Data Contained in Electronic Health Records Associated with Risk of Lung Cancer: A Protocol for a Systematic Review

Lamorna Brown (lb300@st-andrews.ac.uk)
University of St Andrews https://orcid.org/0000-0002-6206-8196
Frank Sullivan
University of St Andrews
Tom Kelsey
University of St Andrews
Utkarsh Agrawal
University of St Andrews

Protocol

Keywords: Lung cancer, overdiagnosis, phenotyping diseases, Electronic Health Records (EHRs), Cochrane library.

Posted Date: September 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-910471/v1

Abstract

Background: Lung cancer is one of the most common and aggressive forms of cancer, resulting in a low survival and high mortality rate. To improve cancer related outcomes, high-risk subpopulations should be identified to reduce overdiagnosis of lung cancer and aid in the implementation of interventions. Electronic Health Records (EHRs) have been effective in identifying cohorts and phenotyping diseases. To identify whether EHR data can be used in risk modelling for lung cancer, this review will seek to identify data features that are contained in EHRs and related to lung cancer.

Methods: A search strategy was developed and then applied to MEDLINE via Ovid, Web of Science, Scopus and the Cochrane library. The titles and abstracts of studies will be identified and screened independently by reviewers. Reviewers will read the full texts of studies that appear to meet the eligibility criteria after initial screening. Articles that meet the criteria at this stage, will have their bibliographies examined for relevant studies. Data extraction will then be performed independently by reviewers and a narrative synthesis will be carried out.

Discussion: While risk factors for lung cancer have been extensively researched, there has to date been no effort to identify whether information that relates to these factors are available in EHRs and can be modelled with. As such, the results of the review will seek to broaden knowledge around the use of EHRs in lung cancer risk modelling and inform researchers of the variables that are available in EHRs.

Registration: PROSPERO CRD42021246781, Registered on 26/04/21.

Background

In 2018, there were estimated 2.09 million lung cancer cases and 1.76 million deaths attributable to lung cancer worldwide [1]. As a result, it is the leading cause of cancer related death in both males and females [2, 3]. While incidence rates in Europe, North America and Oceania show a decreasing pattern for male incidence, it is likely that reported incidence and prevalence rates will not reflect the true extent of the disease due to inadequate ascertainment [4, 5, 6, 7].

The overall survival rate for lung cancer has increased over the past four decades but remains poor (< 21.8%) [8, 9]. Stage at diagnosis remains an important prognostic factor explaining the variation in survival rates across European countries [8, 10]. Table 1 examines the net survival for lung cancer stages in England [11]. The survival rates suggest that individuals diagnosed with early-stage lung cancer have significantly better chances of survival, compared to those diagnosed at a later stage [10, 11, 12]. As there are improved survival rates for those diagnosed at stage I and II, the UK has recommended early-diagnosis of the disease [13].

Table 1
Lung cancer five-year net survival by stage, with incidence by stage (all data: adults diagnosed 2013–2017, followed up to 2018) in England.

Sex	Stage	Number of cases	5 Year Net survival (%)
Female	1	15,219	61.5
	2	5,978	36.4
	3	16,073	13.7
	4	40,874	3.4
Male	1	13,717	51.1
	2	7,540	32.1
	3	19,631	11.6
	4	49,355	2.3

Early-stage diagnosis is challenging given that current approaches to diagnosis rely on patient presentation and symptom appraisal [14]. This may allow for greater delays in diagnosis, as lung cancer doubling times are variable, individuals can normalise non-specific symptoms or attribute symptoms to other conditions [15]. Comorbidities, such as COPD, are common in lung cancer patients which can complicate diagnosis and lead to practitioners missing clinical clues [16]. Other symptoms, such as haemoptysis, are more specific but occur in the later, more acute stages of lung cancer and are related to poor prognosis [17]. To avoid late-stage diagnosis a different approach to lung cancer diagnosis is required, involving the targeting of high-risk populations to enable smoking cessation and screening interventions to be effectively implemented.

Recent randomised trials have found that screening using low dose computer tomography (LDCT) is an effective tool to detect early-stage lung cancer. The 2011 US National Lung Screening trial found a 20% reduction in lung cancer mortality over 5-years [18]. Subsequent trials, such as the NELSON and UK Lung Cancer Screening trial similarly found that those undergoing LDCT scans compared to x-rays had a reduced probability of dying from lung cancer [19, 20, 21]. The Early detection of Cancer of the Lung Scotland (ECLS) trial, also indicated that blood-based biomarkers are effective when used in conjunction with LDCT, significantly reducing late-stage diagnosis and lung cancer mortality [22]. Despite this evidence, screening is yet to be recommended in the UK [23].

