BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers’ comments and the authors’ responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open’s open peer review process please email info.bmjopen@bmj.com
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

<table>
<thead>
<tr>
<th>Journal</th>
<th>BMJ Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>bmjopen-2022-070188</td>
</tr>
<tr>
<td>Article Type</td>
<td>Protocol</td>
</tr>
<tr>
<td>Date Submitted by the Author</td>
<td>16-Nov-2022</td>
</tr>
</tbody>
</table>

Complete List of Authors:

- Wang, Xin; Cancer Hospital Chinese Academy of Medical Sciences, Clinical Trial Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
- Liang, Yi cheng; Cancer Hospital Chinese Academy of Medical Sciences, Thoracic surgery department, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
- Wang, Yuanzhuo; Chinese Academy of Medical Sciences & Peking Union Medical College, School of Basic Medicine
- Meng, Xiangzhi; Cancer Hospital Chinese Academy of Medical Sciences, 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
- Zhou, Boxuan; Cancer Hospital Chinese Academy of Medical Sciences, 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
- Xu, Zhenyi; Harbin Medical University, Department of Epidemiology and Biostatistics
- Wang, Hui; Cancer Hospital Chinese Academy of Medical Sciences
- Yang, Wenjing; Cancer Hospital Chinese Academy of Medical Sciences
- Li, Ning; Cancer Hospital Chinese Academy of Medical Sciences, Clinical Trial Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
- Gao, Yushun; Chinese Academy of Medical Sciences & Peking Union Medical College, 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
- He, Jie; Cancer Hospital Chinese Academy of Medical Sciences, 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

Keywords:

Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Epidemiology < ONCOLOGY, Thoracic surgery < SURGERY
Title page

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

Running title
Chinese Lung Cancer Cohort: a protocol for a real-world study

Xin Wang 1†, Yi cheng Liang 2†, Yuanzhuo Wang 3†, XiangZhi Meng 2, Boxuan Zhou 3, Zhenyi Xu 4, Hui Wang 3, Wenjing Yang 3, Ning Li 1, Yushun Gao 1*, Jie He 2*

Group name on behalf of the CNLCC -study group

Affiliations:
1. Department of Clinical Trials Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 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, Beijing, China.
3. Office for Cancer Diagnosis and Treatment Quality Control, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.
4. Department of Epidemiology and Biostatistics, School of Public Health, Harbin Medical University, Harbin, China.
5. Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing 100005, China.

†These authors contributed equally.

*Corresponding authors:
Prof Jie He, 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, 17 Pan-jia-yuan South Lane, Chaoyang District, 100021, Beijing, China; email: hejie@cicams.ac.cn
Prof Yushun Gao, 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, Beijing, China, 17 Pan-jia-yuan South Lane, Chaoyang District, 100021, Beijing, China; email: ysgaopumc@163.com
Acknowledgements
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).

Data Statement
The dataset will be available from https://www.natdss.cn/#/home

word count: 2455 words

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 multicenter study to explore factors that could influence the outcomes of patients with IA-IIIA NSCLC who underwent curable surgery in real-world scenarios.

Methods and analysis
All patients diagnosed with NSCLC between January 2013 to 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 routine mortality of patients from the Chinese Center 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 11–5-2021, http://www.chictr.org.cn/showproj.aspx?proj=136659.

Keywords: non-small cell lung cancer; nationwide multicenter study; prognosis; online platform

Article Summary
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 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 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, 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 (for example, targeted therapies and immunotherapy) in the past decade6. 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 [VP]), irrespective of epidermal growth factor receptor (EGFR) mutation status.7-11 The 5-year overall survival (OS) rate 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 focus on evaluating prognosis of postoperative NSCLC patients in China were retrospective analyzes with small-scale patients in single center, 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 multi-centre study using original data from the Chinese National Lung Cancer Cohort (CNLCC), in combination with AI-aided electronic medical records processing, and take the first step to retrospectively analyze the outcome and risk factors of patients with NSCLC who underwent curable surgery in China.

