PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Incidence and Survival of Non-small Cell Lung Cancer in Shanghai: A Population-Based Cohort Study
AUTHORS	Fan, Heng; Shao, Zhen-Yi; Xiao, Yuan-Yuan; Xie, Zhi-Hui; Chen, Wen; Xie, Hua; Qin, Guo-You; Zhao, Nai-Qing

VERSION 1 - REVIEW

REVIEWER	Li Tao Cancer Prevention Institute of California, Fremont
REVIEW RETURNED	04-Aug-2015

GENERAL COMMENTS	This is a large population-based study providing insights on the pattern of incidence and survival for NSCLC patients in China. Lung cancer is the most common cancer diagnosed in Chinese men, and the leading cause of cancer related death in the country. Under the circumstances that cancer incidence in China is ferociously rising, it is important, as the authors recognized, to fill the gap in providing key statistics. In the new treatment era with target therapy, it is also important to report on NSCLC separately from previously reported overall lung cancer rates. Major revisions in addition to extensive language polishing and paragraph restructuring are highly recommended for publication of this article. Major comments: 1. Systematic description of data collection and cases ascertainment is very necessary given since no previous publication is cited from this dataset and the definition of a 'mega inpatients database' is not commonly used. The description of 'citizen healthcare information network' 'which covers county-level and above hospital which are qualified to diagnose…' is confusing. The authors have compared their findings with articles based on 'National Central Cancer Registry' and other official reports- it is helpful to explain the difference between this data and data source of other 'official reports'. 2. In the manuscript, description (especially numbers) regarding study cohort construction and explanation/reference for key elements in data constructions, such as 'Shanghai Statistical Yearbook' should be included in the method section. Inclusion/exclusion criteria was only noted for survival analysis, does this mean all NSCLC patients diagnosed between 2011 and 2013 in Shanghai were included in the incidence analysis? It will be very helpful to include a flow chart of cohort establishment and study sample exclusion, with numbers. 3. According to the manuscript, there were 12,996 new cases in

	Shanghai in 2013, but 15,020 cases for year 2011, 2012, and 2013 in total? Also ~70% of the cases were missing on stage and histological subtype, which makes the eligibility of the study questionable. The authors need to justify that this large missing will not drawback the findings of the study. Less than 15% of the incidence cases were included in the survival analysis, which also needs some explanation. 4. Please provide incidence data (maybe in a table) of incidence by age group, sex, stage at diagnosis, and histological subtypes. The short-term survival and lack of generalizability to long-term survival needs to be emphasized when discussing limitations. Minor comments: 1. The authors cited statistics of SEER; further details on the US cancer registry and reference to other SEER-based studies on NSCLC surveillance would be very helpful when there is not sufficient published statistics of incidence/survival in China. 2. Were there any violations of the proportional hazard assumption when using Cox regression model? Given the significant association of stage and survival outcome, the authors might consider using stage as an underlying stratification variable when performing a Cox analysis. 3. A study by Gomez et al. CEBP 2015 compared NSCLC incidence trend for U.S.-born and foreign-born Chinese Americans and found statistically significant decreasing trends of NSCLC among foreign-born males. This is a very strong support for the argument of role of smoking in discussing the incidence in Shanghai for this study. 4. OS was referred first time without an abbreviation (page 7 first paragraph), and 'overall survival' was used many times in the manuscript. 5. Even though data on therapy with epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors, a class of drugs with significant activity in Asian NSCLC patients, is not available, targeted treatment needs to be discussed in the spectrum of the study.

REVIEWER	Anders Mellemgaard Unit of Thoracic oncology Herlev University Hospital Copenhagen, Denmark
REVIEW RETURNED	14-Aug-2015

GENERAL COMMENTS	That the present paper is unique for mainland China not entirely true. Cancer registry's are operative and have been som for many years in Beijing, Cixian, Haining, Haibing etc etc,. In Shanghai, the cancer registry has a good reputation and data from these registries have been included in WHO publications. The paper is however more informative than the annual registry reports. As data is only available from 2013 onwards, time trends are not provided. Can the authors give some indication of the completeness of data, i.e. which cancer cases may not be reported?, only pathology confirmed cases? Surgical rates were high (21%) stage IV. Are these patient with oligometastatic disease? Survival figures are premature as data is very recent and the authors need to state this in their discussion.