Screening tools require a targeted approach, to ensure that resources are allocated to those who are most at risk, reducing overdiagnosis and patient distress [22, 24]. Risk prediction models can identify high-risk populations and risk factors associated with lung cancer. Initial models focused on epidemiological factors, such as smoking and age, and have subsequently been expanded to include clinical assessment

and genetic information to improve prediction [25, 26]. However, there is yet to be consensus on which model should be utilised for screening in the UK, as there has been limited research comparing and externally validating models [26, 27].

The shift towards the widespread use of Electronic Health Records (EHRs) has provided a unique opportunity to examine disease trajectories, clustering and prediction [28, 29]. Studies utilising EHR data can obtain larger sample sizes, utilise data which is arguably more reflective of the general population and explore a greater number of predictors [28]. There are challenges to modelling with EHR data due to poor data quality caused by missing or mis-categorised data [28, 30]. However, a recent risk model for lung cancer produced good discrimination (AUC: 0.88), demonstrating the potential for EHR based lung cancer research [31]. The use of EHRs in epidemiological research has grown but there is at present no review examining whether information that relates to risk factors for lung cancer are present in EHR data. As such, this review will seek to examine data features present in EHRs which provide risk estimates for lung cancer.

Methods

This protocol has been developed according to the Review and Meta-Analysis-Protocols (PRISMA-P) statement (see **Additional File 2**).

Aim

To determine what data features, contained in EHRs, are associated with risk of lung cancer in a current, ever and former smoking population.

The systematic review will focus on identifying data features in EHRs that are associated with incidence of lung cancer. Where appropriate, data on measures of effect will be extracted.

Eligibility Criteria

Studies eligible for inclusion must meet the following criteria:

1. Study designs

As the review is interested in determining data features associated with risk of lung cancer, observational studies such as cohort, case-control, case series, cross-sectional and prospective designs will be included. Systematic reviews are also eligible for inclusion.

2. Participants

Participants included in studies will be current, ever and former smokers as this group is most at risk of developing lung cancer. Ineligible for inclusion are studies that feature participants 18 years or younger or/and that do not include a measure of effect (e.g. risk ratio (RR) and 95% confidence interval (CI)). Data on participants used to model risk must be from electronic health records.

3. Interventions

Studies examining interventions will not be included.

4. Comparators

The comparator group to current/ever/former smokers will be non-smokers or those not diagnosed with lung cancer. Case-control studies include those who do not develop lung cancer as the comparator, or they may compare non-smokers with current or ever smokers.

5. Outcomes

Studies to be considered must contain an estimate of risk for lung cancer as the main outcome being considered e.g., the risk of lung cancer for an individual who is a current smoker and has asthma. The primary outcome will be presented as a measure of effect (i.e. as the RR, hazard ratio (HR), odds ratio (OR), incidence rate ratio (ICR) or standardized incidence ratio (SIR)) for each risk factor and data feature, with the 95% CI.

6. Setting

Studies of any type of setting will be included.

7. Language

International studies will be included but must be in English.

Search strategy

The search strategy will examine the following databases for relevant studies, using the same search string which will be adapted to the database under review: Cochrane library, MEDLINE (ovid), Scopus and Web of Science. The search strategy used for MEDLINE and adapted to other databases is given in **Additional file 2**.

Websites for EHR and administrative databases will be searched for bibliographic lists (e.g., Clinical Practice Research Datalink, www.cprd.com). Furthermore, relevant grey literature will be examined through Open Grey (http://www.opengrey.eu/).

Studies which are eligible for inclusion will then have their bibliographies searched, for additional relevant studies. Content experts may also be contacted for information about other potential ongoing or unpublished studies.

Selection Process

Once the searches in the databases listed above have been undertaken, two reviewers will independently screen the studies by title and abstracts. Studies that appear to meet the inclusion criteria will then have their full texts assessed for eligibility. The reviewers will subsequently determine whether the study can be

included in the review. If the reviewers cannot come to an agreement on a study, a third reviewer will be brought in to determine whether the article is eligible for inclusion.

Data Collection

Two authors will extract information from the studies separately, then compare and discuss results. Forms for the data collection will be piloted on two studies. Covidence will be used for the collection of data [32]. Extracted data will include any demographic or socioeconomic descriptive information. The measures of effect for risk factors and their 95% CI, the methodology of studies, number of participants (and controls for case-control studies), and length of follow up (for cohort) will also be extracted.

