures.[23] Similarly, the lung cancer database project established in 1999 at the NCC Hospital East, Japan also constructed a large-scale cancer registry for lung cancer and integrated data on various factors.[24] However, this is the first study that intends to conduct a population-based, real-world nationwide multicentre study to report the long-term survival outcome with resected stages I–IIIA NSCLC in China. The results will complement and improve the current understanding of the real status of prognosis in patients who underwent curable surgery in China.

An advantage of the study is to explore the feasibility of collecting data with NLP and AI-aided techniques to obtain main treatment outcomes, such as DFS and OS. If EHR can be adequately retrieved and analysed in real-world scenarios, more effective data could be obtained from a massively medical environment.[25] The AI-aided technique could be a technical solution for more consistent and timely data collection.[26] The strength of the CNLCC includes the complete and accurate ascertainment and follow-up of average distributed patients which can counteract selection bias. Moreover, the wide coverage of patients and prospective and efficient design contribute to the establishment and maintenance of a pool of patients with NSCLC in China. It will help as a platform to support further research.

Several potential limitations should be noted. First, the inconsistency and incompleteness of data sources could result in data discrepancy and disparities. Second, potential confounding could come from the referral bias, which may attribute to different medical diagnosis levels across different institutions. In the meantime, blinding is not performed in the retrospective study and attrition may be a source of bias that could not be ignored.

In conclusion, this population-based, nationwide retrospective study will combine epidemiological and clinical data to provide new evidence on a variety of outstanding questions about early and locally advanced NSCLC after surgical treatment.

Current status of the study

Data collection commenced at the end of December 2021 and is ongoing.

Ethics approval and consent to participate

This study has been registered on 11 May 2021, http://www.chictr.org.cn/showproj.aspx?proj=136659. The study has been approved by the medical ethics committee (21/431-3102).

Author affiliations

[1]Clinical Trial Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital Chinese Academy of Medical Sciences and Peking Union Medical College, Chaoyang, China

[2]Department of Thoracic surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Chaoyang, China

[3]School of Basic Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[4]Department of Epidemiology and Biostatistics, Harbin Medical University, Harbin, China

[5]Office for Cancer Diagnosis and Treatment Quality Control, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, People's Republic of China

Contributors JH and YG developed the study idea, concept and design, initiated the collaborative project, developed data collection tools, obtained permissions and funding and drafted the study protocol. XW, YCL and YW were involved in the statistical design, made substantial revisions to the protocol and were involved in the original concept, study design and implementation. JH, YG, XM, BZ, ZX, HW, WY and NL was involved in design and implementation of the study, and revised the protocol. JH and YG is the guarantor.

Funding This work was supported by National Key R&D Program of China (2020AAA0109500), National Natural Science Foundation of China (82188102), the Beijing Municipal Science & Technology Commission (Z191100006619115), R&D Program of Beijing Municipal Education Commission (KJZD20191002302).

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs

Ning Li http://orcid.org/0000-0002-3945-2536
Yushun Gao http://orcid.org/0000-0003-3992-6671