	The authors correctly mentions the limitation of the study. In general the english language is acceptable, but the paper would benefit from a linguistic overlook, correcting the minor errors found.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

We would like to send our sincere thanks to the reviewer for the thoughtful review, patient instructions and constructive suggestions which help us to improve the manuscript and our cognition of the subject. Also, we thank the reviewer for acknowledging the value of our work.

Major comments:

1. Systematic description of data collection and cases ascertainment is very necessary given since no previous publication is cited from this dataset and the definition of a 'mega inpatients database' is not commonly used. The description of 'citizen healthcare information network' 'which covers county-level and above hospital which are qualified to diagnose…' is confusing. The authors have compared their findings with articles based on 'National Central Cancer Registry' and other official reports- it is helpful to explain the difference between this data and data source of other 'official reports'

Response: Thank you for this suggestion. We are sorry for making you confused about the description of the data source. We have rewritten the data source part in methods section and added details about the database (see methods, data source, paragraph 1). Case ascertainment was also added (see methods, data source, paragraph 2). Our data is derived from Shanghai Health Information Network, a non-profit database initiated in 2011 and organized by the Shanghai Municipal Bureau of Health. It integrates the data of Health Information Systems (HIS) from all the comprehensive hospitals and specialist hospitals in Shanghai. "County-level and above hospitals" was an inappropriate translation of a Chinese term which is revised as "comprehensive hospital and specialist hospitals".
In addition, we added comments of the difference between the data and other source in detail in discussion and mentioned the difference concisely in introduction (see the second to the last paragraph of discussion, line 3-8; the second paragraph of introduction, line 3-8).

2. In the manuscript, description (especially numbers) regarding study cohort construction and explanation/reference for key elements in data constructions, such as 'Shanghai Statistical Yearbook' should be included in the method section. Inclusion/exclusion criteria was only noted for survival analysis, does this mean all NSCLC patients diagnosed between 2011 and 2013 in Shanghai were included in the incidence analysis? It will be very helpful to include a flow chart of cohort establishment and study sample exclusion, with numbers.

Response: Thank you for this constructive comment. We are sorry for the vague description. As you suggested, we have rewritten the methods section by adding description, numbers and a flow chart regarding study cohort construction (see methods, case selection and inclusion criteria, Figure 1). Reference for'Shanghai Statistical Yearbook' is added (see methods, data source, paragraph 4, line 4). Incidence analysis only included cases identified in 2013 (n=12,996), the description has been added to paragraph 2 in case selection and inclusion criteria section.

3. According to the manuscript, there were 12,996 new cases in Shanghai in 2013, but 15,020 cases for year 2011, 2012, and 2013 in total? Also ~70% of the cases were missing on stage and histological subtype, which makes the eligibility of the study questionable. The authors need to justify that this large missing will not drawback the findings of the study. Less than 15% of the incidence cases were included in the survival analysis, which also needs some explanation.

Response: Thanks for the reviewer's insightful comments. Since the data source network was established in 2011 and not until 2013 it had covered all the hospitals in Shanghai, most of the patients (n=12,996) identified in the database was diagnosed in 2013. That's the reason why we only use cases diagnosed in 2013 to calculate incidence rate. Also, since the database is newly established, variables like TNM staging and histological subtypes for about 70% patients are not available. However, important variables including diagnosis and demographic information is of good quality and resulted in an age-adjusted incidence comparable to that of other studies. Moreover, the constitutions of gender and TNM staging among cases with non-missing value in our study were consistent with studies of foreign countries. Therefore, we believe the missing will not drawback the findings of the study (see the second to the last paragraph of discussion, line 9-11). Also for controlling the influence of missing data, only patients with known prognostic factors (13.4%) was recruited in survival analysis in order to diminish the selection bias. A limitation of this aspect has been discussed in the revision (see the last paragraph of discussion, line 5-12).

4. Please provide incidence data (maybe in a table) of incidence by age group, sex, stage at diagnosis, and histological subtypes. The short-term survival and lack of generalizability to long-term survival needs to be emphasized when discussing limitations.