Outcomes

The primary outcome is the risk of incidence of lung cancer for risk factors (presented as a measure of effect e.g., RR, HR, OR, ICR or SIR) for each risk factor and outcome of incidence, with the 95% CI.

Risk of bias

For case-control and cohort studies the Newcastle Ottawa scale (NOS) will be used to assess risk of bias [33]. The AXIS tool will be used to assess bias on cross-sectional studies [33]. For other studies the CASP checklists will be used [34]. These will be piloted on a select number of papers initially. Records will be kept on decisions made for data extraction.

Two reviewers will implement the risk of bias forms for studies which have met the inclusion criteria after the full-text articles have been examined. Information collected on risk of bias will be synthesised and tabulated. The results will provide information about the quality of evidence for risk factors which will be examined in the discussion.

Data Synthesis

A narrative synthesis will be carried once the data extraction forms have been completed by the reviewers. A summary of included studies will provide information on the authors, study design, number of study participants, how the studies have recorded smoking behaviour (i.e., smoking status, pack years, duration of smoking etc.), and the measures of effect for the pre-existing data features identified in studies. It is expected there will be some clinical heterogeneity between studies so limitations of the studies will be recorded, extracted and discussed in the paper. Funnel plots will be used to investigate potential publication bias.

Additionally, if the literature supports a statistical combination of results, sensitivity analysis will be performed to assess the included studies, in terms of comparable quantitative information.

Discussion

The risk factors for lung cancer have been extensively researched and documented. Systematic reviews have been carried out on risk models and risk factors that relate to lung cancer prognosis and development in individuals. This review will be the first to examine and synthesize the evidence around whether data on lung cancer risk factors are present in EHRs and can be used to model risk with. This will inform researchers about the non-acute variables which should be included in lung cancer risk prediction models.

Declarations

Acknowledgments

Not Applicable.

Funding

This review has been funded by the Melville Trust for the Care and Cure of Cancer. They have had no role in the development of the protocol.

Availability of data and materials

Not applicable.

Contributions
All authors contributed to the protocol. LB and FS had the original idea, with LB drafting the protocol, design and search strategy. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. World Health Organization. Global Health Observatory. [Internet] 2018. Geneva: World Health Organization; [cited on 16/11/2020]. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer

2. European Respiratory Society. *European Lung White Book* [Internet] 2013. *[cited on 10/12/2020] Available from:* https://www.erswhitebook.org/chapters/
3. World Health Organization. Global Health Observatory. [Internet] 2018. Geneva: World Health Organization; [cited on 16/11/2020]. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer, Accessed on: 16/11/2020.
4. Arnold M, Rutherford MJ, Bardot A, Ferlay J, Andersson TM, Myklebust TÅ, Tervonen H, Thursfield V, Ransom D, Shack L, Woods RR. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. The Lancet Oncology. 2019 Nov 1;20(11):1493-505.
5. Zhang Y, Ren JS, Huang HY, Shi JF, Li N, Zhang Y, Dai M. International trends in lung cancer incidence from 1973 to 2007. Cancer medicine. 2018 Apr;7(4):1479-89.
6. Public Health Scotland. Cancer Statistics [Internet] 2017. [cited on 14/11/2020] Available from: https://www.isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/
7. Coggon, D., Rose, G., Barker, D.J.P., Epidemiology for the Uninitiated, 5th ed., India: BMJ publishing group; 2003
8. Walters S, Maringe C, Coleman MP, Peake MD, Butler J, Young N, Bergström S, Hanna L, Jakobsen E, Kölbeck K, Sundstrøm S. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. Thorax. 2013 Jun 1;68(6):551-64.
9. Arnold M, Rutherford MJ, Bardot A, Ferlay J, Andersson TM, Myklebust TÅ, Tervonen H, Thursfield V, Ransom D, Shack L, Woods RR. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. The Lancet Oncology. 2019 Nov 1;20(11):1493-1505.
10. Lu T, Yang X, Huang Y, Zhao M, Li M, Ma K, Yin J, Zhan C, Wang Q. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. Cancer management and research. 2019;11:943-953
11. Cancer Research UK. Survival [Internet] 2020. [cited on 9/12/2020] Available from: https://www.cancerresearchuk.org/about-cancer/lung-cancer/survival
12. Cancer Research UK. Lung Cancer Survival by Stage. [Internet] 2020. [cited on 10/12/2020]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/survival#heading-Three
13. NICE, 2019, Lung Cancer: Diagnosis and Management. Available at: https://www.nice.org.uk/guidance/ng122/resources/lung-cancer-diagnosis-and-management-pdf-66141655525573, Accessed on 14/12/2020
14. Cunningham Y, Wyke S, Blyth KG, Rigg D, Macdonald S, Macleod U, Harrow S, Robb KA, Whitaker KL. Lung cancer symptom appraisal among people with chronic obstructive pulmonary disease: a qualitative interview study. Psycho- 2019 Apr;28(4):718-725.