Aims
Primary outcomes
1. To analyze the postoperative OS rate and DFS period in real-world Chinese NSCLC patients who underwent curable surgery.
2. To establish and validate the risk prediction models and the factors that estimate the prognosis of the patients after surgery, including demographic characteristics, baseline
3. To describe recurrence rate and outcomes in Chinese NSCLC patients after surgery.

### Exploratory 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 program is initialized and maintained by the China National Cancer Center & the Cancer Institute and Hospital, Chinese Academy of Medical Sciences (CAMS). A multicenter 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 enrollment

This study is a population-based, real-world nationwide multi-center study using original data from post-surgery 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 Center Tumor 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). Considering the average number of patients registered at the CNLCC per year, a total of over 10,000 eligible patients will be enrolled in this study during the study period, after conducting the inclusion criteria.

### Baseline and endpoints

The baseline for surgical 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\textsuperscript{18}) who had undergone complete resection;
3) patients had complete information and routine follow-up;
4) patients without history of other malignancy within five years.

**Follow-up**

The study period is set from the baseline to 5 years after the last patient registration. All patients who undergo 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\textsuperscript{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 tumor information. Basic examinations include the tumor tumour markers (for example, the squamous cell carcinoma antigen (SCC) and cytokeratin 19 fragment (CYFRA21-1), computed tomography (CT) at 3-month or 6-month intervals are routinely evaluated in the first-year post-operation. Further intervals will refer to the guidelines, patient conditions, and the opinion of the attending doctor. Bone scintigraphy, brain CT or Magnetic Resonance Imaging (MRI), and positron emission tomography (PET)-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 artificial intelligence-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 routine mortality of patients from the China centres for disease control and prevention (CDC), which studies the underlying cause of death for all Chinese people. The detail of the parameters is shown in Table 1.

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

<table>
<thead>
<tr>
<th>parameters</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 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); physical examination information; blood routine; blood clotting; liver function; renal function; etiological examinations; important previous medical history; operation history; concurrent conditions.

### Tumor 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.

---

Study records with identifiable numbers of patients will be stored as secure electronic documents in the Datacenter of the Cancer Institute and Hospital, Chinese Academy of Medical Sciences (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 enrollment, data collection and data processing. All cooperation units must be approved by the collaborative committee supervisors before accessing the data. For data sharing, anonymized electronic data will be re-organized and made public to enable secondary research, educational use, and scholarly communication after passing the data masking period.

### Data cleansing and Data management
As stated above, the basic data are collected from patients’ EHR of the registered hospital. The clinical data in each center were main exported from the Electronic Medical Record (EMR), Hospital Information System (HIS), Laboratory Information Management System (LIS), Picture Archiving and Communication System (PACS), 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 (for example, smoking or drinking history, family history of cancer). The data will be transferred to the National Cancer Institute via the virtual private network (VPN) tunnel which is highly secured.
With the help of the senior computer experts in the committee, the algorithm of 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 standardizing. Data from the CDC death databases will be linked to the analyzable 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.

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 summarized 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 applied 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 decentralized 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 cooperating with the EHR. The original manually retracted data selected randomly will run a comparative test with the data extracted by the computer programs. 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.[20]

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

The Data collection will strictly adhere to the 1,964 Declaration of Helsinki. The anonymity and confidentiality of participants are attached with great importance, and the name of the patients are replaced by identified research numbers. The data will be encrypted and stored before transferring to the central server. Only disaggregated data will be used for data sharing and educational usage. 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.

Results of the study will be disseminated publicly through annual reports on a study-specific website. The final report 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

This is the first study that intends to conduct a population-based, real-world nationwide multi-centre study to report the long-term survival outcome with resected stages IA-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 analyzed in real-world scenarios, more effective data could be obtained from a massively medical environment21. The AI-aided technique could be a technical solution for more consistent and timely data collection22. 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 in the Chinese Trial Register (ChiCTR2100052773) on 11–5–2021, http://www.chictr.org.cn/showproj.aspx?proj=136659. The study has been approved by the medical ethics committee (21/431-3102).