Response: Thanks for the reviewer's suggestion. We have provided crude and age-adjusted incidence by age group and sex(see result, Table 1). Regrettably, because of former mentioned missing in stage and histological subtypes, it is hard to calculate incidence rate stratified by these variables. We have mentioned this in discussion of limitation (see the last paragraph of discussion, line 8-9).
We are also terribly sorry that we made a mistake in the first manuscript when calculating age-adjusted incidence rate by adding the weights of Segi's standard population's several age groups instead of weighting them separately. We recognized the mistake when calculating subgroup adjusted incidence rates. Now the overall adjusted incidence rate is 39.05 per 100,000 (former wrong adjusted incidence rate: 28.64 per 100,000). We provide the detail information of calculating the age-adjusted incidence rate here for review:

	Population of Shanghai,			2013 No. of NSCLC cases,2013			
age	overall	male	female	overall	male	female	Segi's weight
0-4	826334	440191	386143	0	0	0	0.120
5-9	659137	354610	304527	0	0	0	0.100
10-14	579970	306542	273427	0	0	0	0.090
15-19	1167867	599652	568215	4	1	3	0.090
20-24	2729503	1395468	1334035	16	8	8	0.080
25-29	2678266	1374582	1303684	35	11	24	0.080
30-34	2216748	1160243	1056505	85	23	62	0.060
35-39	2001315	1068500	932816	205	72	133	0.060
40-44	1955940	1042739	913201	430	160	270	0.060
45-49	1875673	996582	879091	838	377	461	0.060
50-54	1877802	971479	906322	1345	655	690	0.050
55-59	1795187	909238	885949	2391	1253	1138	0.040
60-64	1185752	611718	574033	2510	1355	1155	0.040
65-69	693064	352834	340229	2003	1154	849	0.030
70-74	544831	262113	282718	1384	790	594	0.020
75-79	578228	259258	318970	1058	625	433	0.010
80-84	366171	153227	212943	499	318	181	0.005
>=85	246114	89673	156441	193	101	92	0.005
overall	23977900	12348650	11629250	12996	6903	6093	1.00

According to the reviewer's comments, the limitation of lack of long-term survival was added to the

discussion section (see the last paragraph of discussion, last two sentences).

Minor comments:
1. The authors cited statistics of SEER; further details on the US cancer registry and reference to other SEER-based studies on NSCLC surveillance would be very helpful when there is not sufficient published statistics of incidence/survival in China.

Response: Thanks for the thoughtful suggestion. Details of the SEER registry have been added to the discussion section (see discussion, paragraph 1, line 6-8), noted as "…SEER registry, a population-based national cancer registries covering approximately 28% of the United States (US) population and 50% of Asians in the US". Descriptions of other data sources were also added. Another study based on SEER was referenced in the discussion of survival (2011 overall 1-year survival rate, see survival part of discussion section, paragraph 1, line 8-9). After the manuscript submitted, two studies of lung cancer incidence in China has been published and were referenced in the revision, so no more study from SEER was referenced in discussion of incidence (see discussion, paragraph 2, line 5 and line 11).

2. Were there any violations of the proportional hazard assumption when using Cox regression model? Given the significant association of stage and survival outcome, the authors might consider using stage as an underlying stratification variable when performing a Cox analysis.

Response: Thanks for the comments. We have visually checked the proportional hazard hypothesis with log-log curves. The log-log curves for variables in Cox model (including stages) were parallel. (This method was mentioned in method section, statistical methods, line 15 without showing the log-log curves. The log-log curves for TNM stage is parallel, available on request).
It is better to make staging a stratification variable since the staging affects survival significantly as you have mentioned. Since we attempted to evaluate the prognostic effect of tumor size score controlling regional lymph nodes score when we considered doing Cox regression, we included T,N,M score as potential prognosis factors and excluded staging in the model. However, we used staging as the stratification variable in multivariate Cox model when analyzed the prognostic effect of surgical resection (see the last paragraph of result, table 5).

3. A study by Gomez et al. CEBP 2015 compared NSCLC incidence trend for U.S.-born and foreign-born Chinese Americans and found statistically significant decreasing trends of NSCLC among foreign-born males. This is a very strong support for the argument of role of smoking in discussing the incidence in Shanghai for this study.

Response: Thanks for the reviewer's suggestion. We have revise this part of discussion (see discussion, paragraph 2, line 29-34), and noted in the revision that "The relationship between smoking and lung cancer is also confirmed in the study of Gomez et al. Gomez et al found a significant decline in the incidence of squamous cell lung cancer among foreign-born Chinese Americans from 1900 to 2004, accompanied by a temporal decline in current smoking prevalence among them, while the incidence was stable for adenocarcinoma, which is less closely associated with tobacco smoke than squamous cell lung cancer."(see reference: Gomez SL, Yang J, Lin SW, et al. Incidence Trends of Lung Cancer by Immigration Status among Chinese Americans. Cancer Epidemiol Biomarkers Prev 2015;24:1157-1164)

3. OS was referred first time without an abbreviation (page 7 first paragraph), and 'overall survival' was used many times in the manuscript.