REFERENCES

1 Sung H, Ferlay J, Siegel RL, *et al*. Global cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;71:209–49.
2 Miller KD, Nogueira L, Mariotto AB, *et al*. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin* 2019;69:363–85.
3 Siegel RL, Miller KD, Fuchs HE, *et al*. Cancer statistics, 2022. *CA Cancer J Clin* 2022;72:7–33.
4 Cao W, Chen H-D, Yu Y-W, *et al*. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer Statistics 2020. *Chin Med J (Engl)* 2021;134:783–91.
5 Thai AA, Solomon BJ, Sequist LV, *et al*. Lung cancer. *Lancet* 2021;398:535–54.
6 Hirsch FR, Scagliotti GV, Mulshine JL, *et al*. Lung cancer: Current therapies and new targeted treatments. *Lancet* 2017;389:299–311.
7 Pisters KMW, Evans WK, Azzoli CG, *et al*. Cancer care Ontario and American society of clinical oncology adjuvant chemotherapy and adjuvant radiation therapy for stages I-IIIA Resectable non small-cell lung cancer guideline. *J Clin Oncol* 2007;25:5506–18.
8 Zhi XY, Yu JM, Shi YK. Chinese guidelines on the diagnosis and treatment of primary lung cancer (2015 version). *Cancer* 2015;121 Suppl 17:3165–81.
9 Eberhardt WEE, De Ruysscher D, Weder W, *et al*. 2nd ESMO consensus conference in lung cancer: locally advanced stage III non-small-cell lung cancer. *Ann Oncol* 2015;26:1573–88.
10 Ettinger DS, Wood DE, Aisner DL, *et al*. NCCN guidelines insights: non-small cell lung cancer, version 2.2021. *J Natl Compr Canc Netw* 2021;19:254–66.
11 Kris MG, Gaspar LE, Chaft JE, *et al*. Adjuvant systemic therapy and adjuvant radiation therapy for stage I to IIIA completely Resected non-small-cell lung cancers: American society of clinical oncology/cancer care Ontario clinical practice guideline update. *J Clin Oncol* 2017;35:2960–74.
12 Goldstraw P, Chansky K, Crowley J, *et al*. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth). *J Thorac Oncol* 2016;11:39–51.
13 Colt HG, Murgu SD, Korst RJ, *et al*. Follow-up and surveillance of the patient with lung cancer after curative-intent therapy: diagnosis and management of lung cancer, 3RD Ed: American college of chest physicians evidence-based clinical practice guidelines. *Chest* 2013;143:e454S.
14 Deng XF, Jiang L, Liu QX, *et al*. Lymph node Micrometastases are associated with disease recurrence and poor survival for early-stage non-small cell lung cancer patients: a meta-analysis. *J Cardiothorac Surg* 2016;11:28.
15 Waterhouse D, Lam J, Betts KA, *et al*. Real-world outcomes of Immunotherapy-based regimens in first-line advanced non-small cell lung cancer. *Lung Cancer* 2021;156:41–9.

16 Chen W, Zheng R, Baade PD, *et al*. Cancer statistics in China, 2015. *CA Cancer J Clin* 2016;66:115–32.

17 Zheng R, Zhang S, Zeng H, *et al*. Cancer incidence and mortality in China, 2016. *J National Cancer Center* 2022;2:1–9.

18 Detterbeck FC, Chansky K, Groome P, *et al*. The IASLC lung cancer staging project: methodology and validation used in the development of proposals for revision of the stage classification of NSCLC in the forthcoming (eighth). *J Thorac Oncol* 2016;11:1433–46.

19 Wu Y-L, Planchard D, Lu S, *et al*. Pan-Asian adapted clinical practice guidelines for the management of patients with metastatic non-small-cell lung cancer: a CSCO-ESMO initiative endorsed by JSMO, KSMO, MOS, SSO and TOS. *Ann Oncol* 2019;30:171–210.

20 Walgate R. China SETS up centres for disease control and prevention. *Bull World Health Organ* 2002;80:335.

21 Zeng H, Liu Y, Wang L, *et al*. National cancer data linkage platform of China: design, methods, and application. *China CDC Wkly* 2022;4:271–5.

22 Hernandez-Boussard T, Monda KL, Crespo BC, *et al*. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *J Am Med Inform Assoc* 2019;26:1189–94.

23 Servais EL, Blasberg JD, Brown LM, *et al*. The society of Thoracic Surgeons General Thoracic surgery database: 2022 update on outcomes and research. *Ann Thorac Surg* 2023;115:43–9.

24 Nakaya N, Goto K, Saito-Nakaya K, *et al*. The lung cancer database project at the National cancer center, Japan: study design, corresponding rate and profiles of cohort. *Jpn J Clin Oncol* 2006;36:280–4.

25 van Laar SA, Gombert-Handoko KB, Guchelaar H-J, *et al*. An electronic health record text mining tool to collect real-world drug treatment outcomes: A validation study in patients with metastatic renal cell carcinoma. *Clin Pharmacol Ther* 2020;108:644–52.

26 Cave A, Kurz X, Arlett P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin Pharmacol Ther* 2019;106:36–9.

27 Oken MM, Creech RH, Tormey DC, *et al*. Toxicity and response criteria of the Eastern cooperative oncology group. *Am J Clin Oncol* 1982;5:649–55.