15. Hong JH, Park S, Kim H, Goo JM, Park IK, Kang CH, Kim YT, Yoon SH. Volume and Mass Doubling Time of Lung Adenocarcinoma according to WHO Histologic Classification. Korean journal of radiology. 2021 Mar;22(3):464.
16. Walter FM, Rubin G, Bankhead C, Morris HC, Hall N, Mills K, Dobson C, Rintoul RC, Hamilton W, Emery J. Symptoms and other factors associated with time to diagnosis and stage of lung cancer: a prospective cohort study. British journal of cancer. 2015 Mar;112(1):S6-13.
17. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine. 2011 Aug 4;365(5):395-409.
18. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers JW, Weenink C, Yousaf-Khan U, Horeweg N, van't Westeinde S. Reduced lung-cancer mortality with volume CT screening in a randomized trial. New England journal of medicine. 2020 Feb 6;382(6):503-513.
19. Xu DM, Gietema H, de Koning H, Vernhout R, Nackaerts K, Prokop M, Weenink C, Lammers JW, Groen H, Oudkerk M, van Klaveren R. Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung cancer. 2006 Nov 1;54(2):177-84.
20. Field JK, Duffy SW, Baldwin DR, Brain KE, Devaraj A, Eisen T, Green BA, Holemans JA, Kavanagh T, Kerr KM, Ledson M. The UK Lung Cancer Screening Trial: a pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. Health technology assessment. 2016 May;20(40):1-146.
21. Sullivan FM, Mair FS, Anderson W, Armory P, Briggs A, Chew C, Dorward A, Haughney J, Hogarth F, Kendrick D, Littleford R. Earlier diagnosis of lung cancer in a randomised trial of an autoantibody blood test followed by imaging. European Respiratory Journal. 2021 Jan 1;57(1):1-11
22. Sullivan, F.M. and van Beusekom, M. Early diagnosis of lung cancer in people most at risk. British Journal of General Practice, 2020 Dec;70(701):572-573.
23. Cancer Research UK, Lung Cancer Screening, [Internet] 2020 [cited 25/01/21], Available at: https://www.cancerresearchuk.org/health-professional/screening/lung-cancer-screening#lungscreening0
24. Oudkerk, M., Devaraj, A., Vliegenthart, R., Henzler, T., Prosch, H., Heussel, C.P., Bastarrika, G., Sverzellati, N., Mascalchi, M., Delorme, S. and Baldwin, D.R. European position statement on lung cancer screening. The Lancet Oncology. 2017;18(12):754-766.
25. Marcus MW, Raji OY, Field JK. Lung cancer screening: identifying the high risk cohort. Journal of thoracic disease. 2015 Apr;7(Suppl 2):S156-S162
26. Gray EP, Teare MD, Stevens J, Archer R. Risk prediction models for lung cancer: a systematic review. Clinical lung cancer. 2016 Mar 1;17(2):95-106.
27. Toumazis I, Bastani M, Han SS, Plevritis SK. Risk-Based lung cancer screening: A systematic review. Lung Cancer. 2020 Jul; 147:154-186

28. Goldstein BA, Navar AM, Pencina MJ, Ioannidis J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2017 Jan 1;24(1):198-208.
29. Jensen K, Soguero-Ruiz C, Mikalsen KO, Lindsetmo RO, Kouskoumvekaki I, Girolami M, Skrovseth SO, Augestad KM. Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific reports. 2017 Apr 7;7(1):1-12.
30. Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. BMC medical informatics and decision making. 2017 Dec;17(1):1-2.
31. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, Xia M, Wang O, Liu M, Weng CH, Duong SQ. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine. Journal of Medical Internet Research. 2019;21(5):e13260.
32. Covidence systematic review software [Software]. 2014. Available at: https://www.covidence.org/
33. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?. Military Medical Research. 2020 Dec;7(1):1-11.
34. Critical AppraisalSkills Programme. CASP checklists [Internet] 2019 [cited 05/06/21] Available at: https://casp-uk.net/casp-tools-checklists/

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile2.pdf
- Additionalfile1PRISMAPchecklist.pdf