Response: We are sorry for the error. The mistakes have been corrected in the revision. Other abbreviations have been checked as well.

4. Even though data on therapy with epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors, a class of drugs with significant activity in Asian NSCLC patients, is not available, targeted treatment needs to be discussed in the spectrum of the study.

Response: Thanks for the comment. We have revised this part in discussion and weaken the argument as “Meanwhile, advances in treatment in recent years such as the introduction of target agents and adjuvant chemotherapy after complete resection may improve the survival of NSCLC patients.” (see discussion, survival section, paragraph 2, the last sentence)

Reviewer 2
Thank you for spending time on reviewing our manuscript and for your helpful evaluation and suggestions in terms of design evaluation, methods and results. As your have mentioned, there are national and many local cancer registries existing in China. Different from studies based on these registry data which only reported incidence or/and reported small cell lung cancer together with non-small cell lung cancer, our study identified and analyzed NSCLC cases specifically with more details including treatment and survival analysis. We are appreciate for your reminding and we are sorry for making the description inaccurate. We have revised related sentences in the revision (such as the objectives in the abstract).

1. Can the authors give some indication of the completeness of data, i.e. which cancer cases may not be reported?, only pathology confirmed cases?

Response: Thanks to reviewer for pointing out this important point. Yes, only pathology confirmed data were reported. As your suggested, we have added indication of the completeness of data noted as “This network was initiated in 2011 and it has covered all the comprehensive hospitals and specialist hospitals qualified for cancer diagnosis in Shanghai metropolitan area in 2013” (see method, data source, paragraph 1, line 9-11) and “NSCLC cases were identified using the primary site coding system of the International Classification of Disease for Oncology 3rd Revision (ICD-10) from the World Health Organization and the pathological findings in the medical records. The diagnosis of NSCLC was confirmed by tissue diagnosis.” (see method, data source, paragraph 2)

2. Surgical rates were high (21%) stage IV. Are these patient with oligometastatic disease?

Response: Thank you for this insightful comment. About 20% patients with stage IIIb or stage IV NSCLC underwent surgical resection. We also identified with this higher surgical rate among stage IIIb/IV patients during analysis and then checked the detailed text record of diagnosis and treatment of these patients in our database (there is a variable recording the detailed diagnosis and treatment process since the database is derived from Health Information Systems inside healthcare organizations). We found some of these patients were with oligometastatic disease indeed. However, this variable for many more patients were missing. This is one of the limitations of our studies, and we added it to the limitation in discussion noted as ” In addition, since the network database is newly established, though important variables such as diagnosis and demographic information are available, the TNM classification and histological subtype were still unavailable in several patients.... Missing in more detailed records of diagnosis and treatment also prohibited us from further analysis.” (see the last paragraph in discussion, line 5-10)

3. Survival figures are premature as data is very recent and the authors need to state this in their discussion.

Response: Thank you for this comment. According to your comments, a limitation of this aspect has been discussed in the revision (see the last paragraph of discussion, the last sentence).

4. In general the English language is acceptable, but the paper would benefit from a linguistic overlook, correcting the minor errors found.

Response: Thank you for this suggestion. According to your suggestion, the revision have been polished by an English editing company.

VERSION 2 – REVIEW

REVIEWER	Li Tao Cancer prevention institute of California
REVIEW RETURNED	21-Sep-2015

GENERAL COMMENTS	Thanks for the thorough responses and careful edits to the manuscript by the authors. The reviewer especially appreciate the flow-chart of the cohort construction and the new table 1 for incidence analysis. According to the authors, the data that the manuscript has based on is a novel and comprehensive high-quality system in the metropolitan city of Shanghai, just with very short follow-ups available. would look forward to future publications of this data. A couple minor suggestions: 1. A few p- for differences between non-missings and missings for major patients characteristics will be helpful. 2. typo on page 48, line 14, Gomez et al, 1990-2004, instead of 1900

REVIEWER	anders mellemgaard herlev university hospital copenhagen
REVIEW RETURNED	16-Sep-2015

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.