**Author Contributions**

He Jie & Gao Yushun developed the study idea, concept and design, initiated the collaborative project, developed data collection tools, obtained permissions and funding and drafted the study protocol. Xin Wang, Yi cheng Liang, Yuanzhuo Wang were involved in the statistical design, made substantial revisions to the protocol and were involved in the original concept, study design and implementation. He Jie & Gao Yushun was involved in design and implementation of the study, and revised the protocol. XiangZhi Meng was involved in design and implementation of the study, and revised the protocol. Boxuan Zhou was involved in design and implementation of the study, and revised the protocol. Zhenyi Xu was involved in design and implementation of the study, and revised the protocol. Hui Wang was involved in design and implementation of the study, and revised the protocol. Wenjing Yang was involved in design and implementation of the study, and revised the protocol. Ning Li was involved in design and implementation of the study, and revised the protocol. He Jie & Gao Yushun 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 (Z19110000619115), R&D Program of Beijing Municipal Education Commission (KJZD20191002302).

**Competing interests**

None declared.

**Patient and Public Involvement**

It was not appropriate or possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research

**Patient Consent for publication**
Not Applicable.

**Competing interests**

The authors declare to have no competing interests.

**Figure 1** MAP of distributions of 30 tertiary hospitals in 19 provinces

**Figure 2** Flow chart of the study procedure

**References**


<table>
<thead>
<tr>
<th>Journal:</th>
<th>BMJ Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>bmjopen-2022-070188.R1</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Protocol</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>28-Apr-2023</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Wang, Xin; Cancer Hospital Chinese Academy of Medical Sciences, Clinical Trial Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College</td>
</tr>
<tr>
<td></td>
<td>Liang, Yi cheng; Cancer Hospital Chinese Academy of Medical Sciences, Thoracic surgery department, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College</td>
</tr>
<tr>
<td></td>
<td>Wang, Yuanzhuo; Chinese Academy of Medical Sciences &amp; Peking Union Medical College, School of Basic Medicine</td>
</tr>
<tr>
<td></td>
<td>Meng, XiangZhi; Cancer Hospital Chinese Academy of Medical Sciences, 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</td>
</tr>
<tr>
<td></td>
<td>Zhou, Boxuan; Cancer Hospital Chinese Academy of Medical Sciences, 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</td>
</tr>
<tr>
<td></td>
<td>Xu, Zhenyi; Harbin Medical University, Department of Epidemiology and Biostatistics</td>
</tr>
<tr>
<td></td>
<td>Wang, Hui; Cancer Hospital Chinese Academy of Medical Sciences</td>
</tr>
<tr>
<td></td>
<td>Yang, Wenjing; Cancer Hospital Chinese Academy of Medical Sciences</td>
</tr>
<tr>
<td></td>
<td>Li, Ning; Cancer Hospital Chinese Academy of Medical Sciences, Clinical Trial Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College</td>
</tr>
<tr>
<td></td>
<td>Gao, Yushun; Chinese Academy of Medical Sciences &amp; Peking Union Medical College, 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</td>
</tr>
<tr>
<td></td>
<td>He, Jie; Cancer Hospital Chinese Academy of Medical Sciences, 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</td>
</tr>
<tr>
<td>Primary Subject Heading:</td>
<td>Health informatics</td>
</tr>
</tbody>
</table>
Secondary Subject Heading: Epidemiology, Oncology

Keywords: Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Epidemiology < ONCOLOGY, Thoracic surgery < SURGERY

SCHOLARONE™
Manuscripts
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

Running title
Chinese Lung Cancer Cohort: a protocol for a real-world study

Xin Wang 1†, Yi cheng Liang 2†, Yuanzhuo Wang 5†, XiangZhi Meng 2, Boxuan Zhou 2, Zhenyi Xu 4, Hui Wang 3, Wenjing Yang 3, Ning Li 1, Yushun Gao 1*, Jie He 2*

†These authors contributed equally.

*Corresponding authors:
Prof Jie He, 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, 17 Pan-jia-yuan South Lane, Chaoyang District, 100021, Beijing, China; email: hejie@cicams.ac.cn

Prof Yushun Gao, 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, Beijing, China, 17 Pan-jia-yuan South Lane, Chaoyang District, 100021, Beijing, China; email: ysgaopumc@163.com

Affiliations:
1. Department of Clinical Trials Center, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 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, Beijing, China.
3. Office for Cancer Diagnosis and Treatment Quality Control, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.
4. Department of Epidemiology and Biostatistics, School of Public Health, Harbin Medical University, Harbin, China.
5. Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing 100005, China.
Acknowledgements

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).

Data Statement

The dataset will be available from https://www.natdss.cn/#/home

word count: 2455 words

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 multicenter study to explore factors that could influence the outcomes of patients with IA-IIIA NSCLC who underwent curable surgery in real-world scenarios.

Methods and analysis

All patients diagnosed with NSCLC between January 2013 to 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 routine mortality of patients from the Chinese Center 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 11-5-2021, http://www.chictr.org.cn/showproj.aspx?proj=136659.

Keywords: non-small cell lung cancer; nationwide multicenter study; prognosis; online platform

Article Summary

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 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 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, 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 (for example, 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 [VP]), irrespective of epidermal growth factor receptor (EGFR) mutation status.\(^7\)\(^-\)\(^11\) The 5-year overall survival (OS) rate 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 focus on evaluating prognosis of postoperative NSCLC patients in China were retrospective analyzes with small-scale patients in single center, 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 multi-centre study using original data from the Chinese National Lung Cancer Cohort (CNLCC), in combination with AI -aided electronic medical records processing, and take the first step to retrospectively analyze the outcome and risk factors of patients with NSCLC who underwent curable surgery in China.

Aims

Primary outcomes

1. To analyze the postoperative OS rate and DFS period in real-world Chinese NSCLC patients who underwent curable surgery.

2. To establish and validate the risk prediction models and the factors that estimate the prognosis of the patients after surgery, including demographic characteristics, baseline
health status, tumor disease characteristics, clinical diagnosis, and treatment information.

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 program is initialized and maintained by the China National Cancer Center & the Cancer Institute and Hospital, Chinese Academy of Medical Sciences (CAMS). A multicenter 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 enrollment**

This study is a population-based, real-world nationwide multi-center study using original data from post-surgery 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 Center Tumor 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”, “mapprotools”, “sp”, “cairo”, “RCColorBrewer”, “openxlsx” and “rgdal” of R program (version 4.2.2). Considering the average number of patients registered at the CNLCC per year, a total of over 10,000 eligible patients will be enrolled in this study during the study period, after conducting the inclusion criteria.

**Baseline and endpoints**

The baseline for surgical 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) who had undergone complete resection;
3) Patients had complete information and routine follow-up;
4) Patients without history of other malignancy within five years.

Patient and Public Involvement

It was not appropriate or possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

Follow-up

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

The follow-up mainly relies on the evaluation when patients come back to clinics or readmission. The procedures include findings on clinical examination, results of laboratory investigations, and imaging evaluation, to update the disease status and the tumor information. Basic examinations include the tumor tumour markers (for example, the squamous cell carcinoma antigen (SCC) and cytokeratin 19 fragment (CYFRA21-1), computed tomography (CT) at 3-month or 6-month intervals are routinely evaluated in the first-year post-operation. Further intervals will refer to the guidelines, patient conditions, and the opinion of the attending doctor. Bone scintigraphy, brain CT or Magnetic Resonance Imaging (MRI), and positron emission tomography (PET)-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 artificial intelligence-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 routine mortality of patients from the China centres for disease control and prevention (CDC), an organization dedicated to addressing diseases such as cancer and infectious diseases and public health issues. Previously, NCC and CDC established a cooperation framework, namely National Cancer Data Linkage (NCDL) Platform of China, which was described in the previous study. 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.

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

<table>
<thead>
<tr>
<th>parameters</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sociodemographic characteristics</td>
<td>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.</td>
</tr>
<tr>
<td>Baseline Health Status</td>
<td>Eastern Cooperative Oncology Group performance status (ECOG PS)(^{22}); physical examination information; blood routine; blood clotting; liver function; renal function; etiological examinations; important previous medical history; operation history; concurrent conditions.</td>
</tr>
<tr>
<td>Tumor Disease Characteristics</td>
<td>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;</td>
</tr>
<tr>
<td>Treatment information</td>
<td>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).</td>
</tr>
<tr>
<td>Genetic testing variables</td>
<td>Cancer-associated genotyping results; testing methods; timing and variation of results during preoperative, perioperative, postoperative and after disease advance or recurrence (if exist)</td>
</tr>
<tr>
<td>Disease progression, recurrence and survival information.</td>
<td>Time of disease progress or recurrence; tumour recurrence sequence; corresponding recurrence site; survival status and death time; cause of death.</td>
</tr>
</tbody>
</table>

Study records with identifiable numbers of patients will be stored as secure electronic documents in the Datacenter of the Cancer Institute and Hospital, Chinese Academy of Medical Sciences (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 enrollment, data collection and data processing. All cooperation units must be approved by the
collaborative committee supervisors before accessing the data. For data sharing, anonymized electronic data will be re-organized and made public to enable secondary research, educational use, and scholarly communication after passing the data masking period.

Data cleansing and Data management

As stated above, the basic data are collected from patients’ EHR of the registered hospital. The clinical data in each center were mainly exported from the Electronic Medical Record (EMR), Hospital Information System (HIS), Laboratory Information Management System (LIS), Picture Archiving and Communication System (PACS), 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 (for example, smoking or drinking history, family history of cancer). The data will be transferred to the National Cancer Institute via the virtual private network (VPN) tunnel which is highly secured.

With the help of the senior computer experts in the committee, the algorithm of 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 standardizing. Data from the CDC death databases will be linked to the analyzable 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 summarized 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 applied 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 decentralized 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 cooperating with the EHR. The original manually retracted data selected randomly will run a comparative test with the data extracted by the computer programs. 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 al23.

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

The Data collection will strictly adhere to the 1,964 Declaration of Helsinki. The anonymity and confidentiality of participants are attached with great importance, and the name of the patients are replaced by identified research numbers. The data will be encrypted and stored before transferring to the central server. Only disaggregated data will be used for data sharing and educational usage. 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.

Results of the study will be disseminated publicly through annual reports on a study-specific website. The final report 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 multi-center 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. Similarly,
the lung cancer database project established in 1999 at the National Cancer Center Hospital
East, Japan also constructed a large-scale cancer registry for lung cancer and integrated data on
various factors. However, this is the first study that intends to conduct a population-based,
real-world nationwide multi-centre study to report the long-term survival outcome with
resected stages IA-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 analyzed in real-world scenarios, more effective data could be
obtained from a massively medical environment. The AI-aided technique could be a technical
solution for more consistent and timely data collection. 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 in the Chinese Trial Register (ChiCTR2100052773) on 11–5-
The study has been approved by the medical ethics committee (21/431-3102).

Author Contributions

He Jie & Gao Yushun developed the study idea, concept and design, initiated the collaborative
project, developed data collection tools, obtained permissions and funding and drafted the study
protocol. Xin Wang, Yi cheng Liang, Yuanzhuo Wang were involved in the statistical design,
made substantial revisions to the protocol and were involved in the original concept, study
design and implementation. He Jie & Gao Yushun was involved in design and implementation
of the study, and revised the protocol. XiangZhi Meng was involved in design and
implementation of the study, and revised the protocol. Boxuan Zhou was involved in design and implementation of the study, and revised the protocol. Zhenyi Xu was involved in design and implementation of the study, and revised the protocol. Hui Wang was involved in design and implementation of the study, and revised the protocol. Wenjing Yang was involved in design and implementation of the study, and revised the protocol. Ning Li was involved in design and implementation of the study, and revised the protocol. He Jie & Gao Yushun 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).

**Competing interests**

None declared.

**Patient Consent for publication**

Not Applicable.

**Competing interests**

The authors declare to have no competing interests.

**Figure 1 MAP of distributions of 30 tertiary hospitals in 19 provinces**

**Figure 2 Flow chart of the study procedure**

**References**


